On Data-centric Diagnosis of Aircraft Systems

John Stutz

Abstract-This article discusses results of an investigation into requirements for and problematic aspects of implementing a decision theoretic data-centric approach to the problem of data driven fault detection and diagnosis for aircraft and their subsystem, suitable for offline application to Flight Operational Quality Assurance (FOQA) type data. We discuss fundamental considerations for this type of approach, aspects of diagnosis required for a vehicle health management system, some problematic aspects of the domain, and provide a high level survey of current anomaly detection techniques, considering their suitability for diagnosis. Our principle conclusion is that for datacentric diagnosis, going beyond fault detection and localization requires a mapping from observable symptoms to diagnoses that is not readily available. This suggests a bootstrapping approach involving clustering, outlier detection and expert identification of suspected faults, providing the basis for actual diagnosis.

I. INTRODUCTION

W E are tasked with devising novel decision theoretic fault diagnosis algorithms, suitable for offline application on heterogeneous fleet scale Flight Operational Quality Assurance (FOQA) type data. This is a discussion of the problem, with emphasis on the application domain, how it constrains analytic approaches, and how the fault diagnosis task differs from fault detection.

The primary data types of immediate interest come under the general heading of flight records, conceptually any sequence of observation vectors made over a considerable time. Heterogeneity enters as we anticipate diverse combinations of binary, categorical, ordinal and real valued types, potentially including parsed text records. We currently specifically exclude raw narrative text. We do not require synchronized data recording, nor regular time intervals, only sufficient information that we can reconstruct a current observation vector at any time point in the record. Thus there is a presumption of timely observation updates, particularly that all operator inputs are promptly recorded. We expect to have large quantities of data representing diverse nominal operation modes, and much smaller quantities representing faults that have been previously identified by other means. Not all data sets in which faults occurred will have been identified as such, and these may be presented as nominal sets. Nor will all actual fault types have been previously identified.

Regarding the algorithmic aspect, we are implicitly directed toward data-centric approaches that avoid system modeling on the order of a traditional physics based engineering model. Moderate modeling, on the order of using meta-data to inform the choice of appropriate probability distributions and priors,

Version as of January 12, 2010.

J. Stutz is with NASA Ames Research Center.

Manuscript not yet submitted

is expected to prove useful. But we seek inference systems that are easily adapted to diverse domains without requiring detailed knowledge of the internals of any particular system under study. This will limit diagnostic ability, since lack of an internal system model largely precludes inference regarding the source of faults. Purely data-centric inference systems can detect potential faults as anomalies and localize them with respect to the space of observables. Data-centric inference of causes traditionally requires a "supervised learning" approach based on use of labels attached to fault instances previously identified as such, and included in their training data. When such labeled data does not cover a field, about the best that can be hoped for are probability distributions over the set of previously identified normal modes, known fault labels, and "unknown fault". This implies a boot strapping development: Unknown faults detected and localized in the historic data will need to be referred to domain experts for labeling and then incorporated into the diagnostic model.

Our emphasis on probabilistic inference is a consequence of the decision theoretic objective, which requires the probabilities of possible system states, the set of relevant normal and fault modes, in order to estimate the consequences of alternate decisions. This aspect involves an extension of the usual focus of diagnostics, from simply determining the presence and nature of a most likely fault, to determining the relative probabilities of any plausible faults. These are needed to estimate the consequences, and thus the cost or value of any potential decisions considered in response to the diagnosis. This provides an element of prognosis, but one that lacks an engineering based model-centric approaches' access to the predictive information implicit in an estimate of a system's detailed internal state and potential failure modes. Data-centric prognosis will rely largely on historical information associated with the likely system states, much as medical prognosis has traditionally done.

This decision theoretic approach complicates matters somewhat, requiring our rethinking the application of any diagnostic algorithms that return categorical results. Decision theory *requires* estimated state probabilities of alternate hypotheses regarding system state. A conventional unqualified diagnosis discards the essential estimates, convolving inference and decision, and concealing the assumptions fundamental to each.

We consider data-centric and physics based model-centric approaches as alternatives for diagnosis. These are the extremes along a continuum of emphasis that inevitably involves both data and models. Model-centric fault diagnosis emphasizes the components of a system, their interactions, and their individual and collective fault modes, as constrained by observations of a system instance. Data-centric fault diagnosis emphasizes observational data, preferably dense, redundant, and frequently recorded with respect to (w.r.t.) system failure rates. The data-centric model may be little more than "... these are the data patterns associated with normal modes and known fault modes, and anything else is an unknown fault.". But the criteria for "association" and especially any data preprocessing add additional structure to the often implicit and unacknowledged model underlying any data-centric method.

Meaningful metrics for evaluating diagnostic systems will vary considerably with application domains and specific users. This is one of the major lessons learned from the DX-09 Diagnostic Competition [1] conducted as part of the 2009 International Workshop on the Principles of Diagnosis. There the specification of generic evaluation criteria proved unexpectedly difficult, and in the end was somewhat specific to the competition.

The following section is a brief overview of the basics of Decision theory, and how it affects our choices for fault diagnosis. Section III is a similar overview of user requirements for aircraft health maintenance systems (HMS), and how they affect what we want to achieve. In section IV we review some basic diagnostic methods, with emphasis on some of the problematic areas of the aircraft HMS domain. Section V reviews some of the standard approaches to fault detection with discussion of their application to fault diagnosis. The final section VI develops some thoughts on how our task can be achieved.

II. DECISION THEORY

Normative Decision Theory (DT) attempts to formalize the process of making optimal decisions under uncertainty, by explicitly quantifying the expected values of alternate decisions [2]. DT is based on Bayesian Probability Theory (BPT), using BPT to estimate the probabilities of alternate possible current states of a situation of interest, and to estimate the likely outcomes of decisions given a current state. To this estimation approach, DT adds utility values for the possible alternate outcomes, and marginalizes over the states and outcomes to estimate utility values of alternate decisions.

That's the gist of Decision Theory. It is a simple idea, not easily implemented. The problematic areas are found in several critical underlying assumptions that must be met:

- Specification of *all* potential system states, to the degree of detail needed to estimate state probabilities from system observables.
- Observed data from the current system instance concentrates probabilities on a few system states.
- Knowledge of all possible consequences and their probabilities for any decision that can be made in any system state.
- Numerical utility values for the possible consequences.

In principle, given the above, one can estimate the probability of alternate outcomes of any decision, by conditioning on the observations of a system instance, and then marginalizing over the plausible system states. By factoring in outcome values, and marginalizing over outcomes, one gets the estimated value of each decision under consideration, still conditioned on the current system observations but irrespective of actual system state and decision outcome. In practice, each of the listed elements involves considerable complication for design and validation, even for fairly simple decision problems. These complications are exacerbated when dealing with complex and highly reliable systems like aircraft, where very rare faults can generate situations in which normal decisions can generate catastrophic outcomes involving extraordinary costs. In particular, the possibility of unknown (fault) states largely invalidates any numerical estimates of decision values.

Thus the full normative Decision Theoretic program is not well adapted to full aircraft diagnostic systems. It may be applicable to subsystems that can be described by a limited number of well characterized states, each with clear consequences for any decision. But as such subsystems are combined, the full state space size increases multiplicatively, while subsystem interactions dilute knowledge of decision consequences.

Nevertheless, Decision Theory offers useful ideas and lessons for aircraft scale diagnosis. Particularly critical is the refusal to *decide* on a particular system state, but to keep all under consideration in proportion to their probability. This contrasts to the sequential approach to problem solving, which first decides what problem has occurred and then seeks a solution to that problem. The decision theoretic approach seeks a solution w.r.t. the probability and cost weighted possible consequences of decisions, potentially bringing much more information to the decision.

A second point, not emphasized above, is the value of DT in determining what further information would be of greatest utility in refining knowledge of system state and decision consequences. Decision theoretic query formulation is a matter of current research [3], [4] that needs to be added to our repertoire of diagnostic techniques.

III. USER REQUIREMENTS

Wheeler et al. [5] begin their survey of aircraft health management system's users' objectives with the statement that "One of the most prominent technical challenges to effective deployment of health management systems is the vast difference in user objectives with respect to engineering development.". This reflects both the wide range of users with their divergent operational objectives, and the current shortage of systems able to meet many of those objectives.

Aircraft HMS users span a surprisingly wide range of interests, objectives, and time frames. Some of the principal players and their concerns are:

- Flight Here the emphasis is on safety in operation, by getting critical fault information to the air crew, in good time to respond, without increasing their cognitive burden. Information overload is a very real danger, and false or conflicting alarms will severely reduce system acceptance.
- **Maintenance** is plagued by problematic fault reports, from both crew and current HMS, that when investigated either "cannot be duplicated" or "retest OK". These entail large costs in time and effort, and increase risk of maintenance induced faults. An acceptable HMS must

reduce these. Precise fault location is sought, particularly in distributed system elements like wiring. Ideally an HMS will provide advice on how to verify faults, what tests are needed to distinguish between alternatives, and how to make repairs. Maintenance costs for the HMS itself, particularly for updates and their validation, must also be minimized.

- **Logistics** seeks to minimize overall cost of operations, primarily in terms of minimizing inspection and maintenance frequency and effort, particularly unscheduled maintenance, without increasing risk of failure. Accurate predictions of remaining useful life are desired, to enable condition based maintenance.
- Fleet Management's immediate concerns are operational efficiency, particularly fuel costs. This is also concern with minimizing unscheduled maintenance, preferably by condition based maintenance, and maximizing service life and reconfigurability. Long term goals are to improve designs and qualifications, and to support business and regulatory decisions.

A number of common themes arise, common to most users:

- **Minimize false alarms** This is utterly critical if an HMS is to be effective. A moderate false positive rate may significantly reduce compliance with alarms [6]. Even a low false alarm rate will cause extra cognitive loading due to the need to evaluate the alarm while also dealing with the alarm message.
- Maximally specific fault identification This is essential for efficient allocation of the resources needed to confirm and rectify a fault, and thus central to the concerns of many users.
- Earliest warning of failure Regardless of their primary emphasis, every sector wants the earliest possible warning of developing problems.
- Minimize information overload While human cognitive capabilities can be much enhanced by training in an operational domain, and are remarkably flexible, we are limited in the number and the details of alternatives that we can simultaneously consider. Too much of either and our ability to cope degrades, often well before we realize it, as recent results on multitasking demonstrate [7].
 - Frequent fault messages, or large blocks of alternate possibilities, can quickly overwhelm our ability to keep track. Some degree of prioritizing and filtering will be needed.
 - Conflicting warnings can induce cognitive dissonance. Since alternatives are inevitable, some basis for prioritizing them is essential. Probabilities will provide the basic criteria, but utilities need to be factored in as well.
 - A fault diagnostic system needs to distinguish between critical and non-critical faults, in both immediate and long term time frames, and clearly display the relevant estimates.
 - Any fault monitoring system is itself a source of hardware faults, in its sensors and communication links. When such faults occur they must be identified

as such. Otherwise the user bears the cognitive burden of distinguishing between base system and monitoring system faults.

- Access to auxiliary information A fielded HMS will be far more effective if it interacts with domain systems in a manner that provides access to any kind of information needed to deal with the alarms that it raises. For example, maintenance is confronted with diverse reference sources essential for confirming and repairing any single fault, and would benefit from one point access to all relevant information.
- Ease of use for entire system An inconvenient tool cannot be used efficiently and may simply be ignored. Tool developers tend to forget that others invariably find their systems more difficult to use and less responsive to their needs than the developers anticipated.

Most of the above considerations are directed to full health management systems. But all the above emphasize the need for a fault diagnosis system that provides both probability and utility estimates to support the traditional detection and location phases. Certainty in diagnosis is rare, and sound use of diagnosis in an HMS requires explicit quantification of uncertainties and consequences.

IV. DIAGNOSIS IN GENERAL

Fundamentally, fault diagnosis involves three elements: detection, localization, and identification. A full HMS system adds prognosis, confirmation and correction. Detection is simply recognition that a fault may exist, usually as an anomaly in system description data. Localization involves determining where in the subject system a fault resides, to the degree that the diagnosis system allows this, often only as a subset of anomalous data values. Identification labels a probable fault in a manner that is *informative* to the diagnostic system's *users*. Prognosis attempts to asses a fault's future evolution and severity. Confirmation of an evaluation is generally desired. And finally, there is no point to any of the foregoing if no one attempts to correct or alleviate the fault.

Model centric and data centric diagnostic methods are both traditionally applied to object description vectors or timesliced system vector data. Thus they seek evidence of faults in a vector of observations representing an independent object or at a single point in time. Vectors are generally considered to be independent, conditional on the model, so there is no sense of an evolving state. This vector based view dominates research to the extent that it is often taken as a fundamental assumption, but there are alternatives. Several types of Markovian state estimators attempt to trace changes in the current system state, w.r.t. a sequence of time-sliced observations, by retaining one or more previous state estimates and factoring in inter-state transition probabilities. Most of these vector based techniques are intended for either real or discrete valued data vectors, and require that all attributes be forced into the preferred type. Satisfactory methods for simultaneously incorporating categorical with discrete and real numerical data remain a matter of research [8].

There are also sequential approaches for detection and diagnosis in symbolic data streams [9], [10]. These emphasize

recognition of sequential patterns, in contrast to the traditional vector space patterns. There are two basic approaches: full sequence comparison, and windowing to detect anomalous subsequences in largely normal sequences. A variety of categorical similarity measures [11]–[13] are employed to measure the differences. The primary application areas are in computer transaction records, genome and document classification. Most of the transaction oriented work has dealt with data composed from a fixed set of a few 10s to 100s of symbols. How to best factor numeric data into such a sequential analysis remains an open question.

Extreme model-centric fault diagnosis methods tend to employ an engineering approach, based on a thorough understanding of the components making up a system, how they normally operate and how they can fail, how they are connected and how those connections can fail, how the system is controlled and monitored, and thus how both normal modes and faults will manifest in the observables. Complete knowledge of the system is required, at whatever level of detail is desired for diagnosis. This includes all relevant component failure modes, since failures not modeled cannot normally be identified, even if detected. Efficient propagation of the effects of failures through the model, for comparison against observations, is essential [14], as are efficient algorithms for identifying and rating the most likely faults conditional on current observations. Thus model-centric implementations tend to be very focused on specific systems, for which much detailed information is required, even when created with quite general methods.

Where computational speed is essential, the model-centric approach may resort to model compilation [15] or to model simplification [16]. Simplification is actually a matter of balancing tradeoffs between model detail and performance requirements, and can still require quite sophisticated modeling. Alternately, simplification may be imposed by algorithmic limitations [17] or such hardware limits as memory footprint or a need to implement on field programmable gate arrays(FPGAs).

Extreme data-centric fault diagnosis attempts to identify anomalies without requiring detailed knowledge of system internals. Detailed system knowledge is to be replaced with copious quantities of system observation records, intended to span all normal system operating modes. The immediate objective is then to identify algorithm specific signatures of normal operating modes, on the assumption that faulty modes will distinctly differ from normal ones. This allows for fault detection, and fault localization to the extent of identifying what observations diverge from normal patterns. Fault diagnosis equivalent to that achieved by more modelcentric methods is not possible from system records alone. At very least, diagnosis requires a user's vocabulary of terms associated with fault descriptions.

Data-centric modeling is based on the idea of learning from examples. There are a diversity of considerations to be accounted for in defining such a diagnostic system. First and foremost is what one seeks to obtain from diagnosis and what one has to base diagnosis upon. Here we discuss some major considerations.

A. Learning Patterns

Supervised learning of a classifier, using fully labeled data instances, is the preferred mode of learning for data centric fault diagnosis systems. Fault type labels are eventually required for fault identification, and if possible, they should be introduced as early as possible. Normal operating mode labels will be needed for some approaches, since fault modes can fall between normal modes and within the overall normal region. Traditionally the labeled instances are used to specify labeled regions in the native data space or alternately, a projected feature space such as used with Gaussian Process Regression (section V-A2) or Support Vector Machines (section V-A4). The basic idea is that each mode maps to a consistent region in the data space, modulo some variance, for all relevant observation records. Test instances then get the nearest label, for some specific sense of nearest, or a probability distribution over labels. Classification based approaches are thus well suited to supervised learning. The drawback, of course, is in the difficulty of obtaining accurately labeled training instances, especially fault instances, with suitable coverage and quantity in the aircraft systems domain.

Given the expected dearth of labeled aircraft operating records, both normal and faulty, some form of semi-supervised classification with bootstrapping will likely be needed. This will involve classification over whatever labeled data is available, combined with clustering of unlabeled data. On the assumption that normal modes are dominant, outliers and small cluster instances will be extracted and subjected to expert review. Given labels for these instances, the model is updated and reexamined for borderline instances that still need labeling. On attaining a stable configuration, the model can be used for routine classification. This approach has the operational advantage of requiring only minimal training instance labeling, largely limited to the true faults and the nearby normal instances. As with any clustering system, there may be difficulties due to clusters that do not align well with the properties of interest, and with small ones that get assimilated into large clusters.

B. Models vs. Data

In any estimation problem there is a tension between the degree to which data can be adjusted to fit the mathematical model implicit in a learning approach, the degree to which the model can be adjusted to suit the data, and the mismatch which can be tolerated between them. Additionally, we tend to rely on familiar tools and to interpret any problem in the light what those tools can achieve. This is simply human nature, and not necessarily a problem, if the mismatch is not too great. It does tend to limit what can be learned from any set of observations, and so needs to be guarded against.

Matching models to data is an intellectually more difficult task, demanding knowledge of a range of techniques and what makes them appropriate in different applications. There are a wide variety of basic techniques, and an equally wide range of variations on each. One gets the impression that it is easier to devise a new variation for each new application, than to find an old variation that fits. Chandola et al. [18] has made a valuable start in his survey of the field of anomaly detection techniques, but such surveys can only provide a high level overview. It remains the responsibility of individual developers to determine what will best suit the actual problem at hand and available data.

The cost of adapting models to problem and data come up front, in the time and effort needed to determine what the problem actually is, what data is available to support a solution, and how a solution might be obtained. Potential savings lie in minimizing data conditioning efforts, with consequent loss of information. The payoff lies in getting estimates that address the actual problem, instead of being only incidental to the problem.

C. Data Types

Just as analytic methods tend to diverge in regarding system data as either vector valued instance descriptions or temporal sequences, so they also tend to differ in their handling of categorical (names, symbols, labels, &etc.) and numerical data. An alternate partitioning is into discrete and continuous valued attributes, which emphasizes the somewhat ambiguous nature of ordinals (orderings, counts, &etc.) as a data type. Data modeling techniques tend to be best adapted to one or the other, and may require that one be transformed to the other.

Categoricals: These, when mutually exclusive and exhaustive, fall directly into the provenance of discrete probability theory, with well known modeling techniques. Distance based approaches like the nearest neighbors variants require a scalar magnitude for each component of the difference of two vectors. For ordinary categorical attributes with mutually exclusive values this distance is usually taken as the difference measure: 1 if values differ and 0 if they are identical.

When a categorical attribute's values are not exhaustive, they can be supplemented with "other" or "unknown" and the standard techniques used. With non-exclusive labels, such as one finds in the NASA/FAA Aviations Safety Reporting System (ASRS) classifications, the situation is less clear. There the anomaly category set numbers about 60, we have observed up to 12 assigned to a single report, with a mean count near 2.7.

A tempting approach to handling this situation is to adopt multiple binary one-vs-all models, which can be useful when just seeking to recover the labels. Deciding how many labels to accept could be problematic. The Bayesian equivalent is to replace each label with a binary attribute indicating the label's presence or absence, and to estimate all label probabilities together. This potentially entails a large increase in the attribute space and possible dilution of the labels' information content. Modeling these as independent attributes might not be a problem, but a covariant approach should consider a sparse parameter representation as in section V-A3.

Ordinals & Counts: Plain ordinals are discrete yet have a well defined ordering, while counts add a sense of uniform spacing. Either may be treated as simple categoricals with some, often much, loss of information. Counts that are believed to have been generated by a uniform rate process are well represented by the Poisson distribution, with several derivatives for functions of such rates [19]. For distance based models the difference of two counts is just their numerical difference. That of two ordinals could be taken as the difference of their positions in the ordering, or as specified for categorical differences.

Reals: Continuous or real numbers are the basis of numerical analysis, and the number of ways to view them is correspondingly large. The variations on the Gaussian Normal densities are the canonical choice for representing noisy numerical observations in Bayesian inference. There are many more possibilities [20] including numerous specializations of the Gaussian. Most of these are extremely specific, and so of little interest. But others deserve to be much better known, particularly those devised for directions and for bounded domains, where the standard Gaussian is quite inappropriate.

For distance based models there are a variety of distance measures based on the vector difference. Choice among them is an open questions, but there are usually no difficulties in their application to the data.

Most analytic methods for real number data assume that our values sample the continuous real number line of mathematics. In fact, we normally use floating point numbers, which allow only an infinitesimal sampling of the reals. Sensors often report,, and databases often record "reals", to only a few digits. And some attribute values, flap angle for instance, may normally only vary over a small set of values, despite having a continuous range of possibilities. Thus any real valued attribute requires careful examination of its distribution to determine a sound approach to modeling it.

Heterogeneous Data: Heterogeneous data usually denotes a mixture of discrete and continuous types. It has often been a problem due to difficulties in combining the several types in a well justified manner. Two preprocessing techniques for conversion to uniform type are described here.

Binning of continuous values into discrete categories is popular in some traditions. One such is Naive Bayes classification, where all attributes are normally modeled by independent multinomial distributions. Binning eliminates much of the information in continuous data, retaining only some blocky senses of nearness and ordering. The multinomial model then ignores any residual ordering information. Binning to uniform spacing, population, or similar criteria requires a preprocessing step suitable only for static data. Menzies & Orregio have devised a runtime binning technique suitable for streaming data [21]. This generates non-uniform bins, some of which may eventually be deleted. Thus the representation is dynamic, and requires update of their class models, which is not difficult in their approach to Naive Bayes. Despite the inherent information loss, they report good tracking of concept drift and detection of anomalous events in a variety of data sets [22].

A categorical attribute with n mutually exclusive values can be projected to an (n-1) dimension real valued space, as an n-vertex regular simplex or hyper-tetrahedron, thus preserving a uniform unit distance measure. Multiple attributes are transformed independently. For each such attribute, the data space dimension is increased by the number of values less one. The transformed attributes can be seamlessly merged into an appropriate density model. In preference to forcing our data to fit a model, a probabilistic model can be customized to fit the data. The Naive Bayes approach, which models discrete valued attributes independently, can be extended to handle real valued attributes by marginalizing over subranges centered on the reportable real values. This gives probability masses that are fully commensurate with Naive Bayes' normal multinomial distributions. In addition, covariant attribute subsets can be modeled for either discretes or reals. This approach was implemented in the AutoClass system of Cheeseman et al. [23]. Covariance of discretes with reals remains a problem. The obvious approach of providing a real distribution conditioned on each combination of discretes can require a large number of parameters, and so risks overfitting of training data.

Missing Values: For a variety of reasons, missing values can occur in raw unprocessed data. Unfortunately most data analysis models assume that all values are provided, and in order to proceed, they essentially require that either the missing values be provided or that the attribute be ignored. A pre-analysis data conditioning step may achieve the former, and for some data types this can be done in a reasonably sound manner, as interpolation or regression. Some Bayesian models will permit marginalizing over the attribute's probability distribution conditioned on the other known attributes, but the potential algorithmic costs are high, and may preclude this approach. Even if allowable, the instance information is degraded, and severely so if several attributes have missing values.

With probabilistic models any missing values can be explicitly modeled, by adding a "missing" value for discrete attributes and making real attribute densities conditional on a "known / unknown" added binary. This works well when attributes are modeled independently [23], as in Naive Bayes. But it increases the parameter space somewhat for independently modeled values, and significantly so for covariant models. Additionally, where missing values are common they may dominate the resulting statistics, and a more sophisticated approach may be needed in order to concentrate on those values that are known.

Covariance: Most probability distributions are univariate. Any such can be multiplied to form a joint distribution, under the assumption that the corresponding observations are independent. The several distance based inference approaches make similar assumptions regarding attribute independence or homogeneity. However there is often much to be learned from analysis of attribute covariance. In particular, observations that are strongly influenced by a common causal factor tend to be correlated to some degree. Thus loss of expected sensor correlation may be evidence for loss of function somewhere in the sensing system. Covariance is used in just this manner in the System Invariant Estimator (SIE) of JPL's Beacon-based Exception Analysis for Multimissions (BEAM) system [24]. Since quite different sensor outputs can become correlated via the processes they monitor, correlation monitoring can potentially provide a degree of sensor redundancy without requiring redundant sensors.

The only well known covariant probability distributions are the multivariate multinomial for discrete attributes and the multivariate Gaussian for reals. SIE uses direct computation of pairwise covariance without considering the location aspect provided by the multivariate Gaussian. Special distance measures for correlations have also been devised [25]. The Fisher-Bingham or Kent distribution for covariant directions on the 3D sphere S^2 extends the Gaussian to directional distributions [26].

D. Domain Considerations

The aircraft operations and maintenance domain involves a number of considerations not often encountered in the standard databases commonly used to exercise and test academic inference systems. Size alone is one factor, far from the most significant. These considerations are not wholly unique. Some are shared with most real world operations domains. Emphasis differs between domains, but the necessity of adapting models to suit domain considerations remains.

Transients: In time series data, transient responses to transitions between normal operational modes can be a problem. Transients are almost inevitable in electrical systems, where inductances generate large voltage excursions when circuits are switched on or off. Inertia in mechanical systems can also induce transient responses. Such transients can easily exceed the limits bounding normal steady state modes, so can fall outside of any steady state model, and are potentially identified as anomalies. Thus transients need to be explicitly dealt with, either by preprocessing or by explicit modeling.

Wide band prefiltering of time series will smooth down short duration transients, but only at the expense of lengthening the inter-mode transition time to the order of the filter width. This may leave transition time vectors hanging between normal modes, in an anomalous state. However such smoothed inter-mode transitions might be identified as normal modes themselves, if they are common enough in training data that is deemed normal.

Presence of transients argues for a state inference method that explicitly allows for and accounts for them, and thus one that explicitly models some degree of time wise system evolution, in contrast to simple mode switching. This precludes approaches that treat observation vectors as being independent, conditioned only on the system operating mode, as simple classifiers tend to do. The full system history should not be necessary, but some knowledge of the state estimates over some range of previous observations may be desirable. This suggests a Markovian approach to modeling mode transitions, but we are not aware of any such data-centric approaches.

Sensor Failure: Aerospace engineering lore suggests that in highly engineered systems the sensors are often more likely to fail than the system sensed. So a diagnostic system needs to be able to identify sensor failures, and to distinguish such failures from faults in the underlying system. Furthermore, sensors can fail while continuing to report values that are within their normal range for the current operating mode, so sensor failure diagnosis is more than simply out of range detection.

This implies that no sensor value can be taken as truth. Any sensor value must be verified, presumably by comparing against some other system attributes that tend to confirm or discomfirm correct sensor operation. The classical approach is 3+-fold sensor redundancy, with direct inter comparison, and choice of a non-extreme value. This is expensive, primarily in terms of the extra sensors, communication links, and maintenance added to the system being sensed. However the modeling and decision analysis can be fairly straightforward. An alternate approach is to look for correlations between multiple sensors. Strong correlations can be assumed to result from common causes, particularly if a degree of a correlation persists over operating modes. Thus incorporating correlations into data centric models is of potentially great value for identifying faulty sensors, despite the consequent increases in model complexity.

The probability that sensor faults are more frequent than system faults obliges the diagnostician to differentiate between the two types, and to handle them appropriately. Sensor faults do need to be reported, but they are not in themselves so critical as system faults usually are. The difficulty lies in first distinguishing sensor from system faults, and second, finding some way to accommodate sensor faults in system normal mode models. Even with binary sensors, the combinatorial explosion of alternate sensor failure patterns precludes trying to identify each combination of sensor failures as a normal mode. A semi-naive Bayesian approach might work, considering only independent failures, but even this involves an enormous expansion of possibilities for each normal mode. Nevertheless, a resilient diagnostic system needs to be able to reliably estimate system state in the presence of sensor failures, so this problem must somehow be overcome.

Fleet Variation: As diagnosticians, we would prefer to deal with fleets of nominally identical aircraft as if individuals were largely interchangeable. This would greatly simplify data collection and somewhat simplify modeling, fault detection, and diagnosis. Such an approach may be possible for many subsystems, but less likely so for full aircraft and the more complex and/or highly stressed subsystems like engines and landing gear. There are fundamental and unavoidable reasons for this.

- The more complex a system, the more likely that any two "identical" copies differ in significant ways, even prior to first use.
- Transportation systems operate in a highly variable environment, so can quickly develop quite individual histories, particularly w.r.t. the extreme events that can initiate faults.
- Flight crew operational activities may vary considerably, particularly in stressed situations, despite efforts to standardize operation.
- Maintenance inspection and repair practices may diverge from standards, potentially introducing new faults or altering the signature of normal operating modes.

All of the above imply that complex "identical" transportation systems, and their major components, will have divergent histories, which induce differences in our observations of both normal and faulty operating modes. The degree of such differences remains to be evaluated, and their significance will depend on how the systems and its operating modes are modeled. Thus if we train a subspace localizing algorithm on both individual system and fleet data, alternately for the same operating mode, we expect to find that no individual system distribution quite matches the fleet distribution, while fleet variances are larger and fleet correlations are smaller than those for individual systems. Depending on our goals, this could be sufficient to require that we use individual system instance models for fault analysis. The lesson here is that the degree of fleet variation needs to be either proven negligible or specifically accounted for, before attempting to use a fleet wide analysis approach.

Mode Drift: Any complex systems, particularly those involved with large scale generation and application of power, can be subject to drift in the values of observables associated with any single operating mode. For aircraft, the most obvious causal chains are based on the gradual consumption of fuel, and affect a variety of flight parameters. Any diagnostic system for aircraft flight operations needs to account for mode drifts, either allowing for it by incorporating sufficient leeway in a static mode description, or by updating dynamic mode descriptions. Either way we must deal with the potential problem that fault modes may fall within the extent of such extended normal mode descriptions, and thus not be detected.

Multiplicity of Modes: Multiple operating modes are expected in any non-trivial system, and will be a problem for analytic methods that presuppose binary decision problems. With multi-modal approaches there is always the question of how many modes are to be allowed, or can be allowed without overfitting the estimated model to the training data. This is where Bayesian posterior probability estimation is particularly useful, since the cost in prior probability of the additional parameters required to extend a model will eventually dominate the increased likelihood gained by better fitting the model to the data [23].

Multiple modes with significant mode drift will blur the difference between the two concepts. This will influence how we define both modes and mode drift, possibly on a application specific basis.

V. DATA CENTRIC DIAGNOSIS

Chandola [18] et al. have made a comprehensive survey of data driven techniques for anomaly *detection*, and we follow their organization here. However diagnosis differs from detection in significant ways, so our emphasis differs, with some additions and deletions. The requirement that a diagnosis algorithm shall detect, localize *and identify* most anomalies, preferably with probability estimates for alternative identifications, requires considerable extension of some standard detection techniques and may preclude others.

Fault identification is the key difference distinguishing data centric diagnosis from traditional data centric fault detection. Identification requires knowledge of faults, which will normally be obtained from labeled fault instances. Note that fault labels are not necessarily exclusive: In one set of annotated ASRS reports, we found an average 2.7 fault type labels per instance, with the maximum exceeding 10. This complicates adapting those probabilistic techniques that implicitly assume "mutually exclusive and exhaustive" alternatives .

A. Classification Techniques

Classification based diagnosis attempts to learn a classification model from a set of labeled training data instances, and then to classify test data instances w.r.t. the learned model. The basic assumption is that classes occupy distinct regions in the data space, or some projection thereof. Mutually exclusive classes are normally assumed, but real world labeling may be non-exclusive. Labeled class instances are assumed, both normal modes and anomalies for diagnosis. But there must also be some provision for detecting previously unseen anomalies. Data as instance vectors is traditional, and traditionally classified independently, but machine classification is an old field with many variations, of wide application.

Classification is very often cast as a *decision process*, with logical and normally exclusive outputs. Such classification is of little interest for decision theoretic diagnosis, and then only to the extent that variations can generate class probabilities. Decision process classifiers will not be discussed here, except where they can be upgraded to generate probabilities.

1) Neural Networks (NN): These have a long history of use for classification, having been generally applied to mutually exclusive classes. Training a neural network as a probabilistic classifier for non-exclusive classes requires some variation of traditional techniques. It is likely that this would degrade the NN's ability to distinguish classes. An NN classifier normally has one output per class, and generates a set of output weights for each instance tested. These weights can be treated as likelihoods and normalized into probabilities, when mutual exclusivity is assumed. Outliers are detected as instances that have very low weights at all outputs. Explicitly Bayesian variants are available [27], but Gaussian Processes have preempted these in several ways.

2) Gaussian Processes: Gaussian Process Regression (GPR) assumes a regression function y(x) that is a weighted sum of many, perhaps infinitely many, basis functions: $y(x) = \sum_{i}^{I} w_i \phi_i(x)$. The resulting y(x) is linear in the w_i , and assuming the w_i each have a zero mean Gaussian prior probability distribution, y(x) is also zero mean Gaussian w.r.t. the N training instances. Then, so far as prediction is concerned, the only effect of the basis function set is through the N by N covariance matrix. Conversely, any valid covariance matrix corresponds to some set of basis functions, admitting their implicit use. GPR modeling is then largely a matter of specifying the covariance matrix, optimizing w.r.t. any hyperparameters, and then inverting the covariance for use in prediction.

GPR provides an effective approach to binary classifiers, and has been extended via Monte Carlo or variational methods to multi-class problems [28]. It may also be possible to do multi-classification by applying GPR to the simplex categorical expansion of the class symbol, as described in section IV-C, via multi-dimensional regression. If so, this could provide a direct approach to representing and predicting non-exclusive categories, as points on or within a categorical simplex. However GPR does its separation in the kernel space of the implicit basis functions, which is not directly represented, and so is not accessible for localizing the source of unknown anomalies. Thus GPR is better adapted to detection than diagnosis.

3) Bayesian Classification: This can take a variety of forms depending on the type of data and how class models are mathematically defined. Classes are described by probability distributions over the native data space or some projective space. Most forms are thus vector based, assume independent instances, and mutually exclusive and exhaustive classes. These assumptions can be relaxed at the expense of additional complication.

Naive Bayesian classifiers commonly assume categorical data, with each attribute in each class modeled by an independent multinomial distribution. Classes are learned by accumulating attribute statistics from a set of labeled training instances. Uninformative Dirichlet priors prevent zero probabilities in the distributions. A class's likelihood for a test instance is then the product of the instance's attribute values' probabilities for that class. The class likelihoods are multiplied by the class probability and then L1 normalized to get the class probabilities conditioned on the current mutually exclusive multi-class model. However the class likelihoods can also be informative, either for detecting anomalous instances that do not match well with any class [22], or for detecting concept drift in temporal data [21].

Numerical data is usually incorporated into Naive Bayes by binning each attribute to a categorical replacement. While this shoehorns numericals into the categorical based multinomial model, it loses much of the detail inherent in numerical ordering and differences. Binning usually involves a preprocessing step, but there are techniques for dynamic single pass binning, practical because the classes' corresponding independent attribute statistics can simultaneously be updated as bins are added or merged, without referring to the original data.

Naive Bayes for categoricals can be extended to covariant multinomial Bayes by considering the joint categorical attribute space. Done naively this involves a combinatorial explosion, the per class parameter count going from the sum of attribute value counts to their product. The product space's size can easily exceed the number of training instances available per class, but a class instance count is usually much larger than the number of product cells actually occupied, so sparse parameter representations may allow efficient covariant modeling. An implicit uninformative prior eliminates zero valued probabilities. When this approach is possible, one gains access to inter-class differences that are completely invisible to standard Naive Bayes.

With numerical data, suitable probability distributions, discrete for integers and continuous for reals, are substituted for the multinomial. Naive models retain the independent attribute approach, while covariant models can group attributes that use the same mathematical model, usually the multivariate Gaussian or similar exponentials. The full n attribute covariant Gaussian requires $O(n^2)$ parameters per class, but unlike the covariant multinomial, all values are instantiated. An independent n attribute Gaussian model needs only 2nparameters per class. A full covariant Gaussian model applies the corresponding Mahalanobis distance measure, and so is an exact representation for any distribution corresponding to an affine transformed spherical Gaussian. Thus they make an excellent model for classes that can be described as noisy distorted point distributions. They are less appropriate for more complex distributions, where multiple spherical Gaussians have been used for single class models.

While the Gaussian is the traditional basis of statistical analysis, there are a wide variety of other continuous distributions applicable to numerical data modeling. The von Mises distribution may be considered to represent a Gaussian like distribution over a circular basis, parameterized by a direction and variance. The von Mises-Fisher distribution extends that concept to the S^n hyper sphere with uniform variance, while the Fisher-Bingham allows for covariance on S^2 [26]. The log-normal distribution gives a single bounded Gaussian equivalent, and the log odds normal a double bounded one. Both behave very like a standard normal distribution when the mean is several standard deviations or more away from the bounds, both achieve very large densities as the mean approaches a bound, and both are zero at and beyond their bounds. The potential advantages of using such distributions that match one's meta-knowledge of a data set should be obvious, yet they seem to have been largely ignored in the anomaly detection field. This may be a matter of excessive choice. Wikipedia's continuous probability distributions page [20] currently lists 95 alternatives and is not up to date. Most are extremely specialized, but it will worth one's effort to investigate which best fit reasonable expectations about specific data.

Similar considerations apply to discrete numerical data, such as counts or ordinals, where continuous distributions like the Gaussian are generally inappropriate, and casting to categoricals to suit a multinomial model can destroy much information. The Poisson is the canonical distribution for counts and should always be considered for such attributes, but there are other alternatives.

4) Support Vector Machines: For fault detection one can use semi-supervised one-class SVMs, trained on presumed normal data. With suitable kernels these can learn a complex boundary around the normal region, and any test cases falling outside are declared faults. For fault *diagnosis* a supervised approach is needed, for training standard SVMs on labeled fault classes. Here the binary separation that is fundamental to standard SVMs becomes a problem. One can either train an SVM for each class, in a one against the rest mode, or train on every pair of classes. In both cases the class *assignment* is usually *decided* on a winner take all basis, the first by maximum class weight and the second by majority voting.

There are a number of procedures advocated for converting the results of SVMs and similar *binary decision classifiers* into class probabilities. See the survey by Gebel & Weihs [29] for an introduction. They have since adopted a Dirichlet distribution based approach [30]. A problem with this is the need for $O(n^2)$ binary classifiers, where n is the number of known normal and fault classes to be identified.

Relevance Vector Machines are an SVM inspired variant designed specifically to return probabilities. To achieve this they forego the standard SVMs' guarantee of optimality modulo the choice of kernel function, while remaining binary classifiers trained on labeled data. The probabilities resulting from a set of one against the rest classifiers could reasonably be taken as likelihoods, and normalized to get an overall distribution. Deployed as binary classifiers, the methods of [30] could be used to derive class probabilities. Detecting unknown classes might be problematic.

5) Rule Based Classification: Supervised rule based classification, as decision trees, classification trees or regression trees, has a long and successful history [31], [32]. The basic idea is to recursively partition labeled data into sub-spaces in a way that maximizes the difference between the two part's label statistics. This normally builds a binary tree that terminates in leaves that hold only a single label. Each partitioning is usually done on a single attribute, rarely in the same order on parallel branches. Overfitting is a common problem, and most methods apply a post-partitioning pruning step which removes sparsely populated partitions at the expense of having leaves with mixed labels.

The chief advantage of decision tree classifiers is the straight forward interpretation of the resulting rules. As with most parametric methods, construction is slow while application is quite fast. There is the usual range of variations among implementations, primarily on how to choose which attribute to split upon, and where, for each non-terminal branch. Recent work on minimal trees [33] confirms that often only a fraction of attributes need be considered.

The downside of standard decision trees is that they make a decision at each partitioning, and so provide no measure of how strongly any result should be believed, nor any probabilities for alternative results. Without these, any further inference is conditional on a correct decision, and so is necessarily suspect. However there are variants that do return probabilities, while simultaneously minimizing the overfitting problem [34], which eliminate the principle caveats.

6) k-Nearest Neighbor: K-nearest neighbor is a supervised classification technique that attempts to finesse the representational limitations of the simple mathematical density models by substituting a large labeled set of training instances as a de facto density model. Each subset sharing a label thus defines a non-parametric distribution. For each instance to be classified, the k nearest neighbors are determined, and their labels used to determine the test instance's class, usually by majority vote.

The naive computation time is $O(n_t n_c)$, where n_t and n_c are training and classification data set sizes. Much effort has gone into minimizing this. There is the expected variation in methods for choosing k, measuring distance, and combining distance and labels to determine a class. Independent of simple noise, distance measures are degraded by the presence of irrelevant attributes or attribute scales that are inconsistent with their relevance. Thus both feature selection and rescaling may be an important preliminary aspect of k-Nearest Neighbor classification.

B. Nearest Neighbor Detection Techniques

Nearest Neighbor based anomaly *detection* techniques are non-parametric in the sense of not needing additional parameters to specify a model. The training data provides the model, as an implicit density distribution over the parameter space, and thus is the parameterization. The assumption is that *Normal data instances occur in dense neighborhoods, while anomalies occur far from their closest neighbors.* There are two general techniques:

- Use the distance to the k^{th} nearest neighbor.
- Compute the relative density near an instance.

Conceptually these are roughly equivalent: the distance to the k^{th} nearest neighbor defines a local hypersphere containing only k training instances, and thus maps directly to a local hyper-density. Implementations differ considerably, in how to specify k, how distances or densities are computed, how distances or densities are used to score an anomaly, and how the naive necessity for comparing each test instance against all training instances can be finessed [18]. A standard mode for using nearest neighbor is to seek anomalies by comparing a data set against itself. This is an $O(N^2)$ operation, hence the emphasis on efficient techniques [35].

For anomaly *diagnosis* in a supervised mode, the obvious extension is to combine a k nearest neighbor classifier for normal modes and known anomalies, with a k^{th} nearest neighbor distance or density based outlier detector to catch unknown anomalies. As the operating mode shifts from supervised through semi-supervised to unsupervised, the classifier aspect takes on increasing importance. But since diagnosis requires, at the very least, knowledge of a mapping from symptoms to labels, the objective of unknown anomaly detection will be to locate anomalies for identification by domain experts.

C. Clustering

Clustering is normally done in an unsupervised mode to group instances that are collocated in the native or a projected data space. Thus it seeks to find natural classes, w.r.t. the instance descriptions and a similarity measure. This is an old field, with a wide variety of clustering techniques developed for many application areas. Berkhin's survey [36] of clustering for machine learning lists about 50 named programs. Popular variations are Self-Organizing Maps (SOM), K-Means Clustering, and Probabilistic Clustering via Expectation Maximization (EM) optimization. For unsupervised anomaly detection data is first clustered and then instances are examined for either low degree of cluster membership in any cluster or high membership in small or sparse clusters. In a semi-supervised mode only normal mode data is clustered, then suspect instances are tested, those not falling within a cluster being considered anomalies.

Unsupervised clustering generates clusters that are optimal w.r.t. the clustering algorithm and the data used. Unless both algorithm and data are carefully matched, the clusters may have little correspondence to any properties of interest. Supervised clustering can make use of instance labels in training data to identify an attribute subset, or perhaps a projection, that is optimal to a particular task, modulo the clustering algorithm. This can be done via cross-validation or application of the Bayesian Information Criteria on training data. Supervised clustering on data so selected has given excellent results [37].

Clustering for *diagnosis* must either take note of mode labels in the clustering, thus acquiring a flavor of classification,

or must later learn a mapping from clusters to mode labels. The former might be achieved by partitioning data w.r.t. to mode and clustering within each mode to get "pure" clusters. This is akin to classification with extended classes represented by a sum of distributions. The latter might be achieved using a probabilistic clustering, and associating a distribution over mode labels with each cluster. A new instance's mode probabilities are then computed as the cluster probability weighted sum over the mode distributions. In essence, this defines a multivariate mode probability distribution over the data space that is similar to that implicit in a k-nearest neighbors algorithm.

For data space location based clustering, a common model defines clusters as noisy points in a possibly extended data space, and assigns instances to the nearest cluster, modulo some distance measure. The distance need not be isotropic, and variants on the Mahanalobus are popular: $d_M(x, \mu, \Sigma) = \sqrt{((x-\mu)\Sigma^{-1}(x-\mu))}$, where μ is the cluster center and Σ is the cluster covariance matrix. The Mahanalobus degenerates to the Euclidean distance as $\Sigma \to I$, allows for simple independent attribute scaling when $\Sigma = diag(\sigma)^2$, or for covariant scaling when $\Sigma = diag(\sigma)Cdiag(\sigma)$ for correlation matrix C. The Gaussian radial basis function is a popular variation, with $d_G(x, \mu, \Sigma) = exp(-d_M(x, \mu, \Sigma)/2)$. Normalized to unit mass, these give the Multivariate Gaussian probability model: $P(x|\mu, \Sigma) = (2\pi)^{-n/2} \det(\Sigma)^{-1/2} d_G(x, \mu, \Sigma)$, and a probabilistic interpretation of class membership [23].

D. Classical Statistical Techniques

Classical frequentist statistics suffers from the assumption that randomness is a property of nature that afflicts numerical description of nature, and that the concept of probability is only relevant to such random variables. Conceptually, a random variable is an abstraction of a measurement process, and any set of observed values are a sample from a conceptually infinite population. Thus probabilities are deemed to be long-run relative frequencies representing results of a sampling process that generates random variables. Sampling processes are to be described by sampling distributions, mathematical functions that model the process by defining relative frequencies w.r.t. to some parameters.

Given a sampling distribution with known parameters we can compute likelihoods for any set of observations. Lacking knowledge of one or more parameter values, they are to be estimated from a statistic, e.g. sample mean or variance, computed as a function of the sample values. All such statistics are themselves random variables, but a useful statistic must have its own known sampling distribution, typically a function of the degree of freedom remaining in the sample. Only in such cases is it possible to interpret the significance of a measured sample's statistic. This is expressed in terms of confidence intervals, as the frequency, over many repeats of the data collection process, that the computed intervals would include the true value. In strict formality, nothing can be said regarding the probability of correctness for any particular value, or even that the true value lies within the confidence interval computed from any given sample, although confidence intervals are often informally interpreted so.

The object of inference is generally to determine the degree of support that a set of observations provides to alternate hypothesis, logical statements regarding the nature of the system observed. In this, classical statistics is crippled by the assumption that probabilities are solely a property of random variables. Hypothesis being either true or false, they cannot be random variables. The classical approach is then to consider each hypothesis in turn, choosing a statistic that can be computed from both the observations and from a reference distribution representing many repeats of the data collection process, assuming the hypothesis is true. If the observed statistic falls in a sufficiently unlikely spot on the distribution, the hypothesis is rejected at some degree of confidence. The degree of confidence is essentially the frequency with which repeated experiments would generate more likely values of the statistic. The numerical degree deemed to justify rejection varies considerably between different fields, giving it a certain arbitariness. Failure to reject a hypothesis does not imply it acceptance, merely that the current data is not incompatible at the stated degree of confidence, and says little if anything about its standing w.r.t. other hypothesis.

Given the above, and a viable alternative in Bayesian inference techniques, there seems little point in pursuing the formal methods of classical statistics as a basis for applied inference. The reader who finds this judgment too harsh should study E.T. Jaynes' development of probability as the extension of logic to uncertain situations [38]. Jaynes' earlier paper on "Confidence Intervals vs Bayesian Intervals", reproduced in Rosenkrantz [39], provides a deep analysis of the problems inherent in the classical approach to parameter estimation. For a reasonably balanced exposition of the motivation and application of both approaches to inference in scientific applications, see Gregory [40].

Despite rejecting the methods of classical statistics, it will not do to reject the body of its work. A great many very capable people devoted their careers to wringing useful results out of the only method available to them, and there is much of value in what they achieved. In particular, the classical development of sampling distributions provides the likelihoods that are the hart of Bayesian inference. See Chandola's survey [18], section 7, for an overview and references to the Classical Statistics based anomaly detection literature.

E. Dimension Reduction

This encompasses a variety of techniques that attempt to project high dimension data to a low dimension space, while preserving most of the data's variation, in order to facilitate application of other techniques. Thus these are basically preprocessing techniques. Principal Component Analysis (PCA) is the canonical example, with a number of variations, and very suitable for location based techniques. Non-negative Matrix Factorization (NMF) is another approach, producing non-negative basis and weights interpretable as an additive representation. With a large set of ASRS incident narratives, NMF applied to reducing a bag-of-words parsing has generated reduced basis vectors that clearly and consistently group words into reasonable domain concepts. A simple supervised classifier, applied to the resulting basis and instance weights, does a good job of matching expert fault assignments for intermediate size fault types, but less so for very common or rare types [41].

Generating the projection usually has high computational cost. Applying a fixed projection to new instances is comparatively quick. But some information is inevitably lost in reducing the data dimension. Standard dimension reduction techniques concentrate on identifying and retaining the dominant component's variations. This can pose problems for fault detection and diagnosis in well engineered systems. Low fault probabilities mean low fault frequencies and data variations dominated by the normal operating modes. Thus there is a danger that evidence for faults will be discarded with the noise. A fault pattern will need to be both distinctly different from nearby normal modes, and present in significant quantity to remain separable in the reduced dimension representation. So these approaches currently seem inappropriate for fault detection and diagnosis in the aircraft operations domain.

VI. PROPOSED DIRECTIONS

In taking a strongly data-centric approach we avoid the necessity of expressing a deep and inevitably specific understanding of the system under study, while foregoing the ability to diagnose at that level. We are in some senses reduced to the role of medical diagnosticians of two centuries ago, who could put names to many common problems, and provide effective treatment, without having a fundamental understanding of those problems' causes. But we have the advantages of often extensive sensor sets, diverse sensor types returning quantified results, potentially long term records tracking problem development, machines able to record and organize such data, and algorithms able to monitor and correlate details across entire data sets. Given efficient use of these, a data-centric diagnostic system could be quite sensitive to developing problems, and to identifying known anomaly patterns, even if it knows nothing of the diagnosed system's internals.

However the key element to achieving this will be efficient use of expert knowledge to help organize our understanding of, and provide diagnostic labels for, the fault patterns discovered in our data. Operational data of the sort that we envision using, such as FOQA records, cannot provide either. Annotated operational data sets, identifying fault types to a degree suitable for diagnosis, will be relatively rare since their generation is costly in terms of annotation effort. Thus efficient use of annotation effort dictates that it be concentrated on previously detected fault suspects. So we will seek anomaly detection methods that identify candidate sets for expert analysis. Where annotated data sets already exist, they should inform our evaluation of detection techniques, keeping in mind that they only identify some possible fault types. This stage is a focused knowledge capture effort to obtain the annotated fault instances needed to build operational fault classifiers. Once such data is available we have a variety of techniques for detecting and diagnosing known fault types, and potentially identifying new unknown fault candidates, in both archived and runtime operations data.

Probabilistic classification will be fundamental to applying data-centric diagnosis in operations. An important factor in

improving on current approaches will lie in making maximum use of available meta-data regarding data generation, collection, recording and prior processing. The idea is to model the effects of the data creation process, in a fairly generic way that will be easily adaptable to specific data sets, so as to extract maximum information for our diagnostic inference.

REFERENCES

- [1] T. Kurtoglu, S. Narasimhan, S. Poll, D. Garcia, L. Kuhn, J. de Kleer, A. van Gemund, and A. Feldman, "First international diagnosis competition - DXC'09," in 20th International Workshop on Principles of Diagnosis. DX-09, jun 2009, pp. 383–396. [Online]. Available: http://www2.parc.com/isl/members/lkuhn/paper/2009/dx09-dxc.pdf
- [2] J. O. Berger, Stastical Decision Theory and Bayesian Analysis, 2nd ed., ser. Springer Series in Statistics. Springer, New York, 1985.
- [3] K. H. Knuth, "Lattice duality: The origin of probability and entropy," *Neurocomputing*, vol. 67, no. on Geometrical Methods in Neural Networks and Learning, pp. 245 – 274, aug 2005. [Online]. Available: http://www.huginn.com/knuth/papers/ knuth-neurocomp-05-published.pdf
- [4] K. H. Knuth, P. M. Erner, and S. Frasso, "Designing intellegent machines," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ser. Aip Conference Proc., K. H. Knuth, A. Caticha, J. Julina L Center, A. Giffin, and C. C. Rodriguez, Eds., vol. 954. AIP, jul 2007, pp. 203–211.
- [5] K. R. Wheeler, T. Kurtoglu, and S. D. Poll, "A survey of health management user objectives related to diagnostic and prognostic metrics," in ASME International Design Engineering Technical Conferences (IDETC), 29th Computers and Information in Engineering Conference (CIE). ASME, aug 2009.
- [6] S. R. Dixon, C. D. Wickens, and J. S. McCarley, "On the independence of compliance and reliance: Are automation false alarms worse than misses?" *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 49, no. 4, 2007.
- [7] C. D. Wickens and J. S. McCarley, Applied Attention Theory. CRC Press, 2008.
- [8] A. Zymnis, S. Boyd, and D. Gorinevsky, "Mixed linear system estimation and identification," *Signal Processing*, vol. 90, no. 3, pp. 966–971, mar 2010. [Online]. Available: http://www.stanford.edu/ ~boyd/papers/pdf/mixed_system_est.pdf
- [9] S. Budalakoti, A. Srivastava, and M. Otey, "Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 39, no. 1, pp. 101–113, jan 2009. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp. jsp?tp=&arnumber=4694106&isnumber=4735634
- [10] V. Chandola, V. Mithal, and V. Kumar, "Understanding anomaly detection techniques for symbolic sequences," Computer Science & Engineering, University of Minnesota, technical report 09-001, may 2009. [Online]. Available: http://www.cs.umn.edu/tech_reports_upload/ tr2009/09-001.pdf
- [11] V. Chandola, S. Boriah, and V. Kumar, "Understanding categorical similarity measures for outlier detection," Computer Science & Engineering, U of Minnesota, Technical Report TR 08-008, mar 2008. [Online]. Available: http://www.cs.umn.edu/tech_reports_upload/tr2008/ 08-008.pdf
- [12] E. Rahm1 and P. A. Bernstein, "A survey of approaches to automatic schema matching," *The VLDB Journal*, vol. 10, no. 4, pp. 334–350, dec 2001. [Online]. Available: http://springerlink.metapress.com/content/ y3bavwk2t7328hat/fulltext.pdf
- [13] S. Chapman. Sam's string metrics web page. [Online]. Available: http://www.dcs.shef.ac.uk/~sam/stringmetrics.html
- [14] S. Narasimhan, "Automated diagnosis of physical systems," in http://www.jacow.org/, ICALEPCS-07. Joint Accelerator Conferences Website, oct 2007, pp. 701 – 705. [Online]. Available: http: //ti.arc.nasa.gov/m/project/hyde/ICALEPCS07.PDF
- [15] O. J. Mengshoel, A. Darwiche, K. Cascio, M. Chavira, S. Poll, and S. Uckun, "Diagnosing faults in electrical power systems of spacecraft and aircraft," in *Proceedings of the Twentieth Innovative Applications* of Artificial Intelligence Conference (IAAI-08), Chicago, IL, 2008, pp. 1699–1705. [Online]. Available: https://dashlink.arc.nasa.gov/static/ dashlink/media/topic/IAAI2008.pdf

- [16] D. Luchinsky, V. Osipov, V. Smelyanskiy, D. Timucin, S. Uckun, B. Hayashida, M. Watson, J. McMillin, D. Shook, M. Johnson, and S. Hyde, "Fault diagnostics and prognostics for large segmented SRMs," in *Aerospace Conference*, 2009 IEEE, mar 2009, pp. 1– 8. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp= &arnumber=4839622&isnumber=4839294
- [17] D. Gorinevsky, S. P. Boyd, and S. Poll, "Estimation of faults in DC electrical power system," in 2009 American Control Conference. Control Systems Society & American Automatic Control Council, jun 2009, p. 6p. [Online]. Available: www.stanford.edu/~gorin/papers/ ACC09_ADAPT.pdf
- [18] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," Computer Science & Engineering, U of Minnesota, Technical Report TR 07-017, aug 2007, alternate of ACM1541882. [Online]. Available: http://www.cs.umn.edu/tech_reports_upload/tr2007/07-017.pdf
- [19] anon. Category: Continuous probability distributions. [Online]. Available: http://en.wikipedia.org/wiki/Category:Discrete_distributions
- [20] ——. Category: Continuous probability distributions. [Online]. Available: http://en.wikipedia.org/wiki/Category:Continuous_distributions
- [21] T. Menzies and A. Orrego, "Incremental discretization and bayes classifiers handles concept drift and scales very well," aug 2005, submitted to IEEE Transactions on Knowledge and Data Engineering. [Online]. Available: http://menzies.us/pdf/05sawtooth.pdf
- [22] T. Menzies, D. Allen, and A. Orrego, "Bayesian anomaly detection (BAD v0. 1)," Workshop on Machine Learning Algorithms for Surveillance and Event Detection, ICML-06, jun 2006. [Online]. Available: http://web.engr.oregonstate.edu/~wong/workshops/icml2006/ papers/menzies.pdf
- [23] P. Cheeseman and J. Stutz, "Bayesian classification (AutoClass): Theory and results," in Advances in Knowledge Discovery and Data Mining, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AIAA Press, MIT Press, 1996, pp. 153–180. [Online]. Available: http://ti.arc.nasa.gov/m/project/autoclass/kdd-95.ps
- [24] S. Hayden, N. Oza, R. Mah, R. Mackey, S. Narasimhana, G. Karsai, S. Poll, S. Deb, and M. Shirley, "Diagnostic technology evaluation report for on-board crew launch vehicle," National Aeronautics and Space Administration, Ames Research Center, NASA STI Technical Memorandum NASA/TM-2006-214552, sep 2006. [Online]. Available: http://ti.arc.nasa.gov/m/pub/1218h/1218% 20(Hayden).pdf
- [25] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek, "Robust, complete, and efficient correlation clustering," in *Proc of the 7'th SIAM International Conference on Data Mining*, C. Apte, B. Liu, S. Parthasarathy, and D. Skillicorn, Eds. SIAM, apr 2007, pp. 413–418. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.105.6687&rep=rep1&type=pdf
- [26] J. T. Kent, "Fisher-bingham distribution on the sphere," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 44, no. 1, pp. 71–80, 1982. [Online]. Available: http://www.jstor.org/stable/2984712
- [27] R. M. Neal, Bayesian Learning for Neural Networks, ser. Lecture Notes in Statistics. Springer, New York, 1996, no. 118. [Online]. Available: http://www.cs.toronto.edu/~radford/bnn.book.html
- [28] D. J. C. MacKay, Neural Networks and Machine Learning, ser. NATO ASI Series. Springer, Berlin, 1998, vol. 168, ch. Introduction to Gaussian processes, pp. 133–165. [Online]. Available: http://www.inference.phy.cam.ac.uk/mackay/gpB.pdf
- [29] M. Gebel and C. Weihs, "Calibrating classifier scores into probabilities," in Advances in Data Analysis, ser. Studies in Classification, Data Analysis, and Knowledge Organization, R. Decker and H. J. Lenz, Eds. Springer Berlin Heidelberg, mar 2006, pp. 141–148. [Online]. Available: http://www.springerlink.com/content/u3221j45178h5177/fulltext.pdf
- [30] —, "Calibrating margin-based classifier scores into polychotomous probabilities," in *Data Analysis, Machine Learning and Applications*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. Springer Berlin Heidelberg, mar 2007, pp. 29–36. [Online]. Available: http://www.springerlink.com/content/j6245548078t6k71/fulltext.pdf
- [31] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," Wadsworth International, Tech. Rep., 1984.
- [32] R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kauffman, 1992.
- [33] T. Menzies and Y. Hu, "Just enough learning (of association rules): The TAR2 Treatment Learner," *Artificial Intelligence Review*, vol. 25, pp. 211–299, 2006. [Online]. Available: http://menzies.us/pdf/07tar2.pdf
- [34] P. J. Tan and D. L. Dowe, "MML inference of oblique decision trees," in *Proc. 17th Australian Joint Conference on Artificial Intelligence*, ser. Lecture Notes in Artificial Intelligence (LNAI), no. 3339. Springer-

Verlag, dec 2004, pp. 1082–1088. [Online]. Available: http://www.csse. monash.edu.au/~dld/Publications/2004/Tan+DoweAI2004.pdf

- [35] S. D. Bay and M. Schwabacher, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, L. Getoor, T. E. Senator, P. Domingos, and C. Faloutsos, Eds. ACM, aug 2003, pp. 29–38. [Online]. Available: http://doi.acm.org/10.1145/956750.956758
- [36] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping Multidimensional Data Recent Advances in Clustering*, J. Kogan, C. Nicholas, and M. Teboulle, Eds. Springer Berlin Heidelberg, 2006, pp. 25–71. [Online]. Available: http://www.springerlink.com/content/x321256p66512121/fulltext.pdf
- [37] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," University of Washington, Tech. Rep. TR 380, oct 2000. [Online]. Available: http://citeseerx.ist.psu.edu/ viewdoc/download?doi=10.1.1.74.7229&rep=rep1&type=pdf
- [38] E. T. Jaynes, *Probability Theory The Logic of Science*, G. L. Bretthorst, Ed. Cambridge University Press, 2003.
- [39] —, E.T.Jaynes: Papers on Probability, Statistics and Statistical Physics, ser. Synthese Library, R. D. Rosenkrantz, Ed. Dordrecht, Holland: Reidel, 1983, vol. 158.
- [40] P. C. Gregory, Bayesian Logical Data Analysis for the Physical Sciences. Cambridge University Press, 2005.
- [41] N. Oza, J. P. Castle, and J. Stutz, "Classification of aeronautics system health and safety documents," *IEEE Transactions on Systems, Man, ans Cybernetics - Part C: Applications and Reviews*, vol. 39, no. 6, nov 2009.