Machine Learning for the New York City Power Grid

Cynthia Rudin[†*], David Waltz[*], Roger N. Anderson[*], Albert Boulanger[*], Ansaf Salleb-Aouissi[*], Maggie Chow[‡], Haimonti Dutta[*], Philip Gross[*≀], Bert Huang[*], Steve Ierome[‡], Delfina Isaac[‡], Arthur Kressner[‡], Rebecca J. Passonneau[*], Axinia Radeva[*], Leon Wu[*]

Abstract—Power companies can benefit from the use of knowledge discovery methods and statistical machine learning for preventive maintenance. We introduce a general process for transforming historical electrical grid data into models that aim to predict the risk of failures for components and systems. These models can be used directly by power companies to assist with prioritization of maintenance and repair work. Specialized versions of this process are used to produce 1) feeder failure rankings, 2) cable, joint, terminator and transformer rankings, 3) feeder MTBF (Mean Time Between Failure) estimates and 4) manhole events vulnerability rankings. The process in its most general form can handle diverse, noisy, sources that are historical (static), semi-real-time, or real-time, incorporates state-of-the-art machine learning algorithms for prioritization (supervised ranking or MTBF), and includes an evaluation of results via cross-validation and blind test. Above and beyond the ranked lists and MTBF estimates are business management interfaces that allow the prediction capability to be integrated directly into corporate planning and decision support; such interfaces rely on several important properties of our general modeling approach: that machine learning features are meaningful to domain experts, that the processing of data is transparent, and that prediction results are accurate enough to support sound decision making. We discuss the challenges in working with historical electrical grid data that were not designed for predictive purposes. The "rawness" of these data contrasts with the accuracy of the statistical models that can be obtained from the process; these models are sufficiently accurate to assist in maintaining New York City's electrical grid.

Index Terms—applications of machine learning, electrical grid, smart grid, knowledge discovery, supervised ranking, computational sustainability, reliability

1 Introduction

One of the major findings of the U.S. Department of Energy's "Grid 2030" strategy document [1] is that "America's electric system, 'the supreme engineering achievement of the 20th century' is aging, inefficient, congested, incapable of meeting the future energy needs [...]." Reliability will be a key issue as electrical grids transform throughout the next several decades, and grid maintenance will become even more critical than it is currently. A 2007 survey by the NERC [2] states that "aging infrastructure and limited new construction" is the largest challenge to electrical grid reliability out of all challenges considered by the survey (also see [3]). The smart grid will bring operations and maintenance more online – moving the industry from reactive to proactive operations. Power companies keep historical data records regarding equipment and past failures, but those records are generally not being used to their full extent for predicting grid reliability and assisting with maintenance. This is starting to change. This paper presents steps towards proactive maintenance programs for electrical grid reliability based on the application of knowledge discovery and machine learning methods.

Most power grids in U.S. cities have been built gradually over the last 120 years. This means that the electrical equipment (transformers, cables, joints, terminators, and associated switches, network protectors, relays, etc.) vary in age; for instance, at least 5% of the low voltage cables in Manhattan were installed before 1930, and a few of the original high voltage distribution feeder sections installed during the Thomas Edison era are still in active use in New York City. In NYC there are over 94,000 miles of high voltage underground distribution cable, enough to wrap around the earth three and a half times. Florida Power and Light Company (FPL) has 24,000 miles of underground cable[1] and many other utilities manage similarly large underground electric systems. Maintaining a large grid that is a mix of new and old components is more difficult than managing a new grid (for instance, as is being laid in some parts of China). The U.S. grid is generally older than many European grids that were replaced after WWII, and older than grids in

• * *Center for Computational Learning Systems, Columbia University, 475 Riverside Drive MC 7717 (850 Interchurch Center), New York, NY 10115, U.S.A*

† *MIT Sloan School of Management, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge MA 02139, U.S.A. E-mail: rudin@mit.edu*

‡ *Consolidated Edison Company of New York, 4 Irving Place, New York, NY, 10003, U.S.A.*

≀ *Now at: Google, Inc., 76 Ninth Avenue, New York, NY 10011*

Manuscript received ?; revised ?.

1. http://www.fpl.com/faqs/underground.shtml

Digital Object Indentifier 10.1109/TPAMI.2011.108

places where infrastructure must be continually replenished due to natural disasters (for instance, Japan has earthquakes that force power systems to be replenished).

The smart grid will not be implemented overnight; to create the smart grid of the future, we must work with the electrical grid that is there now. For instance, according to the Brattle Group [4], the cost of updating the grid by 2030 could be as much as $1.5 trillion. The major components of the smart grid will (for an extended period) be the same as the major components of the current grid, and new intelligent meters must work with the existing equipment. Converting to a smart grid can be compared to replacing worn parts of an airplane while it is in the air. As grid parts are replaced gradually and as smart components are added, the old components, including cables, switches, sensors, etc., will still need to be maintained. Further, the state of the old components should inform the priorities for the addition of new smart switches and sensors.

The key to making smart grid components effective is to analyze where upgrades would be most useful, given the current system. Consider the analogy to human patients in the medical profession, a discipline for which many of the machine learning algorithms and techniques used here for the smart grid were originally developed and tested. While each patient (a feeder, transformer, manhole, or joint) is made up of the same kinds of components, they wear and age differently, with variable historic stresses and hereditary factors (analogous to different vintages, loads, manufacturers) so that each patient must be treated as a unique individual. Nonetheless individuals group into families, neighborhoods, and populations (analogous to networks, boroughs) with relatively similar properties. The smart grid must be built upon a foundation of helping the equipment (patients) improve their health, so that the networks (neighborhoods) improve their life expectancy, and the population (boroughs) lives more sustainably.

In the late 1990's, NYC's power company, Con Edison, hypothesized that historical power grid data records could be used to predict, and thus prevent, grid failures and possible associated blackouts, fires and explosions. A collaboration was formed with Columbia University, beginning in 2004, in order to extensively test this hypothesis. This paper discusses the tools being developed through this collaboration for predicting different types of electrical grid failures. The tools were created for the NYC electrical grid; however, the technology is general and is transferrable to electrical grids across the world.

In this work, we present new methodologies for maintaining the smart grid, in the form of a general process for failure prediction that can be specialized for individual applications. Important steps in the process include data processing (cleaning, pattern matching, statistics, integration), formation of a database, machine learning (time aggregation, formation of features and labels, ranking methods), and evaluation (blind tests, visualization). Specialized versions of the process have been developed for: 1) feeder failure ranking for distribution feeders, 2) cable section, joint, terminator and transformer ranking for distribution feeders, 3) feeder MTBF (Mean Time Between Failure) estimates for distribution feeders, and 4) manhole vulnerability ranking. Each specialized process was designed to handle data with particular characteristics. In its most general form, the process can handle diverse, noisy, sources that are historical (static), semi-real-time, or real-time; the process incorporates state of the art machine learning algorithms for prioritization (supervised ranking or MTBF), and includes an evaluation of results via cross-validation on past data, and by blind evaluation. The blind evaluation is performed on data generated as events unfold, giving a true barrier to information in the future. The data used by the machine learning algorithms include past events (failures, replacements, repairs, tests, loading, power quality events, etc.) and asset features (type of equipment, environmental conditions, manufacturer, specifications, components connected to it, borough and network where it is installed, date of installation, etc.).

Beyond the ranked lists and MTBF estimates, we have designed graphical user interfaces that can be used by managers and engineers for planning and decision support. Successful NYC grid decision support applications based on our models are used to assist with prioritizing repairs, prioritizing inspections, correcting of overtreatment, generating plans for equipment replacement, and prioritizing protective actions for the electrical distribution system. How useful these interfaces are depends on how accurate the underlying predictive models are, and also on the interpretation of model results. It is an important property of our general approach that machine learning features are meaningful to domain experts, in that the data processing and the way causal factors are designed is transparent. The transparent use of data serves several purposes: it allows domain experts to troubleshoot the model or suggest extensions, it allows users to find the factors underlying the root causes of failures, and it allows managers to understand, and thus trust, the (non-black-box) model in order to make decisions.

We implicitly assume that data for the modeling tasks will have similar characteristics when collected by any power company. This assumption is broadly sound but there can be exceptions; for instance feeders will have similar patterns of failure across cities, and data are probably collected in a similar way across many cities. However, the levels of noise within the data and the particular conditions of the city (maintenance history, maintenance policies, network topologies, weather, etc.) are specific to the city and to the methods by which data are collected and stored by the power company.

Our goals for this paper are to demonstrate that data collected by electrical utilities can be used to create statistical models for proactive maintenance programs, to show how this can be accomplished through knowledge discovery and machine learning, and to encourage com-

panies across the world to reconsider the way data are being collected and stored in order to be most effective for prediction and decision-support applications.

In Section 2, we discuss the electrical grid maintenance tasks. Section 3 contains the general process by which data can be used to accomplish these tasks. In Section 4 we discuss the specific machine learning methods used for the knowledge discovery process. Section 5 presents the specialized versions of the general process for the four prediction tasks. In Section 6 we give sample results for the NYC power grid. Section 7 discusses the prototype tools for management we have developed in order to make the results useable, and to assist in knowledge discovery. Section 8 presents related work. Section 9 presents lessons learned from the implementation of these systems on the NYC grid.

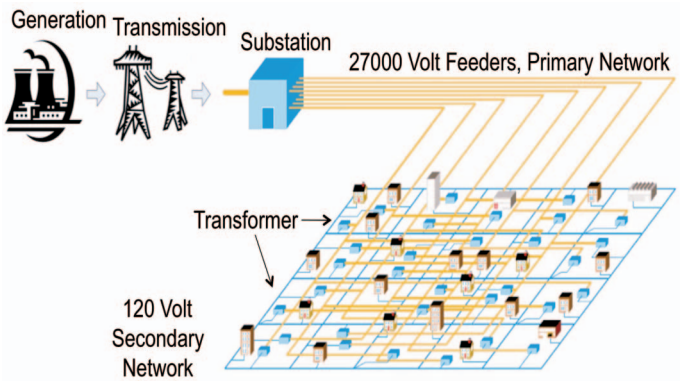

Fig. 1. Typical Electrical Infrastructure in Cities. Source: Con Edison.

2 Proactive Maintenance Tasks

Power companies are beginning to switch from reactive maintenance plans (fix when something goes wrong) to proactive maintenance plans (fix potential problems before they happen). There are advantages to this: reactive plans, which allow failures to happen, can lead to dangerous situations, for instance fires and cascading failures, and costly emergency repairs. However, it is not a simple task to determine where limited resources should be allocated in order to most effectively repair potentially vulnerable components.

In large power systems, electricity flows from source to consumer through transmission lines to substations, then to primary feeder cables ("*feeders*"), and associated cable sections, joints, and terminators, through transformers, and to the secondary (low-voltage) electrical distribution grid (see Figure 1). There are two types of feeders, "*distribution feeders*" and "*transmission feeders*." Our work has mainly focused on distribution feeders (the term "feeder" will indicate distribution feeders), which are large medium to high-voltage cables that form a tree-like structure, with transformers at the leaves. In some cities, these transformers serve buildings or a few customers, and a feeder failure leads to service interruptions for all of these downstream customers. In other cities, including NYC, the secondary cables form a mesh or grid-like structure that is fed by redundant high-voltage feeders, with the goal of continuing service, even if one or more feeders fail. There can be possible weaknesses in any of these components: a feeder may go out of service, the cables, joints and terminators can fail, transformers can fail, and insulation breakdown of cables in the secondary electrical grid can cause failures. In what follows, we discuss how data-driven preemptive maintenance policies can assist with preventing these failures.

2.1 Feeder Rankings

Primary distribution feeder cables are large cables; in NYC they operate at 13,600 or 27,000 volts. They generally lie along main streets or avenues and distribute power from substations to the secondary grid.

A feeder may experience an outage due to a fault somewhere along the feeder, or due to deliberate de-energizing (so maintenance can be performed). If one component, such as a feeder, fails or is taken out of service, this failure is called a "first contingency," and if two components in the same network fail, it is called a "second contingency," and so forth. Loss of a small number of feeders generally does not result in any interruption in customers' electricity service, due to extensive redundancy in the system. (For instance, Con Edison's underground system is designed to operate under second contingency.) However, once one or more feeders in a network are out of service, the remaining feeders and their associated transformers have to pick up the load of the feeders that were disconnected. This added load elevates the risk of failure for the remaining feeders and transformers, and past a certain point, the network will experience a cascading failure, where the remaining components are unable to carry the network's load, and the entire network must be shut down until the system can be repaired.

Each feeder consists of many cable sections (called "*sections*" in what follows); for instance, the average number of sections per feeder in NYC is approximately 150. Each section runs between two manholes, and has "*joints*" at each end. Sections are often made up of three bundled cables, one for each voltage phase. Joints can attach two single cable sections, or can branch two or more ways. Ultimately feeder sections end at transformers that step down the voltage to 120 or 240 Volts needed for the secondary system. Feeder sections are connected to transformers by "*hammerheads*," which are terminators that are named for their distinctive shape. Feeder failures generally occur at the joints or within a cable section. In this subsection, we discuss the problem of predicting whether a given feeder will have a failure (including failures on any of its subcomponents), and in the following subsection, we discuss the prediction of failures on individual feeder components, specifically on

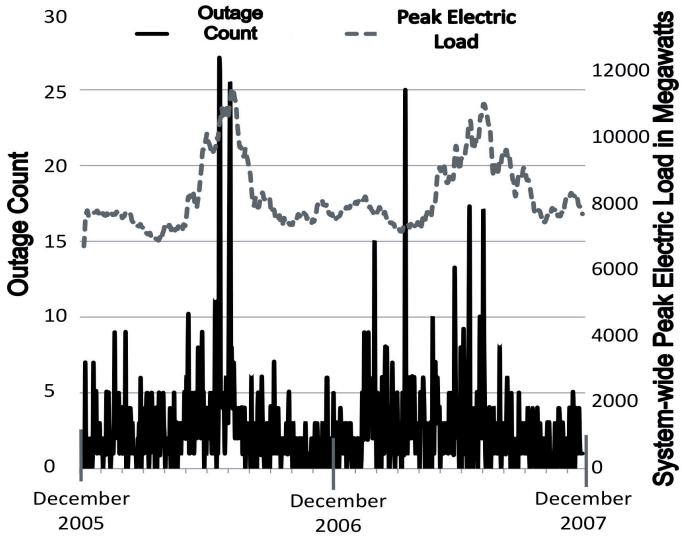

Fig. 2. Number of feeder outages in NYC per day during 2006-2007, lower curve with axis at left, and system-wide peak system load, upper curve at right.

the individual cable sections, joints and hammerheads. We use the results from the individual component failure predictions as input to the feeder failure prediction model.

One kind of joint, the "stop joint," is the source of a disproportionate number of failures. Stop joints connect old "PILC" to modern cables with solid dielectrics. PILC stands for Paper-Insulated Lead-sheathed Cable, an older technology used in most urban centers from 1906 through the 1960's. PILC sections are filled with oil, so stop joints must not only have good electrical connections and insulation (like all joints) but must also cap off the oil to prevent it from leaking. Even though all utilities are aggressively removing lead cable from their systems, it is going to be a long time before the work is complete.[2] For instance, in NYC, the Public Service Commission has mandated that all ∼30,000 remaining PILC sections be replaced by 2020. Note however that some PILC sections have been in operation for a very long time without problems, and it is important to make the best use of the limited maintenance budget by replacing the most unreliable sections first.

As can be seen in Figure 2, a small number of feeder failures occur daily in NYC throughout the year. The rate of failures noticeably increases during warm weather; air conditioning causes electricity usage to increase by roughly 50% during the summer. It is during these times when the system is most at risk.

The feeder failure ranking application, described in Section 5.1, orders feeders from most at-risk to least at-risk. Data for this task include: physical characteristics of the feeder, including characteristics of the underlying components that compose the feeder (e.g., percent of PILC sections); date put into service; records of previous "open autos" (feeder failures), previous power quality events (disturbances), scheduled work, and testing; electrical characteristics, obtained from electric load flow simulations (e.g., how much current a feeder is expected to carry under various network conditions); and dynamic data, from real-time telemetry attached to the feeder. Approximately 300 summary features are computed from the raw data, for example, the total number of open autos per feeder over the period of data collection. For Con Edison, these features are reasonably complete and not too noisy. The feeder failure rank lists are used to provide guidance for Con Edison's contingency analysis and winter/spring replacement programs. In the early spring of each year, a number of feeders are improved by removing PILC sections, changing the topology of the feeders to better balance loading, or to support changing power requirements for new buildings. Loading is light in spring, so feeders can be taken out of service for upgrading with low risk. Prioritizing feeders is important: scheduled replacement of each section costs about $18,000. Feeder failures require even more expensive emergency replacements and also carry a risk of cascading failures.

2. For more details, see the article about replacement of PILC in NYC http://www.epa.gov/waste/partnerships/npep/success/coned.htm

2.2 Cable Sections, Joints, Terminators and Transformers Ranking

In Section 2.1 we discussed the task of predicting whether a failure would happen to any component of a (multi-component) feeder. We now discuss the task of modeling failures on individual feeder components; modeling how individual components fail brings an extra level to the understanding of feeder failure. Features of the components can be more directly related to localized failures and kept in a non-aggregated form; for instance, a feature for the component modeling task might encode that a PILC section was made by Okonite in 1950 whereas a feature for the feeder modeling task might instead be a count of PILC sections greater than 40 years old for the feeder. The component rankings can also be used to support decisions about which components to prioritize after a potentially susceptible feeder is chosen (guided by the results of the feeder ranking task). In that way, if budget constraints prohibit replacement of all the bad components of a feeder, the components that are most likely to fail can be replaced.

For Con Edison, the data used for ranking sections, joints and hammerheads was diverse and fairly noisy, though in much better shape than the data used for the manhole events prediction project we describe next.

2.3 Manhole Ranking

A small number of serious "*manhole events*" occur each year in many cities, including fires and explosions. These events are usually caused by insulation breakdown of the low-voltage cable in the secondary network. Since the insulation can break down over a long period of

MORINO (SPLICER) CLAIMS CONDITION YELLOW F/O 411 W.95 ST.
ALSO————(LOW VOLTAGE TO PARKING GARAGE_———JEC
01/26/00 08:57 MDE.VETHI DISPATCHED BY 71122 01/26/00 09:21
MDE.VETHI ARRIVED BY 23349
01/26/00 11:30 VETHI REPORTS: FOUND COVER ON NOT SMOKING..
SB-110623 F/O 413 W.95 ST..1 AC LEG COPPERED..CUT CLEARED
AND REJOINED....MADE REPAIRS TO DC CRABS...ALL B/O CLEARED
CO = 0PPM —> SB-110623 F/O 413 W.95 ST
01/26/00 11:34 MDE.VETHI COMPLETE BY 23349
**********ELIN REPORT MADE OUT**********MC

Fig. 3. Excerpt from Sample Smoking Manhole (SMH) Trouble Ticket

time, it is reasonable to try to predict future serious events from the characteristics of past serious and non-serious events. We consider events within two somewhat simplified categories: serious events (fires, explosions, serious smoking manholes) and potential precursor events (burnouts, flickering lights, etc). Potential precursor events can be indicators of an area-wide network problem, or they can indicate that there is a local problem affecting only 1-2 manholes.

Many power companies keep records of all past events in the form of trouble tickets, which are the shorthand notes taken by dispatchers. An example ticket for an NYC smoking manhole event appears in Figure 3. Any prediction algorithm must consider how to effectively process these tickets.

2.4 MTBF (Mean time between failures) Modeling

A common and historical metric for reliability performance is mean time between failures (MTBF) for components or systems that can be repaired, and mean time to failure (MTTF) for components that cannot.[3] Once MTBF or MTTF is estimated, a cost versus benefit analysis can be performed, and replacement policies, inspection policies, and reliability improvement programs can be planned. Feeders are made up of multiple components that can fail, and these components can be replaced separately, so MTBF (rather than MTTF) is applicable for feeder failures. When an individual joint (or other component of a feeder) fails it is then replaced with a new one, so MTTF is applicable instead for individual component failures.

In general the failure rate of a component or a composite system like a feeder will have a varying MTBF over its lifetime. A system that is new or has just had maintenance may have early failures, known as "infant mortality." Then, systems settle down into their mid-life with a lower failure rate, and finally the failure rate increases at the end of their lifetimes. (See Figure 4.) PILC can have very long lifetimes and it is hard to determine an end of life signature for them. Transformers do show aging with an increase in failure rate.

3. See Wikipedia's MTBF page:
http://en.wikipedia.org/wiki/Mean_time_between_failures

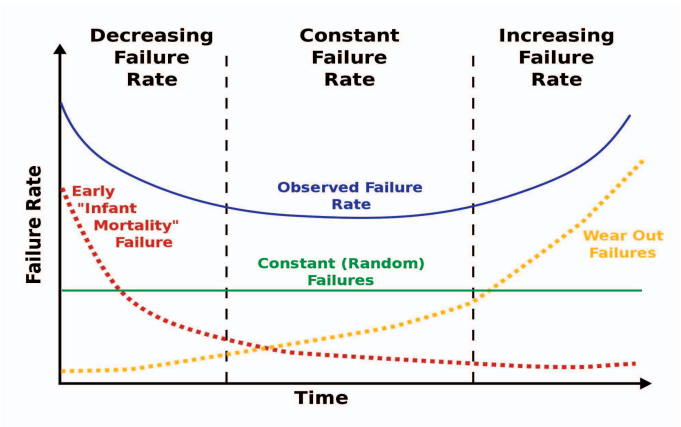

Fig. 4. Bathtub curve. Source Wikipedia: http://en.wikipedia.org/wiki/Bathtub_curve

3 A Process for Failure Prediction in Power Grids

Our general goal is "knowledge discovery," that is, finding information in data that is implicit, novel, and potentially extremely useful [5]. Harding et al. [6] provide an overview of knowledge discovery in manufacturing. The general CRISP-DM framework [7] captures the data processing for (potentially) extremely raw data, however the traditional knowledge discovery in databases (KDD) outline [8] does not encompass this. The general process presented here can be considered a special case of CRISP-DM, but it is outside the realm of KDD.

The general knowledge discovery process for power grid data is shown in Figure 5. The data can be structured text or categorical data, numerical data, or unstructured text documents. The data are first cleaned and integrated into a single database that can be accurately queried. Then one or more machine learning problems are formulated over an appropriate timescale. Ideally, the features used in the machine learning models are meaningful to the domain experts. The parameters in the machine learning algorithm are tuned or tested by cross-validation, and evaluated for prediction accuracy by blind prediction tests on data that are not in the database. Domain experts also evaluate the model using business management tools and suggest improvements (usually in the initial handling and cleaning of data).

The data processing/cleaning is the key piece that ensures the integrity of the resulting model. This view agrees with that of Hsu et al. [9], who state that "... the often neglected pre-processing and post-processing steps in knowledge discovery are the most critical elements in determining the success of a real-life data mining application." Data cleaning issues have been extensively discussed in the literature, for instance in e-commerce [10]. Often, the application of machine learning techniques directly (without the data cleaning step) does not lead to useful or meaningful models. In electrical utility applications, these data can be extremely raw: data can

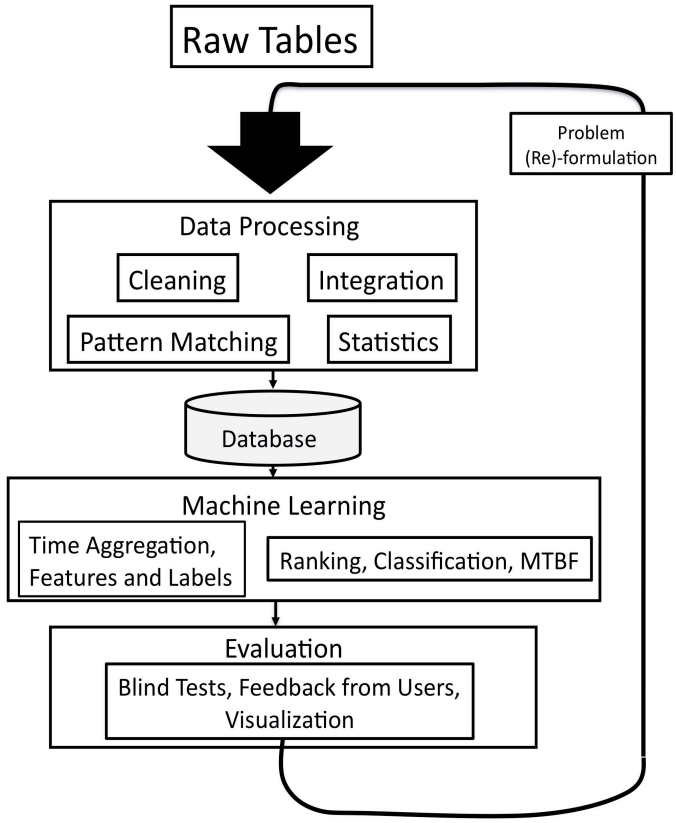

Fig. 5. Process Diagram

come from diverse sources throughout the company, with different schemes for recording times for events or identities of components, they may be incomplete or extremely noisy, or they may contain large numbers of free-text documents (for example, trouble tickets). The data processing step fully defines the interpretation of the data that will be used by the machine learning model. This processing turns historical data from diverse sources into useable features and labels for learning. Data cleaning can include many steps such as pattern matching (for instance, finding regular expressions in structured or unstructured data), information extraction, text normalization, using overlapping data to find inconsistencies, and inferring related or duplicated records. Statistics can be used to assess whether data are missing, and for sanity checks on inferential joins.

An *inferential join* is the process by which multiple raw data tables are united into one database. Inferential joins are a key piece of data processing. An example to illustrate the logic behind using basic pattern matching and statistics for inferential joining is the uniting of the main cable records to the raw manhole location data for the manhole event process in NYC, to determine which cables enter into which manholes. Main cables connect two manholes (as opposed to service or streetlight cables that enter only one manhole). The cable data comes from Con Edison's accounting department, which is different from the source of the manhole location data. A raw join of these two tables, based on a unique manhole identifier that is the union of three fields – manhole type, number, and local 3-block code – provided a match to only about half of the cable records. We then made a first round of corrections to the data, where we unified the spelling of the manhole identifiers within both tables, and found matches to neighboring 3-block codes (the neighboring 3-block code is often mistakenly entered for manholes on a border of the 3 blocks). The next round of corrections used the fact that main cables have limited length: if only one of the two ends of the cable was uniquely matched to a manhole, with several possible manholes for the other end, then the closest of these manholes was selected (the shortest possible cable length). This processing gave a match to about three quarters of the cable records. A histogram of the cable length then indicated that about 5% of these joined records represented cables that were too long to be real. Those cables were used to troubleshoot the join again. Statistics can generally assist in finding pockets of data that are not joined properly to other relevant data.

Data can be either: static (representing the topology of the network, such as number of cables, connectivity), semi-dynamic (e.g., only changes when a section is removed or replaced, or when a feeder is split into two), or dynamic (real-time, with timestamps). The dynamic data can be measured electronically (e.g., feeder loading measurements), or it can be measured as failures occur (e.g., trouble tickets). For the semi-dynamic and dynamic data, a timescale of aggregation needs to be chosen for the features and labels for machine learning.

For all four applications, machine learning models are formed, trained, and cross-validated on past data, and evaluated via "blind test" on more recent data, discussed further in Section 4.

For ranking algorithms, the evaluation measure is usually a statistic of a ranked list (a rank statistic), and ranked lists are visualized as ROC (Receiver Operator Characteristic) curves. Evaluation measures include:

- Percent of successes in the top k%: the percent of components that failed within the top k% of the ranked list (similar to "precision" in information retrieval).
- AUC or weighted AUC: Area under the ROC curve [11], or Wilcoxon Mann Whitney U statistic, as formulated in Section 4 below. The AUC is related to the number of times a failure is ranked below a non-failure in the list. Weighted AUC metrics (for instance, as used the P-Norm Push algorithm [12] derived in Section 4) are more useful when the top of the list is the most important.

For MTBF/MTTF estimation, the sum of squared differences between estimated MTBF/MTTF and true MTBF/MTTF is the evaluation measure.

The evaluation stage often produces changes to the initial processing. These corrections are especially important for ranking problems. In ranking problems where the top of the list is often the most relevant, there

is a possibility that top of the list will be populated completely by outliers that are caused by incorrect or incomplete data processing, and thus the list is essentially useless. This happens particularly when the inferential joins are noisy; if a feeder is incorrectly linked to a few extra failure events, it will seem as if this feeder is particularly vulnerable. It is possible to troubleshoot this kind of outlier by performing case studies of the components on the top of the ranked lists.

4 Machine Learning Methods: Ranking for Rare Event Prediction

The subfield of ranking in machine learning has expanded rapidly over the past few years as the information retrieval (IR) community has started developing and using these methods extensively (see the LETOR website[4] and references therein). Ranking algorithms can be used for applications beyond information retrieval; our interest is in developing and applying ranking algorithms to rank electrical grid components according to the probability of failure. In IR, the goal is to rank a set of documents in order of relevance to a given query. For both electrical component ranking and IR, the top of the list is considered to be the most important.

The ranking problems considered here fall under the general category of supervised learning problems, and specifically supervised bipartite ranking. In supervised bipartite ranking tasks, the goal is to rank a set of randomly drawn examples (the "test set") according to the probability of possessing a particular attribute. To do this, we are given a "training set" that consists of examples with labels:

$$\{(x_i, y_i)\}_{i=1}^{m}, \;\; x_i \in \mathcal{X}, \; y_i \in \{-1, +1\}.$$

In this case, the examples are electrical components, and the label we want to predict is whether a failure will occur within a given time interval. It is assumed that the training and test examples are both drawn randomly from the same unknown distribution. The examples are characterized by features:

$$\{h_j\}_{j=1}^{n}, \;\; h_j : \mathcal{X} \to \mathcal{R}.$$

The features should encode all information that is relevant for predicting the vulnerability of the components, for instance, characteristics of past performance, equipment manufacturer, and type of equipment. To demonstrate, we can have: $h_1(x)$ = the age of component x, $h_2(x)$ = the number of past failures involving component x, $h_3(x)$ = 1 if x was made by a particular manufacturer. These features can be either correlated or uncorrelated with failure prediction; the machine learning algorithm will use the training set to choose which features to use and determine the importance of each feature for predicting future failures.

4. http://research.microsoft.com/en-us/um/beijing/projects/letor/paper.aspx

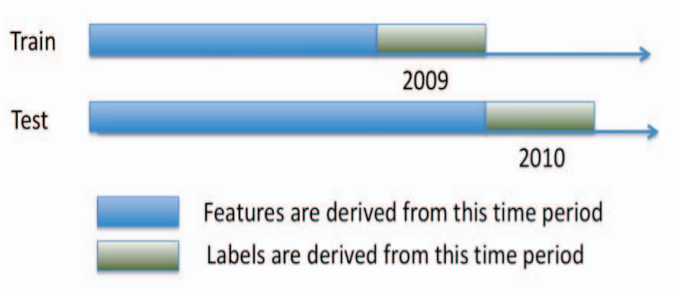

Fig. 6. Sample timeline for rare event prediction

Failure prediction is performed in a rare event prediction framework, meaning the goal is to predict events within a given "prediction interval" using data prior to that interval. There is a separate prediction interval for training and testing. The choice of prediction intervals determines the labels y for the machine learning problem and the features h_j. Specifically, for training, y_i is +1 if component i failed during the training prediction interval and -1 otherwise. The features are derived from the time period prior to the prediction interval. For instance, as shown in Figure 6, if the goal is to rank components for vulnerability with respect to 2010, the model is trained on features derived from prior to 2009 and labels derived from 2009. The features for testing are derived from pre-2010 data. The choice of the prediction interval's length is application dependent; if the interval is too small, there may be no way to accurately characterize failures. If the length is too large, the predictions may be too coarse to be useful. For manhole event prediction in NYC, this time period was chosen to be one year, and time aggregation was performed using the method of Figure 6 for manhole event prediction. A more elaborate time aggregation scheme is discussed in Section 5.1 for feeder failure ranking, where "time shifted" features were used.

The ranking algorithm uses the training set to construct a scoring function, which is a linear combination of the features:

$$f_{\boldsymbol{\lambda}}(x) = \sum_{j=1}^{n} \lambda_j h_j(x),$$

and the examples are rank-ordered by their scores. The ranking algorithm constructs f by minimizing, with respect to the vector of coefficients $\boldsymbol{\lambda} := [\lambda_1, \ldots, \lambda_n]$, a quality measure (a statistic) of the ranked list, denoted $R(f_{\boldsymbol{\lambda}})$. The procedure for optimizing $R(f_{\boldsymbol{\lambda}})$ is "empirical risk minimization" where the statistic is optimized on the training set, and the hope is that the solution generalizes to the full unknown distribution. Particularly, it is hoped that the scoring function will rank the test examples accurately, so that the positive examples are on the top of the list. Probabilistic generalization bounds are used to theoretically justify this type of approach (e.g., [13, 14]).

A common quality measure in supervised ranking is the probability that a new pair of randomly chosen examples is misranked (see [14]), which should be minimized:

$$\begin{aligned}&\mathbf{P}_D\{\text{misrank}(f_{\boldsymbol{\lambda}})\}\\ &\quad:= \quad \mathbf{P}_{\mathcal{D}}\{f_{\boldsymbol{\lambda}}(x_+) \le f_{\boldsymbol{\lambda}}(x_-) \mid y_+ = 1, y_- = -1\}. \quad (1)\end{aligned}$$

The notation $\mathbf{P}_{\mathcal{D}}$ indicates the probability with respect to a random draw of (x_+, y_+) and (x_-, y_-) from distribution $\mathcal{D}$ on $\mathcal{X}\times\{-1, +1\}$. The empirical risk corresponding to (1) is the number of misranked pairs in the training set:

$$\begin{aligned}R(f_{\boldsymbol{\lambda}}) \quad &:= \quad \sum_{\{k:y_k=-1\}} \sum_{\{i:y_i=1\}} \mathbf{1}_{[f_{\boldsymbol{\lambda}}(x_i)\le f_{\boldsymbol{\lambda}}(x_k)]}\\ &= \quad \#\text{misranks}(f_{\boldsymbol{\lambda}}). \quad (2)\end{aligned}$$

The pairwise misranking error is directly related to the (negative of the) area under the ROC curve; the only difference is that ties are counted as misranks in (2). Thus, a natural ranking algorithm is to choose a minimizer of $R(f_{\boldsymbol{\lambda}})$ with respect to $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda}^* \in \text{argmin}_{\boldsymbol{\lambda}\in R^n} R(f_{\boldsymbol{\lambda}})$$

and to rank the components in the test set in descending order of $f_{\boldsymbol{\lambda}^*}(x) := \sum_j \lambda_j^* h_j(x)$.

There are three shortcomings to this algorithm: first, it is computationally hard to minimize $R(f_{\boldsymbol{\lambda}})$ directly. Second, the misranking error $R(f_{\boldsymbol{\lambda}})$ considers all misranks equally, in the sense that misranks at the top of the list are counted equally with misranks towards the bottom, even though in failure prediction problems it is clear that misranks at the top of the list should be considered more important. A third shortcoming is the lack of regularization usually imposed to enable generalization (prediction ability) in high dimensions. A remedy for all of these problems is to use special cases of the following ranking objective that do not fall into any of the traps listed above:

$$R_{\ell g}(f_{\boldsymbol{\lambda}}) := \sum_{\{k:y_k=-1\}} g\left(\sum_{\{i:y_i=1\}} \ell\Big(f_{\boldsymbol{\lambda}}(x_i) - f_{\boldsymbol{\lambda}}(x_k)\Big)\right) + C\|\boldsymbol{\lambda}\|^2, \quad (3)$$

where g is called the *price* function and ℓ is called the *loss* function. $R(f_{\boldsymbol{\lambda}})$ given in (2) is a special case of $R_{\ell g}(f_{\boldsymbol{\lambda}})$ with $\ell(z) = \mathbf{1}_{z\le 0}$ and $g(z) = z$. The objective is convex in $\boldsymbol{\lambda}$ when the exponential loss is used $\ell(z) = e^{-z}$ [14], or the SVM (support vector machine) hinge loss $\ell(z) = (1-z)_+$ [15]; several other convex loss functions are also commonly used. The norm used in the regularization term is generally either a norm in a Reproducing Kernel Hilbert space (for SVMs), which in the simplest case is the ℓ_2 norm $\|\boldsymbol{\lambda}\|_2^2 = \sum_j \boldsymbol{\lambda}_j^2$, or an ℓ_1 norm $\|\boldsymbol{\lambda}\|_1 = \sum_j |\boldsymbol{\lambda}_j|$. The constant C can be set by cross-validation.

Special cases of the objective (3) are: SVM Rank [15] which uses the hinge loss, $g(z) = z$ as the price function, and Reproducing Kernel Hilbert space regularization; RankBoost [14], which uses the exponential loss and no regularization; and the P-Norm Push [12]. The P-Norm Push uses price function $g(z) = z^p$, which forces the value of the objective to be determined mainly by the highest ranked negative examples when p is large; the power p acts as a soft max. Since most of the value of the objective is determined by the top portion of the list, the algorithm concentrates more on the top. The full P-Norm Push algorithm is:

$$\begin{aligned}&\boldsymbol{\lambda}^* \quad \in \quad \text{arg inf}_{\boldsymbol{\lambda}} R_p(\boldsymbol{\lambda}) \ \text{ where}\\ &R_p(\boldsymbol{\lambda}) := \sum_{\{k:y_k=-1\}} \left(\sum_{\{i:y_i=1\}} \exp\Big(-[f_{\boldsymbol{\lambda}}(x_i) - f_{\boldsymbol{\lambda}}(x_k)]\Big)\right)^p.\end{aligned}$$

Vector $\boldsymbol{\lambda}^*$ is not difficult to compute, for instance by gradient descent. The P-Norm Push is used currently in the manhole event prediction tool. An SVM algorithm with ℓ_2 regularization is used currently in the feeder failure tool.

Algorithms designed via empirical risk minimization are not designed to be able to produce density estimates, that is estimates of $P(y = 1|x)$, though in some cases it is possible, particularly when the loss function is smooth. These algorithms are instead designed specifically to produce an accurate ranking of examples according to these probabilities.

It is important to note that the specific choice of machine learning algorithm is not the major component of success in this domain; rather, the key to success is the data cleaning and processing as discussed in Section 3. If the machine learning features and labels are well constructed, any reasonable algorithm will perform well; the inverse holds too, in that badly constructed features and labels will not yield a useful model regardless of the choice of algorithm.

For our MTBF application, MTBF is estimated indirectly through failure rates; the predicted failure rate is converted to MTBF by taking the reciprocal of the rate. Failure rate is estimated rather than MTBF for numerical reasons: good feeders with no failures have an infinite MTBF. The failure rate is estimated by regression algorithms, for instance SVM-R (support vector machine regression) [16], CART (Classification and Regression Trees) [17], ensemble based techniques such as Random Forests [18], and statistical methods, e.g. Cox Proportional Hazards [19].

5 Specific Processes and Challenges

In this section, we discuss how the general process needs to be adapted in order to handle data processing and machine learning challenges specific to each of our electrical reliability tasks in NYC. Con Edison currently operates the world's largest underground electric system, which delivers up to a current peak record of about 14,000 MW of electricity to over 3 million customers. A customer can be an entire office building or apartment complex in NYC so that up to 15 million people are served with

electricity. Con Edison is unusual among utilities in that it started keeping data records on the manufacturer, age, and maintenance history of components over a century ago, with an increased level of Supervisory Control and Data Acquisition (SCADA) added over the last 15 years. While real-time data are collected from all transformers for loading and power quality information, that is much less than will be needed for a truly smart grid.

We first discuss the challenges of feeder ranking and specifics of the feeder failure ranking process developed for Con Edison (also called "Outage Derived Data Sets - ODDS") in Section 5.1. We discuss the data processing challenges for cables, joints, terminators and transformers in Section 5.2. The manhole event prediction process is discussed in Section 5.3, and the MTBF estimation process is discussed in Section 5.4.

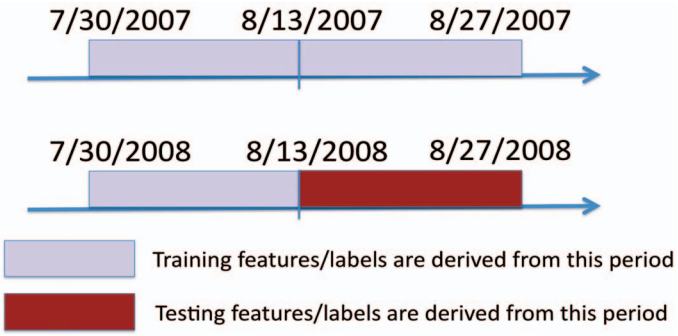

Fig. 7. Example illustrating the training and test time windows in ODDS. The current time is 8/13/2008, and failure data for training was derived from the prediction period of 7/30/2007 - 8/27/2007 and 7/30/2008 - 8/13/2008.

5.1 Feeder Ranking in NYC

Con Edison data regarding the physical composition of feeders are challenging to work with; variations in the database entry and rewiring of components from one feeder to another make it difficult to get a perfect snapshot of the current state of the system. It is even more difficult to get snapshots of past states of the system; the past state needs to be known at the time of each past failure because it is used in training the machine learning algorithm. A typical feeder is composed of over a hundred cable sections, connected by a similar number of joints, and terminating in a few tens of transformers. For a single feeder, these subcomponents are a hodgepodge of types and ages, for example a brand-new cable section may be connected to one that is many decades old; this makes it challenging to "roll-up" the feeder into a set of features for learning. The features we currently use are statistics of the ages, numbers, and types of components within the feeder; for instance, we have considered maxima, averages, and 90th percentiles (robust versions of the maxima).

Dynamic data presents a similar problem to physical data, but here the challenge is aggregation in time instead of space. Telemetry data are collected at rates varying from hundreds of times per second (for power quality data) to only a few measurements per day (weather data). These can be aggregated over time, again using functions such as max or average, using different time windows (as we describe shortly). Some of the time windows are relatively simple (e.g., aggregating over 15 or 45 days), while others take advantage of the system's periodicity, and aggregate over the most recent data plus data from the same time of year in previous years.

One of the challenges of the feeder ranking application is that of imbalanced data, or scarcity of data characterizing the failure class, which causes problems with generalization. Specifically, primary distribution feeders are susceptible to different kinds of failures, and we have very few training examples for each kind, making it difficult to reliably extract statistical regularities or determine the features that affect reliability. For instance, failure can be due to: concurrent or prior outages that stress the feeder and other feeders in the network; aging; power quality events (e.g., voltage spikes); overloads (that have seasonal variation, like summer heat waves); known weak components (e.g., joints connecting PILC to other sections); at-risk topologies (where cascading failures could occur); the stress of "HiPot" (high potential) testing; and de-energizing/re-energizing of feeders that can result in multiple failures within a short time span due to "infant mortality." Other data scarcity problems are caused by the range in MTBF of the feeders; while some feeders are relatively new and last for a long time between failures (for example, more than five years), others can have failures within a few tens of days of each other. In addition, rare seasonal effects (such as particularly high summer temperatures) can affect failure rates of feeders.

We have focused on the most serious failure type for distribution feeders, where the entire feeder is automatically taken offline by emergency substation relays, due to some type of fault being detected by sensors. Our current system for generating data sets attempts to address the challenge of learning with rare positive examples (feeder failures). An actual feeder failure incident is instantaneous, so a snapshot of the system at that moment will have only one failure example. To better balance the number of positive and negative examples in the data, we tried the rare event prediction setup shown in Figure 6, labeling any example that had experienced a failure over some time window as positive. However, the dynamic features for these examples are constructed from the timeframe before the prediction period, and thus do not represent the precise conditions at the time of failure. This was problematic, as the domain experts believed that some of the dynamic data might only have predictive value in the period right before the failure. To solve this problem, we decided to switch to "time-shifted" positive examples, where the positive examples are still created from the past outages within the predic-

tion period, but the dynamic features are derived only from the time period shortly before the failure happened. This allows our model to capture short-term precursors to failures. The features of non-failures (negative examples) are characteristics of the current snapshot of all feeders in the system. Not only does this approach, which we call "ODDS" for Outage Derived Data Sets, capture the dynamic data from right before the failure, it helps to reduce the imbalance between positive and negative examples. Figure 7 shows an example of the periods used to train and test the model.

Another challenge raised by our feeder failure ranking application is pervasive "concept drift," meaning that patterns of failure change fairly rapidly over time, so that a machine learning model generated on data from the past may not be completely representative of future failure patterns. Features can become inactive or change in quality. Causes of this include: repairs being made on components, causing the nature of future failures to change; new equipment having different failure properties than current equipment; and seasonal variation in failure modes (e.g., a greater likelihood of feeder failure in the summer). To address this challenge, ODDS creates a new model every 4 hours on the current dataset. (See also [20, 21, 22].)

An outline of the overall process is shown in Figure 8. A business management application called the Contingency Analysis Program (CAP), discussed in Section 7, uses the machine learning results to highlight areas of risk through graphical displays and map overlays.

As in many real-life applications, our application suffers from the problem of missing data. Techniques such as mean-imputation are used to fill in missing values.

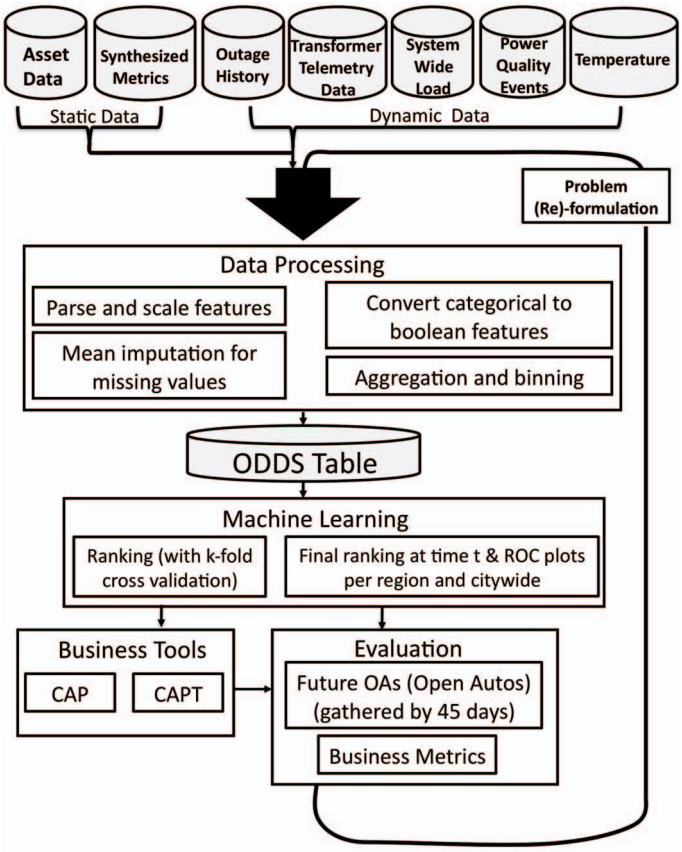

Fig. 8. Process diagram for feeder ranking, using ODDS

5.2 Cables, Joints, Terminators, and Transformers Ranking in NYC

The main challenges to constructing rankings of feeder components overlap somewhat with those faced in constructing rankings for feeders: the use of historical data, and the data imbalance problem.

Ideally, we should be able to construct a consistent and complete set of features for each component and also its connectivity, environmental, and operational contexts at the time of failure. At Con Edison, the cable data used for cable, joint, and terminator rankings resides in the "Vision Mapping" system and are designed to only represent the current layout of cables in the system, and not to provide the layout at particular times in the past. We began to archive cable data starting in 2005 and also relied on other snapshots of cable data that Con Edison made, for example, cable data captured for Con Edison's "Network Reliability Indicator" program that allowed us to go back as far as 2002 configurations.

Generating training data for joints is especially challenging. Joints are the weakest link in feeders with certain heat-sensitive joint types having accelerated failure rates during heat waves. Con Edison keeps a database of feeder component failures called CAJAC. It captures failure data of joints in detail. Con Edison autopsies failed components and the failure reasons they discover are captured in this database. Though the joint failure data are recorded in detail, it is challenging to construct a complete list of the set of installed joints within the grid; the set of installed joints is imputed from the features of the cables being connected. In addition, short lengths of cable, called "inserts," that are sometimes used to make connections in manholes, are not yet captured in the Vision Mapping system, so the number of joints in any manhole can only be estimated in general. Also, the nature of the joint (type of joint, manufacturer, etc.) has had to be inferred from the date of installation. We do this by assuming that the policy in force at the installation date was used for that joint, which allows us to infer the manufacturers and techniques used.

To create the transformer database, several data sources were merged using inferential joins, including data from Con Edison's accounting department, the inspection record database, and the dissolved gas database. Transformer ranking has several challenges. We are working with a transformer population that is actively monitored and aggressively replaced by Con Edison at any sign of impending trouble, meaning that vulnerable transformers that had not failed have been replaced, leading to right censoring (meaning missing

information after a certain time in the life of the transformer). Further, for a transformer that was replaced, it is always a challenge to determine whether a failure would have occurred if the transformer had not been replaced, causing label bias for the machine learning.

As demonstrated for several of the projects discussed here, components that have multiple roles or that act as interfaces between multiple types of components present the challenge of bringing together multiple databases to capture the full context for the component. In order to rank hammerheads, we built a database that joined splice ticket data, cable data, and transformer data, where transformer data itself came from an earlier join of large databases described above.

While working with various data sets involving date-time information, we had to be careful about the meaning of the date and time. In some cases the date entry represents a date when work was done or an event occurred, in other cases, the date is when data was entered into the database. In some instances there was confusion as to whether time was provided in GMT, EST or EDT, leading to some cases where our machine learning systems made perfect predictions, but for the wrong reasons: they learned to detect inevitable outcomes of failures, but where these outcomes apparently predated the outages because of data timing skew.

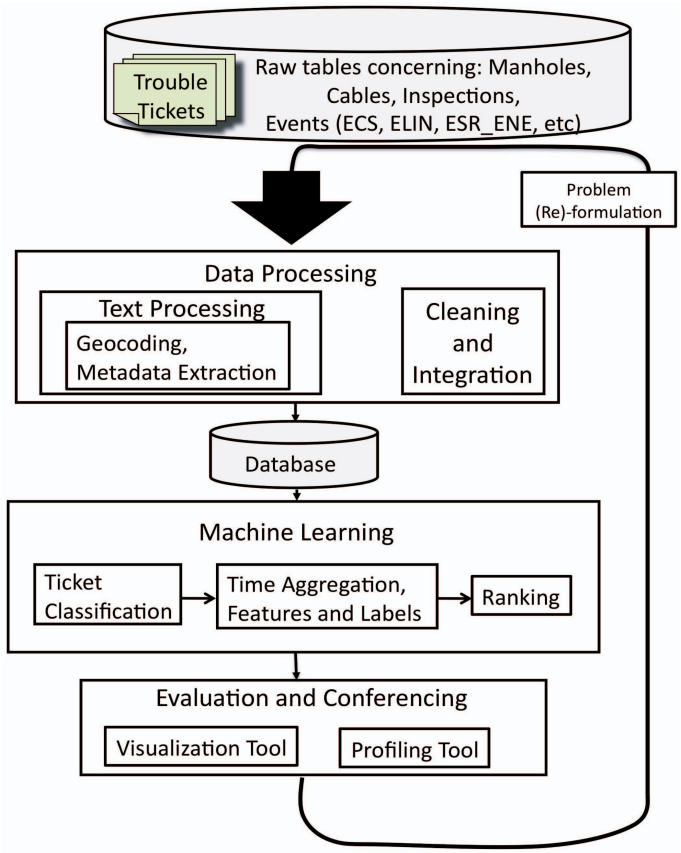

Fig. 9. Process diagram for manhole event ranking

5.3 Manhole Ranking in NYC

One major challenge for manhole event prediction was to determine which of many data sources, and which fields within these sources, to trust; it only made sense to put a lot of effort into cleaning data that had a higher chance of assisting with prediction. The data used for the manhole event prediction process is described in detail in [23], and includes: information about the infrastructure, namely a table of manhole locations and types, and a snapshot of recent cable data from Con Edison's accounting department (type of cable, manholes at either end of cable, installation dates); five years of inspection reports filled out by inspectors; and most importantly, event data. The event data came from several different sources including: ECS (Emergency Control Systems) trouble tickets which included both structured fields and unstructured text, a table of structured data containing additional details about manhole events (called ELIN – ELectrical INcidents), and a table regarding electrical shock and energized equipment events (called ESR_ENE). These data were the input for the manhole event prediction process outlined in Figure 9.

The trouble tickets are unstructured text documents, so a representation of the ticket had to be defined for the learning problem. This representation encodes information about the time, location, and nature (degree of seriousness) of the event. The timestamps on the ticket are directly used, but the location and seriousness must be inferred (and/or learned). The locations of events were inferred using several sources of location information present in the trouble tickets, including a street address (possibly misspelled or abbreviated, e.g., `325 GREENWHICH ST`), structure names typed within the text of the ticket (`S/B 153267`) and structure names sometimes included in the structured fields of three tables (ECS, ELIN, or ESR_ENE). All location information was typed by hand, and these data are very noisy – for instance, the term "service box" was written in at least 38 different ways – and no one source of information is complete. The redundancy in the data was used in order to obtain reliable location information: structure numbers were extracted from the ticket text using information extraction techniques (see Figure 10), then tickets were geocoded to determine the approximate location of the event. If the geocoded address was not within a short distance (200m) of the structure named within the ticket, the information was discarded. The remaining (twice verified) matches were used, so that the ticket was identified correctly with the manholes that were involved in the event.

It was necessary also to determine the seriousness of events; however ECS trouble tickets were not designed to contain a description of the event itself, and there is no structured field to encode the seriousness directly. On the other hand, the tickets do have a "trouble type" field, which is designed to encode the nature of the event (e.g., an underground AC event is "UAC," flickering

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

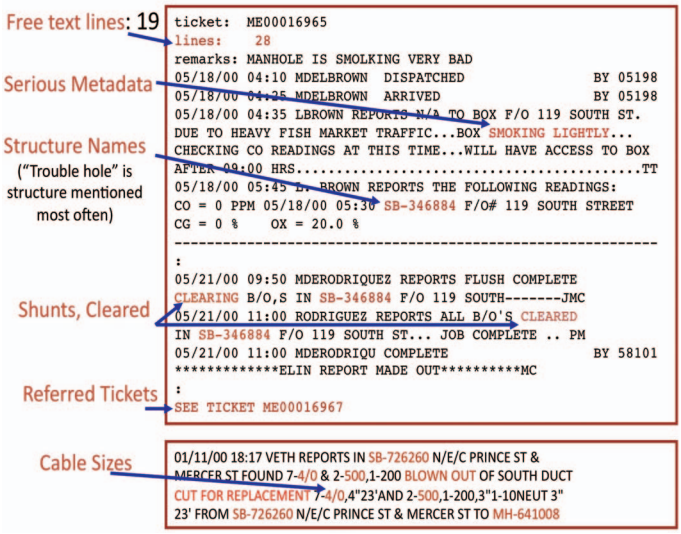

Fig. 10. Ticket processing

lights is "FLT"). Originally, we used the trouble type to characterize the seriousness of the event: the codes "MHX" (manhole explosion), "MHF" (manhole fire), and "SMH" (smoking manhole) were used to identify serious events. However, we later performed a study [24] that showed that the trouble type did not agree with experts' labeling of tickets, and is not a good measure of seriousness. In order to better estimate the seriousness of events, we created a representation of each ticket based on information extracted from the ticket text, including the length of the ticket, the presence of serious metadata (for instance, the term "SMOKING LIGHTLY"), and whether cable sizes appeared in the text (indicating the replacement of a cable). This information extraction was performed semi-automatically using text-processing tools, including the Generalized Architecture for Text Engineering "GATE" [25].

The ticket representation was used to classify the tickets into the categories: serious events, possible precursor events, and non-events. This classification was performed with either a manual, rule-based method or general machine learning clustering methods (k-means clustering). So there are two machine learning steps in the manhole event ranking process: a ticket classification step, and a manhole ranking step.

One challenge faced early on was in choosing the timeframes for the rare event prediction framework. We started originally trying to predict manhole events on a short timescale (on the order of 60 days) based on the domain experts' intuition that such a timescale would yield useful predictions. However, it became clear that manhole events could not easily be predicted over such a short time; for instance if it is known that a manhole event will occur within 60 days after a prior event, it is almost impossible to predict when within those 60 days it will happen. In fact, insulation breakdown, which causes manhole events, can be a slow process, taking place over months or years. A prediction period of one year was chosen for the machine learning ranking task, as illustrated in Figure 6.

The cable data, which is a snapshot at one (recent) point in time, was unified with the other data to construct "static" features and labels for the ranking task. This assumes implicitly that the snapshot approximately represents the number and type of cables over the time period of prediction; this assumption is necessary since the exact state of cables in the manhole at a given time in the past may not be available. However, this assumption is not universally true; for instance it is not true for neutral (non-current carrying, ground) cables at Con Edison, and neutral cable data thus cannot be used for failure prediction, as discussed in [23]. Often, manholes that have had serious events also have had cables replaced, and more neutrals put in; a higher percentage of neutral cables indicate an event in the past, not necessarily an event in the future.

The P-Norm Push (see Section 4) was used as the main ranking algorithm for manhole ranking.

5.4 MTBF Modeling in NYC

It became apparent that to really make our feeder prediction models valuable for proactive maintenance, we had to also produce estimates that allow for an absolute measure of vulnerability, rather than a relative (ranking) measure; many asset replacement decisions are made by assessing how much reliability in days is gained if a particular choice is made (for instance, to replace a PILC section versus another replacement at the same cost).

Machine learning techniques can be used to estimate MTBF. Figure 11 shows the application of one of these techniques [26] to predicting survival times of PILC sections in Queens. This technique can accommodate censored data through inequality constraints in SVM regression. Each row of the table represents one feeder, and each column indicates a time interval (in years). The color in a particular bin gives the count of cable sections within the feeder that are predicted to survive that time interval. That is, each row is a histogram of the predicted MTBF for the feeder's cable sections. The histogram for one feeder (one row) is not necessarily smooth in time. This is because the different cable sections within the feeder were installed at different times (installation not being a smooth function of time), and these installation dates influence the predicted survival interval.

6 Evaluation in NYC

We describe the results of our specific processes as applied to the NYC power grid through the Columbia/Con Edison collaboration. We have generated machine learning models for ranking the reliability of all 1,000+ high voltage (13-27 KV) feeders that form the backbone of the NYC's power distribution system; for each of the ~150,000 cable sections and ~150,000 joints that connect them; for the ~50,000 transformers

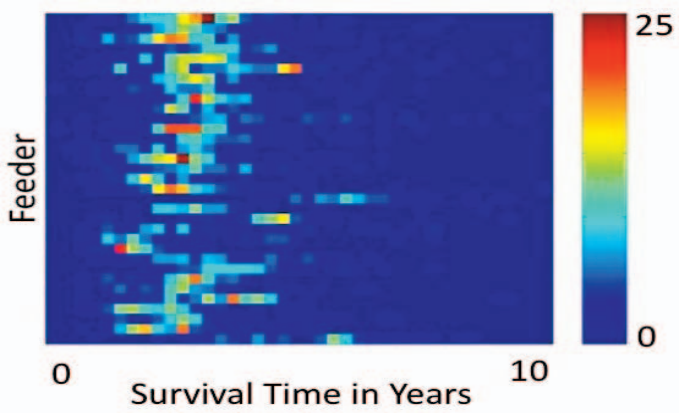

Fig. 11. Predictions from a support vector censored regression algorithm on PILC sections of 33 feeders in Queens.

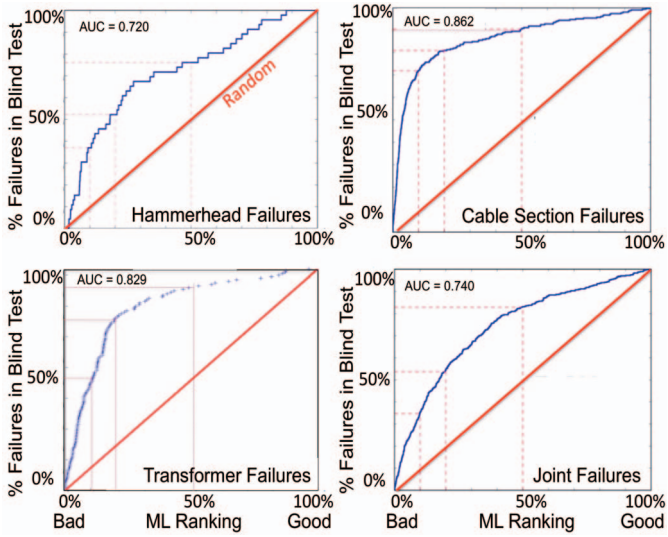

Fig. 12. ROC-like curves from tests of the machine learning ranking of specific components.

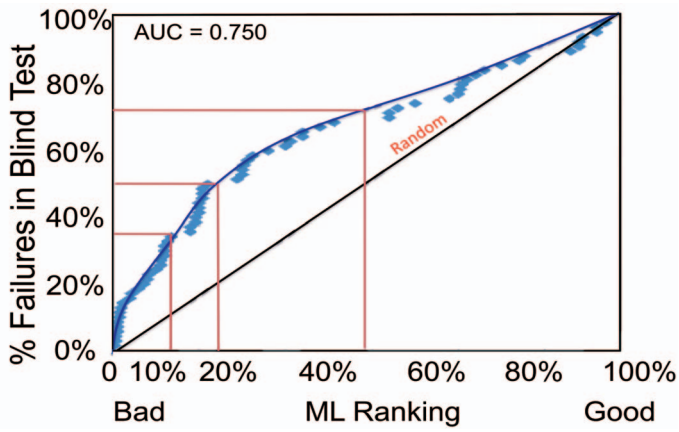

Fig. 13. ROC-like curve for blind test of Crown Heights feeders.

and ∼50,000 terminators that join the transformers to the feeders; and for ∼150,000 secondary structures (manholes and service boxes) through which low voltage (120-240 V) power from the transformers is distributed to buildings in NYC.

Feeder and Component Ranking Evaluation

Our machine learning system for computing feeder susceptibility based on the ODDS system has been online since July 2008, generating a new model every 4 hours. ODDS is driven by the feeds from three dynamic real time systems: load pocket weight,[5] power quality, and outage history. We found that separate training in Brooklyn and Queens, with their 27KV networks, and Manhattan and Bronx, with their 13KV networks, produced better results.

We track the performance of our machine learning models by checking the rank of the failed feeder and the ranks of its components whenever a failure happens. We also compile ROC-like curves showing the components that failed and the feeder that automatically opened its circuit breaker when the failure occurred. These blind tests provide validation that the algorithms are working sufficiently to assist with operations decisions for Con Edison's maintenance programs.

Figure 12 shows the results of a blind test for predicting feeder failures in Crown Heights, Brooklyn, with prediction period from May, 2008 to January, 2009. Figure 13 shows results of various tests on the individual components. At each point (x, y) on the plot, x gives a position on the ranked list, and y is the percent of failures that are ranked at or above x in the list.

We use rank statistics for each network to continually measure performance of the ODDS system. For instance, AUC is reported in Figures 12 and 13. The machine learning system has improved to the point where 60% of failures occur in the 15% of feeders that are ranked as most susceptible to failure. As importantly, fewer than 1% of failures occur on feeders in the best 25% of ODDS feeder susceptibility rankings (Figure 14).

To determine what the most important features are, we create "tornado" diagrams like Figure 15. This figure illustrates the influence of different categories of features under different weather conditions. For each weather condition, the influence of each category (the sum of coefficients λ_j for that category divided by the total sum of coefficients $\sum_j \lambda_j$) is displayed as a horizontal bar. Only the top few categories are shown. For both snowy and hot weather, features describing power quality events have been the most influential predictors of failure according to our model.

The categories of features in Figure 15 are: power quality, which are features that count power quality events

5. Load Pocket Weight (LPW) is a expert-derived measure of trouble in delivering power to the secondary network in localized areas. It is a weighted score of the number of open (not in service) network protector switches, open secondary mains, open fuses, and non-reporting transformers, and other signs of service outage.

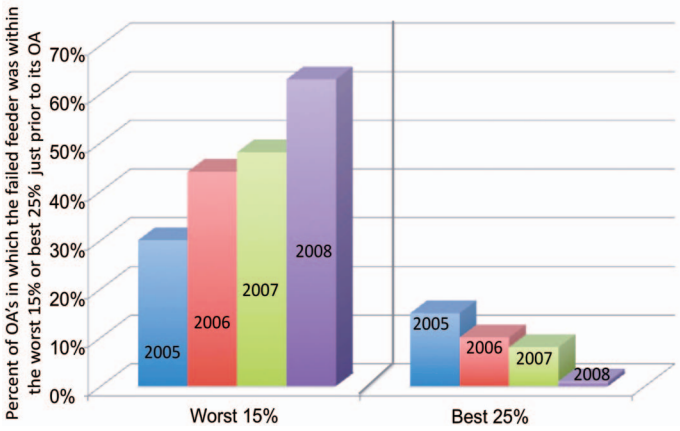

Fig. 14. Percent of feeder outages in which the feeder that failed was within the worst 15% (left) of the ranked list, or best 25% (right), where the predictions being evaluated are those just before the time of failure. The system improved from less than 30% of the failures in the worst 15% in 2005 to greater than 60% in 2008, for example.

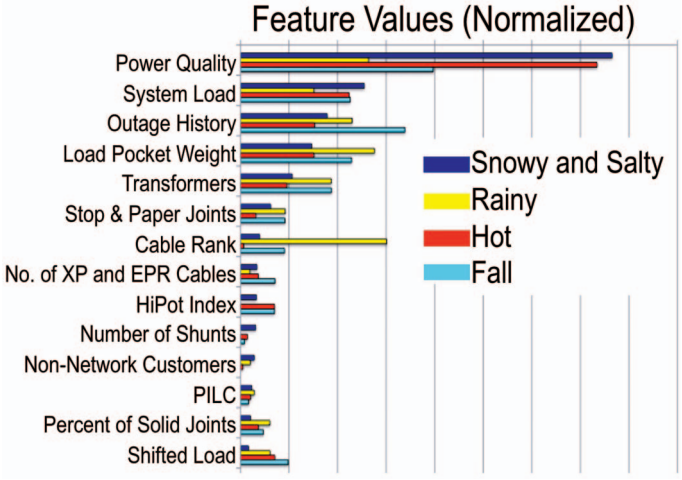

Fig. 15. Influence of different categories of features under different weather conditions. Red: hot weather of August 2010; Blue: snowy January 2011; Yellow: rainy February 2011; Turquoise: typical fall weather in October 2010.

(disturbances) preceding the outage over various time windows; system load in megawatts; outage history, which include features that count and characterize the prior outage history (failure outages, scheduled outages, test failures, immediate failures after re-energization, and urgent planned outage); load pocket weight, which measures the difficulty in delivering power to the end user; transformers, particularly features encoding the types and ages of transformers (e.g., percent of transformers made by a particular manufacturer); stop joints and paper joints, which include features that count joints types, configurations, and age, where these features are associated with joining PILC to other PILC and more modern cable; cable rank, which encodes the results of the cable section ranking model; the count of a specific type of cable (XP and EPR) in various age categories; HiPot index features, which are derived by Con Edison to estimate how vulnerable the feeders are to heat sensitive component failures; number of shunts on the feeder, where these shunts equalize the capacitance and also condition the feeder to power quality events; an indicator for non-network customers, where a non-network customer is a customer that gets electricity from a radial overhead connection to the grid; count of PILC sections along the feeder; percent of joints that are solid joints, which takes into account the fact that joining modern cable is simpler and less failure-prone than joining PILC; shifted load features that characterize how well a feeder transfers load to other feeders if it were to go out of service.

MTBF Modeling Evaluation

We have tracked the improvement in MTBF for each network as preventive maintenance work has been done by Con Edison to improve performance since 2002. To test whether this improvement is significant, we use a nonparametric statistical test, called the logrank test, that compares the survival distributions of two samples. In this case, we wished to determine if the 2009 summer MTBF values are statistically larger than the 2002 summer MTBF values. The performance of the system showed significant improvement, in that there is a less than one in a billion chance that the treatment population in 2009 did not improve over the control population from 2002. In 2009, for example, there were 1468 out of 4590 network-days that were failure free, or one out of every three summer days, but in the 2002 control group, there were only 908 network-days that were failure free, or one out of five summer days, that were failure free. The larger the percentage of network-days that were failure free, the lower the likelihood of multiple outages happening at the same time.

Figure 16 shows MTBF predicted by our model for each underground network in the Con Edison system on both January 1, 2002 (purple) and December 31, 2008 (yellow). The yellow bars are generally larger than the purple bars, indicating an increase in MTBF.

We have performed various studies to predict MTBF of feeders. Figure 17 shows the accuracy of our outage rate predictions for all classes of unplanned outages over a three-year period using a using a support vector machine regression model that predicts feeder MTBF. While the results are quite strong, there are two sources of inaccuracy in this study. First, the study did not model "infant mortality," the increased likelihood of failure after a repaired system is returned to service. This led to an underestimation of failures for the more at-risk feeders (visible particularly in the upper right of the graph). Empirically we observed an increased likelihood of infant mortality for about six weeks following an outage. Second, the study has difficulties handling censored

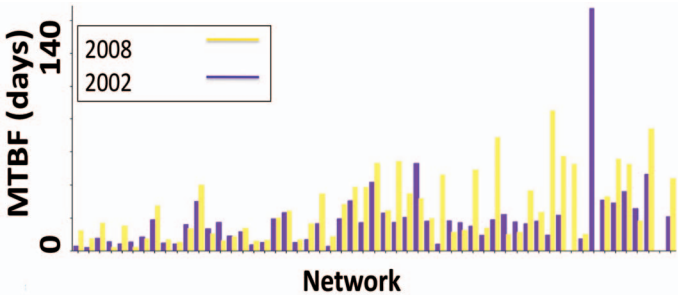

Fig. 16. Linear regression used to determine the Mean Time Between Failures for January 1, 2002 (purple), and December 31, 2008 (yellow) in each underground network in the Con Edison system. Networks are arranged along the horizontal axis from worst (left) to best (right), according to Con Edison's "Network Reliability Index".

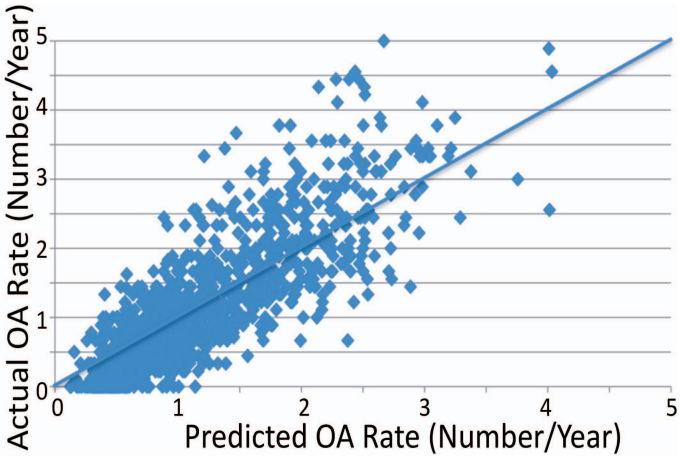

Fig. 17. Scatter plot of SVM predicted outage rate versus actual rate for all classes of unplanned outages. The diagonal line depicts a perfect model.

data. If events are very infrequent, it is not possible for the algorithm to accurately predict their frequency. This right-censoring effect for the low outage rate feeders, due to lack of failures in the three-year observation window, is visible in the lower left of the plot.

Manhole Ranking Evaluation

The most recent evaluation of the manhole rankings was a blind test for predicting 2009 events in the Bronx. The Columbia database has data through 2007, incomplete 2008 data, and no data from 2009 or after. There are 27,212 manholes in the Bronx. The blind test showed:

- the most at-risk 10% (2,721/27,212) of the ranked list contained 44% (8/18) of the manholes that experienced a serious event,
- the most at-risk 20% (5,442/27,212) of the ranked list contained 55% (10/18) of the trouble holes for serious events.

Figure 18 contains the ROC-like curve for the full ranked list.

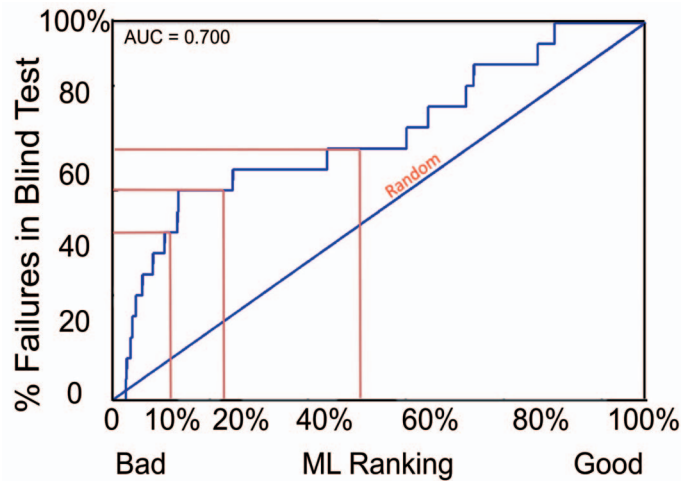

Fig. 18. ROC-like curve for 2009 Bronx blind test of the machine learning ranking for vulnerability of manholes to serious events (fires and explosions).

Before the start of the project, it was not clear whether manhole events could be predicted at all from the secondary data. These results show that indeed manhole events are worthwhile to model for prediction.

7 Management Software

Prototype interfaces were developed jointly with Con Edison in order to make the results useful, and to assist in knowledge discovery.

CAP – Contingency Analysis Program

CAP is a tool designed by Con Edison and used at their main control centers. It brings together information relevant to the outage of a primary feeder cable. When a contingency occurs, Con Edison already has applications in use (integrated into the CAP tool) that preemptively model the network for the possibility of additional feeders failing. These applications determine the failures that could have the worst consequences for the system. Columbia's key contribution to the CAP tool is a feeder susceptibility indicator (described in Section 5.1) that gives the operators a new important piece of information: an indicator of which feeders are most likely to fail next. Operators can use this information to help determine the allocation of effort and resources towards preventing a cascade. The "worst consequences" feeder may not be the same as the "most likely to fail" feeder, so the operator can choose to allocate resources to feeders that are both likely to fail, and for which a failure could lead to more serious consequences. Figure 19 shows a snapshot of the CAP tool interface.

CAPT – Capital Asset Prioritization Tool

CAPT is a prototype application designed by Columbia and Con Edison that offers an advanced mechanism for helping engineers and managers plan upgrades to the feeder systems of NYC. Using a graphic interface,

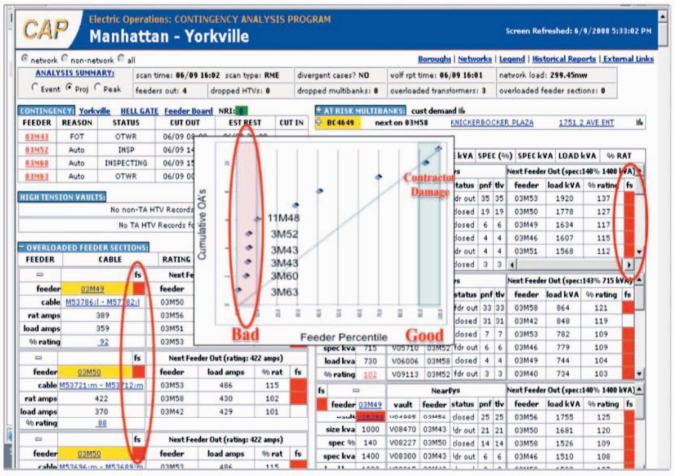

Fig. 19. Screen capture of the Contingency Analysis Program tool during a 4th contingency event in the summer of 2008, with the feeders at most risk of failing next highlighted in red. The feeder ranking at the time of failure is shown in a blow-up ROC-like plot in the center.

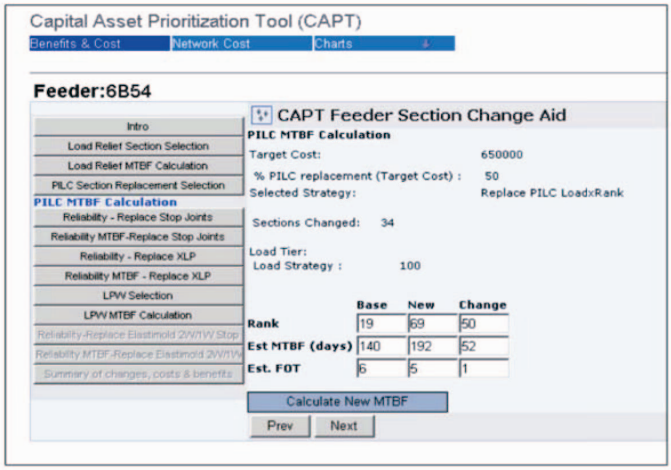

Fig. 20. A screen capture of the Con Edison CAPT evaluation, showing an improvement in MTBF from 140 to 192 days if 34 of the most at-risk PILC sections were to be replace on a feeder in Brooklyn at an estimated cost of $650,000.

shown in Figure 20, users first enter constraints on work they would hypothetically like to do. For instance, users can specify a borough or network, one or more specific feeder sections or type of feeder section, dollar amount to be allocated, etc. CAPT then produces benefit versus cost curves of various replacement strategies with the objective of optimizing "bang for the buck"– the greatest increase in system MTBF for the dollars spent. Such a tool, if proven robust in production tests could become a valuable contributor to capital asset allocations in the future. Typical maintenance plans might attempt to target replacement of at-risk sections, joints, or secondary components. The key components of CAPT include 1) the model (currently an ODDS model along with a regression between SVM scores and observed MTBF) used to estimate MTBF for feeders both before and after any hypothetical changes; 2) the ranked lists for cable sections and joints, based on component rankings, allowing CAPT to recommend good candidates for replacement; and 3) a system that displays, in chart form for the user, tradeoff (Pareto) curves of benefit vs. cost for various replacement strategies (Figure 21).

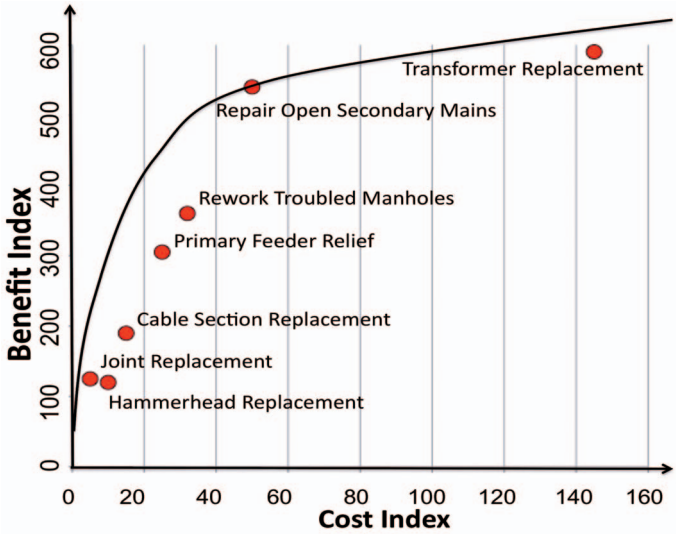

Fig. 21. Example of cost benefit analysis of possible replacement strategies for specific at-risk components analyzed by the machine learning system. The solid line approximates the "efficient frontier" in portfolio management theory.

Manhole Event Structure Profiling Tool and Visualization Tool

We developed several tools that allow a qualitative evaluation of results and methods by secondary system engineers. The most useful of these tools is the "structure profiling tool," (also called the "report card" tool at Con Edison) that produces a full report of raw and processed data concerning any given individual manhole [27]. Before this tool was implemented, an individual case study of a manhole took days and resulted in an incomplete study. This tool gives the reasons why a particular manhole was assigned a particular rank by the model, and allows the vulnerability of a manhole to be roughly estimated at a glance by domain experts. We also developed a visualization tool (discussed in [28]) that uses Google Earth[6] as a backdrop to display the locations of events, manholes and cables. Figure 22 displays two screen shots from the visualization tool.

6. earth.google.com

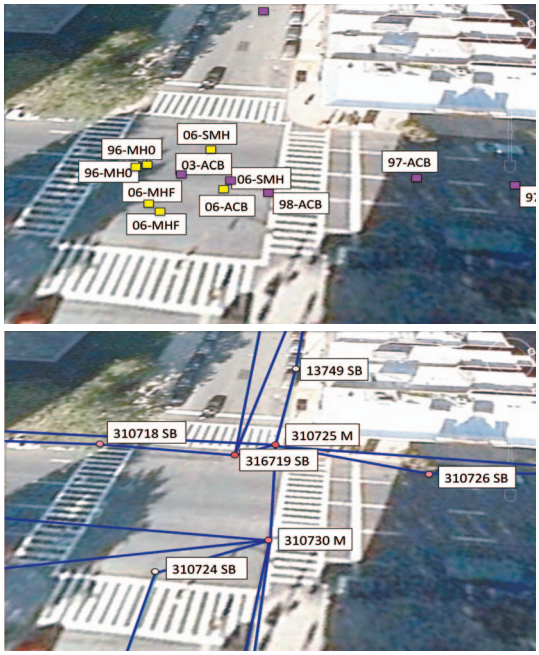

Fig. 22. Images from the manhole events visualization tool, where labels were enlarged for clarity. Top: Geocoded ticket addresses, colored by trouble type. Yellow indicates a serious event type, purple indicates a potential precursor. If the user clicks on a ticket, the full ticket text is displayed. Bottom: Manholes and main cables within the same location, where manholes are colored by predicted vulnerability. Note that a ticket within the top figure does not necessarily correspond to the nearest manhole on the bottom figure.

8 Related Work

Machine learning has been used for applications in *power engineering* since the early days of artificial intelligence, with a surge of recent interest in the last decade. Venues for these works include the 1999 ACAI workshop on Machine Learning Applications to Power Systems (summarized by Hatziargyriou [29]), the proceedings of the yearly International Conference on Intelligent System Applications to Power Systems,[7] and the 2009 Special Session on Machine Learning in Energy Applications at the International Conference on Machine Learning and Applications (ICMLA '09). There are also several books summarizing work on machine learning in power engineering (e.g., [30, 31]). Applications include the prediction of power security breaches, forecasting, power system operation and control, and classification of power system disturbances. The power engineering work bears little similarity to the current work for two reasons. First, much of the power engineering work focuses on specific machine learning techniques, yet for our application the specific machine learning techniques are not the primary reason for success, as discussed earlier. In our applications, the predictive accuracy gained by using a different technique is often small compared to the accuracy gained through other steps in the discovery process, or by formulating the problem differently. The data in power engineering problems is generally assumed to be amenable to learning in its raw form, in contrast with our treatment of the data. The second reason our work is distinct from the power engineering literature is that the machine learning techniques that have been developed by the power engineering community are often "black-box" methods such as neural networks and genetic algorithms (e.g. [32, 33]). Neural networks and genetic algorithms can be viewed as heuristic, nonconvex optimization procedures for objectives that have multiple local minima; the algorithms' output can be extremely sensitive to the initial conditions. Our work uses mainly convex optimization procedures to avoid this problem. Further, "black-box" algorithms do not generally produce interpretable/meaningful solutions (for instance the input-output relationship of a multilayer neural network is not generally interpretable), whereas we use mainly simple linear combinations of features.

We are not aware of any other work that addresses the challenges in mining historical power grid data of the same level of complexity as those discussed here. Our work contrasts with a subset of work in power engineering where data come entirely from Monte Carlo (MC) simulations [34, 35], and the MC simulated failures are predicted using machine learning algorithms. In a sense, our work is closer to data mining challenges in other fields such as e-commerce [10], criminal investigation [36], or medical patient processing [9] that encompass the full discovery process. For instance, it is interesting to contrast our work on manhole events with the study of Cornélusse et al. [37] who used domain experts to label "frequency incidents" at generators, and constructed a machine learning model from the frequency signals and labels. The manhole event prediction task discussed here also used domain experts to label trouble tickets as to whether they represent serious events; however, the level of processing required to clean and represent the tickets, along with the geocoding and information extraction required to pinpoint event locations, coupled with the integration of the ticket labeling machine learning task with the machine learning ranking task makes the latter task a much more substantial undertaking.

9 Lessons Learned

There are several "take-away" messages from the development of our knowledge discovery processes on the NYC grid:

Prediction is Possible

We have shown successes in predicting failures of electrical components based on data collected by a major power utility company. It was not clear at the outset that knowledge discovery and data mining approaches would be able to predict electrical

7. http://www.isap-power.org/

component failures, let alone assist domain engineers with proactive maintenance programs. We are now involved in a Smart Grid Demonstration Project to verify that these techniques can be scaled to robust system use. For example, prior to our successes on the manhole event project, many Con Edison engineers did not view manhole event prediction as a realistic goal. The Con Edison trouble ticket data could easily have become what Fayyad et al. [8] consider a "data tomb." In this case, the remedy created by Columbia and Con Edison involved a careful problem formulation, the use of sophisticated text processing tools, and state-of-the-art machine learning techniques.

Data Are the Key

Power companies already collect a great deal of data, however, if these data are going to be used for prediction of failures, they should ideally have certain properties: first, the data should be as clean as possible, meaning for instance, that unique identifiers should be used for each component. Second, if a component is replaced, it is important to record the properties of the old component (and its surrounding context if it is used to derive features) before the replacement; otherwise it cannot be determined what properties are common to those being replaced.

For trouble tickets, unstructured text fields should not be eliminated. It is true that structured data are easier to analyze; on the other hand, free-text can be much more reliable. This was also discussed by Dalal et al. [38] in dealing with trouble tickets from web transaction data; in their case, a 40 character free-text field contained more information than any other field in the database. In the case of Con Edison trouble tickets, our representation based on the free-text can much more reliably determine the seriousness of events than the (structured) trouble type code. Further, the type of information that is generally recorded in trouble tickets cannot easily fit into a limited number of categories, and asking operators to choose the category under time pressure is not practical. We have demonstrated that analysis of unstructured text is possible, and even practical.

Machine Learning Ranking Methods Are Useful for Prioritization

Machine learning methods for ranking are relatively new, and currently they are not used in many application domains besides information retrieval. So far, we have found that in the domain of electrical grid maintenance, the key to success is in the interpretation and processing of data, rather than in the exact machine learning method used; however, these new ranking methods are designed exactly for prioritization problems, and it is possible that these methods can offer an edge over older methods in many applications. Furthermore, as data collection becomes more automated, it is possible that the dependence on processing will lessen, and there will be a substantial advantage in using algorithms designed precisely for the task of prioritization.

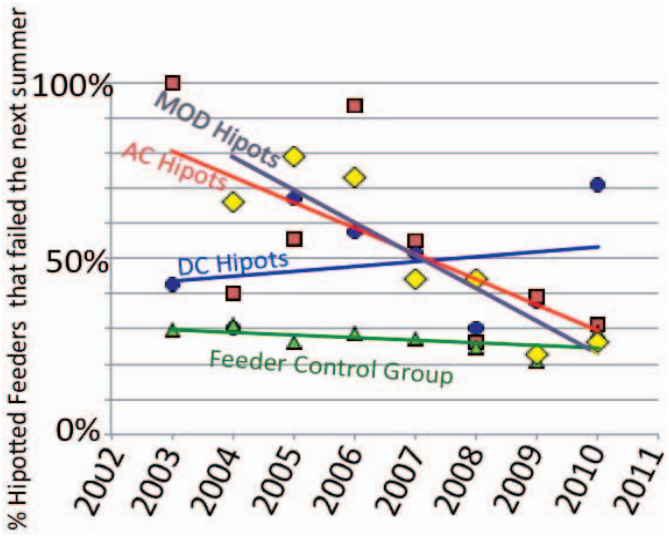

Fig. 23. Overtreatment in the High Potential (HiPot) Preventive Maintenance program was identified by comparing to control group performance. Modified and A/C Hipot tests are now used by Con Edison instead of DC Hipot tests.

Reactive Maintenance Can Lead to Overtreatment

We have demonstrated with a statistical method called "propensity" [39] that the High Potential (HiPot) testing program at Con Edison was overtreating the "patient," i.e., the feeders. HiPot is, by definition, preventive maintenance in that incipient faults are driven to failure by intentionally stressing the feeder. We found however, that the DC (direct current) HiPot testing, in particular, was not outperforming a "placebo" control group which was scored by Con Edison to be equally "sick" but on which no work was done (Figure 23). When a new AC (alternating current) test was added by Con Edison to avoid some of the overtreatment, we were able to demonstrate that as the test was being perfected on the system, the performance level increased and has now surpassed that of the control group. Indeed, operations and distribution engineering at Con Edison has since added a modified AC test that improved on the performance of the control group also. This interaction among machine learning, statistics, preventive maintenance programs, and domain experts will likely identify overtreatment in most utilities that are predominantly reactive to failures now. That has been the experience in other industries, including those for which these techniques have been developed, such as automotive and aerospace, the military, and healthcare.

10 Conclusions

Over the next several decades we will depend more on an aging and overtaxed electrical infrastructure. The

reliability of the future smart grid will depend heavily on the new preemptive maintenance policies that are currently being implemented around the world. Our work provides a fundamental means for constructing intelligent automated policies: machine learning and knowledge discovery for prediction of vulnerable components. Our main scientific contribution is a general process that can be used by power utilities for failure prediction and preemptive maintenance. We showed specialized versions of this process to feeder ranking, feeder component ranking (cables, joints, hammerheads, and transformers), MTBF/MTTF estimation, and manhole vulnerability ranking. We have demonstrated, through direct application to the New York City power grid, that data already collected by power companies can be harnessed to predict, and to thus assist in preventing, grid failures.

Cynthia Rudin is an assistant professor in the Operations Research and Statistics group at the MIT Sloan School of Management, and she is an adjunct research scientist at the Center for Computational Learning Systems, Columbia University. She received a Ph.D. in applied and computational mathematics from Princeton University and B.S. and B.A. degrees from the University at Buffalo.

David Waltz (Senior Member, IEEE) is the director of the Center for Computational Learning Systems at Columbia University, with prior positions as president of NEC Research Institute, director of Advanced Information Systems at Thinking Machines Corp., and faculty positions at Brandeis University and the University of Illinois at Urbana-Champaign. He received all his degrees from MIT. He is a fellow and past president of AAAI (Association for the Advancement of AI), and Fellow of the ACM.

Roger N. Anderson (Member, IEEE) is a senior scholar at the Center for Computational Learning Systems, Columbia University. Roger received his Ph.D. from the Scripps Institution of Oceanography, University of California at San Diego.

Albert Boulanger received a B.S. in physics at the University of Florida, Gainesville, in 1979 and an M.S. in computer science at the University of Illinois, Urbana-Champaign, in 1984. Albert is a senior staff associate at Columbia University's Center for Computational Learning Systems.

Ansaf Salleb-Aouissi joined Columbia University's Center for Computational Learning Systems as an associate research scientist after a postdoctoral fellowship at INRIA Rennes (France). She received M.S. and Ph.D. degrees from the University of Orleans (France) and an engineer degree in computer science from the University of Science and Technology Houari Boumediene (USTHB), Algeria.

Maggie Chow is a section manager at Consolidated Edison of New York. Her responsibilities focus on lean management and system reliability. Maggie received her B.E. from City College of New York and her masters degree from NYU-Poly.

Haimonti Dutta is an associate research scientist at the Center for Computational Learning Systems, Columbia University. She received her Ph.D. degree in computer science and electrical engineering (CSEE) from the University of Maryland.

Philip Gross received his B.S. from Columbia University in 1999 and his M.S. from Columbia University in 2001. Philip is a software engineer at Google.

Bert Huang is a Ph.D. candidate in the Department of Computer Science, Columbia University. He received M.S. and M.Phil. degrees from Columbia University and B.S. and B.A. degrees from Brandeis University.

Steve Ierome received a B.S. in electrical engineering from the City College of New York in 1975. He has 40 years of experience in Distribution Engineering Design and Planning at Con Edison, and 3 years of experience in power quality and testing of overhead radial equipment.

Delfina F. Isaac is a quality assurance manager and was previously a senior statistical analyst in the Engineering and Planning organization at Con Edison. She received both an M.S. in statistics in 2000 and a B.S. in applied mathematics and statistics in 1998 from the State University of New York at Stony Brook.

Arthur Kressner is the president of Grid Connections, LLC. He recently retired from the Consolidated Edison Company in New York City with over 40 years experience, most recently as the director of Research and Development.

Rebecca J. Passonneau is a senior research scientist at the Center for Computational Learning Systems, Columbia University, where she works on knowledge extraction from noisy textual data, spoken dialogue systems, and other applications of computational linguistics. She received her doctorate from the University of Chicago Department of Linguistics.

Axinia Radeva obtained an M.S. degree in electrical engineering from the Technical University at Sofia, Bulgaria, and a second M.S. degree in computer science from Eastern Michigan University. Axinia is a staff associate at Columbia University's Center for Computational Learning Systems.

Leon Wu (Member, IEEE) is a Ph.D. candidate at the Department of Computer Science and a senior research associate at the Center for Computational Learning Systems, Columbia University. He received his M.S. and M.Phil. in computer science from Columbia University and B.Sc. in physics from Sun Yat-sen University.

REFERENCES

[1] Office of Electric Transmission United States Department of Energy and Distribution. "Grid 2030" a national vision for electricity's second 100 years, July 2003.

[2] North American Electric Reliability Corporation (NERC). Results of the 2007 survey of reliability issues, revision 1, October 2007.

[3] S. Massoud Amin. U.S. electrical grid gets less reliable. *IEEE Spectrum Magazine*, January 2011.

[4] M Chupka, R Earle, P Fox-Penner, and R Hledik. Transforming America's power industry: The investment challenge 2010-2030. Technical report, The Brattle Group, Prepared for The Edison Foundation, Washington, D.C., 2008.

[5] William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. Knowledge discovery in databases: an overview. *AI Magazine*, 13(3):57–70, 1992.

[6] J. A. Harding, M. Shahbaz, Srinivas, and A. Kusiak. Data mining in manufacturing: A review. *Journal of Manufacturing Science and Engineering*, 128(4):969–976, 2006.

[7] Ana Azevedo and Manuel Filipe Santos. KDD, SEMMA and CRISP-DM: a parallel overview. In *Proceedings of the IADIS European Conf. Data Mining*, pages 182–185, 2008.

[8] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[9] Wynne Hsu, Mong Li Lee, Bing Liu, and Tok Wang Ling. Exploration mining in diabetic patients databases: findings and conclusions. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 430–436. ACM, 2000.

[10] Ron Kohavi, Llew Mason, Rajesh Parekh, and Zijian Zheng. Lessons and challenges from mining retail e-commerce data. *Machine Learning, Special Issue on Data Mining Lessons Learned*, 57:83–113, 2004.

[11] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997.

[12] Cynthia Rudin. The P-Norm Push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10:2233–2271, October 2009.

[13] Cynthia Rudin and Robert E. Schapire. Margin-based ranking and an equivalence between AdaBoost and RankBoost. *Journal of Machine Learning Research*, 10:2193–2232, October 2009.

[14] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

[15] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2005.

[16] Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems*, volume 9, pages 155–161. MIT Press, 1996.

[17] Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *CART: Classification and Regression Trees*. Wadsworth Press, 1983.

[18] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

[19] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, 34(2):187–220, 1972.

[20] Phil Gross, Ansaf Salleb-Aouissi, Haimonti Dutta, and Albert Boulanger. Ranking electrical feeders of the New York power grid. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pages 725–730, 2009.

[21] Philip Gross, Albert Boulanger, Marta Arias, David L. Waltz, Philip M. Long, Charles Lawson, Roger Anderson, Matthew Koenig, Mark Mastrocinque, William Fairechio, John A. Johnson, Serena Lee, Frank Doherty, and Arthur Kressner. Predicting electricity distribution feeder failures using machine learning susceptibility analysis. In *Proceedings of the Eighteenth Conference on Innovative Applications of Artificial Intelligence (IAAI)*, 2006.

[22] Hila Becker and Marta Arias. Real-time ranking with concept drift using expert advice. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 86–94, 2007.

[23] Cynthia Rudin, Rebecca Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac. A process for predicting manhole events in Manhattan. *Machine Learning*, 80:1–31, 2010.

[24] Rebecca Passonneau, Cynthia Rudin, Axinia Radeva, and Zhi An Liu. Reducing noise in labels and features for a real world dataset: Application of NLP corpus annotation methods. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, 2009.

[25] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL)*, July 2002.

[26] Pannaga Shivaswamy, Wei Chu, and Martin Jansche. A support vector approach to censored targets. In *Proceedings of the International Conference on Data Mining (ICDM)*, 2007.

[27] Axinia Radeva, Cynthia Rudin, Rebecca Passonneau, and Delfina Isaac. Report cards for manholes: Eliciting expert feedback for a machine learning task. In *Proceedings of the International Conference on Machine Learning and Applications*, 2009.

[28] Haimonti Dutta, Cynthia Rudin, Rebecca Passonneau, Fred Seibel, Nandini Bhardwaj, Axinia Radeva, Zhi An Liu, Steve Ierome, and Delfina Isaac. Visualization of manhole and precursor-type events for the Manhattan electrical distribution system. In *Proceedings of the Workshop on Geo-Visualization of Dynamics, Movement and Change, 11th AGILE International Conference on Geographic Information Science*, Girona, Spain, May 2008.

[29] Nikos D. Hatziargyriou. Machine learning applications to power systems. In *Machine Learning and Its Applications*, pages 308–317, New York, NY, USA, 2001. Springer-Verlag New York, Inc.

[30] Abhisek Ukil. *Intelligent Systems and Signal Processing in Power Engineering*. Power Engineering. Springer, 2007.

[31] Louis A. Wehenkel. *Automatic learning techniques in power systems*. Springer, 1998.

[32] A. Saramourtsis, J. Damousis, A. Bakirtzis, and P. Dokopoulos. Genetic algorithm solution to the economic dispatch problem - application to the electrical power grid of Crete island. In *Proceedings of the Workshop on Machine Learning Applications to Power Systems (ACAI)*, pages 308–317, 2001.

[33] Yiannis A. Katsigiannis, Antonis G. Tsikalakis, Pavlos S. Georgilakis, and Nikos D. Hatziargyriou. Improved wind power forecasting using a combined neuro-fuzzy and artificial neural network model. In *Proceedings of the 4th Helenic Conference on Artificial Intelligence, (SETN)*, pages 105–115, 2006.

[34] P. Geurts and L. Wehenkel. Early prediction of electric power system blackouts by temporal machine learning. In *Proceedings of the ICML98/AAAI98 workshop on predicting the future: AI approaches to time series analysis*, pages 21–28, 1998.

[35] Louis Wehenkel, Mevludin Glavic, Pierre Geurts, and Damien Ernst. Automatic learning for advanced sensing, monitoring and control of electric power systems. In *Proceedings of the Second Carnegie Mellon Conference in Electric Power Systems*, 2006.

[36] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang, Yi Qin, and Michael Chau. Crime data mining: a general framework and some examples. *IEEE Computer*, 37(4):50–56, 2004.

[37] Bertrand Cornélusse, Claude Wera, and Louis Wehenkel. Automatic learning for the classification of primary frequency control behaviour. In *Proceedings of the IEEE Power Tech Conference, Lausanne*, 2007.

[38] S. R. Dalal, D. Egan, and M. Rosenstein Y. Ho. The promise and challenge of mining web transaction data. In R. Khatree and C. R. Rao, editors, *Statistics in Industry (Handbook of Statistics)*, volume 22. Elsevier, 2003.

[39] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):45–55, 1983.