Mach Learn (2010) 80: 1–31
DOI 10.1007/s10994-009-5166-y

A process for predicting manhole events in Manhattan

Cynthia Rudin · Rebecca J. Passonneau · Axinia Radeva · Haimonti Dutta · Steve Ierome · Delfina Isaac

Received: 2 October 2008 / Revised: 18 November 2009 / Accepted: 12 December 2009 /
Published online: 28 January 2010

Abstract We present a knowledge discovery and data mining process developed as part of the Columbia/Con Edison project on manhole event prediction. This process can assist with real-world prioritization problems that involve raw data in the form of noisy documents requiring significant amounts of pre-processing. The documents are linked to a set of instances to be ranked according to prediction criteria. In the case of manhole event prediction, which is a new application for machine learning, the goal is to rank the electrical grid structures in Manhattan (manholes and service boxes) according to their vulnerability to serious manhole

Editor: Carla Brodley.

This work was done while Cynthia Rudin was at the Center for Computational Learning Systems at Columbia University.

C. Rudin (✉)
MIT Sloan School of Management, E53-323, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: Rudin@mit.edu

R.J. Passonneau · A. Radeva · H. Dutta
Center for Computational Learning Systems, Columbia University, 475 Riverside Dr., New York, NY 10115, USA

R.J. Passonneau
e-mail: becky@cs.columbia.edu

A. Radeva
e-mail: Axinia@ccls.columbia.edu

H. Dutta
e-mail: Haimonti@ccls.columbia.edu

S. Ierome · D. Isaac
Consolidated Edison Company of New York, 4 Irving Place, New York, NY 10003, USA

S. Ierome
e-mail: IeromeS@coned.com

D. Isaac
e-mail: IsaacD@coned.com

events such as fires, explosions and smoking manholes. Our ranking results are currently being used to help prioritize repair work on the Manhattan electrical grid.

Keywords Manhole events · Applications of machine learning · Ranking · Knowledge discovery

1 Introduction

We describe a knowledge discovery and data mining process developed through a research collaboration between Columbia University's Center for Computational Learning Systems and the Consolidated Edison Company of New York. This collaboration was initiated in order to investigate how machine learning can help to maintain the electrical grid. There are a few hundred manhole events (fires, explosions and smoking manholes) in Manhattan every year, often stemming from problems in the low voltage secondary distribution network that provides power to residential and commercial customers. Our task was to rank the manholes and service boxes in Manhattan's secondary network in order of vulnerability to *serious* manhole events. A measure of the success of our predictive model is that we have been asked to apply it to the other NYC boroughs, and initial results on the first of these (Brooklyn) have been excellent. Among the many obstacles we faced, the most fundamental were our lack of domain expertise, the domain experts' lack of expertise with predictive modeling, limited time to spend with the domain experts, extremely raw and noisy data from disparate sources, a significant amount of subjectivity in the notion of serious event, and most importantly, the lack of a clear, answerable research question. This paper describes how we overcame these obstacles to arrive at a model that was predictive and sufficiently meaningful to the domain experts that it is being used to prioritize repairs to the secondary network.

Manhattan's secondary electrical grid is a network of over 21,000 miles of underground cable that runs along the streets and avenues. The cable can be accessed at approximately 51 thousand manholes and service boxes. (We use the Con Edison terminology "manhole" or "structure" to represent both manholes and service boxes.) As this secondary grid is an evolving system that started over a century ago, the cables are of varying age, quality, and dependability. Insulation breakdown in secondary cables occasionally causes power outages, flickering lights, wire burn-outs, smoking manholes, fires, explosions or other types of outages.

A proactive manhole inspection program was implemented by Con Edison in 2004 as a means to prevent these types of events. Prior to the start of this program, repairs were performed mainly in response to an event or outage. From over four years of inspection reports, a long list of pending repairs and upgrades has now been generated. Since each repair takes some time to perform, it is important to prioritize so that the most vulnerable structures are repaired earlier. The most important goal of the Columbia/Con Edison collaboration has been to assist with this prioritization task.

A large proportion of our data comes from Emergency Control System (ECS) trouble tickets recorded by Con Edison dispatchers over the course of a decade, between the years of 1996 and 2006. The tickets include a free text "Remarks" field of varying length and extremely heterogeneous content. The tickets were not intended to contain a complete description of the event, but were designed instead for keeping track of logistics. A ticket may contain, for instance, a description of repair work related to the event, details about parking in the area, customer contact information, reasons for departure of repair crew personnel, and other things. Only a handful of engineers have experience in interpreting the trouble ticket data, and many tickets are not easy to interpret, even by domain experts.

This paper makes two sorts of contributions, a case study of a new type of real-world machine learning application and a general data mining process for ranking domain entities that are linked to text documents. The process falls mostly into the very general CRISP-DM[1] (Azevedo and Santos 2008) framework of processes for data mining, but does not directly fall into the traditional knowledge discovery in databases (KDD) outline (Fayyad et al. 1996), since KDD does not directly encompass cases where the data is extremely raw (for instance, not even in the form of a database) and the problem is ill defined. In our case, order had to be created from confusion; this meant defining a learnable problem, assembling a database to represent as many potentially relevant characteristics as possible about entities in the domain (here, structures in the secondary grid), and establishing a mechanism for eliciting knowledge from the domain experts ("conferencing"). Once we had a rudimentary problem formulation that had evidential grounding, alongside a database from which to select, refine and mine features, the preconditions for a more traditional KDD process were in place. Given our focus on the problem definition and means to assemble and exploit available knowledge, our case history constitutes a lesson in overall strategy for real world machine learning applications, analogous to the notion of *statistical strategy* used by Hand (1994), rather than a demonstration of specific techniques in machine learning, data reduction, feature selection, or aspects of the mechanics of data mining.

The remainder of the work is organized as follows: in Section 2 we give a brief overview of research in knowledge discovery in databases, and point to applications papers that address problems related to aspects of our work. In Section 3, we outline the data mining process we developed. In Section 4 we itemize the data provided to us for the manhole ranking task. Section 5 provides the statistics behind the first iteration of our model for this problem. That model was not accurate, and did not yet use the free-text of the ECS tickets. Section 6 discusses how the data mining process led to refinements of the preliminary model. Section 7 discusses how features and labels were developed for the manhole ranking problem. Section 8 discusses a means of evaluation for the final model, and Section 9 exemplifies conferencing between scientists and domain experts, followed by a discussion and conclusion.

2 Related work

We address a problem of knowledge discovery, meaning that we aim to find information in data that is implicit, novel, and potentially extremely useful (Frawley et al. 1992). An overview of this type of problem in manufacturing is provided by Harding et al. (2006). In our case, we are using data that was not designed to be used for predictive analysis, but instead for record-keeping. In many industries, databases designed for record-keeping will never be mined, constituting what Fayyad and Uthurusamy (2002) consider "data tombs." These authors suggest that there is knowledge to be gained from analyzing these data: "data tombs represent missed opportunities." However, we argue that the traditional KDD outline is not sufficient for many real-world scenarios and a more general framework is required. For instance, traditional KDD does not address the creation of an initial database. In our process, creation of the database is guided by the problem definition and refinement. In many circumstances, especially those involving domain experts, definition of the problem evolves in increments and thus involves hefty pre/re-processing. Hsu et al. (2000) state that "...the often neglected pre-processing and post-processing steps in knowledge discovery

[1] http://www.crisp-dm.org/.

are the most critical elements in determining the success of a real-life data mining application." In their experience working with doctors on an analytical tool for medical record data, they found that the process of reaching the stage where understandable rules could be generated from an existing database was equally if not more important than the rule generation phase itself. The data they were provided was noisy, and required reconciliation of attributes across databases to support mining for patterns. In contrast, we did not have an existing database but created one from noisy tables and raw text, and we used the discovery of patterns to produce a predictive model. Thus, we consider a broader perspective to the standard knowledge discovery outline that adds an initial stage of developing a problem definition in conjunction with creation of a database, and a method for conferencing with domain experts.

Hsu et al.'s experience is not unique in using domain expertise to guide the development of labels and features for machine learning. Castano et al. (2003) have developed an algorithm for guiding the data gathering efforts of the Mars Rover towards samples that are interesting to domain experts. Kusiak and Shah (2006) developed a system to assist decision makers at a power plant, and domain expertise was used at various stages, including the labeling of examples. Boriah et al. (2008) discuss the problem of detecting land cover change. The team members designed the problem in a way that they themselves were able to function as domain experts in order to evaluate the quality of the model through profiling specific regions. Chen et al. (2004) designed a system for several problems in crime data mining in collaboration with the Tucson police department. In fact, this group handled documents of a similar caliber to our trouble tickets; one of their first tasks was to extract named entities from police reports that are difficult to analyze (due to inconsistent spelling and grammar) using automated techniques. In their work, the creation of the database was aimed broadly at solving several different problems, as oppose to creating and refining the database in order to solve one specific problem, as in our case. In a wide range of natural language processing tasks using machine learning or other techniques, it is frequently necessary to consult domain experts in order to understand the vocabulary of a domain and the conventions for constructing documents and other communications in the domain (Sager 1970; Kittredge 1982; Kittredge et al. 1991; Harris 1982; Grishman et al. 1986). We have designed several tools to facilitate communication of results to domain experts. We argue that conferencing tools should be designed in conjunction with the evaluation stage of the data mining process; this important piece of our process is not subsumed by either KDD or CRISP-DM. A similar point has been made with respect to the problem of applying machine learning to software development: Patel et al. (2008) identified the difficulty developers have in *"understanding" relationships between data and the behavior of statistical machine learning algorithms* as one of three key difficulties in applying statistical machine learning to software development. As we demonstrate, by presenting domain experts with detailed reports ("profiles") concerning individual structures, we elicited feedback that led us to new, more meaningful features.

Trouble tickets (Liddy et al. 2006), maintenance logs (Devaney and Ram 2005), equipment reports (Hirschman et al. 1989), safety reports (Oza et al. 2009) and similar sets of documents have been handled using the methods of natural language processing, knowledge modeling, text classification and machine learning for a wide range of goals. Relatively early work (Hirschman et al. 1989) showed it was possible to handle fragmentary text using full syntactic and semantic parsing, but involved a much smaller dataset than current work on ticket databases. Devaney and Ram (2005) dealt with 10,000 logs, all of which pertain to the same machines. They combined unsupervised text clustering with a domain representation developed in OWL/RDF to classify tickets, then developed a Case-Based Reasoning approach to predict failures. Liddy and her colleagues (Liddy et al. 2006;

Symonenko et al. 2006) developed an application for the same types of trouble ticket data we address, but their goal was primarily to assign trouble tickets with a miscellaneous categorization ("MSE") to a more specific ticket type, and they did not use these tickets to rank structures. The work of Oza et al. (2009) is the only one we have seen that deals with a similarly large dataset, and where disagreements among human experts made it difficult to define document classes, a problem we also faced. They looked at two aeronautics report databases that have a combined size of 800,000 reports. Their reports, unlike Con Edison trouble tickets, generally have a single author, and consist of a readable, discursive narrative. Their end goal was to arrive at a comprehensive, topic-based document classification, whereas our classification task was to scale the severity of events, and we ignored ticket content not relevant to this task. They relied on an existing thesaurus (PLADS) to merge distinct forms of a single term, such as acronyms, abbreviations and phrases, making their documents amenable to a bag-of-words (BOW) document representation, and they used two learning techniques, Support Vector Machines and Non-negative Matrix Factorization. Our early attempts to use BOW features foundered due to the high noise content. In ongoing work, we have used classification methods to generate string normalization rules, making it possible to re-examine the use of BOW representations for text classification and clustering methods.

The features and labels for the manhole event prediction problem were formulated within the framework of rare event prediction. Our evidence indicated that even with "clean" data and an operational definition of "serious" manhole event, predicting these events would not be easy. The number of such events is small–approximately six to nine hundred per year, depending on the definition–compared to the total number of structures in Manhattan, which is around 51K. The framework of rare event prediction is useful for classifying whether events will occur within a pre-determined "prediction period." In our case we aimed to rank instances rather than to classify, which has a secondary effect of offsetting the class imbalance problem. (Essentially this is because examples are not considered relative to any decision boundary, instead they are considered relative to each other.) Other works addressing specifically rare event prediction include a genetic algorithm approach for predicting hardware component failures in the telecommunications industry (Weiss and Hirsh 2000), a system for identifying faulty operating conditions for power plants (Kusiak and Shah 2006), and an association rule approach that characterizes the minority class exclusively before using those rules to discriminate between the classes (Vilalta and Ma 2002).

The bipartite ranking technique used for the prediction problem is based on the bipartite ranking framework defined by Freund et al. (2003), and is designed to concentrate at the top of the list (Rudin 2009). This technique characterizes the relationships between manholes without first estimating the underlying density of the vulnerable and non-vulnerable classes of manholes. Other works adopted density modeling techniques for offline prioritization problems, for instance the prioritization of mutations that cause disease (Jiang et al. 2007), and the prioritization of geographic regions for conservation (Chen and Peterson 2002) and species modeling (Dudík et al. 2007). Note that many other problems dealing with prediction of rare events in continuous time rely on a shorter time scale, including seizure prediction for epileptic patients, or prediction of failures in hard drives (Murray et al. 2005). In contrast, our task is an offline processing problem that uses a long-term history to predict events that may happen several months or even years later.

Our colleagues at the Center for Computational Learning Systems, as far as we know, were the first to establish a modern machine learning framework for applications to the electrical grid (Gross et al. 2006; Becker and Arias 2007), though there is much precedent for other maintenance techniques and statistics for use in power engineering (e.g., monitoring the health of power transformers, Steed 1995; Kirtley et al. 1996). Our colleagues have

been concerned with the short-term prediction problem of ranking electrical components in the primary distribution system, specifically electrical feeders, cables and joints, according to their susceptibility to failure. This application differs dramatically from ours in that feeders have electrical monitors that provide numerical information in real time, for instance, features are based on electrical load simulations and real-time telemetry data. In contrast, our data consists mostly of historical records written mainly in free-text. It is clear when a feeder fails since an outage of that feeder occurs, whereas it is not always clear when a serious manhole event has occurred, as discussed in Sect. 6.1. Formulating the problem definition in the case of feeder failure prediction was significantly more straightforward (though the associated data mining problem was not necessarily any easier). For instance, the static features (based on feeder characteristics) were already assembled into a database by Con Edison, and the features individually are useful predictors.

3 Process for classification of documents to facilitate ranking

Figure 1 illustrates the data mining process. We identify three crucial elements. The first is early development of a skeletal problem formulation that can drive the entire process. The second is assembly of a relational database of information pertinent to the problem through a range of techniques to verify the existing data, and more importantly, to add new structured information. These include Information Extraction (IE) techniques applied to the free text combined with other information sources to make the extracted information more precise. This database is mined using algorithms for ranking. The third is to integrate domain experts into the iterative process. This is accomplished through interactive conferencing tools that afford a transparent link between a predictive model–here, a ranking of structures with respect to their vulnerability to serious events–and descriptive data about the domain entities (Radeva et al. 2009).

Very often, a real world problem presented to researchers by domain experts is cast in such a general way that analysis of the problem cannot begin until the causal relations in the domain are better understood, as we illustrate in Sect. 5. When the domain experts have a clear understanding of at least some of the causal relations, their problem may be amenable to the traditional KDD process, particularly if the available data has already been assembled into a database. In many cases, however, the important entities in a domain have been documented in textual reports, trouble tickets, or other unstructured forms, and not yet mined for useful information. The institutional owner may believe that the document repository contains valuable information pertinent for maintenance or prevention. In such a case, the data mining and modeling tasks cannot proceed until the knowledge potential buried in the document repository has become realized: the knowledge must be formulated. In our experience, pushing for an increasingly precise problem formulation should drive the dual tasks of knowledge formulation and data mining.

The general problem of linking a set of entities to textual reports in order to facilitate ranking the entities according to some performance criterion is one that would apply very generally, as in the following:

– *Manufacturing*: The manager of a factory with hundreds of machines would like to prioritize replacement of non-efficient machines. The efficiency of machines can be categorized using several sources, including past inspections in the form of noisy text documents written by different inspectors in shorthand.

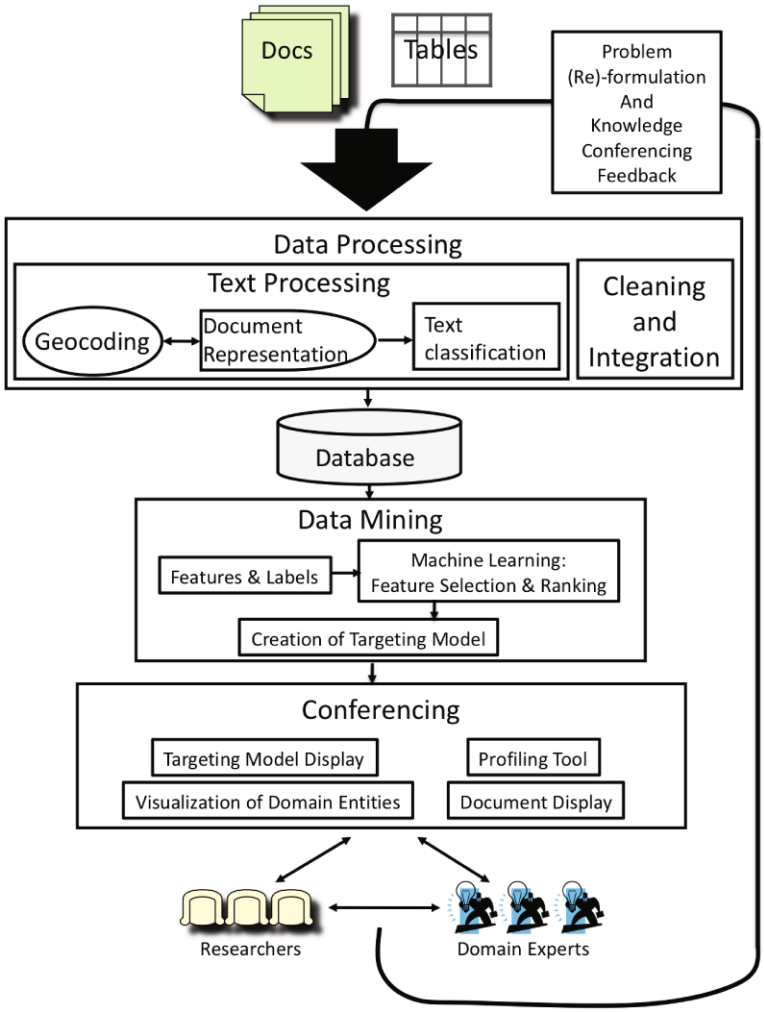


Fig. 1 Knowledge discovery and data mining process for ranking

– *Airlines*: An airline company would like to prioritize replacement of airplanes based on whether the plane is likely to experience a problem based on historical notes of repair crews (e.g., Oza et al. 2009) and characteristics of the planes.
– *Advertising*: A telemarketing company would like to prioritize potential customers based on the probability that the customer will purchase the product. The data may include (noisy) transcripts of phone conversations between telemarketers and potential customers, and demographic characteristics of the customers.
– *Recommender systems*: An Internet company has data from several blogs, and would like to recommend services (blogs, webpages, ads) to those who regularly post on these blogs.

More so than tabular data, documents generated for reporting purposes within commercial institutions typically require interpretation by domain experts, and there may be noise in the form of disagreements among experts, as in the classification of aeronautics safety

reports described by Oza et al. (2009). As we will illustrate in subsequent sections, the ten years of trouble tickets that Con Edison provided us with are an extremely rich but noisy source of data. Misspellings, for example, are so common that we estimate the size of the vocabulary (number of unique terms) could be halved if the spelling were normalized. Yet these trouble tickets became the main source of information for labeling structures as vulnerable or not, and for constructing tabular histories of all documented events each structure had experienced. To a large degree, we relied on document classification based on experts' judgments combined with *Information Extraction (IE)* methods. Information extraction (IE) is defined as the task of organizing and normalizing data taken from unstructured text in order to populate tables in structured databases. A picture of the differences in IE success, depending on the nature of the input documents, is given by the Automatic Content Extraction program National Institute of Standards and Technology, NIST, Information Access Division, ACE. For nearly ten years, NIST has had yearly evaluations of IE systems. ACE results are highly sensitive to the degree of noise in the documents. Performance is higher on written or broadcast news than on usernet newsgroup text or telephone transcripts, and higher on the source language than on machine translation output. Typically, off-the-shelf IE tools perform better on documents written in standard language, and with standard orthography. The noisier the documents, the more that successful IE depends on custom-built pattern-matching procedures, which was the case for the Manhattan trouble tickets.

3.1 Components of the knowledge discovery and data mining process

Problem (re)-formulation: In the abstract, this step involves defining how the objects in the domain are to be represented, the basis for ranking them, and the elimination or reduction of subjectivity in that basis. Our task as defined by Con Edison was to rank structures based on the likelihood of experiencing a serious event. The intuition held by domain experts was that within a relatively short time frame (a few weeks or months), a structure would experience *precursor* events of a less serious nature (such as an interruption in power due to a burnt cable), that could be used to predict more serious events. Defining the problem involved determining a relevant time frame for prediction, and determining how event data should be processed, classified and linked to structures in a way that prediction of serious events would be possible. For example, at the outset, it was not clear to us or to the domain experts whether we should attempt to predict future events on individual structures, or for a small neighborhood of structures.

Document representation: The raw documents must be processed into a form that makes it possible to use the knowledge they contain or constitute. As discussed in Sect. 2, full natural language understanding, even on the fragmentary and telegraphic language found in trouble tickets, can be aimed for but is often not necessary. In our initial work, we gradually defined classes of regular expressions for extracting references to structures, or for assigning document metadata (descriptors that characterize the document), and ultimately migrated to an explicit document annotation language using a text engineering tool (GATE, Cunningham et al. 2002) that allowed us to export document features into a database.

As part of the document representation, there must be some mechanism for associating the documents with the instances to be ranked. In the case of manhole events, the trouble ticket must be associated with the structure that is the "*trouble hole*" for the event, meaning the structure on which the event occurred. Making the link was non-trivial due to noise in the tickets and ticket addresses, and the fact that trouble hole information may appear in one of several different raw tables, as discussed in Sect. 6.3.

Document classification: The problem definition may require documents to be classified into one of several categories. For manhole event prediction, we classified trouble tickets as representing a serious event, a possible precursor to a serious event, or not representing a relevant event. For some applications (though not for manhole event prediction), a classification of the documents into categories based on a word vector or word matrix representation (cf. Oza et al. 2009) may be sufficient. The document classification we arrived at depends on manually formulated rules.

Data integration: Heterogeneous data sources often need to be merged, and significant effort can be required to clean, process and aggregate these data sources. The data may come from different sources and be stored in various formats. For instance, since the Con Edison raw data was mainly entered manually, table joins often led to a loss of a large percentage of records due to inconsistencies in structure names. Noise in the data was a problem that we had to overcome at each stage of processing.

Features and labels: The results of the document classification are used to develop features and labels for the ranking problem. In the case of manhole event prediction, many features are based on a structure's past history of events. The label is whether the structure was the trouble hole for a serious event within a predetermined prediction period, where the seriousness of an event is given by the document classification results.

Conferencing: We use the term "conferencing" to refer to consultation within the research team, and between the research team and the domain experts, in order to affirm or enhance potential patterns discovered in data. This involves discussing results at various stages of the process and developing tools to obtain useful feedback. It is critical that the tools show the connection between the abstract representation of the instances being modeled, the domain entities as known to the domain experts, and the ranking model. The tools we used included a display of our ranked list ("targeting model display"), GATE (Cunningham et al. 2002) for displaying the document annotations and features, a profiling tool that allowed us to view all the known information about any structure in our database (Radeva et al. 2009), and a visualization tool that shows the locations and densities of structures, cables, and events (Dutta et al. 2008).

4 Con Edison data sources

Table 1 lists seven tables of raw data we started with, originating from different departments within Con Edison. The data concerns structures, events, cables, and inspections.

The most important of the raw tables was the Con Edison Emergency Control Systems (ECS) trouble tickets table, containing approximately 1 million tickets over a ten-year period, for all boroughs. It is a rich albeit noisy resource for information about events in the secondary network. Each trouble ticket is a report of an event affecting the New York City electrical distribution system as recorded by a Con Edison dispatcher. The "front" of each ticket has a timestamp (date and time), "trouble type" (type of event, such as manhole fire or smoking manhole), along with several address and cross street fields that are entered manually by the dispatcher. Two sample ECS ticket fronts are shown in Table 2.

The "back" of the ticket is called the ECS "Remarks." Figure 2 shows the ticket number, total number of lines in the Remarks field, and the Remarks of a short ECS ticket. For ease of reference, we have inserted line numbers in the leftmost column. This ticket was briefly related to another ticket (between 16:26 and 19:22 on 01/16/97; see lines 3 and 4) but it represents a distinct event. Appendix B discusses "referred tickets" that (unlike this ticket)

Table 1 Con Edison raw data sources

Data source	Brief description
Structures	A list of all structures (manholes and service boxes) along with their geographic coordinates. There are 51219 structures.
ECS	Trouble ticket database, contains records of past events.
ELIN	Additional details regarding manhole events.
ESR/ENE	Additional details regarding electrical shock and energized equipment events.
Inspections	List of inspections including type of repairs made and types of repairs recommended for each structure.
Property records	Electrical cable information, including the service type (which is either "main," connecting two structures, or "service," connecting a structure to a nearby building), location of the cable, material (copper or aluminum), number of phase and neutral cables, type of insulation, cable size, date of installation.
Vented manhole cover table	Indicates which structures have "new vented" covers, which allow gases to escape more easily. The covers are being replaced as part of an aggressive vented cover replacement program.

Table 2 Partial Sample ECS Ticket "fronts," one for a manhole fire event ("MHF") and one for a manhole explosion ("MHO") note the misspellings and irregularities. The notation N/E/C means "northeast corner"

	First ticket	Second ticket
Ticket number	ME11005661	ME12003775
Borough	M	M
House number	120	N/E/C
East/west/other		
Street name	E.BRAODWAY	GREENWHICH
Street/avenue		ST
Cross street	PIKE ST	CEDAR ST
Actual trouble type	EDSMHF	EDSMHO
Received date and time	2001-03-13 14:30:00-05	2002-03-29 17:40:00-05

do not represent distinct events, and are generally excluded from our analysis. The ticket mentions a specific service box (SB 325681, line 2) as the implicit trouble hole. It also contains a variant of a phrase we have identified as indicating that Con Edison repair crews performed work on the structure ("CLEARED B/O", where "B/O" stands for "burn-out," line 7). Here lines 1–2 (the original "complaint") and 7–8 are free text entered by a dispatcher, and the lines with a date/time prefix (3–6, 9) are automatically generated. A longer sample ECS ticket in Fig. 3 possesses many of the same features. When two domain experts were independently asked whether this ticket represented a serious event, they disagreed. As described in Sect. 6.1, determining which events to consider serious turned out to be difficult for domain experts to articulate in a way we could operationalize.

Mining the Remarks for useful information presented serious obstacles. Because the ECS tickets are created by people working under time pressure, they have the fragmen-

```
Ticket: ME97105931
Lines: 9
Remarks:
1   01/16/97 CABLE DEPT REPORTS: W/S 8 AVE 66’ S/O W.116 ST
2   SB-325681 INSTALLED NEW MAIN NEUTRALS ARE ALIVE.......DR
3   01/16/97 16:26 DUPLICATED TO ME97100590       BY 66920
4   01/16/97 19:22 UNDUPLICATED FROM ME97100590   BY 66920
5   01/16/97 15:00 MDEFELIX DISPATCHED            BY 66920
6   01/16/97 15:45 MDEFELIX ARRIVED               BY 66920
7   FELIX REPORTS IN SB325681 CLEARED B/O HOLE STILL NEEDS FLUSH
8   THIS JOB COMP.
9   01/16/97 19:00 MDEFELIX COMPLETE              BY 66920
```

Fig. 2 Sample ECS remarks

```
Ticket: ME03100287
Lines: 39
Remarks:
 1   MR. ROBERT TOBIA (718)555-5124 - SMOKING. COVER OFF.-RMKS:
 2   01/06/03 08:40 MDETHUILOT DISPATCHED         BY 55988
 3   01/06/03 09:30 MDETHUILOT ARRIVED            BY 55988
 4   01/06/02 09:55 THUILOT REPORTS NO SMOKE ON ARRIVAL. THERE IS
 5   A SHUNT ON LOCATION - SHUNT & SERVICE NOT EFFECTED.
 6   CO = 0PPM 01/06/03 09:45 SB521117 F/O 256-54 W.139 ST.
 7   GAS = 0% , OXY = 20.8% , COVER WITH WAFFLE - SOLID- ON .
 8   REQUESTING FLUSH/ORDERED (#2836).
 9   ******* NO PARKING : TUES. & FRIDAY, 11:30AM - 1PM ****** RV
10   01/06/03 10:45 THUILOT REPORTS BUILDING 260 W.139 ST.
11   COMPLAINED OF LIGHT PROBLEMS. FOUND 1-PHASE DOWN - BRIDGED
12   @ 10:30 ( 2-PHASE SERVICE ) CONSUMER IS CONTENT.
13   ********************** PSC COMPLETED ******************** RV
14   ****************** CHANGE IN PLAN SENT ON ’21’ ********** RV
15   01/06/03 10:45 MDETHUILOT UNFINISHED         BY 55988
16   01/06/03 12:00 MDEFERNAND DISPATCHED         BY 45348
17   01/06/03 12:50 MDEFERNAND ARRIVED            BY 18624
18   01/06/03 18:45 FERNANDEZ REPORTS THAT IN SB-521117 F/O254
19   W139 ST. HE CUT OUT A 3W2W COPPERED JT & REPLACED IT W/
20   A 4W NEO CRAB....BY USING 1 LEG OFF THE 7W FROM THE HE
21   WAS ABLE TO PUSH THE MISSING PHASE BACK TO 260, BRIDGE
22   REMOVED....@ THIS TIME FERNANDEZ REPORTS THERE ARE MORE
23   B/O’S & 2 MORE JTS TO C/O, WILL F/U W/ MORE INFO....TCP
24   21:15 21:10 FERNANDEZ REPORTS THAT HE CUT OUT A 2W2W & 1W1W
25   AC JOINT IN SB-521117 F/O 254 W139 ST. & INSTALLED 2W2W NEO
26   & MADE 1 STRAIGHT SHOT TO PU SERVICES AND MAINS.......TCP
27   ********************#9 TO FOLLOW UP**********************
28   TO INVESTIGATE THE MISSING PHASE FROM THE WEST WHEN PARKING
29   IN OUR FAVOR IN SB-521116F/O 260 W139 ST.................TCP
30   ************************************************************
31   01/06/03 21:26 MDEFERNAND UNFINISHED         BY 18624
32   ==============ELIN REPORT COMPLETED==================GS
33   01/14/03 07:05 SPEC NOTE CHNGD FROM TO B/O   BY 58101
34   ************** INPUTTED INTO SHUNTS & BRIDGES **************
35   **********SEE ME03101307 FOR ADDITIONAL INFO************JFM
36   01/29/03 06:53 MDE.OFFICE DISPATCHED         BY 45348
37   01/29/03 06:53 MDE.OFFICE ARRIVED            BY 43961
38   01/29/03 06:53 MDE.OFFICE COMPLETE           BY 43961
39   02/20/03 12:29 REFERRED TO: MH.INCID EDSSMH FYI BY 22556
```

Fig. 3 Sample ECS remarks

tary language and telegraphic features found to be typical of trouble reports (Linebarger et al. 1988; Hirschman et al. 1989). They contain a very wide range of categories of information, including repair work ("CUT,CLEARED,P/O RETIED DEFECTIVE SERVICE LEG & RESTORED FULL POWER"), parking restrictions in the relevant area of the structure ("-NO PARKING 08:00 TO 18:00 MON TO FRI-"), communications with customers ("HAVE A BAD LANGUAGE PROBLEM...WITH CARETAKER...") and other relevant information ("FOUND EXTENSION CORD TIED TO PIGTAIL AT BASE OF LAMP"). Due to enormous variation in length (1 to 1000 lines), tickets vary in the type and amount of relevant information they contain; for instance, they often do not contain an explicit description of the event itself. The Remarks have a very high rate of formatting idiosyncrasies, word variants and especially misspellings. For these reasons, and because it was not yet clear how to use them, early iterations of the model did not use the Remarks. However, they became essential for our analysis, as they were the only source for determining the "trouble hole" for an event (Sect. 6.3), or for judging the seriousness of an event (Sect. 6.1).

5 Initial problem definition for manhole event prediction

Con Edison has many divisions, each maintaining distinct sources of data pertaining to the electrical grid. Con Edison experts in the research division hypothesized that there was "knowledge" to be gained from analyzing these data, but they had not been mining this knowledge, and it was not clear what mechanism to use in order to find it. In this case, they wanted to "predict serious events," but there was a large gray area around the definition of serious event, and it was not clear how to link events with their locations, or what time frame to use.

At the start of the project, we created a preliminary problem definition in order to test one of the important hypotheses made by domain experts. The strategy adopted was in essence a bare bones version of the process in Fig. 1, where the document representation and text classification steps had been roughly approximated. The hypothesis to be tested was that *serious manhole events sometimes have short-term "precursors" beforehand*. A precursor is a non-serious event that, if we were able to identify it in advance, would lead to a method of preventing serious manhole events. The experts believed there would be a critical time frame of between 3 days and 2 weeks in which serious manhole events were much more likely to follow from precursors. An example of this phenomenon would be a flickering lights (FLT) event followed three days later by a manhole fire in the same area.

Domain experts clarified that serious events are not generally *caused* by improper or incomplete repairs for precursor events. Con Edison crews perform necessary repairs for each event (that is, Remarks terminology "PROBLEM CORRECTED," "SERVICE RESTORED", "CUT, CLEARED AND RETIED"). Instead, an underlying area problem generally causes both the precursor event and the serious event. However, a single repair within a manhole may not completely mitigate weaknesses of that manhole. Repair crews and inspectors cannot see inside the ducts between structures, making it difficult to determine wither a given repair will prevent future problems. Since it is not practical or cost-effective to replace large amounts of cable in response to a localized non-serious event, a cable is only replaced if repair crews can determine that it is a source of the problem. Thus, it is reasonable that a study of past non-serious events would lead to a method for predicting serious events.

We aimed first to test the Con Edison engineers' hypothesis that serious events have precursors, using readily available data, namely the front of the ECS tickets (ignoring completely the ECS Remarks), and the Structures table, which provided a geographic location

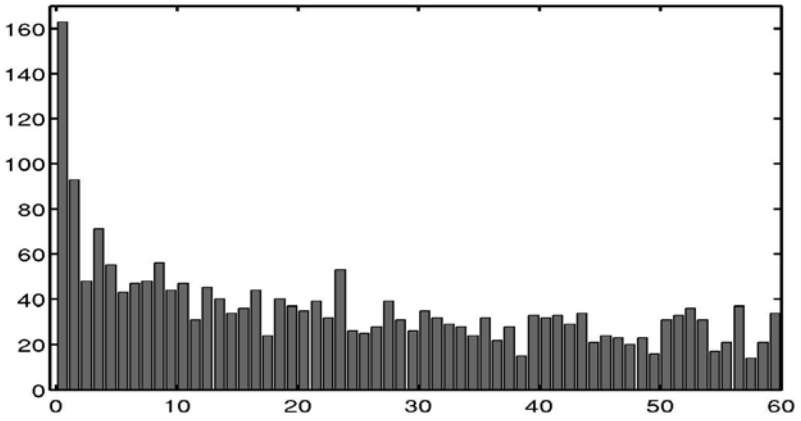


Fig. 4 Histogram of time difference between a precursor ticket and the serious event ticket that follows. Referred tickets are excluded (see Appendix B)

for each structure. At this point, the document representation consisted of an address for the event as geocoded by Columbia Geostan, which is our geocoding module described in Appendix A, along with the trouble type. The ticket classification was performed based solely on the trouble type: if the ticket had an explosion, fire or smoking manhole trouble type (MHX, MHO, MHF, SMH) then the ticket was classified as representing a serious manhole event, all other relevant trouble types were considered as potential precursor events (which we call "burn-outs"), including flickering lights (FLT), no lights (NL), AC burn-out (ACB), side off (SO), and several other trouble types.

In order to link the tickets to their trouble hole(s), we associated each ticket with every structure within 60 meters (using the geocoded addresses for the tickets), which is approximately the size of an intersection. (It is true that this 60 meters generally includes several uninvolved structures and possibly does not include any actual trouble holes.)

Using these basic approximations to the document representation and ticket classification, we tested the hypothesis by looking at pairs of events (classified by trouble type) across time. The goal was to see if there is a "crucial time frame" between a precursor and a serious event. Unfortunately we were not able to validate the domain experts' hypothesis: if a precursor is followed by a serious event within 60 meters, it is not easy to predict when this event is likely to occur. Figure 4 shows a histogram of the time difference between a precursor and a serious event that follows. It is often the case that a precursor takes place simultaneously with a serious event (corresponding to the peak at 0), however, there is no point at which a significant drop occurs, indicating that it is only slightly more likely for the event to occur after a few days than after a few weeks or months.

Since we had evidence suggesting that short-term prediction efforts would not yield satisfying results, we decided instead to consider the longer term. Domain experts advised us that the physical processes causing insulation breakdown can take place over several months or years so indeed it is reasonable to test for long term predictive factors. Con Edison called this the "hotspot" theory, and the evidence we produced to support this theory was perhaps the major turning point in the project.

We observed that most serious events have at least one precursor: of the 6,670 serious event tickets, 5,458 of them were preceded by at least one burn-out within the last 3 years (again excluding referred tickets, see Appendix B). We formally define a *true precursor* to

be a burn-out that is followed by a serious event ticket within 60 meters and 60 days. Other burn-outs are *false precursors*. We observed that true precursors are much more likely to have at least one prior burn-out in the long term history than false precursors (93% vs. 86%; $p \approx 0$ [2]), and true precursors are much more likely to have at least one prior manhole event in the long term history than false precursors (57% vs. 42%; $p \approx 0$). Further, the probability that a burn-out will be a followed by a serious event is significantly larger if it has a past history of events (6.5% vs. 2.9%; $p \approx 0$). Thus, predicting whether a potential precursor will be followed by a nearby manhole event requires a long term history of events in the area, rather than a short term one.

We built an initial model based on history of events in the local area. Instead of setting the prediction period to be a small time window as the domain experts initially suggested, we agreed to a window of one year. Each structure (manhole or service box) corresponds to one instance in the supervised ranking problem associated with a given time frame. In the initial model, the label for the structure was determined by whether the structure was within 60 meters of a serious event within a given year. The features were based on the long term history of prior events in the area. Since the long term history features performed fairly well, the outcome was a model that predicted fairly well, but the results were not sufficiently good that they could be used for prioritizing repairs. The model had some drawbacks: its predictions were very smooth geographically due to each event being associated with every structure within 60 meters; structures in the same area could not be at extreme ends of the ranked list. In the refinement of the model, we derived a more exact match of events to structures, and used cable and inspection features to make the model more targeted. The intuition of the Con Edison experts was that a structure with no past history of events and very few cables is less vulnerable than an adjacent structure with many cables and events. The simple geographic model could not test this intuition, but the more targeted model described below can do this. In fact, we have since found that nearby structures can vary greatly in vulnerability to events.

Despite its drawbacks, the initial model described above contained all the basic elements of our final model. It used a basic document representation with a link between documents to structures (via geocoding results), a preliminary classification of tickets based on the ticket trouble type, and features representing the past history of structures that captured the "hotspot" theory.

6 Refining the process

We aimed to improve several steps in our process, specifically the document representation, document classification, and the addition of features from other sources. This section focuses on how these parts were strengthened as we incorporated information from the ECS Remarks.

6.1 Ticket classification

The tickets must be classified into three classes: serious events, non-serious events, and non-events. The serious and non-serious events determine the labels and some of the features for the ranking task. For the initial ranking model we discussed in Sect. 5, we used only the

[2]Two proportion z-test: $z > 13$.

trouble type to define which events were serious; however, for instance, smoking manholes (trouble type "SMH," of which there are several thousand) can be very serious, or not serious at all. A more elaborate definition would rely heavily on information extracted from the Remarks.

We developed a working operational definition of a serious event, implemented as a binary function: it assigns a "seriousness" score of 1 if the event is serious, 0 otherwise. We also defined a "candidate precursor" scoring function that sorts non-serious events into those that should be considered as candidate precursor events and those that should be excluded from the model. We refined both scoring functions by eliciting direct feedback from domain experts, and by asking experts to classify tickets, as described below and in our other work (Passonneau et al. 2009). The implementation of this seriousness measure for tickets provided our model with a dramatic increase in accuracy.

The main attributes used in the two scoring functions are the trouble type, various thresholds on the number of free text lines, whether the ticket appears in ELIN, whether the ticket mentions at least one structure, whether the ticket mentions cable sizes specific to the secondary system, and whether the ticket contains serious event metadata. Each ticket was either filtered out before applying the scoring functions, or was assigned to the serious category, or if not serious was assigned to the class of candidate precursor events, or was excluded from the model. We elicited feedback from domain experts on the combinations of attributes used in both scoring functions. But we found that we gained most from a qualitative analysis of a human labeling task, where experts labeled 171 tickets as either representing a serious event, a candidate precursor, or a ticket to be excluded from consideration. After sorting tickets by structure and by time (producing histories for a given structure), 171 tickets were randomly selected from histories with a greater proportion of tickets with serious trouble types (MHX, MHF, MHO and SMH). The results indicate the difficulty of our learning task: the two experts had only modest levels of agreement with each other. We evaluated the agreement among the two labelers using Krippendorff's (1980) Alpha, an interannotator agreement coefficient similar to Cohen's (1960) Kappa. It measures the degree of agreement above chance, given the distribution across the three categories. The agreement among the experts was alpha $= 0.47$, or less than halfway between random behavior (alpha $= 0$) and perfect agreement (alpha $= 1.0$).

After a second pass of labeling in which we had experts adjudicate cases they disagreed on, we used the human labeled data to refine our seriousness score by noting characteristics distinguishing the three categories of tickets on this sample, then confirming the reliability of the characteristics in larger samples. This led to the serious metadata described below in Sect. 6.2.

We used the 171 tickets labeled by the domain experts to evaluate the accuracy and precision of our seriousness score. Our final seriousness score had an accuracy of 91% and precision of 81%. A low baseline accuracy can be computed by taking the trouble type alone as the criterion for seriousness; if we exclude tickets the humans disagreed on, and those that they agreed should be excluded from the dataset, trouble type alone has a rather low accuracy of 38%. During development, we found that improvements in the scoring function as measured against these 171 tickets tracked improvements in the machine learning performance. In sum, subject matter experts could not provide criteria for an operational definition, but the sorting they performed for us proved to be a powerful method for eliciting implicit criteria that we could make explicit.

6.2 Improving the document representation

In this section, we focus on improvements to the document representation leading to improvement on the ticket classification task. Ticket characteristics that replicated expert behavior include ticket length and several types of metadata. One important observation was that experts sometimes excluded tickets that did not represent a distinct non-trivial event, such as tickets with extremely short Remarks (or referred tickets). Thus, in order to filter particular types of noise from the tickets, as well as to classify tickets as serious or potential precursors, we needed to determine threshold "lengths" for the ticket, and also we needed to define metadata to extract from the Remarks.

Free-text lines: In many language processing tasks, length (of documents, of words) provides useful classification features. If the Remarks are long, it generally indicates that substantial repair work had been performed. In order to make the document length more representative of the quantity of information, we eliminated strings of punctuation symbols (often used as line separators) and other noise, and tagged each line to indicate whether it consisted of "free-text" (text entered manually by a dispatcher; e.g., lines 1, 2 and 7 of Fig. 2), or automatically generated lines (e.g., lines 3 through 6 of Fig. 2).

Metadata: In document repositories of all sorts, metadata identifies characteristics of the documents to support tasks such as classification, search or browsing. A library catalog is an example of metadata about holdings, such as author, publication date, genre or topic. Here we use metadata to classify ECS tickets into the three classes discussed above. However, the combination of a large, domain-specific vocabulary with the property that most "words" of any length have *numerous variant spellings* (see Fig. 5) presented an obstacle to extracting more than a few types of metadata.

We assigned four categories of metadata where tickets refer to (1) secondary cable sizes, (2) actual work performed on the structure, (3) indicators that the event was serious, and (4) structure upgrades. For each type of metadata, we collected a set of patterns consisting of a sequence of one or more terms that need not be contiguous. The less constrained the pattern, the more likely it will apply spuriously, thus most of the patterns we have collected are relatively specific, and must apply within a single line (60 characters). The term "cut and rack" indicates that cables in the structure should be made parallel to minimize congestion.

Table 3 lists one or more classes of patterns that must be present for each type of metadata to be assigned to a ticket. The presence of a specific metadata pattern can be context dependent. For example, phrases such as `FOUND MH SMKG HEAVILY` or `STRUCTURE SMOKING LIGHTLY ON ARRIVAL` are reliable indicators that a manhole or other structure was smoking. However, collecting a set of such phrases that apply across the board to all types of tickets is not possible. For example, a report from a customer that a structure is smoking is not necessarily reliable: in some cases, the ticket later indicates that a crew member on site was unable to find evidence of smoke. A report from a fire department crew member who is on site is more likely to be reliable.

Two metadata features that provide complementary information are "cable sizes" and "work performed." Both provide an indicator that repair work took place in a structure, a precondition for classifying a ticket as a relevant secondary event; 40% of all tickets have one or both types of metadata. To maximize the precision of our metadata, we collected distinct sets of line-based patterns, depending on several factors (trouble type, number of lines, presence in ELIN, etc.).

1. Enormous length variation: 1–552 lines
   For the trouble types we investigate here, the maximum length is 552; for other trouble types and other boroughs, the range is greater.
2. Interleaving of manually/automatically entered text: 0 to 380 lines are free text (0%–69% of the entire ticket)
3. Fragmentary and telegraphic language: `OPENED M/L/S TO DROP BLDG AFTER DAMAGED WAS DONE.`
   There is a high frequency of standardized abbreviations (e.g., `B/O`), abbreviations spontaneously generated by the operator (e.g., `COMP` for complete, as in line 8 of Fig. 2), and omission of function words (e.g., prepositions) and punctuation.
4. Specialized terminology: `CRAB, TROUBLE HOLE`
5. Specialized meanings for familiar terms: `BRIDGE, LEGS`
6. Line breaks within words: `AFFECTE/ D`
7. Large vocabulary of approximately 91 K unigram types (distinct alphanumeric sequences; not counting distinct numeric sequences, e.g., structure numbers, dates)
   a) Single most frequent: by ($N = 554{,}614$)
   b) Many singletons ($N = 53{,}647$): arrival, back-feed, loadlugger, . . .
8. Numerous variants per unigram (typically misspellings); e.g., Barricaded
   `BARRICADE | BARRIC | BARRICA | BARRICAD | BARRICADED | BARRICADES | BARRICADING | BARRICATED | REBARRICADE | REBARRICADED | RE-BARRICADED | BARRICADED | BARRICADES | BARRICADE | BARACCADED | BARRACADED | BARACADED | BARICADED | BARRICAD | BARRACCADED | BARRACADE | BARRACADES | BARRICADS | BARACADES | BARACADE | BARICADE | BARRICADING | BARRIACDED | BARRICATED | BARICADES | BARRCADED | BARRICDES | BARRIACADED | BARRICDED | BARRICAED | BARACCADES | BARICADING | BARRACIDED | BARRICACDED | BARRICADEED | BARRICARED | BARRICEDED | BARRICIADED | BARRRICADED | BARRRICADES | BERRICADED | BAARICADED | BARACADDED | BARACCADE | BARBARAICAD | BARICADS | BARICAEDS | BARICATED | BARRACADING | BARRACCAEDED | BARRACDE | BARRACEDED | BARRCADE | BARRCADES | BARRCD | BARRIACADE | BARRIACDE | BARRIACDES | BARRICACADE | BARRICADDED | BARRICADEING | BARRICADSE | BARRICAQDED | BARRICCADED | BARRICCADES | BARRICD'S | BARRICED | BARRIOCADES | BARRRICADE | BRRICADED`

Fig. 5 Characteristics of ECS Remarks field

Table 3 Metadata Pattern classes. "CFR" means "cut for replacement"

Metadata type	Pattern classes	Used to indicate
Cable sizes	`500`, `4/0`	Mention of secondary cable sizes
Work performed	`SHUNT`, `CLEARED`, `CFR`	Ticket where work was performed
Serious event	`FIRE`, `BLOWN`, `SMOKE`, `WIPEOUT`	Ticket describes a serious event
Structure upgrade	`CUT & RACK`	Structure upgrade to be performed

6.3 Finding the trouble hole

One of the main improvements we made to the refined model was to improve the link between documents and instances for the ranking task. During the initial phase of this project (discussed in Sect. 5) we were reluctant to use trouble hole information from the Remarks for several reasons: first, we had difficulty interpreting the Remarks (for instance, finding trou-

ble hole information manually); second, the noise in the Remarks led us to believe that the trouble hole information would not be easy to extract and would not necessarily be accurate; third, we did not expect the Remarks to be comprehensive in recording trouble hole information. However, as it turned out, the ECS Remarks is often the only place to find exact trouble hole information, as in line 7 of Fig. 2: `FELIX REPORTS IN SB325681 CLEARED B/O HOLE STILL NEEDS FLUSH`. Our information extraction code extracts structure information from the ECS Remarks for manholes, service boxes and vaults. This is a non-trivial information extraction task, for instance, the term "Service Box" is represented in at least 38 different variations across ECS Remarks.[3] Our approach was to over-extract structure information and then prune: due to the possibility that the wrong structure number appears within the ticket, we kept only records such that the structure's physical location is at most 200 meters from Columbia Geostan's location for the ticket. Our code extracted structure information from 53.77% (33,194 out of 61,730) of the ECS tickets for the secondary grid in Manhattan. In other words, almost half the tickets are potential noise, since they did not tell us anything about a specific structure. If a structure is mentioned more frequently than any other structure in the text of the ECS Remarks for the event, it was assigned as a trouble hole for the event. We estimate that for the tickets where we have identified a trouble hole, the accuracy of our identification exceeds 87%, based on performance against a known subset.

Additional trouble hole information was incorporated from specialized databases including ELIN, ESR/ENE, and the front of the ECS ticket when relevant.

6.4 Integration of data from other sources

Part of the knowledge discovery process is working with domain experts to obtain the most useful data, that is, data that after cleaning, is useable. There are several departments at Con Edison that keep track of different types of cable data; for instance only some of the departments have moderately reliable cable installation dates and only some of the departments have data that can be better matched to structures. As part of the knowledge discovery process, we went through many attempts to find ways that cable data could be used, based on the strong intuition of the Con Edison experts. One of our initial (failed) attempts to link structures to events was by connecting the address of the event to its closest structure on the grid, namely the structure connected by service cable. Unfortunately, as we discussed earlier, it was an untrue assumption that the nearest structure to an address is the trouble hole for the event, and it was another untrue assumption that the service cable records could be reasonably matched to an address and a structure. The most success we have had with cable data was to use "Property Records" data, which is kept by Con Edison's accounting department. These cable records have at least moderately accurate installation dates, and are able to be joined noisily to structures. Unfortunately this comes at the expense of having to estimate the conductor material (aluminum vs. copper). A raw join of the property records table to the structures table accounts for approximately half of the cables. After brute force cleaning and pattern matching, we were able to match ∼3/4 of the cables to the structures. The effort made to include cable data paid off dramatically, in that features related to the

[3] `SB, S, S/B, S.B, S?B, S.B, S.B., SBX, S/BX, SB/X, S/XB, /SBX, S\BX, S.BX, S.BX, S?BX, S BX, SB X, S B/X, S/B/X, S.B.X, S/BXX, S/BVX, S BOX, S/BOX, S.BOX, S,BOX, S,BOX, S-BOX, XBOX, SVBX, SERV BX, SERV-BOX, SERV.BOX, SERV,BOX, SERV/BOX, SERV/BOXC, SERVICE BOX.`

number of cables are some of the best predictors of serious events, which conforms to the intuition of the experts.

We had a similar experience obtaining and cleaning inspection and vented cover data, although the payoff from these data was not as great as that from cable data for several reasons, including the newness of the inspection and vented cover programs.

7 Structure ranking

Rare event prediction tasks are usually formulated as classification problems. In this case, we chose a *bipartite ranking* formulation, as defined by Freund et al. (2003). The formulation requires a set of examples with binary labels (the same setup as for binary supervised classification tasks). The bipartite ranking algorithm produces a scoring function that minimizes a chosen objective, and the examples are rank-ordered by the values of the scoring function. In this case, the objective is essentially a weighted version of the area under the ROC curve (AUC) (Bradley 1997) that favors the top of the ranked list. The resulting algorithm, which is described in earlier work (Rudin 2009), performs better than others we have tested on the features and labels discussed below, including support vector machine classifiers and pruned decision trees.

Formally, given examples $\{(x_i, y_i)\}_{i=1\ldots m}$, where $x_i \in \mathcal{X}$ (the set of manholes) with labels $y_i \in \{-1, +1\}$ (not vulnerable/vulnerable), and features $\{h_j\}_{j=1\ldots n}$, $h_j : \mathcal{X} \to \mathcal{R}$, the P-Norm Push algorithm (Rudin 2009) minimizes $R_p(\lambda)$, where λ is a vector of coefficients that defines the scoring function f:

$$R_p(\lambda) := \sum_{k:y_k=-1} \left(\sum_{i:y_i=1} e^{-[f(x_i)-f(x_k)]} \right)^p, \quad \text{where } f(x) = \sum_j \lambda_j h_j(x).$$

The value of p is chosen by the user, depending on how much it is desired to concentrate at the top of the list. When $p = 1$, the algorithm is equivalent to the commonly used RankBoost algorithm (Freund et al. 2003), which does not concentrate at the top of the list, and $R_1(\lambda)$ is a convex proxy for the AUC.[4] For large p, $R_p(\lambda)$ is related to objectives for information retrieval that focus mainly on the top of the list. Using large p had the effect of modeling the more vulnerable structures in greater detail.[5]

7.1 Labels

The label y_i is $+1$ if the structure i was the trouble hole for a serious event during the time period specified for prediction. For instance, if we are trying to predict events in 2005, the

[4] In the case of no ties, the misranking error relates to the AUC as follows, where $\mathcal{I} = \{i : y_i = 1\}$ and $\mathcal{K} = \{k : y_k = -1\}$:

$$(1 - \text{AUC})|\mathcal{I}||\mathcal{K}| = \sum_{i\in\mathcal{I}} \sum_{k\in\mathcal{K}} \mathbf{1}_{[f(x_i)<f(x_k)]} \leq R_1(\lambda).$$

[5] A related algorithm that concentrates at the top of the list, called the IR-Push (Rudin 2009), was also used here and is derived similarly to the P-Norm Push. The IR-Push optimizes:

$$R_{gIR}(\lambda) := \sum_{i:y_i=1} \ln\left(1 + \sum_{k:y_k=-1} e^{-[f(x_i)-f(x_k)]}\right), \quad \text{where } f(x) = \sum_j \lambda_j h_j(x).$$

.

label y_i is $+1$ if the structure i was the trouble hole for a serious event during 2005. There is no "ground truth" for this supervised ranking problem since the labels depend on the result of both the ticket classification task and the trouble hole estimation. Hence we used experts' judgments as the ground truth for a blind prediction test as discussed in Sect. 8.

7.2 Features

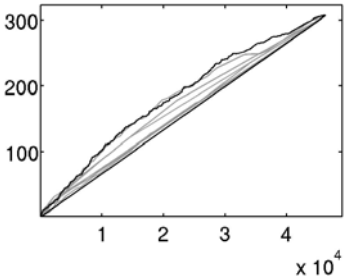
We have developed approximately one hundred features falling into three categories. The first and third category were derived relative to the time frame of prediction. If we aimed to predict serious events in 2005, the features were derived from records prior to 2005. The categories are:

- Features based on past events, specifically, based on a history of ECS tickets in the nearby area. For example, one important feature is $h(x_i) =$ the number of events within the most recent 3 years for which the ith structure was a trouble hole. Another feature is the number of past precursor events within a 60 meter radius of the structure.
- Features based on cable data. Some of the cable features include the number of phase main cables, phase service cables, number of cables with specific cable size, total number of cables in structures that are within 60 meters.
- Features based on inspection history. Features include an indicator for whether the structure has been inspected, the number of Tier 1 repairs completed during past inspections, and whether a cut and rack is possibly pending (which is a Tier 2 suggested repair). We have not found the inspection features to be useful for prediction, in the sense that their individual performance is about as bad as a random guess for future prediction of serious events. We believe this is because there are only a few years worth of inspection reports (the inspection program started in 2004) and also these data are very noisy.

Since we only had a snapshot of the cables in Manhattan at a single recent point in time, the cable features were derived at the time of the snapshot, under the assumption that the snapshot approximately represents the state of the grid in the past. This is not an unreasonable assumption for phase cables (current carrying cables), but we have found that it is likely to be an unreasonable assumption for neutral cables since cable replacement bundles tend to have larger number of neutrals. This was discovered as a result of investigating a high correlation between structures having manhole events and structures possessing a larger number of neutral cables. Neutral cables are thus not used to derive our present set of features. A similar anti-correlation occurs for aluminum cables, which are replaced with copper cables after an event. At this point, we do not know whether a structure had an aluminum cable at the time of an event unless it has not been replaced.

A combination of basic feature selection methods were used to streamline the model. Following the terminology of Kohavi and John (1997) (also see the review of Guyon and Elisseeff 2003), we implemented a filter method that ranks individual features according to performance and relevance criteria, where the AUC was used as the performance criterion. Features categorized as irrelevant possessed a correlation between the feature and labels that was determined not to be causal (e.g., number of neutral cables, number of aluminum cables, number of inspections). Features eliminated due to the performance filter criterion included the inspection features and the features based on past electric shock (ESR) and energized equipment (ENE).[6] When enough irrelevant features were eliminated so that the algorithm

[6]We already knew that the inspection data were noisy, but the finding about ESR/ENE data was puzzling. To clarify, Con Edison engineers explained that many of the ENE events are not Con Edison responsibility,



(a) ROC Curves: True Positives vs. False Positives

Feature	AUC
1	0.5590
2	0.5159
3	0.5457
4	0.5171
5	0.5840
6	0.5159
7	0.5135

(b) AUC Values

Fig. 6 (**a**) Testing ROC Curves for Model and Individual Features. The model was trained on data through 2005 and tested on 2006. The *darker curve* is the ROC for the model, *lighter curves* are for the individual features. (**b**) AUC values for the features, which were used in the feature selection process

could be run repeatedly, we adapted a simple wrapper method for backwards elimination of features. Features were individually eliminated based on whether they increased the accuracy on the training set.

The model that has been most successful, as judged by a combination of predictive power, sparsity in the number of features, and a sufficiently meaningful collection of features, contains the following features:

1. **Total Mentions**: The number of past events (either precursor events or serious events) in which the structure was mentioned in the Remarks. The structure is not necessarily the trouble hole for the event.
2. **Total Mentions Recently**: The number of events (either precursor events or serious events) in which the structure was mentioned in the Remarks within the past 3 years.
3. **Number of Times Structure was the Trouble Hole**: The number of past events (either precursor events or serious events) in which the structure was a trouble hole, as determined by the method discussed in Sect. 6.3.
4. **Number of Times Trouble Hole Recently**: The number of events (either precursor events or serious events) in which the structure was a trouble hole within the past 3 years.
5. **Number of Main Phase Cables**.
6. **Number of Service Phase Cables**.
7. **Number of Service Cables Installed Between 1960 and 1969**.

All features were normalized to the interval $[0, 1]$. In order to test the model, we trained it to predict serious events in 2005 and measured its performance on events in 2006. The Mann-Whitney U-test performed on individual features yields p-values below 10^{-4} for both training and testing, and the AUC values for the features (with respect to the labels) are shown in Fig. 6. It is worth noting that the 1960's service cable feature, which has the lowest AUC of the features used in the model, may be a proxy for aluminum cables; however, preliminary results indicate that this feature does not appear to perform as well in Brooklyn. However, all of the other features perform well in Brooklyn.

The main phase cable feature performs better than the other features towards the middle and bottom of the list, but the features based on past mentions in ECS tickets perform better at the top of the list.

and that ECS did not distinguish between those two categories; in fact, it is logical that there is no apparent correlation of ESR/ENE's with manhole events.

The ranking model was combined with a correction for factors that are not statistically predictive due to the nature of the data, yet are important to domain experts. These factors include the presence of aluminum cables, average cable age, inspection results such as cut and racks and solid manhole covers. The correction was included mainly to avoid the possibility of low-ranked high-vulnerability structures, that is, structures that have no statistically valid factors to warrant a high rank, but instead possess qualities that may cause a domain expert to consider the structure as potentially vulnerable.

8 Results of blind prediction test

Since the ECS tickets have no ground truth interpretation, we worked with the domain experts to derive an acceptable method of evaluation for the model. We chose a *blind* evaluation, in which a years' worth of ECS data were withheld from our database and classified manually by experts. The experts identified a small subset of trouble holes for serious events (ELIN events, only fires and explosions, no smoking manholes) in 2007. We did not independently verify that these tickets represent serious events. There were 44 trouble holes provided by the experts. Out of the top 5000 structures in our ranked list, 9 of them had events in 2007. This means that the top 10% of structures contained 20% of the events. The probability of a result this good or better via random guessing is approximately 2.4%.[7] Furthermore, out of the top 1000 structures, 5 of them had events in 2007. This means that the top 2% of the ranked list contained 11% of the trouble holes for serious events. The probability of randomly achieving a result this good or better is on the order of a tenth of a percent. Again, there is no gold standard, but we believe these results are extremely good given the difficulty of this modeling problem. Furthermore, this would not be an easy modeling problem even with "clean" data; many of the trouble holes for serious events exhibit no warning signs or precursors.

9 Conferencing tools

The communication gap between the domain experts at Con Edison and the team of Columbia scientists was not bridged overnight. Particularly, it took a long time for us to comprehend the experts' overall perspective on the state of the power grid and how the data should be interpreted with respect this view. At the same time, we needed to convey our methods well enough to the Con Edison experts in order to elicit the most helpful feedback. We discuss two of the most useful conferencing tools built for the project: the structure profiling tool, and the visualization tool.

9.1 Structure profiling tool

The structure profiling tool developed for the project (Radeva et al. 2009) generates an automatic summary of raw and processed data. The aim is to display everything a domain expert might use to manually evaluate the vulnerability of a given structure.

[7] This p-value is given by a sum over the right tail of the hypergeometric distribution $(9, \ldots, 44)$ with population size 51219, containing 44 manholes with a serious event and $n = 5000$ observations. Similarly the probability of achieving 5 or more successes within the top thousand is 0.0016.

Table 4 Profile of ECS tickets that explicitly mention SB 116741. Columns are: ticket number, date received, trouble type and ELIN trouble type if the ticket appeared in ELIN, trouble hole (indicator "x" appears for tickets where this structure is the trouble hole for the event), total number of free-text lines in the ECS Remarks, indicator for the presence of structure upgrade metadata, indicator for work performed metadata, and an indicator for serious event metadata

Ticket	Date	Type/ELIN	TH	FLines	C&R	Shunt	Meta
ME051004556	2005-03-02	SIP	(*)	3	*		
ME041018969	2004-12-22	ACB	*	11			
ME041006012	2004-03-18	SMH/SMH	*	51			*
ME041005915	2004-03-17	UDC	*	10		*	
ME031002011	2003-02-10	SMH/SMH	*	3	*	*	*
ME031000147	2003-01-03	SMH/SMH	*	34		*	*
ME021015893	2002-12-25	SMH/SMH	*	27	*	*	*
ME021011838	2002-09-21	SO	*	28		*	*
ME001000528	2000-01-21	MHO/MHO		21		*	
ME991005175	1999-04-24	ACB		10		*	
ME981006807	1998-08-05	LV	*	94			

Consider a structure on the upper east side of Manhattan, denoted by SB 116741. It was ranked within the top 500 structures out of 51219 total structures in Manhattan (which is the top 1% of the ranked list), for prediction of a manhole event in 2007. The structure experienced a smoking manhole event in 2008, meaning the high rank was a relevant prediction. Several factors contributed to its high rank: the history of past events that we will discuss, the large number of cables (only 2% of service boxes have more cables), and the inspection results that suggest future repairs.

Table 4, which was obtained from the structure profiling tool, provides a summary of the tickets that mention SB 116741, and Table 5 provides excerpts from these tickets' Remarks showing the context in which the structure was mentioned. In these excerpts the trouble hole information and relevant metadata ("`CLEARED`," "`BLOWN`," "`SMOKING`," "`C&R`") are highlighted. The trouble hole information in the Remarks is noisy: the first ticket wrongly lists *MH* 116741 as the trouble hole (luckily we were able to link the ticket with SB 116741 via a different source). Viewing the tickets associated to a particular manhole, as in Table 4, also helped us to convey to domain experts that there are structures with an impressively long history of past events.

SB 116741 allows us to demonstrate the noisiness of the inspections data. SB 116741 needs a cut and rack, as discussed in three tickets, including one companion "SIP" ticket for an inspection report. However, the inspections summary in Table 6 shows that the checkbox for "cut and rack" was not checked in any inspection report (including the one corresponding to the SIP ticket as shown in Tables 4 and 5). Further, we have no record of whether the cut and rack was actually performed. We suspect this type of noise is a reason why the inspection data did not produce the kind of powerful predictive features as, for instance, those obtained from cable data.

Due to the enormous scale of the Manhattan grid, its gradual growth over time, and the reactive cable replacement policy, there may be several different generations of cables within the same manhole. Table 7 shows the profile of cables we were able to match to a manhole in Chelsea, MH 465022, which was ranked 20th in the full list, mainly because of its large number of cables (in the top 0.4% of all structures). The oldest cable in the structure was installed in the 1920's, and many cables were replaced in 2004 (most likely due to an

Table 5 Excerpts of remarks data for SB 116741. Most of these quotes indicate that repair work was performed. Note that F/O means "front of," and "C-F-R" means "cut for replacement." "Flush" refers to the structure being cleaned out by a special flush truck to allow a repair crew member to enter the structure

ME051004556	VERA, EMP109302, REPORTS NEEDED REPAIRS WERE FOUND DURING INSPECTION OF MH116741 ON 12/23/04. CUT & RACK REQUIRED AND MAIN NEEDS TO BE CUT FOR REPLACEMENT.
ME041018969	12/22/04 11:15 D.WOODS/UG-SPLICER REPORTS FOUND MULTIPLE B/O'S IN SB-116741 F/O 2-4 E 99ST.AKA 1151 PARK AVE.ALSO HAVE BURNT NEUTRALS ON A 8W GOING EAST.---------->VF
ME041006012	03/18/04 08:20 FD#508 REPORTS SMOKING MANHOLE, FD ON LOC. FIRE DEPT AT THIS LOCATION WITH 2 STRUCTURES SMOKING HEAVY..... FOUND SB-116739 F/O# 1-7 E 99 ST AND SB-116741 F/O# 2-4 E 99
ME041005915	MAIURO ALSO REPORTS S/B116741 F/O 2-4 E 99ST HAS TROUBLE ON SECONDARY MAINS... THIS COULD BE THE TROUBLE FOR 6 E 99ST.....JFM
ME031002011	02/10/03 08:49 FDNY REPORTS AT E.99ST & 5AVE A SMOKING MANHOLE.....AA A 500 MAIN GOING NORTH. CLEARED B/O . SERVICE O.K. ATT. AREA 3 HOLE NEEDS C&R
ME031000147	01/03/03 VELEZ REPORTS THAT SB-116741 F/O 2 E99 ST.IS A 3 COVER GRTG. SMOKING & BARRICADED...CO READINGS AS FOLLOWED:
ME021015893	12/26/02 03:15 PREZUTO REPORTS FOUND SB116741 F/O 2-4 E 99 ST SMOKING LIGHTLY, COVER ON VENTED, WILL BARRICADE : FOUND MULTI B/O ON SERV. TO BUILDING 1151. CHECKED BASEMENT AND FOUND ALL READI NGS NORMAL. IN S/B #116741 CUT AND CLEARED 1-B/O ON 1-DC SERV ICE LEG GOING TO BUILDING. ALSO CUT AND CLEARED 1-AC SERV.
ME021011838	FERRARO ALSO REPORTS WOODS FOUND A BLOWN LIMITER & MULTIPLE B/O'S IN SB-116741 F/O 2-4 E.99 STREET...WOODS REPLACED BLOWN LIMITER & IS IN THE PROCESS OF CLEARING B/O'S.....VR ==========ATTENTION UNDERGROUND========== K.FERRARO NO.9 O/S REPORTS SB-116741 F/O 2-4 E.99 STREET HAS TO BE CUT & RACKED A.S.A.P. DUE TO 5W5W CRABS WIHT BLOWN LIMITERS & MAIN THAT ARE NOT RACKED PROPERLY(SPAGHETTI HOLE)
ME001000528	NYPD STATES A MANHOLE COVER IS MISSING ON 99TH ST BTWN 5TH A V & MADISON AVE. POLICE UNIT STANDING BY : TRBL SB-116739 OPP 2-4 E.99ST....3-COVER GRATING WITH ALL COVERS IN STRUCTURE.....ALSO SB-116741 F/O 2-4 E.99ST COVER AJAR.....SKREPKOWICZ REPORTS FOLLOWING "CO" READINGS: : ALSO IN SB-116741 F/O 2-4 E.99 ST ALL B/O CLEARED ------MB
ME991005175	ALSO HAS C-F-R 3-500NL,2-4/0,4"38 FROM SB-116739 F/O 1-7 E.99 ST TO SB-116741 F/O 2 E.99 ST .....
ME981006807	PF MARCADO ELEC PH 718-999-6401 ---ONLY 190 VOLTS ON 1 PHASE --AC ELEVATORS ---- COMPUTER RUN NEEDS FULL VOLTAGE---ETS,ES : 08/06/98 01:00 BRONSON REPORTS FLUSH IN PROGRESS IN SB116741 UNDERGROUNDIN PROCESS OF TIEING OPEN MAINS.....

Table 6 Inspection profile for SB 116741. The first column provides the date of inspection. The T1A and T1B columns list whether any Tier 1A or Tier 1B repairs were needed during the inspection. Cut & Rack is a Tier 2 suggested repair, meaning cables need to be cut and made parallel. Main and Service replacements are also Tier 2 suggested repairs

Date	T1A	T1B	Cut&rack needed	Main replacement	Service replacement
2004-12-22	F	T	F	F	F
2004-12-23	F	F	F	T	F
2005-03-16	F	F	F	F	F
2005-04-21	F	F	F	F	F
2005-04-26	F	F	F	F	F
2005-05-05	F	F	F	F	F
2005-05-06	F	F	F	F	F

underground direct current "UDC" event that year). This type of structure illustrates our earlier point that a completed repair within a structure does not imply immunity to future events.

Structure profiles of highly ranked structures revealed two general categories of structures: those with a long history of serious events such as SB 116741, and structures with a large number of cables (including older cables) such as MH 465022. We did not find structures exhibiting mainly older cables and a lot of past serious events. This is not mysterious: structures with many past serious events have cable replacements and thus mostly newer cables.

9.2 Visualization tool

A method for viewing data and results was an essential tool for gaining intuition regarding the underlying geospatial trends, judging the success of our machine learning models, determining density-of-event estimates and identifying hotspots visually. For instance, this tool was helpful for viewing the structures with respect to the surrounding buildings, and in terms of the cable layout along the streets. A preliminary version of our visualization work was presented by Dutta et al. (2008).

The visualization tool, which is made available to users via a website that connects to our server, interfaces with our database and with Google Earth locally. The architecture is illustrated in Fig. 7. To use the tool, users first need to specify a region to be displayed using their local version of Google Earth. Following this, the server connects to our database and retrieves the information to be displayed, specifically the tickets, structures, and cables within the region. The user is provided with a file containing the display that opens in Google Earth.

Figure 8 shows events, cables and structures near MH 465022. The top image shows geocoded tickets, which are colored based on their trouble type (yellow for manhole events, purple for burn-outs). Users can click on a yellow or purple dot and read the Remarks for the corresponding ticket. For tickets that are geographically close together (or at the same address), the locations of the dots have been slightly jittered to make them all visible. The bottom image shows the structures colored by rank and the main cables as lines between structures. The color of the line represents the number of main cables connecting the two structures, where green indicates fewer cables and blue indicates more cables. The structures are colored red to white from most vulnerable to least vulnerable. Figure 8 immediately demonstrates why it was important to create a more targeted model than our initial

Table 7 Profile of cables matched to MH 465022. Each row is part of a bundle of cables. The first two columns indicate which structures are connected by the bundle ("From" and "To" assigned arbitrarily). The third column is the number of cables in the part of the bundle, followed by the cable size (500, 4/0, 200, 350), the installation year, the type of insulation, and whether the cable is a main or service cable, where main cables connect two structures, service cables connects one structure to a building. MH 465022 has all main cables. The codes for the insulation material is as follows: "RN" stands for RUBBER & NEOPRENE, "RL" for RUBBER & LEAD, "PL" for PAPER & LEAD, and "BB" for BARE & BARE, which is the insulation type listed for neutral cables. All other cables are phase cables

To structure	From structure	#Cables	Size	Installed	Insulation	Main/serv
M465022	SB45080	1	4/0	1939	BB	Main
M465022	SB45080	6	4/0	1950	RN	Main
M465022	SB45080	2	4/0	1950	BB	Main
M465022	M46498	2	4/0	2004	BB	Main
M465022	M46498	3	500	2004	RN	Main
M465022	SB46501	6	4/0	1958	RN	Main
M465022	SB46501	3	500	2008	RN	Main
M465022	SB46501	2	4/0	2008	BB	Main
M465022	SB46501	2	4/0	1958	BB	Main
M465022	M45050	2	4/0	1961	BB	Main
M465022	M45050	2	4/0	1961	BB	Main
M465022	M45050	6	4/0	1961	RN	Main
M465022	M45050	6	4/0	1961	RN	Main
M553721	M465022	3	500	2004	RN	Main
M553721	M465022	1	4/0	1939	BB	Main
M553721	M465022	2	4/0	1953	BB	Main
M553721	M465022	2	4/0	2004	BB	Main
M553721	M465022	3	500	2004	RN	Main
M553721	M465022	2	4/0	1997	BB	Main
M553721	M465022	1	4/0	1939	BB	Main
M553721	M465022	2	4/0	2004	BB	Main
M553721	M465022	2	4/0	2004	BB	Main
M553721	M465022	3	500	1997	RN	Main
M553722	M465022	1	4/0	1939	BB	Main
M553721	M465022	3	500	2004	RN	Main
M553722	M465022	1	4/0	1950	BB	Main
M553721	M465022	3	500	2004	RN	Main
M553721	M465022	2	4/0	2004	BB	Main
M553721	M465022	6	4/0	1953	RN	Main
M465022	M46498	3	250	1928	PL	Main
M465022	SB45080	6	4/0	1939	RL	Main
M465022	SB45080	3	200	1931	RL	Main
M465022	M45110	3	4/0	1940	RL	Main
M465022	M45110	1	350	1940	RL	Main
M553721	M465022	6	4/0	1939	RL	Main
M553722	M465022	6	4/0	1950	RL	Main
M553721	M465022	6	4/0	1939	RL	Main
M553722	M465022	6	4/0	1939	RL	Main
M553722	M465022	3	4/0	1939	RL	Main

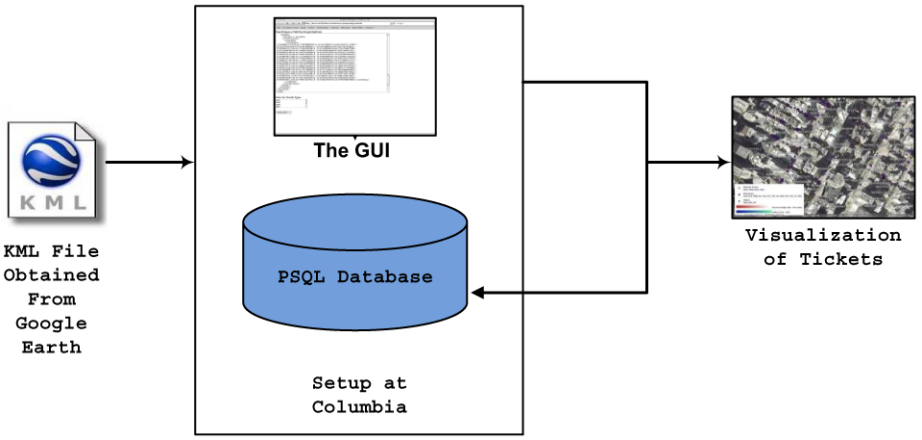


Fig. 7 Architecture of the visualization tool. First, Google Earth 4.2 (beta) is used locally to generate a Keyhole Markup Language (KML) file representing a polygon. After the KML file arrives at our website, Java servlets that reside on our Apache Tomcat 5.5 server connect to our PostgreSQL 8.2 database and retrieve ECS tickets, structures, and cables within the polygon. Finally, the information is displayed using Google Earth. In order to accomplish this, a .zip file is downloaded to the default directory set by the browser. The zipped directory contains images of dots needed for the display, KML files specifying the location of each event, a KML file specifying the locations of the structures, and a KML file specifying the cable locations

geographic model; in fact there are structures in very close proximity where one structure has a high rank and another has a low rank.

When viewing the top image in Fig. 8, note that the ECS ticket location, which is a street address that has been geocoded (possibly with noise), can be many meters from the trouble hole for that event. In other words, a cluster of events at one intersection may not necessarily implicate a structure at that intersection.

10 Conclusion

We have demonstrated the application of statistical machine learning techniques to a real world problem where the available knowledge sources were extremely raw. This led to the development of a knowledge discovery and data mining process for general problems of this type, where classification of text documents facilitates ranking of domain entities. We started this project with a large quantity of disparate, noisy (and at the time, unintelligible) data, with no guarantee or clear indication that this data could be useful for prediction, and a murky goal of predicting "serious" events.

In our favor, we had a multidisciplinary team and the benefit of domain experts who were committed to the project and open to providing feedback by means of our conferencing tools. The final targeting model, developed after many iterations of the knowledge discovery and data mining process, was predictive as well as meaningful to the domain experts. Statistical results on a blind test, plus results from the conferencing tools, indicate that our model is able to pinpoint vulnerable structures, and thus to make useful recommendations. The targeting model we produced is currently being used to prioritize future inspections and repairs in Manhattan's secondary electrical grid, is being evaluated for use in Brooklyn, and is being extended to all other boroughs.

Fig. 8 *Top*: ECS tickets located near MH 465022. *Bottom*: Structures and Cables near MH 465022. A *white arrow* indicates the position of MH 465022

Our experience shows that researchers who are naive about a given domain can assemble a useful database from extremely raw data when the problem definition drives the process of knowledge formulation, when the researchers and domain experts find successful methods of communicating about the entities in the domain, and when experts' intuitions are thoroughly tested in order to separate the wheat from the chaff. This bodes well for the prospect of increasing collaboration between academia and industry.

Acknowledgements This work was sponsored by a grant from Con Edison. We would like to thank the following members of the Columbia CCLS: Roger Anderson, Dave Waltz, Nandini Bhardwaj, Diego Fulgueira, Ashish Tomar, Zhi An Liu, Nancy Burroughs-Evans, Daniel Alicea, Hatim Diab, Sam Lee, Fred Seibel, Charles Collins, Sara Stolbach, and Jawwad Sultan. From Con Edison, we would like to thank Serena Lee,

Robert Schimmenti, Tracy Cureton, Won Choe, Artie Kressner, Maggie Chow, Aaron Prazan and Stavros Livanos. Thanks to the reviewers and editor for their time, effort and helpful comments.

Appendix A: Geocoding

"Columbia Geostan" is our geocoding system for ECS tickets. It provides a latitude and longitude for each ticket. The geocoding process is non-trivial due to misspellings and irregularities such as missing house numbers and Con Edison-specific terminology such as "opp#250" (for opposite house 250), "int" (intersection), "s/e/c" (southeast corner), and ranges of house numbers or streets, for instance, "181-182 St." Columbia Geostan semi-automatically standardizes the ECS address fields, allowing us to use a standard geocoding service (such as Google Earth's free service; Google Earth 2009) to obtain geographic coordinates for each address. In the cases where the house number is useless or missing, we use the intersection of the street and cross street; this occurs commonly, in fact the intersection is used for approximately 15% of tickets.

The four main stages of Columbia Geostan are: cleaning the address and cross street fields using specialized regular expressions; sending two queries to Google Earth's geocoder for each ticket (one for the street address and one for the intersection with the cross street); parsing the html output; and scoring each suggestion using an independent scoring criterion. The scoring criterion takes into account the match between the query we sent and Google's response, and prefers exact street addresses to intersections. The suggestion receiving the highest score is selected as the final address for that ticket.

This system yielded a latitude and longitude for approximately 97% of ECS ticket addresses in Manhattan; for the tickets in Fig. 2, Columbia Geostan system provides the corrected addresses: '120 E Broadway' with latitude = 40.713882 and longitude = −73.992441, and 'Cedar St. & Greenwich St.' with latitude = 40.7097100 and longitude = −74.0128400.

Appendix B: Referred tickets

Multiple ECS tickets can refer to the same event, because there can be multiple reports made to the Con Edison call center regarding the same event. Since we wish to consider distinct events (rather than distinct tickets), it is essential to understand the relationships between these tickets. We define a "referred ticket" as an ECS ticket that refers to another ECS ticket (which we call the "lead ticket"). The lead ticket contains information describing the event, while a referred ticket simply cross references the lead ticket and contains little or no information about the event. In order to find candidate referred tickets, we search for

```
Ticket: ME03114094
Remarks:
09/01/03 12:15 FDNY/242 REPORTS SMOKING MANHOLE F/O 2236
7AVE. CREW REQUESTED. _____________________________   =-->
TD
09/01/03 17:15 DUPLICATED TO ME03114093 BY 13151
------------- ELIN COMPLETE-----------------
REFER TO ME03114093 FOR INFO
09/01/03 17:19 REFERRED TO: MH.INCID EDSSMH FYI BY 13151
09/01/03 22:36 COMPLETED WITH JCRF SCREEN BY 13151
```

Fig. 9 Sample ECS Remarks for a referred ticket

synonyms for "refer to ticket" such as "duplicated to ticket" or "see ticket." However, a ticket might reference another ticket because the events are distinct but related, often with the same keyword phrases. We do not want to exclude related tickets. The logic to identify genuine referred tickets, such as the one in Fig. 9, depends on many factors such as the length of the ticket, context, and internal corrections to lines within the ticket, etc.

Our code finds that 7503 of the total 61,730 tickets are referred. Of the referred tickets we've found, 88.11% (6264 out of 7109) of them have a time difference of less than 1 day between the referred ticket and corresponding lead ticket, and 97.55% (6935 out of 7109) have a time difference of less than a week from the lead ticket.

References

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. In *Proceedings of the IADIS European conf. data mining* (pp. 182–185).

Becker, H., & Arias, M. (2007). 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 '07)* (pp. 86–94). New York: ACM.

Boriah, S., Kumar, V., Steinbach, M., Potter, C., & Klooster, S. A. (2008). Land cover change detection: a case study. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'08)* (pp. 857–865). New York: ACM.

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

Castano, R., Judd, M., Anderson, R. C., & Estlin, T. (2003). Machine learning challenges in Mars rover traverse science. In *Workshop on machine learning technologies for autonomous space applications, international conference on machine learning*.

Chen, G., & Peterson, A. T. (2002). Prioritization of areas in China for the conservation of endangered birds using modelled geographical distributions. *Bird Conservation International*, *12*, 197–209.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: a general framework and some examples. *IEEE Computer*, *37*(4), 50–56.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). 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'02)*.

Devaney, M., & Ram, A. (2005). Preventing failures by mining maintenance logs with case-based reasoning. In *Proceedings of the 59th meeting of the society for machinery failure prevention technology (MFPT-59)*.

Dudík, M., Phillips, S. J., & Schapire, R. E. (2007). Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, *8*, 1217–1260.

Dutta, H., Rudin, C., Passonneau, R., Seibel, F., Bhardwaj, N., Radeva, A., Liu, Z. A., & Ierome S, Isaac, D. (2008). 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.

Fayyad, U., & Uthurusamy, R. (2002). Evolving data into mining solutions for insights. *Communications of the ACM*, *45*(8), 28–31.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*, 37–54.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: an overview. *AI Magazine*, *13*(3), 57–70.

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, *4*, 933–969.

Google Earth (2009). http://www.earth.google.com.

Grishman, R., Hirschman, L., & Nhan, N. T. (1986). Discovery procedures for sublanguage selectional patterns: initial experiments. *Computational Linguistics*, 205–215.

Gross, P., Boulanger, A., Arias, M., Waltz, D. L., Long, P. M., Lawson, C., Anderson, R., Koenig, M., Mastrocinque, M., Fairechio, W., Johnson, J. A., Lee, S., Doherty, F., & Kressner, A. (2006). Predicting electricity distribution feeder failures using machine learning susceptibility analysis. In *Proceedings of*

the eighteenth conference on innovative applications of artificial intelligence IAAI-06, Boston, Massachusetts.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, *157*(3), 317–356.

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

Harris, Z. (1982). Discourse and sublanguage. In Kittredge, R., & Lehrberger, J. (Eds.) *Sublanguage: studies of language in restricted semantic domains* (pp. 231–236). Berlin: de Gruyter.

Hirschman, L., Palmer, M., Dowding, J., Dahl, D., Linebarger, M., Passonneau, R., Lang, F., Ball, C., & Weir, C. (1989). The PUNDIT natural-language processing system. In *Proceedings of the annual AI systems in government conference* (pp. 234–243).

Hsu, W., Lee, M. L., Liu, B., & Ling, T. W. (2000). 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 '00)* (pp. 430–436). New York: ACM.

Jiang, R., Yang, H., Zhou, L., Kuo, C. C. J., Sun, F., & Chen, T. (2007). Sequence-based prioritization of nonsynonymous single-nucleotide polymorphisms for the study of disease mutations. *American Journal of Human Genetics*, *81*(2), 346–360.

Kirtley, J. Jr., Hagman, W., Lesieutre, B., Boyd, M., Warren, E., Chou, H., & Tabors, R. (1996). Monitoring the health of power transformers. *IEEE Computer Applications in Power*, *9*(1), 18–23.

Kittredge, R. (1982). Sublanguages. *American Journal of Computational Linguistics*, 79–84.

Kittredge, R., Korelsky, T., & Rambow, O. (1991). On the need for domain communication knowledge. *Computational Intelligence*, *7*(4), 305–314.

Kohavi, R., & John, G. (1997). Wrappers for feature selection. *Artificial Intelligence*, *97*(1–2), 273–324.

Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*. Beverly Hills: Sage.

Kusiak, A., & Shah, S. (2006). A data-mining-based system for prediction of water chemistry faults. *IEEE Transactions on Industrial Electronics*, *53*(2), 593–603.

Liddy, E. D., Symonenko, S., & Rowe, S. (2006). Sublanguage analysis applied to trouble tickets. In *Proceedings of the Florida artificial intelligence research society conference* (pp. 752–757).

Linebarger, M., Dahl, D., Hirschman, L., & Passonneau, R. (1988). Sentence fragments regular structures. In *Proceedings of the 26th association for computational linguistics*, Buffalo, NY.

Murray, J. F., Hughes, G. F., & Kreutz-Delgado, K. (2005). Machine learning methods for predicting failures in hard drives: a multiple-instance application. *Journal of Machine Learning Research*, *6*, 783–816.

National Institute of Standards and Technology (NIST), Information Access Division (ACE) Automatic Content Extraction Evaluation. http://www.itl.nist.gov/iad/mig/tests/ace/.

Oza, N., Castle, J. P., & Stutz, J. (2009). Classification of aeronautics system health and safety documents. *IEEE Transactions on Systems, Man and Cybernetics, Part C*, *39*, 670–680.

Passonneau, R., Rudin, C., Radeva, A., & Liu, Z. A. (2009). 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)*.

Patel, K., Fogarty, J., Landay, J. A., & Harrison, B. (2008). Investigating statistical machine learning as a tool for software development. In *Proceedings of ACM CHI 2008 conference on human factors in computing systems (CHI 2008)* (pp. 667–676).

Radeva, A., Rudin, C., Passonneau, R., & Isaac, D. (2009). Report cards for manholes: eliciting expert feedback for a machine learning task. In *Proceedings of the international conference on machine learning and applications*.

Rudin, C. (2009). 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.

Sager, N. (1970). The sublanguage method in string grammars. In R. W. Ewton Jr. & J. Ornstein (Eds.), *Studies in language and linguistics*, University of Texas at El Paso (pp. 89–98).

Steed, J. (1995). Condition monitoring applied to power transformers-an REC view. In *Second international conference on the reliability of transmission and distribution equipment* (pp. 109–114).

Symonenko, S., Rowe, S., & Liddy, E. D. (2006). Illuminating trouble tickets with sublanguage theory. In *Proceedings of the human language technology/North American association of computational linguistics conference*.

Vilalta, R., & Ma, S. (2002). Predicting rare events in temporal domains. In *IEEE international conference on data mining* (pp. 474–481).

Weiss, G. M., & Hirsh, H. (2000). Learning to predict extremely rare events. In *AAAI workshop on learning from imbalanced data sets* (pp. 64–68). Menlo Park: AAAI Press.