Vision:
The Comprehensive Test Ban Treaty (CTBT) is receiving renewed attention in the Obama administration, in light of this administration’s long range objective of steering towards a nuclear weapon free world. While the current International Data Center (IDC) has successfully detected several low magnitude explosions, opportunities exist to further enhance its detection and localization capabilities, drawing upon the more advanced tools and techniques for large scale data analysis blossoming in the private/public sector.

Over the years, the IMS sensor data volume has grown to tens of terabytes, and will continue to grow as more receiver stations come on line. Exploiting this valuable repository of historical data to extract better sensor models, earth models, phase classifiers, and association algorithms could prove extremely valuable especially as the amount of data begins to exceed lower threshold occurrences. Innovative approaches for automatically mining large databases are providing a valuable service for applications ranging from test and image search to outlier detection in financial transactions and tracking and analysis. These approaches have matured from the purely empirical to machine learning based and statistically driven algorithms, adapted to the physics of the problem. Furthermore, methods for extracting extremely weak signatures from noisy data and improving the accuracy of anomaly characterization are also becoming more robust; these could also be brought to bear to enhance IDC effectiveness for qualifying small magnitude events.

As such, we believe that it is timely to formulate a strategic R&D thrust focused at the means of handling present and future IMS sensor data sets, taking into account the state-of-the-art, the potential for advances in areas such as data structures and query techniques, the shaping of sensor data for exploitation, and emerging methodologies for pattern discovery. The long term goal of this effort will be to assist the CTBTO analyst in making robust, expedient decisions, aided by a historical perspective, in the face of rapidly growing multi-sensor information and the increasing need for more accurate and timely feedback for event characterization.

To facilitate such an activity, we propose to define a virtual development environment vDEC (Virtual Data Exploitation Center) which will (1) connect experts from different disciplines with the IDC framework, to assess and develop upgrades for the current data processing infrastructure, with a view to lowering the cost and improving the accuracy and efficiency of data analysis; (2) enable a pathway for new technology insertion and uptake within the CTBTO operational environment; and (3) aid in constructing data exploitation architectures and system solutions for enhancing the fidelity of time critical on-site inspections (OSI).

Mission:
The objectives of this Center will be to link together via the world-wide web computer scientists and statisticians from academia, industry and stake holder agencies, with domain experts, to focus on the data processing challenge. Current analyses rely upon slow, expensive and often error-prone manual processing of automated front end procedures. vDEC’s goal will be to evolve robust, efficient automated methodologies for waveform analysis and network processing, in an environment where new and innovative ideas can be tested and applied in context and with minimal delay.

vDEC will exploit Open-IDC defined data sets for algorithmic development and prototyping, consistent with current standards for data formats and interfaces. The Center will simultaneously champion industry partnerships so as to draw in methods and best practices from within the private sector, as well as define a conduit for technology migration and long term maintenance for IDC infrastructure enhancements. The goal will be to evolve and drive a phased roadmap which progressively integrates modules into the data center processing chain. Projects will be selected on the basis of their usefulness to the IDC over a rolling 3-5 year time horizon.

Technical Portfolio:
The following enumerates some of the research themes, in varying degrees of maturity, which could be exploited for CTBFT verification.

Probabilistic Classification and Inferenceing:
The current IDC software architecture automatically isolates and labels detections, assigns phases, and identifies events, creating an output bulletin of seismic activity. An automated analysis identifies parameters, and then meticulously check each event to remove redundancies and false positives and insert missed events. This discrepancy between the automated bulletin and human review could be substantially reduced by incorporating modern machine learning methods for phase/event classification which utilize historical data from past analyst review bulletins to provide learning sets for training classifiers. A number of prior-data driven techniques have evolved over the years to address such probabilistic classification challenges, including kernel methods, nonparametric methods, clustering and decision trees. These remain candidates for seismic event extraction.

Drawing conclusions from IDC data sets currently relies upon a series of pipelined stages, where distinct software modules perform tasks based upon locally derived information. This has the advantage of flexibility and simplifies modular implementations. However, incorrect front end labeling can propagate through the chain to produce erroneous end results with no recourse to midcourse correction. Thus, tasks such as automated detection of weak sub-threshold signals, or cooperative correction of signatures based upon multiple sensors remain a human responsibility. The question to be asked is whether this sequential aspect of current processing can be converted into a bidirectional flow via probabilistic inference techniques, so that information at later stages about hypothesized events can effectively revise earlier detection and classification decisions. Emergent hierarchical methods provide the possibility of drawing conclusions based upon all the evidence available, not just the local context, by consistently updating the uncertainties in different information sources. The time may be right to explore some of these avenues for qualifying event signaturs.

Seismic Data Storage and Search:
The IDC currently stores seismic data in a relational database (e.g. Oracle, MySQL) where records are indexed and queried by a predetermined set of standard attributes (schema) such as the station name, number of channels and time. Relational databases however become impractical as the data size grows, especially when more complex attributes are desired, since computing those indices would require examining all the records in the database, an increasingly time consuming task. To address this challenge, data intensive businesses such as Google and Yahoo! have developed efficient schema-neutral distributed storage paradigms, hosting the data on thousands of nodes and moving the computation closer to each node, and relying upon redundancy to protect against transient faults or hardware failures. In addition, there is a wide database research community exploring alternate methods for spatial and temporal data structure and layout optimization for rapid access and query. These skills can be leveraged to assist in IDC archival data search and content extraction.

Physics-Based/Earthly Empirical Models:
Production of high quality ground truth for seismic characterization requires using state-of-the-art supercomputing to calculate the geology of the transmission medium, and even then, the geographic sensor coverage in many areas is insufficient to either high fidelity 3D models or comprehensive empirical calibration of travel times. Stochastic techniques are emerging which incorporate physics based constraints of seismic travel times to curb the incoming data sets and improve predictions. These methods can prove extremely valuable for tomographic inversion and event location as well as for providing more accurate training data for classification and inferenceing.

In addition to the above, as more diverse sensors come on line, the opportunity for automatically federating and fusing data sources to establish event correlations becomes available. Such data fusion can not only add the inferenceing process but also serve to update existing models for phase/event classification, thereby further enhancing the validity of the final output. These techniques could prove extremely valuable for OSI where time to solution is of the essence.

vDEC will champion and harness innovations in data exploitation to improve the effectiveness of CTBFT verification.

Continuous Station Self Calibration and Self Diagnosis:
Today’s station-specific “recipes” and error masking rules require manual design on a pre-station basis, and are not adaptive to changing monitoring conditions, resulting in several stations being effectively non-functional on an ad hoc basis.

To remedy this, one approach may be to implement algorithms for fault checking and self error correction at the collection nodes, so as to filter out random faults and sensor drift as well as keep the data processing pipeline (and by default, a human analyst) informed of upstream station uncertainties. In addition, techniques which provide feedback to the collection network based upon the validity of end results can provide a valuable learning tool for continuous sensor calibration. These methods are widely deployed in machine learning and robotics communities for mapping and localization and could prove useful even in the vDEC domain. Alternatively, it may be as or more effective to calibrate sensor parameters based upon data from other stations; this would require no manually defined ground-truth events, as long as a reasonable fraction of stations were operational.