3.2 APPLICATION OF DECISION SUPPORT METHODS TO WEATHER SENSITIVE OPERATIONS

Richard Domikis*, Jessica L. Scollins, Michael Glanzmann, and Leonard Bisson

The Boeing Company, Mission Systems Division, Springfield, VA 22153

ABSTRACT

While the science of weather is an evolving and challenging subject, the effective application of weather data is equally challenging and often where the true value of weather information and decision aids provides significant benefit.

As with any solution, once deployed and effective, there is a desire to maintain and even improve operational savings. As the weather data industry has become more commercialized, one result is a competitive and option-rich environment from which prospective weather data users can select. A key factor in selecting and using vendor data is ensuring operational advantages can be realized.

This paper describes a proof-of-concept project we have recently completed. In this example we have applied data mining techniques to improve operational performance of an industrial system that uses multi-vendor frequent weather data for current and next day decisions. The results from this initial analysis are encouraging. We have found areas where marked improvements appear possible as well as interesting weather vendor specific trends and nuances that can be avoided to use to the customer's advantage.

1. INTRODUCTION

Data Mining (DM) is typically described as computer-assisted data analysis applied to enormous sets of data with the goal of extracting knowledge from the data. More simply, it is the process of searching for relevant patterns. Results from data mining can, if provided, be key factors in future decisions. DM extends techniques such as statistical analysis, information retrieval, and known phenomenon modeling with machine learning and pattern recognition to gain knowledge over a particular data set [1].

Data mining tools help discover patterns that ultimately are used to predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions.

Pattern recognition technologies augment classical statistical and mathematical techniques.

Data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed. For businesses, data mining is used to discover patterns and relationships in the data in order to help make better strategic and tactical decisions.

Data mining can help spot trends in data that can help improve decisions and avoid errors based on previously "unexpected" events [2].

Data mining is unique compared to other methods of investigation. DM seeks out hidden information that experts may miss because it lies outside their specialty or expectations. Other investigative techniques usually rely on a Subject Matter Expert (SME) to prove a suspected hypothesis. This SME is searching for specific evidence to support or refute the hypothesis. However, with data mining, analysis is often run without a hypothesis and the findings are completely outside of the typical search realm and often outside common suspicions. The goal is to extract subtle but significant data patterns not evident during human analysis alone.

Without DM technology it would be extremely difficult and require massive amounts of manpower, time and money to perform equivalent analysis on large or diverse data sets. Time and effort required, in addition to potential human error, imply DM is not only beneficial but necessary for certain problems.

There are numerous DM methods such as artificial neural networks, decision trees, rule induction, genetic algorithms, and nearest neighbor classifications. Artificial neural networks are non-linear predictive models that often learn through training on similar or sample data. These DM methods resemble biological neural networks in structure. Decision trees are tree-shaped

Corresponding Author Address: Richard Domikis, The Boeing Company Mission Systems; email: richard.domikis@boeing.com

structures that represent sets of decisions that generate rules for the classification of a dataset. Rule induction is the extraction of useful if-then rules from data based on statistical significance. Genetic algorithms are optimization techniques based on the concepts of genetic combination, mutation and natural selection. Finally, nearest neighbor is a classification technique that classifies each data point based on other data points most similar to it in an historical context [3].

Just as there are many methods used in data mining, there are many tools that facilitate the creation and operation of these DM methods. Such tools include SPSS' Clementine©, and similar tools from SAS, Oracle, and IBM. There also exists a CRoss Industry Standard Process for Data Mining called CRISP-DM. CRISP-DM is a data mining process model that describes commonly used approaches that expert data miners use to tackle problems [4].

Many industries, if not all, would benefit from the application of data mining tools especially if done while following the CRISP-DM process. This paper focuses on the weather community and the value of data mining on industries that use weather data in everyday operations. Currently, in the weather community, data mining is used in areas such as frontal forecasting, severe weather precursor event analysis, cumulous cloud detection, and mesocyclone signatures. Weather is often viewed as a special parameter since its importance in a variety of seemingly unrelated industries challenges common understanding and solutions.

For example, data mining has been successfully used by this team to discover sources of unexpected next-day temperature variations. In the energy or power industry weather plays a significant role as an input to load forecasting and thus power-generation decisions. Weather also plays a role in the transportation industry, airlines and rail systems make critical decisions everyday where weather variables are a critical input. The U.S. government makes decisions such as ordering ships to sea when a port is threatened with severe weather. In every case, decisions require support and DM provides actionable knowledge that allows faster and better decisions to be made. This paper discusses the application of data mining using a proven process on weather data in support of industry and decision support operations.

2. CRISP-DM PROCESS

CRISP-DM is a CRoss Industry Standard for Data Mining which was launched in September 1996 and was founded by three of the earliest companies to utilize data mining for business: Daimler-Benz, SPSS (then ISL), and NCR. The CRISP-DM premise was to create a standard process that was non-proprietary, application & industry and tool neutral. This enabled the data mining process to focus on business issues, create a framework for guidance, and create templates for analysis from past proven experiences. The importance of developing a standard process in the data mining industry was to give data mining users a model that could be used on any data mining project. The process model was developed by an interest group of over 200 participants from diverse industries that included technology, financial, and retail companies. These participants presented their views on data mining from project experience and a common system model (CRISP-DM) was created. The CRISP-DM process contains six steps: problem understanding, data understanding, data preparation, modeling, evaluation, and deployment. Our team groups these six steps into three easy to remember phases: Understanding, Design, and Use [5]. Figure 1 shows the CRISP-DM methodology.

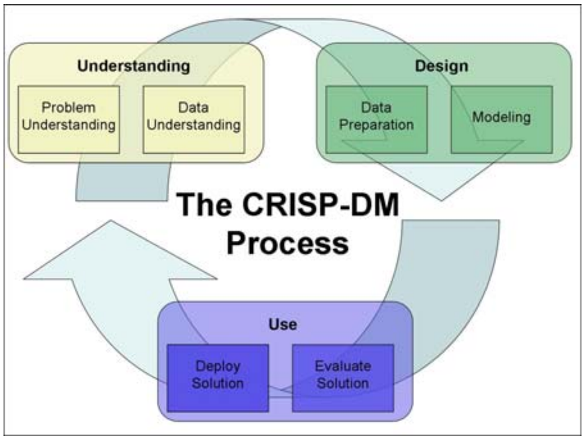


Figure 1 The CRISP-DM Methodology

2.1 *Phase 1: Understand*

The Understand phase is characterized by two steps: Understanding the Problem and Understanding the Data. The goal in this phase is to become familiar with the problem domain as well as the targeted areas for improvement. Understanding the Data consists of gathering available data and understanding the sources and format of the data. The following sections discuss these steps in greater detail.

2.1.1 Understanding the Problem

In order to effectively apply data mining to a problem, the problem must be understood. While this concept seems obvious, it is amazing how many times research is poorly focused. The first two steps in the CRISP-DM process focus on understanding; understanding the problem and understanding the data. It is important and often more difficult than many suspect, to stay focused on providing improvements that can be effectively applied to operations. Users know what they currently do and often believe they know what areas need to be improved. Those are important but not conclusive facts that must be gathered and considered during this period of understanding. In this phase of the process the goal is to understand what is being done as well as what has been tried or excluded. A fundamental goal of data mining is to find patterns that improve success over conventional methods. Success must be understood and not artificially limited to an intermediate variable. In one of our studies, the customer initially stated that next day temperature was an area that required improvement. While this belief was valid, during our discussion it was determined that true benefit was in reducing the number of next day temperature aberrations. The eventual success criteria was not forecasting next day temperature or just improving the average, it was forecasting large aberrations. The value of interacting with the customer was that it gave an accurate understanding of where improvements would help operations.

2.1.2 Understand the Data

Similar to understanding the problem is the need to understand the data. This includes the data currently used, data previously used, data not used and even includes “local” data viewed as not relevant. In many cases, data mining finds relevant patterns in "non-relevant" data. More data is always better, but this doesn't mean that all of the data will be used. During this step the team should seek to identify all the locally relevant sources of data. In the case of data that is currently or likely to be used, understanding extends to the actual content of the data. The team should seek to answer questions such as: Is the data complete? What period does the data cover? Does the data contain errors or omissions? These are all questions that help evaluate the data’s potential value and support the next step in the design phase activity called data preparation. Understanding data also extends to data sets outside the domain of the current user. Once the problem is understood, and the current and local data are understood there are at least two other data sources that should be considered. First look to the current users for ideas and "hunches". In many cases the end users will quietly state an important observation such as "I'm not sure but it seems like this data is always bad on normal days”. In one of our studies that clue resulted in the discovery of a data source that you would not want to use in an answer but provided a fine indicator of future aberrations. Subject Matter Experts (SMEs) are the second external source of data. By considering SMEs and theories on the fringe of the domain it is possible to find additional supporting patterns. Much of research is tied up in long periods of data collection and confidence building. While that is good and proper science, data mining is not necessarily looking for explanations it is looking for patterns hence theories are often viable even if the SMEs aren't yet sure of their validity or cause.

2.2 Phase 2: Design

In the design phase the key is to obtain and prepare data as well as select and train appropriate DM methods. These activities are often iterative and that is one reason why understanding the data and what is available is so important. The current solution is a starting point for the design phase, the goal is not to improve what is being done but improve decision support at the end of the operational chain. This is why factors beyond the current solution are an important part of this process. CRISP-DM is designed to provide a process that results in improved value in the form of better knowledge based decisions. Artificially limiting the design to only what is currently done implies the solution is contained within that data and that there is enough data to find it. That is not a success path for data mining but a path that is often followed [6].

2.2.1 Preparing the data

During the Understand phase, an initial set of data sources were identified and chosen for use. Preparing data is complex because of the variety of forms of data ownership: the data may be owned by the system, be accessed externally or created by the system. You must decide if you are going to use all the data, some of the data, certain vendor data, etc. The chosen data must be consolidated and parameterized and checked for formatting equivalency such as units, notation, number of data points, etc. For example, in areas where you may be missing data or are sure you have areas with "bad" data values you may choose to set those values to represent the mean of all "good" values for that parameter, or you may choose to exclude that time step of data completely. Once you have correctly formatted data it is ready for modeling. During the model step that follows, data may be added, removed or altered hence data preparation is an iterative process. Only iteration and experience will help to decide the correct choices of data and data preparation.

2.2.2 Modeling the problem:

The Modeling step is where most believe the value of DM resides. There is some validity in this belief due to the fact that it is in this area where the benefits of data mining, such as pattern recognition and discovery, first appear. Modeling is an activity that looks at the needs, data, and current approach in order to select the methods most likely to improve decisions and operations. It does not imply the creation of artificial or substitute behaviors. With a strong starting point the goal is to select directions in both the data and methods that will yield improvements. There is no hard or fast rule for method selection. In part method selection is based on the source of the undiscovered pattern. This is the very root of DM, to find a pattern one must look at the data in the context of improving a parameter. There are clues and prior knowledge that may suggest one analysis method is typically better than another. Sparse or robust data may encourage or discourage certain methods. In general the goal is to find rules to group data into one or more of the following pattern areas: Predict, Classify, Segment, Associate, Sequence and Detect Outliers. Each of these areas, coupled with types and volumes of data, will suggest more and less successful DM methods. To accommodate specific and formatting requirements of these techniques, data must frequently be recycled back into the preparation step. A flexible and robust tool that accommodates numerous analysis methods is the key to success. SPSS' Clementine is one of these tools and offers users a wide variety of methods that can be quickly applied and altered.

2.3 Phase 3: Use

Use is the last phase in the CRISP-DM process. The Use phase is made up of two stages: Evaluation and Deployment. The Evaluation stage involves inspecting the results/patterns obtained from the Modeling stage and evaluating those results against the specified needs. Up until this point, there is little to confirm that the DM approach will produce actionable knowledge. DM must provide better information in a form and timeframe that allows it to be effectively used. Use includes both test and operational deployment.

2.3.1 Evaluation

As with the previous phases, the third and last phase of CRISP-DM also contains two steps: testing the concept and deploying the solution. In the Evaluation step, data has been analyzed and a model or models that produce potential benefits are evaluated. This is one of the most critical stages in the entire CRISP-DM methodology since a thorough evaluation of not only the model(s), but also every step leading up to the construction of the model(s) must occur. To save time and ensure success the results must be confirmed and evaluated against historical and operational data whenever possible. In the simplest of terms when using data to recognize patterns there is always a risk the pattern is merely an artifact of that data set and not indicative of the real world. If you were to build an "averaging engine" and confirm it worked on a data set then a subsequent test with the same data does not rule out model errors. Here again robust tools help a great deal, the ability to sequester data away from training early so to preserve data for testing is an important and often overlooked aspect in DM tools.

Before the deploying the solution step can occur, it must be confirmed that targeted operations have been adequately supported and that any additional issues have been addressed. By testing against the end user needs the deployment of a viable and beneficial solution is ensured. At the completion of this segment, the critical decision must be made as to whether to completely, partially or not deploy the solution. In cases where the decision is not to deploy the solution often the decision is to spend more time improving the model or finding additional data sources.

2.3.2 Deployment

The title of this step can be a little misleading. Although deploying the solution is the final step of the final phase of the CRISP-DM methodology, the deployment of the DM model is typically not the end of the project. The most important aspect of the Deployment step is to implement a solution that is easily understandable, usable, and most importantly repeatable. This is a point that must be emphasized because either partially implemented or completely implemented solutions are often the inputs to other data mining activities. It is crucial to document details and procedures not only about the model(s) properties but also how to use them in the future. The primary strength of DM is rapid pattern recognition and obtuse pattern recognition when conventional means have reached the point of diminishing returns. Unlike conventional statistical analysis, DM discovers patterns that are both continuous and fleeting. Data mining cannot be viewed as a static or stable solution. Some patterns will be stable and other patterns will change over time. This implies that improvements are at times dynamic and must be maintained. For example, in one project we found differences and potential improvements in vendor data. Since this multi-source data was not generated by the user, it is reasonable to expect that the sources of the data are continuously seeking to improve “their” data and hence business. In this study, we found a poor data source that was a great indicator of aberrations in other data sources. To expect a vendor to continue to provide such a poor source is risky but not to take advantage of it while it exists is inefficient. Deployments must include operational trending, self assessments and safety valves that ensure when models trend away from performance the team knows and can respond. Ideally, deployed solutions accommodate these trends adaptively, but there are limits when patterns or sources change.

3. PATHFINDER STUDY

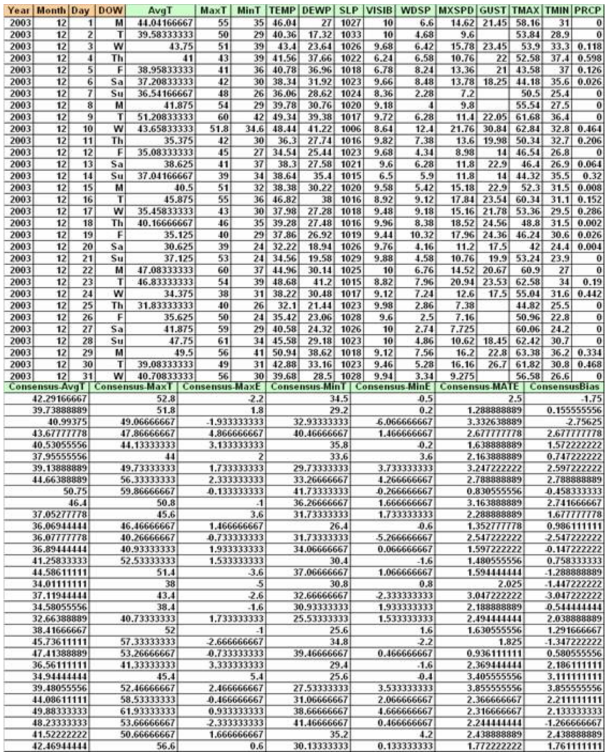
This team has worked a number of pathfinder studies to determine the magnitude of potential benefits to demonstrate that data mining is relevant and applicable to real-world weather users. While these studies are far from complete or exhaustive each has shown clear improvements that can be refined and deployed in operations. In the following study the team applied SPSS’ Clementine data mining tool suite to improve next day temperature forecasts with a key focus on identifying and avoiding large error days (aberrations). Two years of historical data was used to conduct this analysis. The intent was to explore how data mining techniques can supplement and enhance next day temperature forecasts.

3.1 Understanding

During the Understand phase, it was determined that outlier days, those when the temperature forecast was in error by several degrees, tended to be the most important. As a result, we established a success criteria based on an ability to predict 24 hours in advance when these high error days would occur. This became the pathfinder study for this customer and to explore the applicability of data mining to this data. For the Data Preparation step, we consolidated historical database files, derived additional fields including vendor bias and forecast error magnitude, and collected additional meteorological data to supplement the error modeling. Upon completion of the Data Preparation step, Boeing then input the data into several classification models to determine patterns and forecast the magnitude of the next day temperature errors. The descriptive statistics and data mining models were created using a professional COTS software package called Clementine, created by SPSS, Inc. We have also applied other statistical analysis tools to investigate trends and patterns. This limited effort confirmed there are areas that will yield improvements and automation through data mining efforts at the operational level.

3.2 Design

Prior to generating descriptive statistics on the temperature forecasts and error tendencies, the historical data first had to be prepared for analysis. The first step was to consolidate the hourly-daily temperature and other relevant data. Next, several additional fields from the existing metrics, such as vendor bias, the forecast error binning (e.g. high, medium, or low), and prior day histories, etc. were created. Based on visual inspection and binning of known data, logical groupings within the error distribution were generated, 0-2 ^{o}F was designated as low error, 2-4 ^{o}F as medium error, and greater than 4 ^{o}F as high error. These bins account for both data patterns as well as operational value. Thousands of bins or bins in tenths of a degree might be supported by data but would not lend themselves to operational use. The key was to analyze and design with end-use in mind. Finally, additional meteorological parameters (e.g., dew point, wind speed, weather, etc.) for the area of analysis were independently collected from NOAA’s National Climatic Data Center’s (NCDC) Global Summary of Day database to further supplement the error modeling. Figure 2 below provides examples of the consolidated data used in this DM analysis.

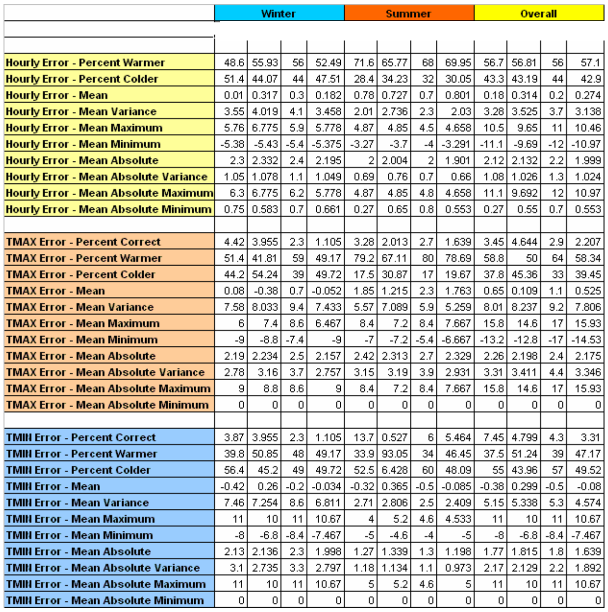


Year	Month	Day	DOW	AvgT	MaxT	MinT	TEMP	DEWP	SLP	VISIB	WDSP	MXSPD	GUST	TMAX	TMIN	PRCP
2003	12	1	M	44.04166667	55	35	46.04	27	1027	10	6.6	14.62	21.45	58.16	31	0
2003	12	2	T	39.58333333	50	29	40.36	17.32	1033	10	4.68	9.6		53.84	28.9	0
2003	12	3	W	43.75	51	39	43.4	23.64	1026	9.68	6.42	15.78	23.45	53.9	33.3	0.118
2003	12	4	Th	41	43	39	41.56	37.66	1022	6.24	6.58	10.76	22	52.58	37.4	0.598
2003	12	5	F	38.95833333	41	36	40.78	36.96	1018	6.78	8.24	13.36	21	43.58	37	0.126
2003	12	6	Sa	37.20833333	42	30	38.34	31.92	1023	9.66	8.48	13.78	18.25	44.18	35.6	0.026
2003	12	7	Su	36.54166667	48	26	36.06	28.62	1024	8.36	2.28	7.2		50.5	25.4	0
2003	12	8	M	41.875	54	29	39.78	30.76	1020	9.18	4	9.8		55.54	27.5	0
2003	12	9	T	51.20833333	60	42	49.34	39.38	1017	9.72	6.28	11.4	22.05	61.68	36.4	0
2003	12	10	W	43.65833333	51.8	34.6	48.44	41.22	1006	8.64	12.4	21.76	30.84	62.84	32.8	0.464
2003	12	11	Th	35.375	42	30	36.3	27.74	1016	9.82	7.38	13.6	19.98	50.34	32.7	0.206
2003	12	12	F	35.08333333	45	27	34.54	25.44	1023	9.68	4.34	8.98	14	46.54	26.8	0
2003	12	13	Sa	38.625	41	37	38.3	27.58	1021	9.6	6.28	11.8	22.9	46.4	26.9	0.064
2003	12	14	Su	37.04166667	39	34	38.64	35.4	1015	6.5	5.9	11.8	14	44.32	35.5	0.32
2003	12	15	M	40.5	51	32	38.38	30.22	1020	9.58	5.42	15.18	22.9	52.3	31.5	0.008
2003	12	16	T	45.875	55	36	46.82	38	1016	8.92	9.12	17.84	23.54	60.34	31.1	0.152
2003	12	17	W	35.45833333	43	30	37.98	27.28	1018	9.48	9.18	15.16	21.78	53.36	29.5	0.286
2003	12	18	Th	40.16666667	46	35	39.28	27.48	1016	9.96	8.38	18.52	24.56	48.8	31.5	0.002
2003	12	19	F	35.125	40	29	37.86	26.92	1019	9.44	10.32	17.96	24.36	46.24	30.6	0.026
2003	12	20	Sa	30.625	39	24	32.22	18.94	1026	9.76	4.16	11.2	17.5	42	24.4	0.004
2003	12	21	Su	37.125	53	24	34.56	19.58	1029	9.88	4.58	10.76	19.9	53.24	23.9	0
2003	12	22	M	47.08333333	60	37	44.96	30.14	1025	10	6.76	14.52	20.67	60.9	27	0
2003	12	23	T	46.83333333	54	39	48.68	41.2	1015	8.82	7.96	20.94	23.53	62.58	34	0.19
2003	12	24	W	34.375	38	31	38.22	30.48	1017	9.12	7.24	12.6	17.5	55.04	31.6	0.442
2003	12	25	Th	31.83333333	40	26	32.1	21.44	1023	9.98	2.86	7.38		44.82	25.5	0
2003	12	26	F	35.625	50	24	35.42	23.06	1028	9.6	2.5	7.16		50.96	22.8	0
2003	12	27	Sa	41.875	59	29	40.58	24.32	1026	10	2.74	7.725		60.06	24.2	0
2003	12	28	Su	47.75	61	34	45.58	29.18	1023	10	4.86	10.62	18.45	62.42	30.7	0
2003	12	29	M	49.5	56	41	50.94	38.62	1018	9.12	7.56	16.2	22.8	63.38	36.2	0.334
2003	12	30	T	39.08333333	49	31	42.88	33.16	1023	9.46	5.28	16.16	26.7	61.82	30.8	0.468
2003	12	31	W	40.70833333	56	30	39.68	28.5	1028	9.94	3.34	9.275		56.58	26.6	0

Consensus-AvgT	Consensus-MaxT	Consensus-MaxE	Consensus-MinT	Consensus-MinE	Consensus-MATE	ConsensusBias
42.29166667	52.8	-2.2	34.5	-0.5	2.5	-1.75
39.73888889	51.8	1.8	29.2	0.2	1.288888889	0.155555556
40.99375	49.06666667	-1.933333333	32.93333333	-6.066666667	3.332638889	-2.75625
43.67777778	47.86666667	4.866666667	40.46666667	1.466666667	2.677777778	2.677777778
40.53055556	44.13333333	3.133333333	35.8	-0.2	1.638888889	1.572222222
37.95555556	44	2	33.6	3.6	2.163888889	0.747222222
39.13888889	49.73333333	1.733333333	29.73333333	3.733333333	3.247222222	2.597222222
44.66388889	56.33333333	2.333333333	33.26666667	4.266666667	2.788888889	2.788888889
50.75	59.86666667	-0.133333333	41.73333333	-0.266666667	0.830555556	-0.458333333
46.4	50.8	-1	36.26666667	1.666666667	3.163888889	2.741666667
37.05277778	45.6	3.6	31.73333333	1.733333333	2.288888889	1.677777778
36.06944444	46.46666667	1.466666667	26.4	-0.6	1.352777778	0.986111111
36.07777778	40.26666667	-0.733333333	31.73333333	-5.266666667	2.547222222	-2.547222222
36.89444444	40.93333333	1.933333333	34.06666667	0.066666667	1.597222222	-0.147222222
41.25833333	52.53333333	1.533333333	30.4	-1.6	1.480555556	0.758333333
44.58611111	51.4	-3.6	37.06666667	1.066666667	1.594444444	-1.288888889
34.01111111	38	-5	30.8	0.8	2.025	-1.447222222
37.11944444	43.4	-2.6	32.66666667	-2.333333333	3.047222222	-3.047222222
34.58055556	38.4	-1.6	30.93333333	1.933333333	2.188888889	-0.544444444
32.66388889	40.73333333	1.733333333	25.53333333	1.533333333	2.494444444	2.038888889
38.41666667	52	-1	25.6	1.6	1.630555556	1.291666667
45.73611111	57.33333333	-2.666666667	34.8	-2.2	1.825	-1.347222222
47.41388889	53.26666667	-0.733333333	39.46666667	0.466666667	0.936111111	0.580555556
36.56111111	41.33333333	3.333333333	29.4	-1.6	2.369444444	2.186111111
34.94444444	45.4	5.4	25.6	-0.4	3.405555556	3.111111111
39.48055556	52.46666667	2.466666667	27.53333333	3.533333333	3.855555556	3.855555556
44.08611111	58.53333333	-0.466666667	31.06666667	2.066666667	2.366666667	2.211111111
49.88333333	61.93333333	0.933333333	38.66666667	4.666666667	2.316666667	2.133333333
48.23333333	53.66666667	-2.333333333	41.46666667	0.466666667	2.244444444	-1.266666667
41.52222222	50.66666667	1.666666667	35.2	4.2	2.438888889	2.438888889
42.46944444	56.6	0.6	30.13333333	0.133333333	1.772222222	1.761111111

Figure 2 Snapshot of Consolidated 2 Yr Analysis Data

Descriptive statistics were generated for each data source (Vendor A, Vendor B and Vendor C) as well as consensus to assess overall forecast performance and to look for tendencies and biases. Figure 3 depicts a summary of the descriptive statistics for hourly mean relative and absolute error (the daily mean of hourly errors) and the mean relative and absolute error in the min and max daily temperature forecasts. The statistics are broken out by winter versus summer since these appeared more susceptible to load volatility and would offer the most operationally useful insights.



	Winter				Summer				Overall			
Hourly Error - Percent Warmer	48.6	55.93	56	52.49	71.6	65.77	68	69.95	56.7	56.81	56	57.1
Hourly Error - Percent Colder	51.4	44.07	44	47.51	28.4	34.23	32	30.05	43.3	43.19	44	42.9
Hourly Error - Mean	0.01	0.317	0.3	0.182	0.78	0.727	0.7	0.801	0.18	0.314	0.2	0.274
Hourly Error - Mean Variance	3.55	4.019	4.1	3.458	2.01	2.736	2.3	2.03	3.28	3.525	3.7	3.138
Hourly Error - Mean Maximum	5.76	6.775	5.9	5.778	4.87	4.85	4.5	4.658	10.5	9.65	11	10.46
Hourly Error - Mean Minimum	-5.38	-5.43	-5.4	-5.375	-3.27	-3.7	-4	-3.291	-11.1	-9.69	-12	-10.97
Hourly Error - Mean Absolute	2.3	2.332	2.4	2.195	2	2.004	2	1.901	2.12	2.132	2.2	1.999
Hourly Error - Mean Absolute Variance	1.05	1.078	1.1	1.049	0.69	0.76	0.7	0.66	1.08	1.026	1.3	1.024
Hourly Error - Mean Absolute Maximum	6.3	6.775	6.2	5.778	4.87	4.85	4.8	4.658	11.1	9.692	12	10.97
Hourly Error - Mean Absolute Minimum	0.75	0.583	0.7	0.661	0.27	0.65	0.8	0.553	0.27	0.55	0.7	0.553
TMAX Error - Percent Correct	4.42	3.955	2.3	1.105	3.28	2.013	2.7	1.639	3.45	4.644	2.9	2.207
TMAX Error - Percent Warmer	51.4	41.81	59	49.17	79.2	67.11	80	78.69	58.8	50	64	58.34
TMAX Error - Percent Colder	44.2	54.24	39	49.72	17.5	30.87	17	19.67	37.8	45.36	33	39.45
TMAX Error - Mean	0.08	-0.38	0.7	-0.052	1.85	1.215	2.3	1.763	0.65	0.109	1.1	0.525
TMAX Error - Mean Variance	7.58	8.033	9.4	7.433	5.57	7.089	5.9	5.259	8.01	8.237	9.2	7.806
TMAX Error - Mean Maximum	6	7.4	8.6	6.467	8.4	7.2	8.4	7.667	15.8	14.6	17	15.93
TMAX Error - Mean Minimum	-9	-8.8	-7.4	-9	-7	-7.2	-5.4	-6.667	-13.2	-12.8	-17	-14.53
TMAX Error - Mean Absolute	2.19	2.234	2.5	2.157	2.42	2.313	2.7	2.329	2.26	2.198	2.4	2.175
TMAX Error - Mean Absolute Variance	2.78	3.16	3.7	2.757	3.15	3.19	3.9	2.931	3.31	3.411	4.4	3.346
TMAX Error - Mean Absolute Maximum	9	8.8	8.6	9	8.4	7.2	8.4	7.667	15.8	14.6	17	15.93
TMAX Error - Mean Absolute Minimum	0	0	0	0	0	0	0	0	0	0	0	0
TMIN Error - Percent Correct	3.87	3.955	2.3	1.105	13.7	0.527	6	5.464	7.45	4.799	4.3	3.31
TMIN Error - Percent Warmer	39.8	50.85	48	49.17	33.9	93.05	34	46.45	37.5	51.24	39	47.17
TMIN Error - Percent Colder	56.4	45.2	49	49.72	52.5	6.428	60	48.09	55	43.96	57	49.52
TMIN Error - Mean	-0.42	0.26	-0.2	-0.034	-0.32	0.365	-0.5	-0.085	-0.38	0.299	-0.5	-0.08
TMIN Error - Mean Variance	7.46	7.254	8.6	6.811	2.71	2.806	2.5	2.409	5.15	5.338	5.3	4.574
TMIN Error - Mean Maximum	11	10	11	10.67	4	5.2	4.6	4.533	11	10	11	10.67
TMIN Error - Mean Minimum	-8	-6.8	-8.4	-7.467	-5	-4.6	-4	-5	-8	-6.8	-8.4	-7.467
TMIN Error - Mean Absolute	2.13	2.136	2.3	1.998	1.27	1.339	1.3	1.198	1.77	1.815	1.8	1.639
TMIN Error - Mean Absolute Variance	3.1	2.735	3.3	2.797	1.18	1.134	1.1	0.973	2.17	2.129	2.2	1.892
TMIN Error - Mean Absolute Maximum	11	10	11	10.67	5	5.2	4.6	5	11	10	11	10.67
TMIN Error - Mean Absolute Minimum	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3 Descriptive Statistics from Multiple Sources

In addition to generating a baseline statistical analysis for the two year period, we also developed an example, illustrative data mining model to predict day-ahead temperature forecast error (categorized as either "high", "medium", or "low" in magnitude) based on vendor forecasts and the other meteorological parameters mentioned before and confirmed with truth – known after the fact.

A C5.0 model was chosen primarily since the team believed it would present patterns it finds as human intelligible “rules” and thus would offer the easiest interpretation and a “white-box” view into the model for this study. Other models are easily implemented by substituting a regression or neural network node for the C5.0 node and re-executing the model stream. The goal was to apply and review findings in this pathfinder not to optimize DM benefits. Early analysis suggested that seasonal behavior was not consistent across sources; some performed better in the summer others in the winter. This is a classic pattern discovery. Once discovered many would say, “Of course”, yet prior to discovery the consensus did not account for this behavior. As mentioned earlier the cause of these discrepancies and patterns are not always understood and certainly not guaranteed to persist. Vendors will change solutions and behaviors often without notice.

While this can be viewed as a risk in DM solutions it is also a benefit. A deployed solution can be viewed also as a monitor of change.

Figure 4 depicts one resulting C5.0 data modeling stream in Clementine used to predict the consensus forecast error magnitude (“ConsensusBin”) for summer. In this case, we decided to partition the source data using a 50-50 split when running the model, meaning 50% of the dataset was randomly chosen for training the model while the other 50% was sequestered for testing. A 50-50 partition was used to keep the model from being over-trained and allow test on data with known truth. Weather solutions have the benefit of high availability of historical data (known truth). DM can be applied on a variety of data, many of which have unclear truth. Truth, even after the fact is challenging to confirm because outside influences as well as the discovered pattern may be affecting behavior. In this example we know what the actual value was within a known threshold. As long as we work within that accuracy threshold, DM applied to weather is one of the stronger uses of DM.

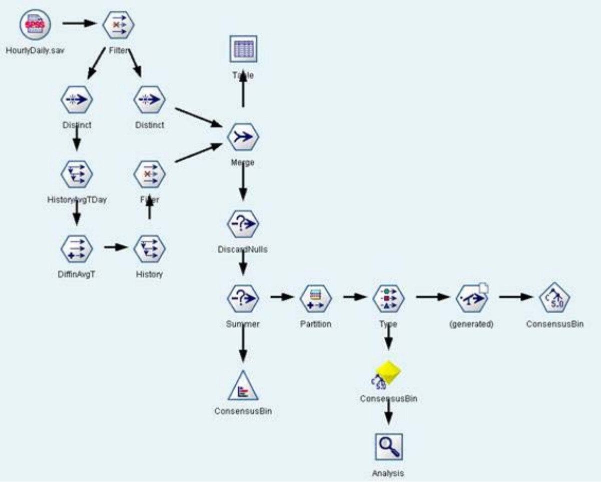
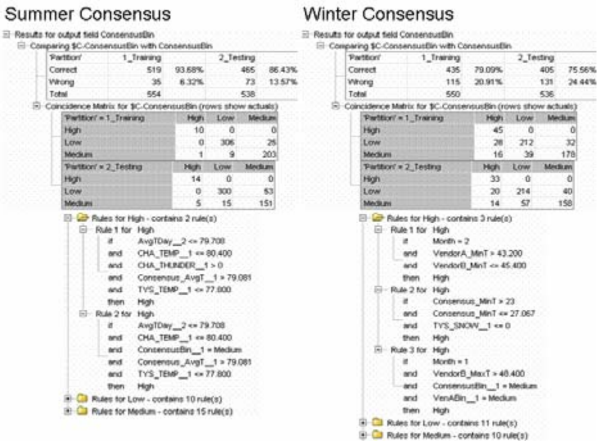


Figure 4 Clementine C5.0 data model stream predicting consensus temperature forecast error (“ConsensusBin”) for summer

3.3 Use

Figure 5 depicts model results summarizing overall performance based on processing data with known truth. For summer, the model correctly predicted the next day error category in 86% of the cases. For winter, the model correctly predicted the next day error category in 75% of the cases. The coincidence matrices show the breakdown for each of the error categories, with columns showing predictions and rows the actual. To illustrate, for the summer test data the model correctly predicted 14 of 14 high error cases (100%), 300 of 353 medium error cases (85%), and 151 of 171 low error cases (88%). It should be noted that the total number of cases exceeds the total number of days. This is due to the fact the database was artificially boosted to create a larger sample size since there are only two seasons worth of data to model with for summer and winter, respectively. This demonstrates the need and value of “deep” data sources when possible.



Summer Consensus

Results for output field ConsensusBin

Comparing $C-ConsensusBin with ConsensusBin

'Partition'	1_Training		2_Testing	
Correct	519	93.68%	465	86.43%
Wrong	35	6.32%	73	13.57%
Total	554		538	

Coincidence Matrix for $C-ConsensusBin (rows show actuals)

'Partition' = 1_Training	High	Low	Medium
High	10	0	0
Low	0	306	25
Medium	1	9	203
'Partition' = 2_Testing	**High**	**Low**	**Medium**
High	14	0	0
Low	0	300	53
Medium	5	15	151

```
Rules for High - contains 2 rule(s)
    Rule 1 for  High
        if      AvgTDay__2 <= 79.708
        and     CHA_TEMP__1 <= 80.400
        and     CHA_THUNDER__1 > 0
        and     Consensus_AvgT__1 > 79.081
        and     TYS_TEMP__1 <= 77.800
        then    High
    Rule 2 for  High
        if      AvgTDay__2 <= 79.708
        and     CHA_TEMP__1 <= 80.400
        and     ConsensusBin__1 = Medium
        and     Consensus_AvgT__1 > 79.081
        and     TYS_TEMP__1 <= 77.800
        then    High
Rules for Low - contains 10 rule(s)
Rules for Medium - contains 15 rule(s)
```

Winter Consensus

Results for output field ConsensusBin

Comparing $C-ConsensusBin with ConsensusBin

'Partition'	1_Training		2_Testing	
Correct	435	79.09%	405	75.56%
Wrong	115	20.91%	131	24.44%
Total	550		536	

Coincidence Matrix for $C-ConsensusBin (rows show actuals)

'Partition' = 1_Training	High	Low	Medium
High	45	0	0
Low	28	212	32
Medium	16	39	178
'Partition' = 2_Testing	**High**	**Low**	**Medium**
High	33	0	0
Low	20	214	40
Medium	14	57	158

```
Rules for High - contains 3 rule(s)
    Rule 1 for  High
        if      Month = 2
        and     VendorA_MinT > 43.200
        and     VendorB_MinT <= 45.400
        then    High
    Rule 2 for  High
        if      Consensus_MinT > 23
        and     Consensus_MinT <= 27.067
        and     TYS_SNOW__1 <= 0
        then    High
    Rule 3 for  High
        if      Month = 1
        and     VendorB_MaxT > 48.400
        and     ConsensusBin__1 = Medium
        and     VenABin__1 = Medium
        then    High
Rules for Low - contains 11 rule(s)
Rules for Medium - contains 10 rule(s)
```

Figure 5 Error model results for Consensus – summer and winter

Figure 5 also lists the rules the model identified for predicting consensus high error days. For summer, Rule 1 states that if the actual average daily temperature for the two days prior (AvgTDay_2) was less than or equal to 79.7 degrees and the prior day average temperature (CHA_TEMP_1) was less than or equal to 80.4 degrees and the location experienced thunderstorms on the previous day (CHA_THUNDER_1 > 0) and the consensus average forecast temperature the previous day was greater than 79 degrees and the average actual temperature at another location for the previous day was less than or equal to 77.8 degrees then the next-day error will be high.

3.4 Results/Observations

The following are some high-level observations mined from both the descriptive statistics and the error modeling analysis. A more thorough study will undoubtedly yield more detailed observations than those noted here.

1) Temperature forecast errors and variances are higher in winter vs. summer for all sources used, suggesting temperature forecasting is more difficult and variable in the winter. This greater variability and forecast challenge would seem to be confirmed by the fact that there are roughly three times the number of high error days in winter and given that the error prediction model didn't perform as well on the whole at predicting error magnitudes in the winter (76% consensus) vs. the summer (86% consensus). In particular, the error model had a difficult time predicting error magnitude for one data source in the winter (only 68%), though this number should be taken conservatively since that data source, in the winter months, had the least number of samples from which to train the model and base the analysis.

2) The daily maximum temperature tends to be more significantly over-forecast in summer vs. winter for all sources. This suggests the sources tend to err more cautiously on the high-side in the summer. We cannot determine if this bias is intentional but can be acted on as a bias and removed.

3) One source, unlike the other two, grossly under forecasted the minimum daily temperature 93% of the time in the summer, meaning that it forecasted low temperatures that were warmer than actual low temperature. Again, this is a tendency that once known can be biased out and monitored.

4) The consensus forecast generally outperforms each of the individual sources in the summer and winter, both in terms of absolute accuracy and variance. This confirms typical designs of a multi-source input to weather forecast. This phenomenon is to be expected with exceptions in certain events where one or more of the sources may outperform the consensus on a case by case basis. See observation #7 for further elaboration.

5) The C5.0 model's ability to predict the error magnitude is similar for all sources as well as consensus, plus or minus a couple of percentage points. Interestingly, the model correctly predicted almost all of the high error cases for all sources and consensus in both summer and winter. The only exception being the model incorrectly predicted a medium error for one source on three occasions when it was actually high. Overall, if correct, this would suggest the model is significantly better at identifying unique patterns for high error days than for medium or low error days. Regardless, this extreme high predictive accuracy for the high error days is an initial finding and requires further efforts and analysis.

6) Source data is critical to data mining. The sample size in this pathfinder study was artificially boosted to create bigger samples from which the model could more easily identify and generalize patterns. Only having two seasons of historical data available challenged this kind of data analysis. Ideally, we would like to see at least 3-4 seasons' worth of data from which to train the model and at least 1-2 seasons' worth to independently test and validate. Since this was not possible given the limited archive of data, we augmented the sample with advice and assistance from SPSS. The drawback to data set creation, however, is that one may inadvertently create artificial biases in the data itself. To help mitigate this problem, we chose to use a very conservative data partitioning constraint when training the model, the 50-50 split mentioned earlier. Normally, an 80-20 train-to-test ratio is used if the sample size is large, so a 50-50 split by comparison is very stringent. Nevertheless, as data is continually archived creating more robust and reliable predictive models using data mining will become increasingly achievable. We encourage weather data users to continue archiving data and even expand the number of parameters saved to allow for the greatest possible latitude for the types of predictive models it may wish to employ later. There is also an opportunity to place prototype analysis tools side-by-side with operational tools allowing weather data users to observe and gain confidence in these potential improvements prior to use operationally.

7) The intent of this study was not to develop a predictive model or final set of rules that could be implemented operationally. Rather, it was to show an example of how these techniques can provide real value to existing operational data and capabilities given more detailed analysis with data mining tools. The model described here used multiple separate temperature forecast sources as inputs (i.e. three unique sources) along with other parameters to predict next-day temperature error. In that context, it could be easily used as guidance for selecting the optimal model for any given day or hour.

8) The nature of this approach is intended to be a high-level "meta model", that is to say a collective model based on outputs from multiple existing models. Therefore, this approach is in no way

disruptive to current operations as it conceptually sits on top of existing data and model outputs while residing on the side as a “monitor” physically separated and unobtrusive to other processes and data flows. In the same manner, it could easily accommodate and benefit from future models if inserted. In fact, the nature of the data mining meta-modeling approach is such that it would benefit from more model inputs so it can easily evolve and adapt with the system as a whole with minimal operational impacts. We will discuss this meta-model concept and a potential application in the following section.

4. META-MODEL DATA MINING POTENTIAL APPLICATION

There are certainly other dramatic and important weather applications that can be improved with data mining. Our previous pathfinder study determined that DM has benefits that can be realized operationally. It also identified a clear ability to be applied as a higher level meta-model over multiple sources. This meta-model application of DM is significant and the basis of our next area or research. Data mining cannot be looked at as simply an improved model or better statistical average. A significant portion of its strength resides in the fact that it finds patterns indiscriminately. There is no preconception of source or cause. That feature enables research in several ways. First, the ability to take existing models and determine how to use them as a more effective ensemble is a powerful tool. Ensembles are not uncommon, but dynamic and DM based ensembles are at best rare. As previously described in our pathfinder study we made marked improvements in aberration detection this alone justifies data mining in that study. In addition to improved composite behavior we also discovered single source and multi source phenomenon. In many cases these phenomena can be provided to the source model group to investigate and potentially improve their models. There are times when the cause(s) of the phenomena discovered cannot be changed, such as slow moving and fast moving filters, in these cases simply recognizing the behavior is valuable. All of this can be accomplished with little or no insight into the source models.

What follows is an example from our current research. We hope this example inspires others to actively pursue data mining in the weather community.

The current forecast of hurricane track, intensity, point of landfall and confidence corridor run an interesting parallel to our pathfinder study and forms the basis of our next area of DM research. There are multiple sources of data all of which may be improved based on currently undiscovered patterns. While each of these models will improve and seek better performance and acceptance there is a possibility the sum of these models is greater than the whole. This could result in more accurate hurricane track and intensity forecasts.

Clearly there is an abundance of data related to Hurricane forecasts to be utilized with data mining. This includes track models, intensity models and post event error analysis. Our concept is to apply the Clementine data mining tool suite in a similar fashion to our pathfinder study. The potential exists for both individual storm and seasonal improvements using data mining techniques.

Using the CRISP-DM process, our first step is to understand the problem. For this pathfinder we would look to improve the error associated with the forecasted hurricane tracks. Tables 1 and 2 show the average errors associated with the track models during the 1996-97 hurricane seasons.

Table 1: Average Errors (nm) of the Early Track Models for 1996-97 Atlantic Tropical Cyclones [6]

	Forecast Interval (hr)				
Model	**12**	**24**	**36**	**48**	**72**
CLIPER	51	103	161	220	351
NHC90	46	85	129	180	285
BAMS	61	114	168	222	336
BAMM	49	91	133	177	268
BAMD	47	88	132	183	293
LBAR	41	75	111	159	284
GFDI	42	69	98	128	200
No. Cases	346	310	279	255	207

Table 2: Average Errors (nm) of **CLIPER** and the Late Track Models for 1996-97 Atlantic Tropical Cyclones [7]

	Forecast Interval (hr)				
Model	**12**	**24**	**36**	**48**	**72**
CLIPER	51	104	166	237	408
AVN	56	98	139	178	248
NOGAPS	57	81	107	126	193
UKMET	57	92	136	165	244
GFDL	44	70	96	120	178
No. Cases	93	88	77	67	51

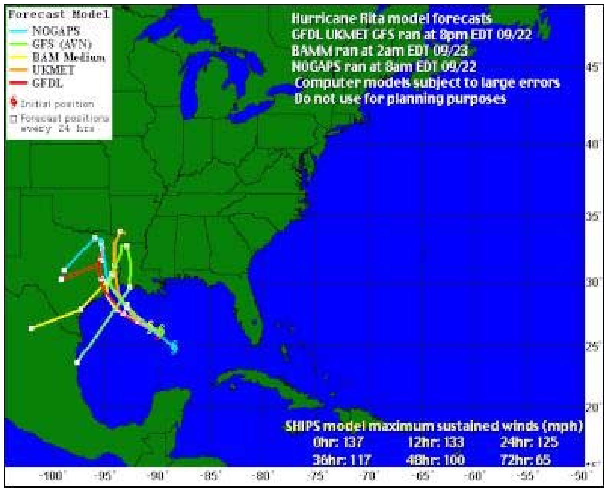

Forecast Model
NOGAPS
GFS (AVN)
BAM Medium
UKMET
GFDL
Initial position
Forecast positions every 24 hrs
Hurricane Rita model forecasts
GFDL UKMET GFS ran at 8pm EDT 09/22
BAMM ran at 2am EDT 09/23
NOGAPS ran at 8am EDT 09/22
Computer models subject to large errors
Do not use for planning purposes
SHIPS model maximum sustained winds (mph)
0hr: 137
12hr: 133
24hr: 125
36hr: 117
48hr: 100
72hr: 65
45°
40°
35°
30°
25°
20°
-100°
-95°
-90°
-85°
-80°
-75°
-70°
-65°
-60°
-55°
-50°

Figure 6 Computer Generated Forecast Models for Hurricane Rita [8]

Figure 6 shows the computer model predicted tracks for hurricane Rita. The differences in the computer generated track models are apparent.

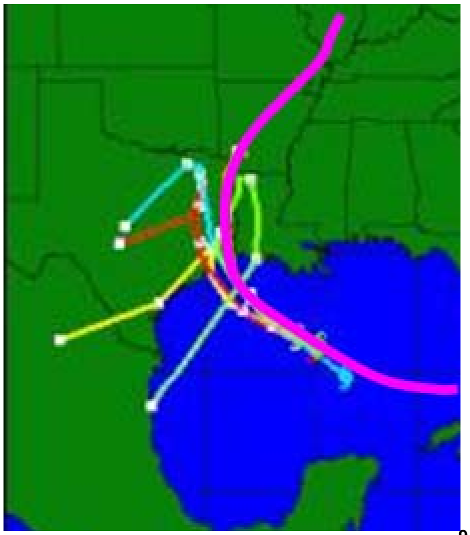


Figure 7 Hurricane Rita Actual Track (Pink) [9]

Figure 7 shows Rita's actual track (pink) overlaid on the forecasted model tracks. At different times in the hurricane's lifecycle some of the prediction models are more accurate than others. This observation leads us to believe there is potential for data mining to uncover hidden patterns that would dramatically improve the overall track forecast.

5. CONCLUSION

The fact that this year the weather community predicted an equally severe hurricane season as last year's record season has not gone unnoticed. The fact that this past season was not even close to last season's severity is also obvious. Explanations don't matter, it was a flawed forecast and these errors must be remedied in the future.

Some people think the weather folks have it easy, if you're right your brilliant and if you're wrong it was an unforeseeable aberration. To a degree this is true, some aberrations cannot be foreseen at least not until more data sources are used or existing data is used better. The reality is a great deal of data, weather data as well as other types of data, are not used. This isn't a result of negligence or lack of vision but a result of a challenging technical trend.

The weather community is experiencing an exponential growth in available data. This growth in available data is challenging systems and solutions which are based on conventional concepts. Almost all recognize that processing power increases very quickly. While the continued trend of Moore's law can be debated we are still in an area of processing growth. Another "law", Kryder's law, suggests storage devices (hard drives) are benefiting from exponential increases in data density with falling cost per unit of storage. These two laws have a corollary; we'll call Noah's Law. Noah's law says the result of improved processing coupled with decreasing storage costs is leading to an increase in the amount of data created. Figure 8 suggests existing systems will be overwhelmed by a rapid increase in data. The antiquated concept of "throw more iron at it" doesn't work. The problem is rooted in the fact that new systems and new products use the latest, high performance hardware. In past solutions, keeping up meant that the entire systems had to improve or drop data. While this works, to some degree, complex systems with large numbers of sources quickly cause this solution to fail. Some systems will fail obviously, and others quietly while unused data falls on the floor until it reaches congressional inquiry.

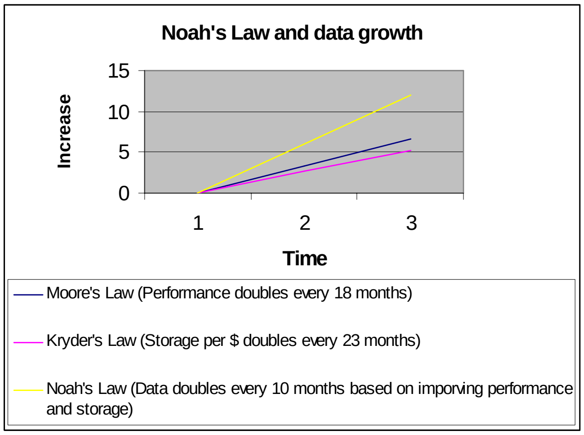


Figure 8 Noah's Law and Data Growth

We cannot change the fact that performance will improve and storage will become larger and less expensive, nor should we. We should take every advantage we can get from processing and storage. What we can change is how we think about data and knowledge. Data and specifically patterns in data result in knowledge. Data is ultimately not intended to be used completely. Like a sheet of aluminum the raw material provides the source from which the product is created. What cannot happen is to allow a process to waste data or create products inefficiently. Tools such as data mining offer the improved efficiency required in this increasingly data rich environment.

6. REFERENCES

[1] Data Mining. Wikipedia the Free Encyclopedia. 2 Nov. 2006 <http://en.wikipedia.org/wiki/Data_mining>.

[2] Data Mining. Ed. Doug Alexander. University of Texas. 2 Nov. 2006 <http://www.eco.utexas.edu/~norman/BUS.FOR/course.mat/Alex/#3>.

[3] IBID

[4] Cross Industry Standard Process for Data Mining. CRISP-DM Orginization. 2 Nov. 2006 <http://www.crisp-dm.org/Process/index.htm>.

[5] IBID

[6] IBID

[7] Summary of the NHC/TPC Tropical Cyclone Track and Intensity Guidance Models. Ed. National Hurrican Center. NOAA National Weather Service. 2 Nov. 2006 <http://www.nhc.noaa.gov/modelsummary.shtml#TABLE1>.

[8] IBID

[9] Hurricane Rita model forecasts. Ed. National Hurrican Center. NOAA National Weather Service. 2 Nov. 2006 <http://www.scoop.co.nz/stories/images/0509/2f6bf7fd304a16b7591a.jpeg>.

[10] IBID