AFRL-IF-RS-TR-2004-323
Final Technical Report
November 2004

REAL-TIME POPULATION HEALTH DETECTOR

General Dynamics

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. M165

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2004-323 has been reviewed and is approved for publication

APPROVED: /s/
RAYMOND A. LIUZZI
Project Engineer

FOR THE DIRECTOR: /s/
JAMES A. COLLINS, Acting Chief
Information Technology Division
Information Directorate

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** November 2004	**3. REPORT TYPE AND DATES COVERED** **Final Sep 01 – May 04**

4. TITLE AND SUBTITLE

REAL-TIME POPULATION HEALTH DETECTOR

5. FUNDING NUMBERS

C - F30602-01-C-0138
PE - 62301E
PR - BIOS
TA - 00
WU - 03

6. AUTHOR(S)

Gene E. McClellan, General Dynamics
Mark Musen, Stanford University

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

General Dynamics
1400 Key Blvd, Suite 100
Arlington VA 22209

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/IFTB
525 Brooks Road
Rome NY 13441-4505

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

AFRL-IF-RS-TR-2004-323

11. SUPPLEMENTARY NOTES

AFRL Project Engineer: Raymond A. Liuzzi/IFTB/(315) 330-3577

12a. DISTRIBUTION / AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

The objective of the Bio-ALIRT is to evaluate the utility of non-traditional data sources to enable early detection of potential bio-terrorist attacks. Early detection will permit prompt intervention and appropriate treatment, potentially saving large numbers of lives and avoiding widespread infection among military and civilian populations. General Dynamics (then Veridian Systems Division), in cooperation with Stanford University, performed research and evaluated bio-terrorism detection services in the Hampton Roads area of Virginia. The effort resulted in development of analytical technology and a prototype Real-time Population Health Detector (RPHD) application relying on non-traditional data collected from a variety of sources to detect aggregate changes in people’s behavior in a community that indicate the early outbreak of illness. This report examines the problem of detecting “anomalous signals” in a mixed signal background, consisting of periodic and random contributions; and which mixtures are posited to exhibit, to a certain extent, what might be called a state of “statistical stability” for significant periods of time-at least until some “signal” of interest obtrudes. The context is, of course, that in which we monitor “non-traditional” indicators for the potential early manifestation of responses to disease progression in a given population. To meet the complex operational and research needs of these surveillance applications, this report also describes a system called BioSTORM (Biological Spatio-Temporal Outbreak Reasoning Module. BioSTORM is a computational framework that provides run-time mediation between data sources and problem solvers with the goal of meeting the performance and flexibility demands of emerging disease surveillance systems.

14. SUBJECT TERMS

Bio-Terrorism, Early Detection, Knowledge-Base Technology, Statistical Methods, Disease Outbreak

15. NUMBER OF PAGES 33

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

List of Figures

1 Introduction

This report is the final technical report from the General Dynamics-Stanford Team for Defense Advanced Research Projects Agency's (DARPA) Bio Advanced Leading Indicator Recognition Technology (Bio-ALIRT) program after early termination of the program by the Government. The General Dynamics-Stanford Team was funded by Contract F30602-01-C-0138 through the Air Force Research Laboratory (AFRL). According to DARPA’s instructions to document final status, this report describes the final technical product and the data collections status as of the end of the program.

1.1 Contract Objectives

The objective of the Bio-ALIRT program and this contract is to evaluate the utility of non-traditional data sources to enable early detection of potential bio-terrorist attacks. Early detection will permit prompt intervention and appropriate treatment, potentially saving large numbers of lives and avoiding widespread infection among military and civilian populations.

General Dynamics (then Veridian Systems Division), in cooperation with Stanford University, won a competitive DARPA contract awarded through the Air Force Research Laboratory (AFRL) to research and evaluate bio-terrorism detection services in the Hampton Roads area of Virginia. The effort resulted in development of analytical technology and a prototype Real-time Population Health Detector (RPHD) application relying on non-traditional data collected from a variety of sources to detect aggregate changes in people’s behavior in a community that indicate the early outbreak of illness.

2 Final Summary

2.1 Outbreak Detection Methods

2.1.1 Anomaly Detection by Linear Mean-square Forecasting Using Kalman Filters

We are concerned with the problem of detecting “anomalous signals” in a mixed signal background, consisting of periodic and random contributions, and which mixtures are postulated to exhibit, to a certain extent, what might be called a state of “statistical stability” for significant periods of time, at least until some “signal” of interest obtrudes. The context is, of course, monitoring “non-traditional” indicators for the potential early manifestation of responses to disease progression in a given population (e.g., over-the-counter drug sales).

A scheme based on a *forecasting* model endeavors to tell us whether the most recent conditions are sufficiently different from those predicted by a model of expectation, such that we may classify these conditions as “anomalous”. *Expectation* models classify observations by their degree of novelty or surprise.

Expectation-based detection methods use linear mean square forecasting on observed history of data (or transformed data) to yield optimal predictions for data not yet observed.

2.1.1.1 Schematic of Forecast-based Anomaly Detection System

The process of time-series forecasting that we seek to apply to the anomaly detection problem consists in three steps:

1. Model selection
2. Model fitting
3. Model application: Forecasting and inference

The very act of choosing the structural form above is our rudimentary foray into the field of “model selection.” That is, we assert that the dynamics we seek to characterize are adequately “modeled” in general by a structural system of this type. This amounts to a dynamic, adjustable local linear trend and level being fit to the data. Figure 1 details this process of model selection and model fitting.

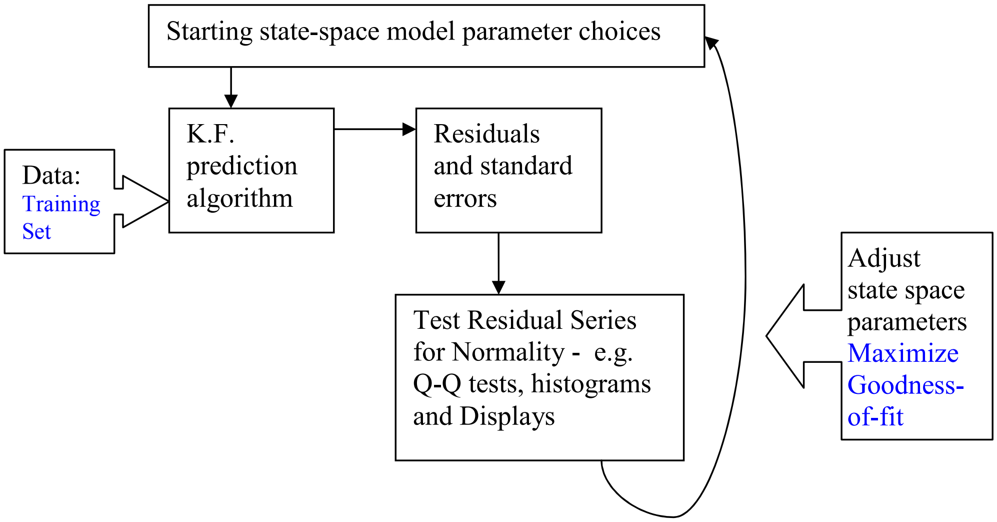

Starting state-space model parameter choices
K.F. prediction algorithm
Data: Training Set
Residuals and standard errors
Test Residual Series for Normality - e.g. Q-Q tests, histograms and Displays
Adjust state space parameters Maximize Goodness-of-fit

Figure 1. Flow diagram of the model fitting process for Forecast-based Detection.

After fitting, the forecasting model is used on multiple streams of data; to produce outputs consisting of "forecasts" (expected observation conditional on past) and "residuals" (observed difference between forecast and actual observation when it becomes available). In addition to the "forecasts" and the "residuals," summary "diagnostic markings" are produced that condense the results of statistical tests applied to the series of forecasts and residuals.

By using the series of "diagnostic markings" alone, it is possible to explain whether or not and, if so, *when* a given *single* series of data exhibits dynamics that are markedly different from the dynamics exhibited during an historical "representative period." This period was taken to be representative of the "normal state of affairs" and, as such, was used to fit the forecasting models themselves.

Of the several types of "diagnostic markings", each carries with it different information and allows one to infer whether the series of data exhibits anomalous dynamics with respect to three distinct explanatory dimensions. This is done by considering any one or a combination of the following three numerical "tests" performed upon the series of observations, forecasts and forecast residuals:

a. A test which determines whether the latest observation is large in respect to the running one-step forward forecast of expectation.

b. A test which determines whether there is an anomalous trend, or drift in the recent sequence of observations inconsistent with our previous understanding of the stable 'state of affairs', as represented by the dynamics of the fitted forecast model..

c. A test which determines whether the forecast value for the subsequent day (or several days) is large with respect to a fixed or a-priori magnitude, representative of epidemiological conditions warranting intervention.

Test **a** evaluates the extremity of the latest forecast residual in a statistical prediction interval, that is a by-product of the optimized forecasting scheme. Test **b** comprises a class of tests which to one extent or another evaluate the auto-correlation of the residual series, and also under the "null" hypothesis indicate that the dynamics in the present period are close to that of the model fitting period. These residuals should have statistically insignificant autocorrelation. Test **c** evaluates the magnitude of the forecast in light of a standard of *practical significance* (as opposed to *statistical significance*) such as may be developed from an epidemic model which gauges the effect of a particular hypothetical epidemic upon the data aggregates in question.

Tests for residual autocorrelation and / or drift away from the expected "zero-mean" condition include direct inference on the sample auto-correlation function; as well as so-called "Cumulative Sum" (CUSUM) tests on the integrated residual series. The forecasting and anomaly inference process is displayed in Figure 2.

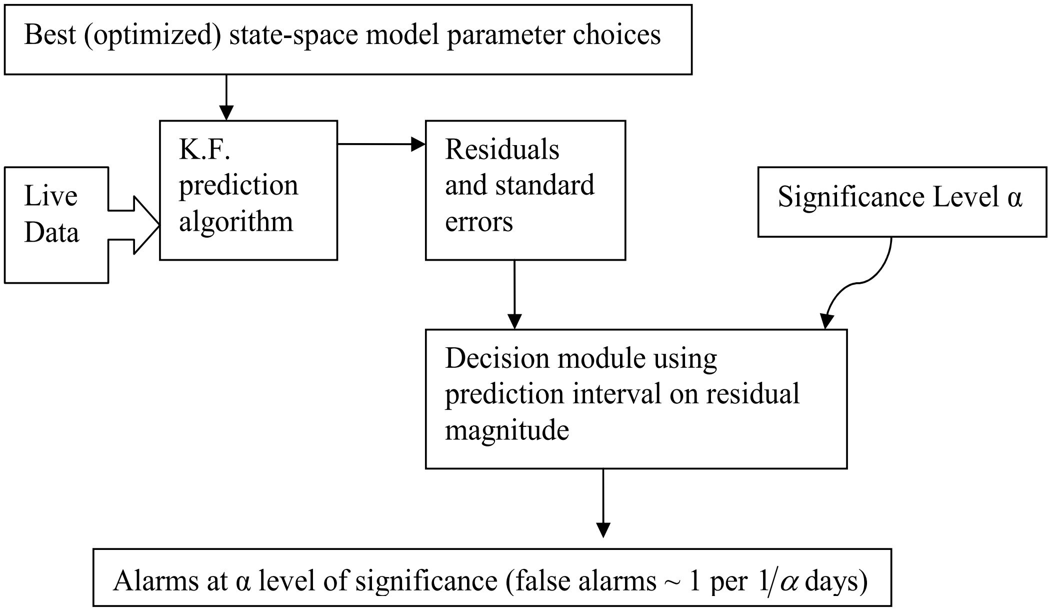

Best (optimized) state-space model parameter choices
Live Data
K.F. prediction algorithm
Residuals and standard errors
Significance Level α
Decision module using prediction interval on residual magnitude
Alarms at α level of significance (false alarms ~ 1 per 1/α days)

Figure 2. Flow diagram for the forecast and inference anomaly detection process.

2.1.1.2 State Space Time Series Linear Model Formulation

A stochastic observed time-varying quantity y_t of dimension N may be modeled as a non-stationary random process comprising the concomitant dynamic evolution of an unobserved set of stochastic state variables (by a linear Markov process) and an additional linear transformation. A decomposition of a random process of this sort is sometimes referred to as an unobserved components model or an unobserved components decomposition of the variability in a random process. It affords a high degree of flexibility and may be especially well-suited to certain problems where prior information may be brought to bear in the selection of components and their dynamic behavior. These models may be expressed in general as a stochastic difference equation as:

$y_t = Hx_t + \eta_t$ Observation equation

$x_t = \Phi x_{t-1} + \nu_t$ State Transition Equation

If the column vector $\bar{x}$ is of dimension $M \times 1$ then the matrix Φ must be dimension $M \times M$. If the observation vector is of size N then the matrix H must be size $N \times M$.

The observation error η_t and the state disturbance ν_t are mutually uncorrelated Gaussian random variables of zero mean, zero serial correlation (white-noise) with stipulated covariance matrices:

$$D(\bar{\eta}_t) \equiv E\left(\eta_t \, \eta_t^T\right) = \Omega_t$$

$$D(\bar{\nu}_t) \equiv E\left(\nu_t \, \nu_t^T\right) = \Psi_t$$

It is assumed that the matrices H, Φ, Ω, Ψ are known for all times *t=1,2,…,n* and that an initial estimate of the state at time *t=0*, x_0 , is available, and an initial of the covariance matrix of this state is also available at time zero: $D(\bar{x}_0) \equiv E\left(x_0^J x_0^K\right) = P_0$. In fact, these matrices are *free parameters* (sometimes called model hyperparameters) and must be estimated using a maximum likelihood procedure applied to a suitable “training set”. That is to say, the model must be *fit* to representative data *before* it is utilized in a prediction of forecasting scheme.

We denote the “information set” I_t at time *t* to be the set of all observations at times $t = 1,...,t$ that is:

$$I_t \equiv \{y_1, y_2, ..., y_{t-1}, y_t\}$$

At each time *t,* the *Kalman Filter recursions* (see Durbin, J. and S.J. Koopman, *Time Series Analysis by State Space Methods*, Oxford University Press, 2001) provide estimates of the state-vector distributional mean and covariance, *conditional on the accrued information set*, in two substeps.

1.) $\mu_{t|I_{t-1}} = E\left(x_t | I_{t-1}\right)$ and $P_{t|I_{t-1}} = D\left(x_t | I_{t-1}\right)$

2.) $\mu_{t|I_t} = E\left(x_t | I_t\right)$ and $P_{t|I_t} = D\left(x_t | I_t\right)$

The first sub-step produces estimates of the distributional parameters for time *t* conditional on information up through time *t-1*.

The second sub-step produces "corrected" estimates of the distributional parameters at time *t* conditional on information up through time *t* (including the latest observation, at time *t*).

The quantities $\mu_{t|t-1}$ and $P_{t|t-1}$ (suppressing the symbol *"I"* for the obvious shorthand) are *"optimal one step ahead forecasts*" of the distributional parameters for the state.

Since by construction, the expected value and covariance of the observation at any time are conditional on information up until *t-1*:

$$E(y_t \mid I_{t-1}) = E(Hx_t \mid I_{t-1}) = H\mu_{t|t-1}$$

$$D(y_t \mid I_{t-1}) = HP_{t|t-1}H^T + \Omega_t$$

Supposing that we regard the information set I_{t-1} as representative of a state of statistical (and actual) stability, we may, using the optimal estimators of the expectation and dispersion on the right hand side of this equation, evaluate the "extremity" of a new observation y_t with reference to the multivariate Gaussian distribution which may be denoted symbolically as:

$$N\left(H\mu_{t|t-1}, HP_{t|t-1}H^T + \Omega\right)$$

2.1.1.3 Statistical Inference on Prediction Residuals

We proceed at every time-step to evaluate the extremity of the newest observation y_t by examining the associated prediction error e_t in reference to the putatively stable statistical model consistent with the information set I_{t-1}: that is, with the multivariate Gaussian distribution:

$$e_t \sim^{(?)} N\left(0, HD(x_t \mid I_{t-1})H^T + \Omega\right) = N(0, S)$$

For vector processes, this relationship must be regarded as a multivariate hypothesis test for the extremity of the residual vector e_t. As such, the standard inferential tests involve interpreting the size of the quadratic form $e^T S^{-1} e$ with respect to the requisite tail probability of Fishers *F*-distribution with the number of degrees of freedom. This becomes asymptotically a Chi-squared statistic after the prediction scheme has been run through a moderate number of time-steps. For scalar processes, the squared residual magnitude e^2 is put to test in respect to the student t-distribution; or the standard unit normal distribution after a moderate number of time-steps have been counted.

2.1.1.4 Local Level Structural Model

Raw count data is subjected to a normalizing (i.e., "Box-Cox") transformation that invariably has the effect as well of de-coupling the variability of the typical weekly periodicity from the mean (seasonal) level in the data. This transformation is often logarithmic.

The typical seven-day periodicity (evident in most clinical and drug purchase data) is then removed by either (a) subtractive removal of a "mean" weekly cycle, estimated from a trailing window of 60 to 90 days; or (b) by lagged 7-day differences. In what follows, the variable denoted will denote the cyclically adjusted transformed data. It is this variable on which anomaly detection by monitoring of linear mean-square forecast residuals is predicated.

The so-called "local level structural model" given below:

$$y_t = \mu_t + \varepsilon_t \qquad \varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right)$$

$$\mu_t = \mu_{t-1} + \eta_t \qquad \eta_t \sim N\left(0, \sigma_\eta^2\right)$$

The increments ε, η are regarded as each being drawn from serially uncorrelated zero-mean Gaussian distributions; whose respective variances ($\sigma_\varepsilon^2, \sigma_\eta^2$) are unchanging in time—i.e., each constitutes a sequence of "independent identically distributed random variables".

This model contains two parameters: $\sigma_\varepsilon^2, \sigma_\eta^2$, or "hyper-parameters" as they are often called in the literature (for the technical reason that the time-varying *state* term in the structural model μ is itself conceptualized as a time-varying distributional *parameter* to be estimated).

These parameters $\sigma_\varepsilon^2, \sigma_\eta^2$ must be *estimated* by some sort of fitting procedure using a segment of observed data that we may treat as representative of the process we seek to model under stable, normal conditions.

All model fitting approaches require an initial segment of data, which will be treated as the "training set" and, as such, must be assumed to be of stable statistical character. This character must be taken as *representative of the future regime of stability that we seek to monitor*.

2.1.1.5 Maximum Likelihood Estimation

The most conceptually straightforward approach to model fitting would be the so-called "exact Maximum Likelihood" approach. One simply uses the KF algorithm as a one-step ahead forecast for the initial segment ("training set") and computes a goodness-of-fit statistic based on the set of forecast residuals (observed value minus predicted value) that accrue. For every possibly choice of the parameters $\sigma_\varepsilon^2, \sigma_\eta^2$ there will be a corresponding measure of the goodness of fit. The standard probability arguments motivate a likelihood statistic that will be proportional to any reasonable such measure of goodness of fit. Using some multidimensional optimization algorithm, one ultimately selects an optimal set of parameters $\sigma_\varepsilon^2, \sigma_\eta^2$, that set which maximizes the overall goodness of fit (that maximizes the statistical likelihood of the parameter set given the training data set).

For our simple scalar structural model above, the exact form of the log-likelihood is obtained via the sequence of one-step ahead forecast errors from the Kalman recursions:

$$e_t = y_t - H\mu_{t|t-1}$$

The log-likelihood then follows by treating the sequence $\{e_1, e_2, ..., e_t\}$ as consisting of independent identically distributed zero mean Gaussian deviates:

$$\log L = -\frac{t}{2}\log 2\pi - \frac{1}{2}\sum_{i=2}^{t}\left(\log S_i + \frac{e_i^2}{S_i}\right)$$

This quantity is minimized with respect to the two free parameters σ_ε^2 and σ_η^2 , although to satisfy the positivity constraint it is usual to apply numerical minimization algorithms to the *logarithms* of the two variance parameters.

(This expression for the likelihood is evaluated beginning after the first time-step.) This is a technical point related to the problem of developing proper initial conditions for the state vector in the absence of prior knowledge. For the simple local level model we utilize the so-called "diffuse prior" initialization for the state statistics, and the likelihood function beginning after the first datum known as the "diffuse log-likelihood".

For more complex vector state models, the vector analogue of the log-likelihood function follows from the multivariate normal distribution applied to independent vector residuals, thus the functional form brings in the inverse and the determinant of the forecast covariance matrix S. The series is initialized using the so-called "diffuse prior" values, y_1 and σ_ε^2 for the state and its variance, respectively.

2.1.2 Expectation Model Versus Signal Model Detection Schemes

Monitoring "non-traditional" indicators for the potential early manifestation of responses to disease progression in a given population (e.g., over-the-counter drug sales) relies on three main questions:

1. What are the salient dynamical and statistical properties of the class of "signals of interest" itself?
2. What are the salient dynamical and statistical properties of the statistically stable state itself?
3. Which dynamical and statistical properties of the statistically stable state are those we expect to be most blatantly "interrupted" by the presence of a signal of interest?

In our application, treatment of the first question requires a basic phenomenological description of the manner in which (first) a population of a stipulated geographic and demographic character is exposed to a disease source or vector of stipulated type, virulence and span; and (second) the manner in which the progression of symptoms ultimately manifests in behavioral response in the aggregate, such as an uncommonly large aggregate increase in the propensity to purchase certain palliative Over The Counter (OTC) remedies. Simply put, this is a question of stipulating a signal or class of signals which we will treat as representative of how an "outbreak" manifests in the given data source.

Treatment of the second question follows from a sufficiently broad and deep investigation of the historical record of such data sources. We seek to monitor a division of those sources into putatively "stable" and "unstable" sections and a decomposition of the variability of those stable sections into various "explainable" sources, each factor being associated with a statistically stable random residual component. Simply put, this is a question of stipulating the typical state of affairs when an "outbreak" does not manifest in the given data source. It goes without saying that we may find true "outbreaks" in the historical record which are nevertheless not representative of the types of (hypothetical) outbreaks we actually wish to study, in light of question number one.

Treatment of the third question would constitute an elucidation of the significant differences (phenomenological or quantitative) between the statistically stable state and that state characteristic of a stipulated "signal" or signals representative of the class of "outbreaks".

In what follows we illustrate how various methods of treating univariate time series data treat this problem of signal detection line up under two somewhat different focii. What turns out to be the case is that the class of such discrimination methods we work with, all of which are built out of the same components (linear stochastic models and linear filters), can be divided into two classes. They divide according to whether they, in effect, address themselves primarily to question 2 above; or primarily to question 1.

Those methods of discrimination that address themselves primarily to question 2 are rooted in the expectation that a faithful (and optimal) modeling of the "stable state of affairs" will provide sufficient discriminatory power against the unusual state of affairs and will be characterized by the presence of any one of various "outbreak conditions" that one is seeking to detect. Methods that are so devised essentially correspond to statistical tests designed to reject a "null hypothesis" against a non-specific alternative, although in fact they depend upon certain built-in assumptions about the lay of the land of the entire class of alternatives. Simply put, these methods assume that no matter how the outbreak is manifest it is likely to be grossly different from the stable state of affairs in at least some quantitative dimension.

Those methods of discrimination that involve more explicit assumptions about the nature of the signal of interest (and thus address themselves to that extent to question 1) utilize some quantitative of a phenomenological model bearing on the shape and dimension of the signals of interest. These methods treat the detection problem by alerting us when such signals are present and doing nothing otherwise.

Now it should be noted that the use here of the term "signals of interest" and "presence" or "absence" is somewhat more general than the commonplace sense of a feature of a particular shape and duration appearing in a graph of signal level versus time. When we speak of "characteristic features" we may refer to any quantitative aspect either in the time or frequency domain that is typically associated with the state of affairs we seek to detect.

For lack of any better terminology, we refer to detection schemes that begin with a model that addresses question 2, as “Expectation model detection schemes”. Detection schemes that address question 1 directly, we will refer to as “Signal model detection schemes”. Of course a given detection scheme may, in the end, involve algorithms of both kinds.

A scheme based on an expectation model endeavors to tell us whether the most recent conditions are sufficiently different from those predicted by a model of expectation so that we may classify these conditions as “anomalous”. Expectation models classify observations by their degree of novelty or surprise.

A scheme based on a signal model, tells us whether the most recent conditions are sufficiently similar to those that would be associated with one of various conditions of interest (“outbreak scenarios”) to merit the announcement of an alarm.

Alarms generated by a given expectation model do not necessarily correspond to conditions of interest. That is to say, these alarms are statistically significant events, but they are not necessarily practically significant. Only an implicit or explicit standard developed out of some signal model can determine practical significance.

On the other hand, alarms generated by a particular “signal-based” model are not necessarily entirely contained in the set of statistical alarms generated by some particular “expectation-based” model. For the signal-based model may be much more sensitive to a special class of events than any garden variety expectation-based detector that does not take into account salient (e.g., time domain or frequency domain) features of the signal itself.

2.1.2.1 Expectation-Based Models: Linear Mean-square Forecasting

Expectation-based detection methods use linear mean square forecasting on observed history of data (or transformed data) to yield optimal predictions for data not yet observed.

This is accomplished by fitting to historical data linear stochastic models of the form:

$$a_0 y_t + a_1 y_{t-1} + ... + a_M y_{t-M} = b_0 \varepsilon_t + b_1 \varepsilon_{t-1} + ... + b_N \varepsilon_{t-N}$$

where y_t represents the datum observed at time t and the ε_t constitutes an (unobservable) sequence of independent identically distributed Gaussian random variables with common variance σ^2.

The working assumption is that our data constitute a linear mixture of linear random processes possessing continuous spectra with the possibility of narrow band spectra at special frequencies, associated with 7-day cycles. These processes may be modeled as either stationary or evolutionary. So long as the evolution can be embedded in a linear Markov transition model of some higher dimension, evolutionary non-stationarity in the spectrum can be viewed as merely the projection of a stationary, higher-dimensional process onto a lower dimensional subspace. The linear forecasting theory for stationary processes is applied to the higher-dimensional state vector.

Gaussianity, or the possibility of a non-linear transformation toward Gaussianity is taken as precondition for much of this work, although an extended theory for a time series drawn from exponential families (specifically Poisson distribution) can be applied in the situation where the raw event counts are especially few and far between.

2.1.2.1.1 The Kalman Filter: Linear Forecasting Models

The Kalman Filter algorithm is the fundamental unifying tool whereby any linear dynamical random process of arbitrary (finite) order can be cast in the form of a first-order vector "state-space" linear recursion. Any of the linear difference equation stochastic models in the Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) formalism are easily cast in such first-order form. For example, the forecast functions for linear stochastic models, conditional expected values given past history and their associated precision, are precisely the quantities that the Kalman Algorithm updates recursively. The Recursions provide one with a unified scheme for fitting the forecasting models by direct maximization of the likelihood function for the prediction residuals. In short, model parameter fitting by least squares is made transparent for a very wide class of models, as is the scheme for forward prediction, after the model has been fit to historical data.

Furthermore, linear models that are not among the basic class of the ARIMA linear difference equation forms are just as easily incorporated in the state space formalism and in principle just as readily fit to the data. In particular dynamic versions of ARIMA predictive models are quite easily written out as state-space vector models. The procedure for fitting these models by likelihood optimization relies on the Kalman recursions in a manner formally indistinguishable from its application to the more standard class of static Box-Jenkins models. The computational burden for these models ultimately rests upon a nonlinear optimization procedure that must find the optimum of a function of more and more variables with the practical limitation being set by the ineluctable fact that models with more parameters require greater and greater amounts of data to fit within the bounds of any meaningful precision at all.

Since our methodological basis for anomaly detection is tied to having a fitted model and monitoring measures of model lack-of-fit, the ability to iteratively fit models within a class, or swap model specification between classes and re-fit, etc., it is of unquestioned utility that model fitting and prediction be cast in a unifying formalism. This is what the Kalman Filter Recursions provide.

It is important, of course, to bear in mind, that the Kalman Filter device is not synonymous with the forecasting model and the Kalman filter is an algorithm for facilitating the fitting of a particular model and the forecasting process based upon that model.

2.1.2.1.2 Remarks on Non-stationarity

Standard forecasting methodology (e.g., Box-Jenkins and Structural Modeling) includes of necessity that all persistent non-stationary features be properly accounted for in the modeling procedure. These effects (e.g., day of week) are forecastable and are accounted for in design of standard linear prediction models. They can be "modeled-away" in a preliminary stage by direct estimation of the germane characteristics (e.g., phase, amplitude); or they can be included as parametric entities within a time-series model, all the parameters of which are fitted in one fell swoop by likelihood or least-squares estimation. In the canonical Box-Jenkins analysis, for instance, seasonal or other cyclical effects are removed by preliminary lagged differencing, as are obvious linear or polynomial trends. The remainder is treated as a random process with stationary mean, subject to fitting by low-order ARIMA parametric models. We are building a modular forecasting system in which each of these various (and in the end not dissimilar) treatments of non-stationarity (of the mean) can be applied if desired towards the end of facilitating comparison and contrast. Non-stationarity of the variance on the other hand is to a large extent handled by preliminary non-linear transformation (i.e., optimal Box-Cox) of certain cyclic or otherwise volatile data. This effectively decouples the local mean from the local variance and typically has, as a consequence, the effect of stabilizing the variance of such series.

Non-stationarity in the underlying data generating process is handled by adaptive modeling. Either a fixed model is regularly updated or refit (according to decision criterion bearing on the appearance of serial correlation in residuals) or a model is fit that is a dynamic generalization of the, e.g., fixed coefficient ARIMA stationary models. Such models are easily accommodated within the framework of fitting and forecasting using the Kalman recursions.

2.1.2.1.3 Multivariate Residual Diagnostics as a Detection Device for Vector Data

Multiple univariate forecast models are fit to separate streams of data as detailed above. These multiple forecasting models produce parallel outputs consisting of "forecasts" (expected observation conditional on past) and "residuals" (observed difference between forecast and actual observation when it becomes available). The individual streams of "forecasts" and the "residuals" may be used to generate parallel streams of summary "diagnostic markings" (as detailed above) that condense the results of statistical tests applied to the series of forecasts and residuals.

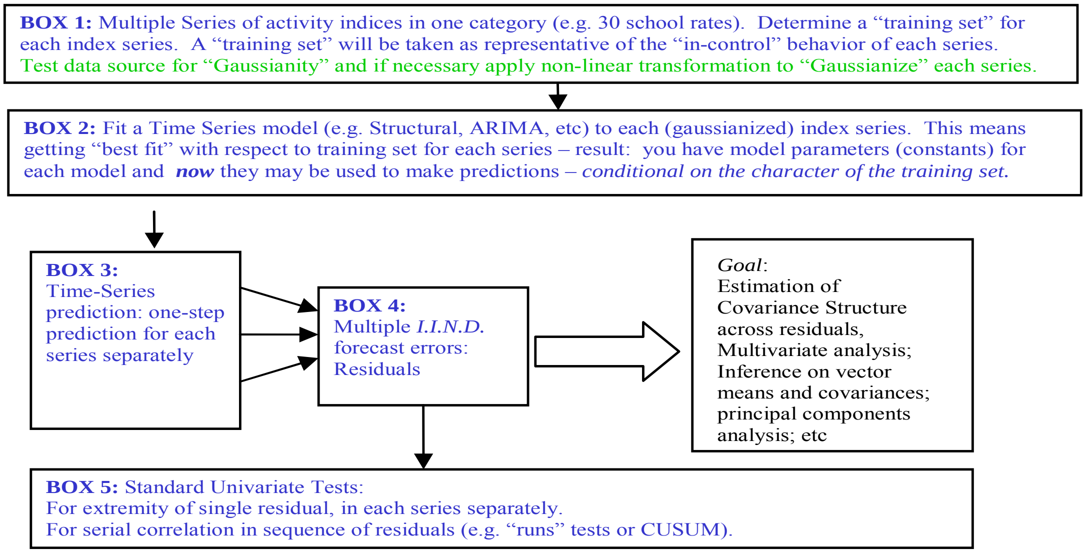
While it is possible to interpret the multiple streams of forecast residuals in a manner entirely analogous to that by which a single stream is tested for anomaly, considerable information is discarded if the multivariate structure of these residuals is not utilized. For example, one can easily imagine a situation in which, for a fixed level of significance, no single residual series is alarmingly "large" at a given time (i.e. all series pass Test **a**); while the set of residuals considered *in the aggregate* clearly exhibits a common anomalous jump. The ways in which statistical anomalies manifest themselves are considerably more diverse in *vector* as opposed to *scalar* data.

Thus, our detection scheme follows the vector of multiple forecast residuals. This vector of random variables with each component individually enjoying the property, under conditions of statistical stability, of having no significant serial correlation comprises the basis for all subsequent analysis. The application of various canonical multivariate statistical tests to the vector sequence of parallel forecast residuals may be construed as a system of "multiple decision fusion," but one which avoids problems of multiple significance testing (and adventitiousness) by treating all tests as firmly grounded in the assumption of multivariate Gaussian statistics.

The types of tests one performs on such sets of individually serially uncorrelated residuals fall into two categories: (a) tests to evaluate the likelihood that the vector of forecast residuals, at any given time, has moved “far” from its expected value, namely the zero vector; (b) tests to evaluate the likelihood that the correlation structure *amongst* the set of residuals has changed significantly, when compared between two consecutive periods or with respect to a past period.

In effect, one may ask two kinds of questions about a vector of random quantities: (a) has the magnitude of the vector changed, and (b) has the relationship between the components of the vector changed? The first of these two questions, as expressed by a canonical statistical test on our vector of forecast residuals, takes the form of the canonical “Hotelling T^2” test on the vector magnitude of an ostensibly zero-mean vector (if the components are statistically independent this test reduces to a standard χ^2 inference on the Euclidean vector magnitude squared). The second of these two questions takes the form of a panoply of tests applied to the sample covariance (or correlation matrix) formed locally by the recent history of forecast residuals. The practical interpretation of correlation-based tests is that we seek to know (1) whether pairs of data which typically vary together no longer do and (2) whether pairs of data which typically do not vary together have begun to do so. In our context where multiple streams are often aggregates of, say, sales volume for a particular medicine at various nearby geographical locations, this analysis takes on the obvious interpretation where we may discover that geographically separated sources abruptly fall under the influence of an anomalous and possibly worrisome common cause. Multivariate tests for changes in the correlation structure of the sequence of forecast residuals are best handled by principal components analysis (based on singular value decomposition of the correlation or covariance matrix). First, sample covariance matrices will be singular if the number of data streams is larger than the number of days comprising one’s local time-scale of analysis, thus the inverse covariance matrices required in the Hotelling test cannot be used directly. Secondly, changes in the correlation structure are often seen directly as a shift in the structure of the sample covariance matrices, whether as a change in the number of principal components accounting for a fixed fraction of total variance, or as a change in the direction of the principal eigenvectors corresponding to those components.

The interface between separate fitting of multiple, univariate time-series models and multivariate analysis of the vector of model residuals is indicated in Figure 3.



BOX 1: Multiple Series of activity indices in one category (e.g. 30 school rates). Determine a "training set" for each index series. A "training set" will be taken as representative of the "in-control" behavior of each series. Test data source for "Gaussianity" and if necessary apply non-linear transformation to "Gaussianize" each series.

BOX 2: Fit a Time Series model (e.g. Structural, ARIMA, etc) to each (gaussianized) index series. This means getting "best fit" with respect to training set for each series – result: you have model parameters (constants) for each model and ***now*** they may be used to make predictions – *conditional on the character of the training set.*

BOX 3: Time-Series prediction: one-step prediction for each series separately

BOX 4: Multiple *I.I.N.D.* forecast errors: Residuals

Goal: Estimation of Covariance Structure across residuals, Multivariate analysis; Inference on vector means and covariances; principal components analysis; etc

BOX 5: Standard Univariate Tests:
For extremity of single residual, in each series separately.
For serial correlation in sequence of residuals (e.g. "runs" tests or CUSUM).

Figure 3. Schematic of Multivariate Forecast Residual based Detection

2.1.2.2 Signal Based Detection

The signal-based detection scheme will typically consist of a linear filter tuned to admit a relatively narrow range of signals believed to be the manifestations of events of interest in the data being monitored. Such a scheme can be thought of as a simple band-pass or matched-filter (in the frequency domain) or as a correlation detector in the time domain. Obviously such a specification requires a more or less explicit commitment to a stipulated class of signals thought to correspond to events of interest.

In the case of outbreak detection in so-called "non-traditional" data (ancillary to clinical data) a fairly simple signal detector might be thought of as a type of "inertial discriminator", based on a straightforward low-pass filtered version of the data. The low pass filtered signal obviously tracks fluctuations in the data that are no more rapid than some specified cutoff. The cutoff is chosen so that fluctuations admitted by the filter are stipulated as not alarming. The mean squared amplitude (over some suitable averaging period) represents the signal power associated with events that fluctuate more violently than those we choose to characterize as "normal". The dividing line between the two regimes must be set to represent the wavelength of a fluctuation phenomenon that one believes should typically not be present in the data under "non-outbreak" conditions. This detector is sensitive, thus, to abrupt changes (more rapid than some derived threshold) since high frequency content and the size of the numerical rate of change go hand-in-hand.

Such a detection scheme is built simply by tuning a low-pass filter, running data through the filter, and monitoring the average envelope amplitude of the residual series. It is perhaps most sensible to use a bank of such filters with increasing cutoff frequencies so that results can be interpreted with respect to a series of hypothetical signal classes, each of greater and greater spectral concentration at shorter wavelengths.

Cross-correlations between pairs of such “residual processes” may be monitored over a suitable time period that progresses along by regular steps. If cross-correlations between pairs of such residuals processes peak at some lag, then this constitutes empirical evidence that one series has “predictive” value with respect to the other, at least with respect to the class of signals admitted by a particular filter band pass.

2.1.2.3 Comparison: “Signal Based” and “Expectation Based” Detectors

Note that the signal detectors can be employed without the need for an extensive model specification and fitting phase. It is not necessary to characterize to any degree of precision the dynamics of the normal or “calm” historical data except by stipulating some class of signals that are expected not to be present except under outbreak conditions.

The “expectation-based” detector, on the other hand, operates by fitting an optimal time series model to calm historical data. Such a time-domain fit is equivalent to a parametric frequency-domain representation of the power spectrum (as a rational function) of calm historical data. An outbreak is declared post-hoc whenever new data is in effect anomalous with respect to that spectral characterization.

In situations where we are more or less certain that a certain class of fluctuation phenomena is invariably associated with “outbreaks”, it is probably simplest to use a signal-based detector to discriminate alarming data from calm data.

In situations where the scope of signals of interest is too broad (or one simply is averse to such explicit commitment), an expectation based anomaly detection scheme is merited. It must be understood, however, that achieving an excellent parametric time series fit to historical data does not necessitate that the resulting forecasting device is sensitive to any particular class of signals of interest.

In classical terms, when the class of "alternative hypotheses" to a statistical test is extremely broad, the sensitivity of that test to any particular single event within that alternative class is low. Only by bringing into the model specification some description of the signals of interest can the sensitivity of an expectation based detector be improved. This is a dual optimization problem.

2.2 BioSTORM: A Computational Framework for Deploying Problem Solvers

To meet the complex operational and research needs of these surveillance applications, we have developed a system called BioSTORM (Biological Spatio-Temporal Outbreak Reasoning Module, Buckeridge, 2004). BioSTORM is a computational framework that provides run-time mediation between data sources and problem solvers with the goal of meeting the performance and flexibility demands of emerging disease surveillance systems. The system has the following goals:

1. To acquire and curate data from traditional and non-traditional sources.
2. To develop a knowledge-based infrastructure to integrate and experiment with alternative data sources and Problem Solving Methods (PSMs).
3. To develop new PSMs for temporal and spatial analysis.

The system has five principal components:

1. BioSAIL, an ontology for integrating data source.
2. A Data Broker that provides a software library for integrating multiple data sources that have been described using BioSAIL.
3. A Mapping Interpreter, which is a tool for integrating problem solvers in a deployed system and for tailoring data sources to the needs of individual PSMs
4. A control structure that is used to efficiently deploy analytic methods; and
5. A set of analytic techniques, which is built from a library of statistical and knowledge-based PSMs for analyzing surveillance data streams.

2.2.1 BioSAIL: An Ontology for Integrating Data Sources

Public health surveillance data are diverse and usually distributed in various databases and files with little common semantic structure. Thus, they can be difficult to integrate. In the absence of integrated data, it is next to impossible to consistently apply reusable analytic methods. A Protégé-2000 based ontology called BioSAIL tackles this issue. It attempts to provide a domain-independent semantic structure for raw data to assist in the integration of disparate data sources. BioSAIL provides a hybrid approach to data integration that combines the semantic rigor of creating a global ontology with the flexibility and level of detail that comes from devising customized data source specific local ontologies. Our general approach is to enable data to be self-descriptive by associating them with a structured context. We have defined a general structure for describing these contexts, which forms a template ontology. The template ontology acts as a meta-model, providing a consistent structure within which detailed descriptions of different data sources and their data can be constructed. While our template remains domain-agnostic, it allows detailed descriptions of individual data sources to be constructed in terms of separate domain ontologies. BioSAIL's goal is to provide tools for rapidly describing an extremely diverse data in a coherent manner that can facilitate reasoning on that data. We have used BioSAIL to develop a description of a number of data sources using the Centers for Disease Control (CDC) ontology as the underlying domain ontology. BioSAIL will provide the basis for describing all data sources that will be used in this proposed research.

2.2.1.1 Using BioSAIL to Describe a Data Source

When describing a new data source using BioSAIL, a developer fills in the template by choosing specific attributes and attribute values from a predefined taxonomy. This template-directed process allows users to create a customized local model that shares a common structure, space of attributes, and set of possible attribute values with all other similarly created models. Any system that can process our template ontology and attributes can access enough relevant context to reason successfully about data from sources described therewith.

The template ontology provides the structure within which a user details a description of context at both the level of data sources and individual data elements with those sources. A user describes the context of data from a particular data source by filling in the template slots with relevant details. This process requires creating instances of classes from the template ontology and choosing specific metadata and semantic attributes to fill in the slots of those instances. At the data source level, the ontology provides a taxonomy of attributes grouped into general categories from which users choose when filling in these slots. To describe individual data elements, we adopted Logical Identifier Names and Codes (LOINC) terminology. The LOINC scheme, which is used to contextualize results reported by clinical laboratories, describes a piece of data along five major semantic axes. We have generalized the LOINC axes from their specific role in reporting clinical lab results to a generic set of descriptors for many different types of reported data. Therefore, we have a systematic approach to constructing models of data and provides consistent syntax and semantics among models created with the template. Because of the shared structure of the template and its defined vocabulary of possible attributes, it is thus possible to rationally reason about specific data in each model. For example, a system can use the taxonomic relationships between metadata attributes in the contexts of different pieces of data to data to infer the relationship between those data themselves. We used the template to create a detailed knowledge base for syndromic surveillance data integration and analysis by incorporating the disease ontology that we developed with the CDC.

2.2.2 Data Broker: A Software Library for Data Integration

The Data Broker is a software library that uses the BioSAIL ontology to allow problem solvers to read data from a number of sources transparently at run-time. The data broker queries the BioSAIL ontology describing a particular data source and constructs a stream of uniform data objects that conform to the ontology. It then supplies those objects to the invoking PSMs. The Mapping Interpreter takes this stream of objects and reduces, abstracts, and transforms it to a format meaningful and manageable for each problem solver. To work with the Mapping Interpreter, each problem solver must publish an input-output ontology describing the data that it wishes to receive. Thus, each problem solver receives a customized set of data objects and can ignore the original sources or formats of data. In effect, BioSAIL provides the ontology, and the Data Broker operationalizes it.

2.2.3 Mapping Interpreter: An Ontology for Unifying Problem Solvers

BioSAIL provides a consistent mechanism for describing data sources, however, public health surveillance methods that operate on these data may have many different input requirements. Methods may require data at different temporal or spatial granularities. For example, some methods may work on data at reasonably high levels of abstraction; whereas other methods may work directly with relatively raw data. Given the diversity of input requirements, we require a standard mechanism to map data to the requirements of the various methods at run time. This mapping must operate at both the data and the ontology level.

We designed an ontology of the generic mapping relations that can occur between the input requirements of a method and corresponding domain constructs, a mapping ontology. This ontology defines the types of transformations applied to domain concepts and attributes to match the method's input–output terms. These transformations range from the simple renaming of domain elements to corresponding terms in the method ontology, to the composition of lexical or numerical expressions to match method terms. For instance, when configuring a propose-and-revise PSM for prediction of ribosome conformations, we created a set of expressions to map an array of possible ribosomal-unit positions in the domain ontology into an incremental constraint-fix expression in the method ontology. After creating a knowledge base of such domain-to-method mappings, domain instances must be translated into a set of input instances for the method. A mapping interpreter applies the mapping relations to the domain instances to produce the method instances. The Virtual Knowledge Base Constructor developed in our laboratory is one such mapping interpreter, and it can process any knowledge base compatible with Open Knowledge Based Connectivity Protocol (OKBC).

The Mapping Interpreter provides a controlled set of possible mapping relations between concepts and attributes of data source ontologies (e.g., BioSAIL) and a problem solvers input specifications. For each category of data, specific mapping relations define the transformation of data elements into runtime inputs of problem solvers. The mapping relations can specify data-abstraction rules that may differ. For example, the relation can specify that data be aggregated at different spatial or temporal granularities. Types of statistical abstraction, such as aggregation and normalization, can also be selected. The Mapping Interpreter processes the mapping relations for each data group and each problem solver, and then generates streams of individual events that are ready to be processed by the problem solvers. New problem solvers must adhere to an existing input specification or publish an explicit new input specification. If they publish a new input specification, a new set of mapping relations must be created.

The Data Broker and Mapping Interpreter thus provide run-time data integration in three stages:

1. The Data Broker uses BioSAIL metadata describing the low-level datum classes to retrieve sets of data from raw databases when requested by PSMs. At present,

relational databases and simple flat files are supported, but other formats, such as XML, are planned.

2. The Data Broker then formats and groups the data as specified by BioSAIL and packages the data with the appropriate BioSAIL contexts to create semantically meaningful data objects.
3. Next, the Mapping Interpreter supplies this data to invoking PSMs recast to conform to the input requirements of the each PSM. Data context information from the knowledge base is used to inform the mapper which mappings are valid and meaningful to apply. As a result, PSM are fed with streams of semantically consistent data that conform to their input requirements and can thus make meaningful comparisons of disparate data sources. Together the Data Broker and the Mapping Interpreter provide a semantic wall between analytic methods and the raw data.

2.2.4 Control Structure

BioSTORM uses a Linda-based model to provide a distributed architecture for problem solver deployment. Linda provides a model that enables users to create parallel programs that run on a wide variety of platforms. It is based on logically global, associative object memory space called a tuple space. Linda provides interprocess communication and synchronization facilities via the insertion and removal of tuples from this tuple space. It effectively implements parallelism with a small number of simple operations on the tuple space to create and coordinate parallel processes. Linda lends itself to producer-consumer distributed data structure algorithms, where processes generate data to be consumed by other processes. Thus, this model is very much suited to problem solver deployment for biosurveillance.

We have used the JavaSpaces implementation of the Linda model to implement BioSTORM. Because of data throughput limitations of JavaSpaces (and many Linda-based implementations), we have developed a hybrid Linda/relational model to provide efficient data flow in a deployed system. In this model, data are conceptually grouped into bundles based on shared semantic properties. Semantic markers are then used to tag these bundles. Data may be bundled spatially, temporally, or based on other semantic properties. For example, a bundle may consist of all 911 calls for a particular ZIP code for a particular day. Raw data are associated with these markers using a knowledge base to describe the location and layout of the actual data. When read or write operations are invoked by problem solvers, the semantic markers associated with the bundles are transferred through the JavaSpace tuple space. Actual transfer of the bundled data itself is performed though an Open Database Connectivity/Java Database Connectivity (ODBC/JDBC) bridge using the layout and location information associated with the marker tuple. An analogous process applies when data are inserted into tuple space by problem solvers. BioSTORM completely hides the details of this process from invoking problem solvers. All JavaSpaces control and synchronization functionality is retained with the superior performance of relational databases used in place of relatively inefficient JavaSpaces object storage. Hence, this approach can scale to deal with large data sets.

Using the Data Broker and Mapping Interpreter, it integrates disparate, possibly distributed, data sources into semantically uniform data streams, maps these streams to multiple problem solvers, and deploys these problem solvers to conduct surveillance. It is responsible for invoking problem solvers when data arrive for the problem solvers to process and for feeding new arriving data to them through both the Data Broker and the Mapping Interpreter. Its overall task is to unify the Data Broker, Mapping Interpreter, and problem solvers into a coherent, efficient runtime system. It provides a modular framework to enable concurrent application and structured evaluation of multiple analytic methods.

2.2.5 Analytic Techniques: Methods for Analyzing Surveillance Data Streams

Existing surveillance algorithms also have varying data requirements. Some algorithms operate at the population level For example, whereas others work at the level of individuals, many algorithms expect time series data at varying granularities, and certain algorithms require spatial data at several levels of aggregation. Unless we provide more intelligent integration methods that can make correlations between different kinds of data, and that can aggregate and abstract data into information at the population, spatial, temporal levels, even ostensibly integrated data are useless to public health specialists who need to make sense out them. Hence, the integration task must be informed by the type of analyses that are to be performed on the integrated data and must be able to produce integrated data at many levels of abstraction.

Many of the statistical techniques traditionally employed in public health systems can be applied to syndromic surveillance; however, because of the range of nonspecific signals that may have to be considered, more abstract qualitative knowledge-based techniques will also be required. The problem is that statistical methods operate on low-level data, while reasoning with knowledge operates on data at a higher level of abstraction. For example, statistical time series methods operate on raw disease counts to detect an abnormality, while humans reason about increasing trends or sudden spikes in disease. It is, therefore, difficult to incorporate important qualitative knowledge into statistical analyses. However, if low-level data can be abstracted to higher-level concepts, then qualitative knowledge can be used to reason about the concepts. These techniques entail the application of surveillance knowledge to concepts abstracted from low-level data. It requires explicit modeling of knowledge that is relevant to both the surveillance and reasoning processes plus the development of methods for abstracting higher-level concepts from low-level data. In addition to enabling the integration of knowledge with data from multiple sources, a knowledge-based approach also facilitates system modification, enhances portability, and allows users to better understand system function and results. A knowledge-based approach to epidemic surveillance can address many of the shortcomings of the current surveillance process.

Driven by this knowledge-based approach, we have developed a practical framework for detection algorithms that is based on their information requirements and functional characteristics. This framework consists of a decomposition of the surveillance analysis task into sub-tasks and an ontology of surveillance analysis methods that automate these sub-tasks. All surveillance analysis methods we have defined share a common input and output specification, or method ontology. The method ontology makes explicit the data requirements of a method and, thereby, enables data sources to be mapped to analytic methods using the Data Broker and Mapping Interpreter, and facilitates the interoperation of analytic methods by identifying appropriate interactions between methods and data types. The goal is to facilitate identification of methods suitable for a specific analysis, and to support mapping of data and knowledge to an analytic method. The overall goal of this framework is to establish a methodology for incorporating existing surveillance algorithms into our system and to establish the knowledge requirements of these methods.

We have used Unified Problem-solving Method Language (UPML) to extend this model. The Intelligent Broker for Reasoning on the Web (IBROW) consortium defined a declarative language for describing knowledge-based problem solving components, the UPML. The design goal of UPML is to provide human-understandable method descriptions, while including formal specifications that allow for the organization of PSMs in libraries and for their automatic indexing and retrieval. UPML is more than a method-description language; it is also an architecture for modeling the other building blocks of an intelligent system, tasks and domain models. The connection of all three kinds of components is modeled by specific constructs. These constructs are bridges that express the relationship map between two kinds of components (similar to our domain-to-method mappings) and refiners that represent the stepwise adaptation of a component.

2.3 Data Acquisition

The thrust during final months has been to establish a transition path for data collection activities initiated under the Bio-ALIRT program. Our goal is to ensure the local community will have data sources available when it undertakes a more complex approach to syndromic surveillance.

2.3.1 General Dynamics Data Source Activities

This section is a summary of the data status and new progress that General Dynamics Advanced Information Systems (GDAIS) has made in acquisition of data in the Hampton Roads area.

2.3.1.1 Data Collection

GDAIS has completed data transfer agreements with the following organizations:

- **Navy Public Works Center Norfolk**
    - To receive data by email on the number of hours worked and number of hours sick leave for civilian employees in each of over 60 cost centers located in installations in the Hampton Roads area,
    - To obtain wastewater chemistry and flow data regularly.
- **Navy Exchange Command**
    - To receive by email information on sales of over 900 line items together with prices and total foot traffic at 11 locations in the Hampton Roads area.
- **Hampton Roads Sanitation District**
    - To receive by email every 15 minutes information on pressure and flow in the potable water system.
- **Norfolk Hospital Center**

  - To receive by email daily the number of arriving and departing cars parking in the transient parking structure.

- **Norfolk Area Military Treatment Facility Patient Files**
  - GDAIS received historic CHCS data from all military treatment facilities in the Hampton Roads region under a Data Use Agreement with Walter Reed Army Institute of Research (WRAIR).
- **National Retail Data Monitor (NRDM)**
  - With project partner University of Pittsburgh to obtain retail sales information from available outlets in the Hampton Roads region. Completed security arrangements for regular data transfer.
- **Army Air Force Exchange System (AAFES)**
  - To receive from AAFES retail sales data for 600 categories of items at installations in the Hampton Roads Region.
- **Riverside Health System (RHS)**
  - RHS for daily transfer of data from the organization's Ask-A-Nurse information system including reason for call, disposition, and both demographic and geographic information about the caller.
- **Chesapeake Mosquito Control Commission**
  - To receive by email episodic information on Mosquito, Infected/Dead Birds, Infected Horses.
- **Cough Detection**
  - Detection software prototype completed and successfully tested with recording from Navy classroom. No ongoing collection.
- **Humidity and Pollen Counts**
  - Humidity and pollen counts collected from Pollen.com daily.

2.3.1.2 ART*Epi*™ or BASIS Installations

- Naval Medical Center Portsmouth (NMCP) – installed.

- GDAIS has completed an agreement to install ART*Epi*™ at Riverside Health System in Newport News.

2.3.1.3 MOUs in Place—With Potential for Further Data Collection

- The Fire Department of the City of Norfolk has agreed in principle to provide 911 dispatch data, but implementation is delayed until they resolve a staffing shortfall.

2.3.2 Stanford Data Activities

Stanford's variety of data sources from the Santa Clara Department of Public Health (DPH) include:

- Bioterror syndrome tally data from Emergency Medical Service (EMS) responses to 911 calls.
- Bioterror syndrome tally sheets from multiple Emergency Room (ER) and call center locations.
- Animal death and injury reports.

Stanford continues to work with Veterans Administration clinical data from the Palo Alto location in Santa Clara County. Stanford performed a significant amount of comparative analyses of the Santa Clara 911 data and 911 data previously received from the San Francisco Department of Public Health (DPH). Stanford collaborated with Santa Clara County DPH to improve the quality of collected data including defining bio-terror syndromes. No new data were acquired from the San Francisco DPH.

2.4 Transition

GDAIS has worked with the Virginia Department of Health (VDH) epidemiologist for the Eastern Region to explore funding opportunities to support storage and analysis of traditional and non-traditional data in the Hampton Roads community.

2.5 Summary of Problems or Areas of Concern

Because the Government has discontinued the program, General Dynamics and Stanford University have notified various data providers that it no longer makes sense to pursue their data collection under this program.

3 Report Preparation

Report prepared by:

Gene McClellan

General Dynamics Program Manager

(703) 516-6204

gene.mcclellan@gd-ais.com

Mark Musen

Stanford University Program Manager

(650) 725-3390

musen@stanford.edu