Sudarshan S. Chawathe's Publications

2023-02-11

Notes:

- Most recent version: `http://chaw.eip10.org/pubs/`.
- A PDF version suitable for offline viewing or printing: `pubs.pdf`.
- Bibliographic files for your citation pleasure:
  - BibTeX: `pubs.bib`
  - RIS: `pubs.ris`
  - EndNote: `pubs.end`
- Validation checks:
  - XHTML: `https://validator.w3.org/check?uri=http://chaw.eip10.org/pubs/`
  - CSS: `https://jigsaw.w3.org/css-validator/validator?uri=http://chaw.eip10.org/pubs/`
- Older list with more details of earlier publications: `http://chaw.eip10.org/opubs/`.
- Missing abstracts, keywords, and other details to be added soon.

- ***Pragmatic Domestic Electrical Load Disaggregation***. Sudarshan S. Chawathe. In Proceedings of the 13th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2022), Las Vegas, Nevada, March8–11 2023. To appear.

  This paper studies methods for determining the major electrical loads that contribute to aggregate electric energy consumption for a household or similar unit. It provides an alternate formulation of a well studied general problem and a framework and prototype implementation to address it. The focus is on methods that do not require any instrumentation or data beyond hourly (or similar low frequency) records of aggregate energy consumption, as is often easily available from power utility companies due to the increasing prevalence of smart meters. As well, the focus is on pragmatic approaches that are likely to provide useful information for typical household electricity consumption in contrast to methods more suited to industrial environments. Another notable feature is that disaggregation is performed not centrally at the utility company or similar entity with data from a large number of households but instead in a distributed and independent manner at each household. This feature provides two key benefits: (1) It permits the injection of information known to a household but not (easily) by others in order to simplify the problem. (2) It provides better privacy protections for such data.

  ▷ Non-Intrusive Load Monitoring (NILM); Electrical Load Disaggregation; Smart Meters; Data Integration. Data Visualization.

- ***Classification of Small Molecules Regulating Circadian Rhythm***. Sudarshan S. Chawathe. In Proceedings of the 13th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2022), New York, NY, October26-29 2022.

  The process of drug discovery using in silico methods often produces datasets with a very large number of attributes (fields) per instance (record). Automated classification of such data on properties such as toxicity provides significant benefits for drug design but must cope effectively with the large number of attributes and the relatively small number of instances. This paper studies this problem in the context of a dataset, from prior work, used to discover promising small molecules for controlling circadian rhythm in humans. By identifying a suitable small subset of the attributes that are effective for this classification

task, experimental results indicate accuracies that compare very favorably with prior work on the same data.

▷ Small Molecules; Scientific Data; Data Integration; Data Visualization; Classification.

- ***Automated Determination of Mushroom Edibility Using an Augmented Dataset***. Sudarshan S. Chawathe. In IEEE World AI IoT Congress (AIIoT 2022), Seattle, Washington, June6–9 2022.

  This paper studies methods and datasets for automated classification of mushrooms as edible or poisonous based on easily observable properties such as colors, textures, and dimensions of mushroom parts. The focus is on data-intensive methods that build upon recent work that has led to an augmented database of mushroom features. This dataset is studied in detail with the goal of explicating properties and easing further use of the dataset by others. The merit of the database features for the classification task is quantified using several metrics. Results quantify the accuracy and efficiency of classification using all and only a few of the features.

  ▷ Mushroom Database; Classification and Taxonomy; Scientific Data; Data Integration; Machine Learning.

- ***Optical Features for Automated Determination of Agricultural Product Varieties***. Sudarshan S. Chawathe. In IEEE World AI IoT Congress (AIIoT 2022), Seattle, Washington, June6–9 2022.

  This paper studies methods to determine varieties of agricultural specimens using features extracted from optical images generated by low-cost commodity hardware and simple, efficient algorithms. It presents a framework for this and some related tasks of agricultural informatics, with a focus on data-intensive aspects. It describes a system implementation that permits such data to be iteratively and interactively explored and studied while also permitting efficient programmatic access. The core classification problem of determining a raisin variety is studied experimentally and the quantitative results are competitive with prior work. Some of the methods generate simple, human-understandable classifiers, of which a few examples are presented. Data exploration and visualization is implemented using self-organizing maps (SOMs) and several examples of useful visualizations are described.

  ▷ Agricultural Informatics; Data Exploration and Visualization; Self-Organizing Maps (SOMs); Classification; Machine Learning.

- ***Predicting Bicycle Package Delivery Demand Using Historical Spatiotemporal Data***. Sudarshan S. Chawathe. In Proceedings of the 12th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2022), Las Vegas, Nevada, January26-29 2022.

  The primary task addressed by this paper is the prediction of current, or near future, demand for package deliveries at a location using spatiotemporal historical records for that location and for others near it. This work adopts a data-driven approach and describes methods for exploring and visualizing such datasets in order to gain a better understanding of the domain and to select appropriate specific methods for tasks such as demand prediction and location identification. As a concrete example, the paper uses such a dataset recently provided by the Pedal Me service in London.

  ▷ Demand Prediction; Package Deliveries; Logistics; Visualization; Data Exploration; Data Science.

- ***Classification of Dry Beans Using Image Features***. Sudarshan S. Chawathe. In Proceedings of the 12th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2021), New York, NY, December1-4 2021.

This paper presents human-understandable methods for automated classification of dry beans using features extracted from optical images. It presents a detailed study of these features in the context of classification by examining their merits and the effect of using a reduced feature set. It also presents the results of constructing self-organizing maps (SOMs) for these features. An important result is that classification limited to human-understandable methods for this task does not incur any penalty in accuracy and comes with the benefit of significantly lower computational costs. Another result is that SOMs applied to this data provide a useful visualization that invites further study.

▷ Agricultural Informatics; Dry Beans; Classification; Self-Organizing Maps (SOM).

- ***Inferring Human Activity Using Wearable Sensors***. Sudarshan S. Chawathe. In Proceedings of the 12th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2021), New York, NY, December1-4 2021.

  This paper presents methods that use data from wearable sensors, such as those found in low-cost commodity hardware, to infer the human activity (such as reading or walking) corresponding to the sensor readings. A related task is the identification of individuals based on the same data. The classification accuracy of the methods used in this work is higher than earlier work using the same dataset. Further, a significant reduction in the number of sensor data streams produces only a very small impact on this accuracy, which is a feature of practical significance due to implications for network bandwidth and energy budgets in such systems.

  ▷ Human Activity Recognition (HAR); Wearable Sensors; Sensor Data; Classification; Machine Learning.

- ***Human Identification by Gait Using Body-Worn Sensors***. Sudarshan S. Chawathe. In Global Conference on Artificial Intelligence and Applications (GCAIA 2021), Jaipur, Rajasthan, India, September8–10 2021.

  This paper studies methods for identifying human individuals and gender using gait-related features as measured by sensors worn on the body. A recently published dataset due to prior work is used to study the effectiveness of well established and efficient methods for such identifications. The dataset is based on experiments with 16 participants wearing sensors that are part of a widely used gait-sensing platform. The accuracies of the best of these methods compare favorably with those reported by prior work. Since the records in the dataset are characterized by a very large number of fields (323 attributes per record), methods for attribute selection are of particular interest and are also studied. The underlying implementation is briefly described, with a focus on some data management challenges posed by the large number of attributes. A notable result is that prediction accuracies of several competitive methods are not diminished even when the number of attributes is reduced very drastically using attribute-selection methods based on metrics such as ReliefF and Symmetrical Uncertainty.

  ▷ Human Gait; Body-Worn Sensors; Data Management; Classification; Machine Learning.

- ***Using Data from In-Vehicle Recommender Systems to Predict Traveler Characteristics***. Sudarshan S. Chawathe. In Global Conference on Artificial Intelligence and Applications (GCAIA 2021), Jaipur, Rajasthan, India, September8–10 2021.

  In-vehicle recommender systems may be used to present travelers with offers (such as coupons) customized by location, history, and other contextual information. Such systems both utilize and augment a dataset that records which offers are accepted and under what contextual conditions. This paper studies the use of such datasets to make predictions on whether a coupon presented to a traveler with some known characteristics and in a certain context relative to travel parameters is likely to be accepted. It also studies the use of such data

to infer traveler characteristics based on coupon acceptance and related data. This work emphasizes the use of simple and understandable (explainable, for humans) models whose examination is likely not only to provide greater confidence in predictions but also to permit design of offers customized to elicit desired responses and information from travelers. Using a recently published dataset due to prior work, these methods are studied experimentally both quantitatively and qualitatively (by examining a few concrete models).

▷ In-Vehicle Recommender Systems; Intelligent Transportation Systems; Data Science; Classification; Machine Learning.

- ***Epidemiological Spatiotemporal Data Exploration and Prediction***. Sudarshan S. Chawathe. In IEEE World AI IoT Congress (AIIoT 2021), Seattle, Washington, May10–13 2021.

  This paper addresses epidemiological spatiotemporal datasets such as those reporting the number of cases of infectious diseases over time and by geographical location. It studies methods for exploratory data analysis as well as prediction of future cases based on prior data. It emphasizes methods that provide explainable predictions, such as those based on rules and decision trees. These methods are studied in the context of a recently published dataset of weekly Chickenpox cases in Hungarian counties over a 10-year period. As noted in prior work, this dataset exhibits several features, such as seasonality and heteroskedasticity, that make the prediction task especially challenging. This paper describes some results of an experimental study of both the exploratory and predictive aspects.

  ▷ Spatiotemporal Data; Data Exploration; Self-Organizing Maps (SOMs); Prediction; Machine Learning.

- ***Explainable Predictions of Industrial Emissions***. Sudarshan S. Chawathe. In IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS 2021), Toronto, Canada, April21–24 2021.

  Predictive emission monitoring systems for gas turbines are important in the power generation industry. A key task in this context these systems is the prediction of flue gas emissions using process and environmental measurements that are easier to obtain. This paper presents methods for such predictions with an emphasis on explainability. A notable result is that despite the potential restrictions imposed by this emphasis, the numerical accuracy compares very favorably with prior work that uses models that are more difficult to explain.

  ▷ Predictive Emission Monitoring Systems; Exhaust Emissions Prediction; Gas Turbines; CO; NOx; Machine Learning.

- ***Data Structures for Ordered Short Character-Sequences***. Sudarshan S. Chawathe. In Proceedings of the 11th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2021), Las Vegas, Nevada, January27-30 2021.

  A lexicon, or dictionary of key-value pairs, is a general abstraction that is widely used in diverse areas of computer science, notably compilers and database systems. The primary operations of interest on such lexicons are membership testing and extraction of a value associated with a key appearing in the lexicon. This paper focuses on the special case of ordered lexicons with keys that are short sequences of characters. An important motivating application is the representation of the large and growing lexicon of emoji in the Unicode standard. It presents space-efficient data structures for some specialized but practically significant cases. In particular, the methods take advantage of contiguous sequences of keys in the lexicon to yield a very highly compressed representation while maintaining efficiency in lookup operations.

  ▷ Data Structures; Lexicons; Unicode; Emoji.

- ***Analyzing Auction Data for Anomalous Bidding***. Sudarshan S. Chawathe. In Proceedings of the 11th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2020), New York, NY, October28-31 2020.

  Online auctions as exemplified by sites such as ebay.com are responsible for very large volumes of transactions and monetary value. Their growth has also led to a growth in fraudulent activities in these markets. This paper studies transaction data from such auctions with the goal of using it to detect anomalous and potentially fraudulent bidding. To that end, it explores several approaches based on classification, clustering, and visualization. The quantitative results signal very high accuracy in classification but their promise is tempered by some limitations of the experimental dataset. Clustering and visualizations using self-organizing maps (SOMs) is found to be more effective for this data than clustering using more conventional methods such as k-means. In particular, the SOMs reveal several interesting relationships among the dataset's attributes and their correlations to anomalous bidding.

  ▷ Online Auctions; Fraud Detection; Classification; Clustering; Visualization; Machine Learning; Self-Organizing Maps (SOMs).

- ***Estimating Predicate Selectivities in a NoSQL Database Service***. Sudarshan S. Chawathe. In Proceedings of the 11th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2020), New York, NY, October28-31 2020.

  An estimate of the number of items in a database that satisfy an equality or range predicate is useful for several tasks, such as cost-based query optimization, provisioning of system resources, and determining the financial costs of using database services. In a traditional database system, such estimates are computed and used internally by the system and have been well studied. In contrast, such estimates have not received much attention in the context of a cloud-based database service, where they must be computed by the application that uses the service using only the limited features of the interface provided by the service. This paper motivates and formulates the selectivity-estimation problem for database services. It describes the characteristics of this problem that distinguish it from the analogous problem in traditional database systems. It outlines some subproblems and methods to address them. It provides a method for estimating selectivities based on random sampling along with some experimental results.

  ▷ Cloud Databases; Database Services; Cost Estimation; DynamoDB.

- ***Mining Bike-Share Data***. Sudarshan S. Chawathe. In Proceedings of the IEEE International Smart Cities Conference (IEEE ISC2 2020), September28 2020.

  This paper studies methods for processing bike-share datasets for the purpose of extracting information that can assist riders, bike-share program designers, city planners, and others. Bike-share datasets describe how shared bicycles are used in an urban environment. They vary considerably in composition and coverage but typically include information such as the locations (bicycle racks) of origin and destination, timestamps, and identifiers for bicycles and riders. This paper provides methods for visualizing such data in a manner that distills useful patterns and for using the data to predict usage. In order to overcome the difficulty in generating meaningful clusters using conventional methods, it presents a novel method of clustering that uses graph condensations. It describes an experimental study of these methods using a publicly available dataset from a popular bike-share program.

  ▷ Bikeshare; Transportation; Data Analysis.

- ***Using Accelerometers in Mobile Phones to Estimate Blood Alcohol Levels***. Sudarshan S. Chawathe. In Proceedings of the IEEE International Smart Cities Conference (IEEE ISC2 2020), September28 2020.

This paper studies methods for determining the blood alcohol content of individuals by using data from commodity accelerometers in mobile phones carried on person. A significant challenge is that such data is very noisy and often irregular (many large gaps) as well. This paper provides a detailed analysis of a recently released dataset of accelerometer traces and associated readings of transdermal alcohol content (TAC). It describes a set of features extracted from the raw accelerometer traces that are effective for the task of determining TAC levels. It presents results of an experimental study of regression methods that use these features to predict TAC levels from accelerometer traces as well as of classification methods that predict whether the person carrying the mobile phone has TAC levels above given thresholds.

▷ Alcohol Consumption; Accelerometers; Regression; Classification.

- ***Diagnostic Classification Using Hepatitis C Tests***. Sudarshan S. Chawathe. In IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS 2020), Vancouver, Canada, September9–12 2020.

  This paper describes methods for automated classification of individuals by Hepatitis C medical category using data from a series of commonly used diagnostic tests. The methods are evaluated experimentally using a publicly available dataset from prior work. The accuracy of some methods compares favorably with similar results reported in prior work. In addition to quantitative results on prediction accuracy, training and testing times, and model sizes, the paper includes a detailed look at some concrete representative classifiers generated by a few of the competitive methods, permitting a human domain expert to further study the models and classifiers.

  ▷ Medical Informatics; Classification.

- ***Detecting Physical Activities Using Body-Worn Accelerometers***. Sudarshan S. Chawathe. In IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS 2020), Vancouver, Canada, September9–12 2020.

  This paper addresses the task of using data from accelerometers attached to a person's body to determine the kind of physical activity being performed by that person. The activities of interest are routine ones such as sitting, walking up a flight of stairs, walking, and jogging. The paper describes methods for segmenting the time-series data from accelerometers and for extracting features that are effective for determining activities when used in conjunction with well established classification algorithms. These methods are implemented in a prototype that is used to evaluate their effectiveness on a publicly available dataset of tagged accelerometer traces. The prototype also provides intuitive visualizations of the accelerometer traces, allowing a human expert to gain a better understanding of both the dataset and the predictions from the classifiers. Although the methods in this paper use fewer and simpler features extracted from the raw accelerometer data, they provide higher accuracies when compared to those reported in prior work on the experimental dataset.

  ▷ Human Physical Movement; Activities of Daily Living; Accelerometers; Classification.

- ***Characterizing Shoulder Implants in X-Ray Images***. Sudarshan S. Chawathe. In Global Conference on Artificial Intelligence and Applications (GCAIA 2020), Jaipur, Rajasthan, India, September8–10 2020.

  This paper studies methods for characterizing shoulder implants in X-ray images of the shoulder, upper arm, and chest region. The task of characterizing in this context entails sub-tasks such as detecting the presence of an implant, segmenting it from the rest of the image, determining its orientation relative to the image and other features such as bones, detecting shape features of the implant, and determining properties of the implant (e.g., manufacturer, model). The task is complicated due to the proximity and similarity of bones

and other objects as well as due to potentially low image contrast, spurious edges, and other artifacts. This paper describes the challenges and outlines and evaluates solutions using a recently published dataset of 597 X-ray images of shoulder implants.

▷ Medical Imaging; Medical Informatics; Image Processing; Classification.

- ***Human-Understandable Classifiers for COPD From Biosensor Data***. Sudarshan S. Chawathe. In Global Conference on Artificial Intelligence and Applications (GCAIA 2020), Jaipur, Rajasthan, India, September8–10 2020.

  This work addresses the task of analyzing data from a biometric sensor operating on saliva samples in order to predict the sample-donor's status in regard to COPD (Chronic Obstructive Pulmonary Disease). It emphasizes the use of human-understandable classification methods, such as those based on a small number of rules. Using recently published data, it studies the characteristics of such biosensor data and presents some concrete results in that context. It also summarizes the results of an experimental evaluation of such methods on this dataset.

  ▷ Medical Diagnostics; Classification; Medical Informatics; Machine Learning.

- ***Index-Selection for Minimizing Costs of a NoSQL Cloud Database***. Sudarshan S. Chawathe. In Proceedings of the 17th International Conference on Economics of Grids, Clouds, Systems and Services (GECON 2020), Izola, Slovenia, September15-17 2020. Springer LNCS.

  The index-selection problem in database systems is that of determining a set of indexes (data-access paths) that minimizes the costs of database operations. Although this problem has received significant attention in the context of relational database systems, the established methods and tools do not translate easily to the context of modern non-relational database systems (so-called NoSQL systems) that are widely used in cloud and grid computing, and in particular systems such as DynamoDB from Amazon Web Services. Although the index-selection problem in these contexts appears simple at first glance, due to the very limited indexing features, this simplicity is deceptive because the non-relational nature of these databases and indexes permits more complex indexing schemes to be expressed. This paper motivates and describes the index-selection problem for NoSQL databases, and DynamoDB in particular. It motivates and outlines a cost model to capture the specific monetary costs associated with database operations in this context. The cost model has not only been carefully checked for consistency using the system documentation but also been verified using actual usage costs in a live DynamoDB instance.

  ▷ Cloud Computing; Cost Model; Index Selection; DynamoDB; NoSQL Databases; Physical Database Design.

- ***Organizing and Compressing Collections of Files Using Differences***. Sudarshan S. Chawathe. In Proceedings of the 24th International Database Engineering and Applications Symposium (IDEAS 2020), Incheon/Seoul, South Korea, August12-18 2020. ACM.

  A collection of related files often exhibits strong similarities among its constituents. These similarities, and the dual differences, may be used for both compressing the collection and for organizing it in a manner that reveals human-readable structure and relationships. This paper studies methods for such organizing and compression of file collections using differences and presents the results of an experimental evaluation on a well known public dataset.

  ▷ File Collections; Differencing; Compression.

- ***Mining Frequent Differences in File Collections***. Sudarshan S. Chawathe. In Proceedings of the Ninth IEEE International Workshop on Data Integration and Mining (DIM-2020), Las Vegas, Nevada, August11-13 2020. IEEE. In conjunction with IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI 2020).

Collections of textual files, or documents, with substantial inter-document similarities are common in diverse domains. A practically significant class of such similarities, and the dual differences, are well characterized by edit scripts, or colloquially diffs, that use a simple sequence model for documents. The study of such diffs provides valuable insights into the inter-document relationships within a collection and can guide data integration within and across collections. This paper describes a framework for such study that is based on frequently occurring inter-document differences. It motivates and defines a general problem of mining frequent differences and outlines some specific instances. It presents the design and implementation of a prototype system for interactively discovering and visualizing frequent differences. A notable feature of this method is its use of difference-components, or deltas, to bootstrap the discovery of interesting structure in file collections. The paper describes a preliminary experimental evaluation of the method and implementation on a widely used corpus of file-collections.

▷ File Collections; Differencing; Data Mining; Data Integration.

- ***Efficient File Collections for Embedded Devices***. Sudarshan S. Chawathe. In Proceedings of the 8th Workshop on Communications in Critical Embedded Systems (WoCCES 2020), Rennes, France, July7 2020. IEEE.

  This paper studies methods for efficiently transferring and storing collections of related files in embedded devices and other environments with limitations on storage, network, and energy use. Files in collections based on purpose (e.g., system configurations) or other aspects often exhibit substantial inter-file similarities. These similarities may be used to achieve significant reductions in the network resources required for transferring or updating the collection, as well as for the storage resources required on the embedded devices on which it is stored.

  ▷ Embedded Devices; File Collections; Compression.

- ***Rice Disease Detection by Image Analysis***. Sudarshan S. Chawathe. In Proceedings of the 10th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2020), Las Vegas, Nevada, January6-8 2020.

  This paper provides a method for automatically classifying diseases in rice plants by analyzing photographs of rice leaves. The method uses image processing algorithms to detect leaves and likely disease-induced lesions in the leaves. Next, several attributes are computed based on the dimensions of leaves and lesions, the numbers and shapes of lesions, as well the color characteristics of lesions and intact portions of leaves. These attributes are used to build classification models using well established algorithms. The method is evaluated using a publicly available database of rice leaf images.

  ▷ Rice Disease; Rice Leaf; Image Processing; Classification; Machine Learning.

- ***Topic Analysis of Climate-Change News***. Sudarshan S. Chawathe. In Proceedings of the 10th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2020), Las Vegas, Nevada, January 2020.

  This paper explores the application of computational methods to the analysis of the large and growing corpus of news articles and related data on climate change. Topics are analyzed using Latent Dirichlet Allocation and methods customized to specific news sources that take advantage of keywords and other metadata that may be present. Results of this method on news articles drawn over several months are presented.

  ▷ Climate Change; News; Topic Modeling; Machine Learning

- ***Cost-effective data-collection systems for citizen science***. Sudarshan S. Chawathe. In Proceedings of the Acadia National Park Science Symposium, Schoodic Education and Research Center, Acadia National Park, Maine, October24 2019.

Citizen science efforts often include data collection by volunteers. Computerizing such data collection provides several benefits, including improved data consistency, shorter time from collection to use, and immediate feedback to the data collectors. Implementing such a computerized data collection system is often challenging because it is difficult to accurately estimate the level of participation and, therefore, the required load-handling capacity. Overestimating the capacity results in unnecessary infrastructure costs while underestimating it leads to sluggish or failed systems. The so-called serverless or cloud based systems are attractive in this context because they permit the apparent (paid) infrastructure to scale with load. Determining cost profiles of different designs in this environment and, therefore, selecting a suitable one are challenging tasks that are addressed by this work.

- ***Data Modeling for a NoSQL Database Service***. Sudarshan S. Chawathe. In Proceedings of the 10th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2019), New York, NY, October10-12 2019. Columbia University.

  Cloud-hosted NoSQL database services, such as AWS DynamoDB, offer significant advantages, including low start-up costs, high performance and availability, wide scalability, and ease of deployment and management. These advantages have led to their rapid adoption and growth. However, the data storage, querying, and modification features supported by such NoSQL services are very rudimentary in comparison with those of relational and object database systems. Further, data modeling decisions made to map application requirements to the supported NoSQL model have very significant impact on not only performance but also financial cost incurred in using the services. Unlike the well developed body of work for relational- and object- database design, there is a great dearth of systematic procedures for NoSQL database design. This paper addresses this design problem by providing methods that map standard data models to the typically idiosyncratic and rudimentary models supported by NoSQL database services, using AWS DynamoDB as a specific instance.

  ▷ Cloud Computing; NoSQL; Databases; Data Modeling.

- ***Using Historical Data to Predict Parking Occupancy***. Sudarshan S. Chawathe. In Proceedings of the 10th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2019), New York, NY, October10-12 2019. Columbia University.

  This paper describes methods that use historical data on the rates of occupancy of car parking facilities to predict future occupancy rates. The methods are evaluated using a publicly available dataset of car park occupancy rates. The results suggest that a usable level of prediction accuracy may be achieved using only a modest amount of data that is easy to gather using current technologies.

  ▷ Intelligent Transportation Systems; Smart Cities; Car Parking; Regression; Machine Learning.

- ***Trusted Remote Function Interface***. Mark E. Royer and Sudarshan S. Chawathe. In Proceedings of the 10th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2019), New York, NY, October10-12 2019. Columbia University.

  The Trusted Remote Function Interface (TRFI) is a small library that exposes services via a REST API to allow function execution with scientific programming languages. Functional units are uploaded to a remote server using the provided REST API. The API stores registered functions for later execution. Maintaining code using this technique allows clients to repeatedly execute functions without having the native language, typically Octave or Python, installed on the client machine. A common problem in scientific applications is the requirement for a program to interface with scientific scripting languages. Typically, this is not a straightforward approach for accomplishing the data exchange and subsequent function execution on that data from popular languages such as Java or JavaScript. This task is extraordinarily cumbersome if the interpreter, used by the scientific programming language,

is not installed locally. By separating the function signature from the underlying implementation, and providing a uniform REST API, the TRFI library allows function interfacing in two ways. First, direct interfacing by using the equipped Java library. Second, the more common scenario is interfacing remotely by deploying the library using a JAX-RS compatible web server. The result of the TRFI library's design and the provided REST API is the facilitation of code interoperability and reuse for scientific applications.

▷ Java; REST; Octave; Python; function interoperability; data exchange

- ***Unformatted, Certified Scientific Objects***. Mark E. Royer and Sudarshan S. Chawathe. In Proceedings of the 10th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2019), New York, NY, October10-12 2019. Columbia University.

  We present an approach for scientific data management systems to apply certificates to scientific objects, which are typically unformatted datasets, to facilitate analysis by climate scientists. For a program to process data, the program requires cleaned data in a form that supports automatic manipulation. Most systems require that data must adhere to a specific format to achieve that goal. The technique described in this paper takes the opposite approach; instead, any dataset may be imported and manipulated in the system. But upon initial import, however, only a subset of system functions may work with any given dataset. As the data is refined and transformed by system functions, more functions may become compatible. Certificates are associated with objects that pass constraint validation within the system to ensure that they conform to function requirements. The attached object constraints represent invariant properties of the object, which may be used by functions in the system as function preconditions. Furthermore, the functions defined in the system may associate certificates with the newly generated results. Certificates related to function results are effectively function postconditions, which in turn are used to associate certificates with the objects generated in the system. Additionally, attached object certificates reflect the refinement of data into a more pristine version. This paper describes the technique for modeling and enforcing the constraints for data scientists that have similar requirements.

  ▷ Data analysis; Constraints; Data processing

- ***Indoor-location classification using RF signatures***. Sudarshan S. Chawathe. In Proceedings of the 17th IEEE International Symposium on Network Computing and Applications (IEEE NCA 2019), Cambridge, MA, September26-28 2019.

  Indoor localization using radio-frequency signals in the 2.4 GHz band is attractive due to availability of low-cost commodity WiFi hardware. However, using such signals for localization is challenging due to signal-propagation complexities such as multipath, fading, and shadowing. This paper describes a method for classifying indoor locations using frequency-domain signatures of RF signals. The method is evaluated using a publicly available dataset of detailed signal measurements in a real environment.

  ▷ indoor localization; radio-frequency signals; classification; machine learning

- ***Cost-Based Query-Rewriting for DynamoDB (Work in Progress)***. Sudarshan S. Chawathe. In Proceedings of the 17th IEEE International Symposium on Network Computing and Applications (IEEE NCA 2019), Cambridge, MA, September26-28 2019.

  DynamoDB is a popular NoSQL database service that permits queries in a restrictive but useful query language. The metered costs (which translate to financial costs) of executing such queries are measured in units of provisioned capacity or number of requests. Costs of equivalent queries may differ by orders of magnitude but the onus of choosing a low-cost equivalent query is on the service's client and must be performed by query rewriting. This paper formulates this query-rewriting problem for DynamoDB and outlines methods for choosing low-cost equivalent queries.

▷ DynamoDB; query evaluation; cost estimation; databases; NoSQL; cloud computing.

- ***Hand Gestures from Low-Cost Surface-Electromyographs***. Sudarshan S. Chawathe. In 2019 IEEE National Aerospace and Electronics Conference (NAECON), July15-19 2019. 71st annual conference.

  Low-cost and commodity off-the-shelf surface electromyographs (sEMGs) may be used for unobtrusive detection of human hand gestures. Although these EMG signals are not as detailed as conventional ones, an experimental investigation of feature engineering and classification reveals that they can yield accurate hand gesture information.

  ▷ hand gesture detection; surface electromyograph; sEMG; EMG; COTS; sensors; classification; machine learning

- ***Ultrasonic Flowmeter Diagnosis by Classification***. Sudarshan S. Chawathe. In 2019 IEEE National Aerospace and Electronics Conference (NAECON), July15-19 2019. 71st annual conference.

  Modern ultrasonic flowmeters provide routine diagnostic information that may be used to infer their health. This inference task is modeled as a classification problem and studied experimentally using a publicly available dataset. A few classifiers, such as Bayesian Networks, provide good accuracy and also suggest relationships among the diagnostic variables.

  ▷ ultrasonic flowmeter; diagnostics; classification; machine learning

- ***Computational Analysis of Climate-Change Discourse in News and Social Media***. Sudarshan S. Chawathe. In Proceedings of the 27th Annual Harold W. Borns Symposium, May 2019.

  The study of topics that frame the discourse of climate change in news and social media is useful for understanding media and public perceptions of the field and its recent developments. Computational methods for topic modeling, syntactic analysis, and guided data exploration may be applied to readily available big-data streams to extract topics and related information in near-real time.

- ***Ice core dating integration in the Climate Data Workbench***. Mark E. Royer, Sudarshan S. Chawathe, Andrei V. Kurbatov, and Paul A. Maywewski. In Proceedings of the 27th Annual Harold W. Borns Symposium, May 2019.

  We present the software integration of ice core dating tools to the Climate Data Workbench (P301 system). The implementation allows researchers to use different annual indicators in ice core time series in order to develop and apply time scales. During the creation of the time scale, an interpolated, dated version of the actively investigated core is presented to the researcher in real-time.

- ***Condition Monitoring of Hydraulic Systems by Classifying Sensor Data Streams***. Sudarshan S. Chawathe. In Proceedings of the 9th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2019), Las Vegas, Nevada, January 2019.

  Condition-based maintenance (CBM) of hydraulic systems requires methods for condition monitoring: Sensors installed in a hydraulic system for this purpose generate streams of real-time data that must be analyzed to accurately characterize the health of the system. Prior work has developed an experimental hydraulic system with such an installation and yielded a public dataset of sensor readings with associated values of condition variables that quantify the system's health. This paper presents classification-based methods for inferring these condition variables from the sensor data streams. These methods significantly improve on the classification accuracy reported in prior work on this data. Further, this accuracy is maintained even when the number of sensor-based attributes used as input is substantially reduced.

  ▷ condition monitoring, condition-based maintenance, hydraulic systems, sensors, classification

- ***Recognizing Human Falls and Routine Activities Using Accelerometers***. Sudarshan S. Chawathe. In Proceedings of the 9th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2019), Las Vegas, Nevada, January 2019.

  Detecting falls and other mishaps using data from sensors worn by individuals is an important task with applications in healthcare. A related task is using such sensor data to detect routine activities of daily living. This paper models such detection of falls and routine activities as a classification problem. Using a publicly available dataset of real accelerometer traces generated by participants performing intentional falls and other activities, the efficacy and performance of several classifiers are studied experimentally.

  ▷ fall detection, activities of daily living, accelerometers, sensors, classification

- ***Clustering Blockchain Data***. Sudarshan S. Chawathe. In Olfa Nasraoui and Chiheb-Eddine Ben N'cir, editors, Clustering methods for Big Data Analytics: techniques, toolboxes and applications, chapter 3. Springer, 2019.

  Blockchain datasets, such as those generated by popular cryptocurrencies Bitcoin, Ethereum, and others, are intriguing examples of big data. Analysis of these datasets has diverse applications, such as detecting fraud and illegal transactions, characterizing major services, identifying financial hotspots, and characterizing usage and performance characteristics of large peer-to-peer consensus-based systems. Unsupervised learning methods in general, and clustering methods in particular, hold the potential to discover unanticipated patterns leading to valuable insights. However, the volume, velocity, and variety of blockchain data, as well as the difficulties in evaluating results, pose significant challenges to the efficient and effective application of clustering methods to blockchain data. Nevertheless, recent and ongoing work has adapted classic methods, as well as developed new methods tailored to the characteristics of such data. This chapter motivates the study of clustering methods for blockchain data, and introduces the key blockchain concepts from a data-centric perspective. It presents different models and methods used for clustering blockchain data, and describes the challenges and some solutions to the problem of evaluating such methods.

- ***The Tiny Java Library for Maintaining Model Provenance***. Mark E. Royer and Sudarshan S. Chawathe. In Proceedings of the 9th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2018), New York, NY, November 2018. Columbia University.

  We present a small library for maintaining the provenance of objects in a software model called The Tiny Java Library for Maintaining Model Provenance (TJLP). A unique characteristic of the library is that it may be applied to existing software models with minimal modification. The library allows the software developer to introduce the ability to move back (undo) and forward (redo) through an object's instance history with minimal code modification. The requirement is that the model implements the Model interface. Finally, methods that are considered critical in the object's provenance are adorned with an Undoable annotation. The code necessary to maintain the object's history is automatically inserted into the critical, undoable-method bytecode when the class definition is loaded by an extended class loader. The states of the model objects are preserved both in memory and on disk to accommodate various computer system configurations. The library performs well for small to medium size models using the default settings, but it may be customized in order to perform better with larger models, especially if the model size approaches the RAM of the underlying computer system.

  ▷ Java annotations, data provenance, bytecode injection

- ***A software workbench for studying past climate***. Mark E. Royer, Sudarshan S. Chawathe, and Andrei V. Kurbatov. In Proceedings of the Acadia National Park Science Symposium, Bar Harbor, Maine, October20 2018.

The study of past climate enables a better understanding of present and future climate conditions. However, directly measured data for temperature and other climate variables is available for only the recent past (a few hundred years). Study of climate in the more distant past, from centuries to millennia before present, requires the use of indirect methods which use other variables as proxies. Chief among such methods is the use of data derived from ice cores. Analyzing such ice-core data in order to gain insights into past climate is a complex task that requires data from diverse sources to be combined, transformed, and visualized in multiple and often novel ways. In the past, such analysis was often performed using an ad hoc collection of software tools, such as spreadsheets and plotting programs. There are two primary reasons why this past approach to analyzing data is no longer effective: First, recent technological advances in the physical and chemical processing of ice cores to extract measurements have resulted in orders-of-magnitude increase in the volume of data. Not only does this volume of data render some software tools inoperable but also it makes it difficult for a human to interpret data visually. Second, and more important, ad hoc application of multiple tools to analyze data, even when it produces usable results, typically leaves no systematic record of the precise sequence of transformations that yield a data product, such as a chart of temperature over time, from the original data sources. The P301 project addresses these shortcomings of prior data analysis methods by providing an interactive, graphical software workbench with a few notable features in this context: First, it can analyze even the largest ice-core datasets available today, and more, in interactive times (a few seconds at most). Second, it permits a scientist to interactively use, define, and compose software tools for analyzing data in diverse and powerful ways. Third, all transformations of both tools and data are automatically recorded by the system in a manner that permits examination, study, transformation, and workflow management.

- ***Monitoring IoT networks for botnet activity***. Sudarshan S. Chawathe. In Proceedings of the 17th IEEE International Symposium on Network Computing and Applications (IEEE NCA 2018), Cambridge, MA, November 2018.

  The Internet of Things (IoT) has rapidly transitioned from a novelty to a common, and often critical, part of residential, business, and industrial environments. Security vulnerabilities and exploits in the IoT realm have been well documented. In many cases, improving the security of an IoT device by hardening its software is not a realistic option, especially in the cost-sensitive consumer market or in legacy-bound industrial settings. As part of a multifaceted defense against botnet activity on the IoT, this paper explores a method based on monitoring the network activity of IoT devices. A notable benefit of this approach is that it does not require any special access to the devices and adapts well to the addition of new devices. The method is evaluated on a publicly available dataset drawn from a real IoT network.

  ▷ Internet of Things (IoT), botnets, network monitoring, machine learning

- ***Analysis of Burst Header Packets in Optical Burst Switching networks***. Sudarshan S. Chawathe. In Proceedings of the 17th IEEE International Symposium on Network Computing and Applications (IEEE NCA 2018), Cambridge, MA, November 2018.

  Optical Burst Switching (OBS) networks provide a practical alternative to optical packet switching and optical circuit switching by separating control information from the primary data, sending the former on a separate control channel. However, this separation also renders OBS networks susceptible to a denial- or degradation-of-service attack (intentional or otherwise) when the data provisioned by a header packet on the control channel does not materialize. This paper addresses the problem of detecting and characterizing such problems and describes a method based on monitoring network traffic on the control and data channels. The method is evaluated on a publicly available dataset.

  ▷ optical burst switching, quality of service, machine learning. classification

- ***Indoor Localization Using Bluetooth-LE Beacons***. Sudarshan S. Chawathe. In Proceedings of the 9th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2018), New York, NY, November 2018. Columbia University.

  > Persons and devices in indoor environments, such as office buildings, may determine their location using Bluetooth LE beacons, such as iBeacons. Some number of these beacons are distributed over the environment of interest and their identifiers and locations are broadcast widely. The vector of received signal strengths from all these beacons may be intuitively expected to correlate well with location in the physical environment. However, the complexities of Bluetooth signal propagation in environments with obstructions and channels (walls, furniture, ducts, etc.) make it difficult to compute locations in this manner from only the signal values and known locations of beacons. Instead, a data-driven approach that uses a training set composed of observed signal strength vectors at known locations is more effective. This paper studies such methods using a publicly available dataset obtained by collecting training data in an academic building.

  ▷ indoor localization, beacons, Bluetooth-LE, iBeacons, machine learning, data-driven methods

- ***Classifying Self-Care Activities of Children and Youths with Disabilities***. Sudarshan S. Chawathe. In Proceedings of the 9th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2018), New York, NY, November 2018. Columbia University.

  > The classification of functioning and disabilities in children and youths is an important task that informs healthcare. The ICF-CY (International Classification of Functioning, Disability, and Health in Children and Youth) provides a standard framework for such classification. Occupational therapists use the ICF-CY in conjunction with observations of the routine activities performed by a child (such as eating, toileting, washing) to determine a suitable diagnostic group. This paper presents a method for assisting occupational therapists and others in this task using machine learning. The method is studied experimentally using a publicly available dataset of self-care activities.

  ▷ ICF-CY, self-care activities, physical and motor disability, classification, medical informatics

- ***HDFJavaIO: A Java library for reading and writing Octave HDF files***. Mark E. Royer and Sudarshan S. Chawathe. In Proceedings of the 9th IEEE Annual Ubiquitous Computing, Electronics, and Mobile Communication Conference (IEEE UEMCON 2018), New York, NY, November 2018. Columbia University.

  > Scientific Java programs often need to interact with specialized programming environments, such as Octave and Matlab, that focus on numerical computations. This paper presents the HDFJavaIO library that allows Java programs to interact with Octave using Hierarchical Data Format 5 (HDF5) files, which are commonly used in the scientific community for working with large data sets. Because features of HDF5 files include almost all of the features of NetCDF, this library and method may also be used to create data files that can be used with NCL scripts and other applications that use these large-data formats without the need for further modifications by Java application developers. This paper presents the relevant details of the Octave HDF5 file format and the Java techniques used to build the data interchange library. It also presents the results of an experimental analysis of the library's performance and its comparison to existing approaches.

  ▷ Java, Octave, HDF5, NetCDF, Hierarchical Data Format, data interchange

- ***Recognizing Activities of Daily Living Using Binary Sensors***. Sudarshan S. Chawathe. In Proceedings of the IEEE International Conference on Universal Village (IEEE UV 2018), Cambridge, MA, October 2018. MIT.

Activities of Daily Living (ADLs), or a person's routine activities of self-care, are important factors influencing the feasibility of home health care or aging in place for many individuals. Automated, sensor-based recognition of such activities affords home stay, greater independence and privacy, and improved quality of life to individuals who would require stay in a supervised or medical facility. This paper describes a data-driven framework for the design and deployment of such an automated system for activity recognition using simple, unobtrusive, and privacy-friendly binary sensors. It presents the results of an experimental study, with both numerical and qualitative observations, of this framework on a publicly available real dataset.

- ***Analysis of Sparse Roadway Trajectories***. Sudarshan S. Chawathe. In Proceedings of the IEEE International Conference on Universal Village (IEEE UV 2018), Cambridge, MA, October 2018. MIT.

  Recent technological advances enable the gathering of extensive data on vehicular trajectories of large numbers of travelers at an unprecedented level of detail. Such trajectory datasets provide a wealth of information for purposes such as urban planning, carpool formation, and public-transportation design. This paper describes methods for analyzing and visualizing such data with an emphasis on sparse-traffic environments. It outlines the needs of applications in this domain and presents methods for clustering trajectories and for visualizing the results. The methods are evaluated by an experimental study on a publicly available dataset from real travelers.

- ***Monitoring Blockchains with Self-Organizing Maps***. Sudarshan S. Chawathe. In Proceedings of the 2018 International Workshop on Privacy, Security and Trust in Computational Intelligence (PSTCI 2018), New York, NY, August 2018. IEEE TrustCom-2018.

  Blockchains such as those used by the Bitcoin and Ethereum cryptocurrencies provide a global, observable record of all transactions and associated data. Analyzing blockchain data is useful for tasks such as detecting fraudulent activities, studying the use and growth of the system, and understanding its levels of anonymity and traceability. Such analysis is challenging due to the high volume and rapidly changing characteristics of popular blockchains. In particular, online (soft real-time) analysis of blockchains requires methods that adapt organically to changes in the data. This paper describes such a method based on self-organizing maps and reports on experiments using the Bitcoin blockchain data.

- ***Improving Email Security with Fuzzy Rules***. Sudarshan S. Chawathe. In Proceedings of the 2018 International Workshop on Privacy, Security and Trust in Computational Intelligence (PSTCI 2018), New York, NY, August 2018. IEEE TrustCom-2018.

  Phishing and other malicious email messages are increasingly serious security threats. An important tool for countering such email threats is the automated or semiautomated detection of malicious email. This paper reports work on using fuzzy rules to classify email for such purposes. The effectiveness of a fuzzy rule-based classifier is studied experimentally on a real dataset and compared with results for other classifiers, including those based on crisp rules and decision trees. The human-readability and editability of the classifiers produced by these methods is also studied.

- ***A Low-Overhead Scalable Data-Collection Service***. Sudarshan S. Chawathe. In Proceedings of the Borns Symposium, May 2018.

  We study the large-scale soft-realtime distributed collection, analysis, and reporting of data, emphasizing low-cost, low-overhead solutions that scale gracefully as usage varies over several orders of magnitude.

- ***A Tiny Java Library for Maintaining Model Provenance***. Mark E. Royer and Sudarshan S. Chawathe. In Proceedings of the Borns Symposium, May 2018.

We present a lightweight Java library that simplifies maintenance of the provenance of software object models. The implementation is based on annotations that are interpreted by an extended class loader to inject the Java bytecode to enable model maintenance.

- ***A New Approach for Ultra-High-Resolution Ice Core Data Processing***. Heather Clifford, Nicole Spaulding, Mark Royer, Sharon Sneed, Elena Korotkikh, Michael Handley, Andrei Kurbatov, Sudarshan Chawathe, Pascal Bohleber, Michael McCormick, Alexander More, Christohper Loveluck, and Paul Mayewski. In Geophysical Research Abstracts. EGU General Assembly, volume 20, Vienna, Austria, April 2018.

  Ice core archives provide the most direct and detailed evidence of past climate and atmospheric conditions. How- ever, the resolution of traditional ice core sampling methods limits the scope of information that can be extracted from the ice regarding meteorological events (e.g., dust storms, volcanic eruptions, anthropogenic emissions) that are captured at interannual to sub-annual scales. Using laser ablation inductively coupled mass spectrometry (LA- ICP-MS), a novel ultra-high-resolution multi-element sampling method for ice cores, we recovered the highest- resolution continuous glacio-chemical record yet from an ice core, measuring close to 5 million samples from 40 meters of core. This unique record was compiled using samples from the 2013 Colle Gnifetti ice core, located in the Swiss-Italian Alps. Here we present the first results from a new approach to high-resolution ice core data analysis through a new array of statistical tools, data processing algorithms and statistical machine learning tools adapted for ice core data sets. Our new data processing framework is designed to detect, extract and synthesize environmental signals from ultra-high-resolution glaciochemical time series in concert with more traditional ice core sampling data to further refine paleoenvironmental signals. The authors gratefully acknowledge the Climate Change Institute at the University of Maine, funding from grant AC3862 of the Arcadia Fund and NSF grant PLR-1443306.

- ***Java unit annotations for units-of-measurement error prevention,***. Mark E. Royer and Sudarshan S. Chawathe. In Proceedings of the 8th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2018), pages 858–864, Las Vegas, Nevada, January 2018.

  This project is a Java library for representing measurement units that provides easier avoidance and detection of a significant source of errors in scientific code. The technique uses the Java virtual-machine's class-loading extensions and annotations with run-time retention policies to enforce units conformance and conversion at run time. Analysis of the Java bytecode is performed at run time (or possibly compile time) to check conformance and conversion of unit-annotated types.

- ***Lexical Text Segmentation Using Dictionaries***. Sudarshan S. Chawathe. In Proceedings of the 8th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2018), pages 56–62, Las Vegas, Nevada, January 2018.

  Text segmentation refers to the task of partitioning text into disjoint segments based on some matching and optimization criteria. Examples include partitioning text into words, graphemes, and phonemes. The problem is especially challenging when the language does not require spaces, punctuation, or other simple separators; when segments may be combined in nontrivial ways; and in the presence of errors in transcription or recognition. This paper focuses on a purely lexical method of segmentation: Text is segmented using only a dictionary of known words along with a compatible cost function. No grammatical or other higher-level knowledge is used. The method uses efficient algorithms for multiple-string matching, such as the classic Aho-Corasick algorithm, to yield significant improvements in running time when compared with a simpler dynamic programming algorithm. An experimental study compares the running times of the dictionary-based and dynamic programming algorithms.

- ***Compact Representations of Character-Sets***. Sudarshan S. Chawathe. In Proceedings of the 8th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2018), pages 49–55, Las Vegas, Nevada, January 2018.

  Text segmentation refers to the task of partitioning text into disjoint segments based on some matching and optimization criteria. Examples include partitioning text into words, graphemes, and phonemes. The problem is especially challenging when the language does not require spaces, punctuation, or other simple separators; when segments may be combined in nontrivial ways; and in the presence of errors in transcription or recognition. This paper focuses on a purely lexical method of segmentation: Text is segmented using only a dictionary of known words along with a compatible cost function. No grammatical or other higher-level knowledge is used. The method uses efficient algorithms for multiple-string matching, such as the classic Aho-Corasick algorithm, to yield significant improvements in running time when compared with a simpler dynamic programming algorithm. An experimental study compares the running times of the dictionary-based and dynamic programming algorithms.

- ***Annotating Unit Functions in the Climate Data Workbench***. Mark Royer, Sudarshan S. Chawathe, Andrei V. Kurbatov, and Paul A. Mayewski. In Proceedings of the Borns Symposium, April 2017.

  We describe a method for representing measurement units for the Climate Data Workbench, providing easier avoidance and detection of a significant source of errors in scientific code. Our method uses the Java virtual-machine's class-loading extensions, and annotations with runtime retention policies, to enforce units conformance and conversion at runtime.

- ***Functional-programming with Generic Mapping Tools (fGMT)***. Sudarshan S. Chawathe. In Proceedings of the Borns Symposium, April 2017.

  We describe fGMT, a functional interface to the very popular GMT collection of mapping and plotting tools. Our implementation uses scsh Scheme and is designed to permit incremental building of higher-level interfaces that incorporate domain-specific knowledge.

- ***The P301 Web API***. Mark Royer, Sudarshan S. Chawathe, Andrei V. Kurbatov, and Paul A. Mayewski. In Proceedings of the Borns Symposium, April 2016.

  The P301 Web API is a RESTful interface that allows P301 users to share data that have been uploaded to the P301 system. The system supports accessing data in JavaScript Object Notation (JSON) and Extensible Markup Language (XML) formats, which helps to facilitate the development of Web-based applications. A variety of queries for accessing the data in the system allows for flexibility in client system designs.

- ***Toward a Domain-Specific Language for Patterns in Ice-Core Data***. Sudarshan S. Chawathe. In Proceedings of the Borns Symposium, April 2016.

  We describe a language for expressing simple patterns in time series data derived from ice-cores and similar sources. Such patterns use simpler features mapped to tokens by an earlier phase of analysis. In turn, they allow more complex features to be expressed and analyzed.

- ***Interactive Exploration of Time Lines from Ice Core Data Sets***. Sudarshan S. Chawathe. In Proceedings of the Borns Symposium, April 2015.

  Time lines are derived from ice core data typically by counting layers or peaks in sequences of measured values. This work (in progress) explores the extent to which automation and interactive exploration may assist this task.

- ***Deploying a Multi-Interface RESTful Application in the Cloud***. Erik Albert and Sudarshan S. Chawathe. In Proceedings of the 6th International Conference on Data Management in Cloud, Grid and P2P Systems (Globe-13), Prague, Czech Republic, August 2013.

This paper describes the design, implementation, and deployment of an application server whose primary infrastructure is an elastic cloud of servers. The design is based on the Representational State Transfer (REST) style, which provides significant benefits in a cloud environment. The paper also addresses implementation issues within a specific cloud service and highlights key decisions and their effect on scalability and cost. Finally, it describes our experiences in deploying a widely used platform with both Web and mobile client interfaces and its ability to cope with load spikes while maintaining a low quiescent cost.

- ***Fast Fingerprinting for File-System Forensics***. Sudarshan S. Chawathe. In In Proceedings of the 12th annual IEEE Conference on Technologies for Homeland Security (HST), pages 591–596, Waltham, Massachusetts, November 2012.

  An important method used to speed up forensic file-system analysis is white-listing of files: Well-known files are detected using signatures (message digests) or similar methods, and omitted from further analysis initially, in order to better focus the initial analysis on files likely to be more important. Typical examples of such well-known files include files used by operating systems, popular applications, and software libraries. This paper presents methods for improving the effectiveness and efficiency of such signature-based white-listing during file-system forensics. One concern for effectiveness is the resilience of the white-listing method to an adversary who has complete knowledge of the method and who may make small, inconsequential changes to a large number of well-known files on a target file-system in order to overload the analysis and thereby practically defeat it. Another concern is the ability to detect near-matches in addition to exact matches. Efficiency refers to primarily the rate at which a target file system may be processed during analysis; preparation-time, or indexing, efficiency is a lesser concern as that computation may be performed during non-critical times. Our work builds on techniques such as locality-sensitive hashing to yield an effective filter for further analysis tools.

- ***Managing Diverse Data Sets Using P301***. Mark Royer, Sudarshan S. Chawathe, Andrei V. Kurbatov, and Paul A. Mayewski. In Proceedings of the 20th annual Harold W. Borns, Jr. Symposium, Orono, Maine, April 2012.

  The integration and analysis of data sets from diverse sources provides scientists with an opportunity to gain insights that are not apparent from the individual data sets or sources. For many sources, improving technology and other factors have resulted in a very rapid growth in both the volume and the diversity of data. This wealth of data has the potential for significant scientific breakthroughs. However, this potential is difficult to realize unless there is a systematic and effective method for managing this data. The methods used by researchers in the past typically do not scale up to current and anticipated levels of data volume and diversity. The P301 project addresses this problem with the goal of accelerating the data flow from data sources to research results. Below, we outline one aspect of this work: Managing the syntactic and semantic consistency of data using an interactive framework that eases the task of importing, cleaning, analyzing, and visualizing data, and of recording such data transformations and results using histories and certificates.

- ***Deploying a Highly Scalable Web Application in the Cloud***. Erik Albert and Sudarshan S. Chawathe. In Proceedings of the 20th annual Harold W. Borns, Jr. Symposium, Orono, Maine, April 2012.

  The 10Green Web application integrates air quality data from diverse sources and provides an intuitive interface that summarizes this information in a manner accessible to scientists and non-scientists alike. From a Computer Science perspective, this application presents interesting challenges in both the back end (e.g., data integration and analysis, maintainability) and the front end (e.g., Web-based visualization, interactive response times, and portability across very diverse client architectures). Here, we focus on scalability and outline the implementation aspects that allow the application to scale from a few hundred users to hundreds of thousands of concurrent users at low cost.

- ***A REST Framework for Dynamic Client Environments***. Erik Albert and Sudarshan S. Chawathe. In Erik Wilde and Cesare Pautasso, editors, REST: From Research to Practice, chapter 10. Springer, 1st edition, August 2011. ISBN 978-1-4419-8302-2.

  > The REST Framework for Dynamic Client Environments (RFDE) is a method for building RESTful Web applications that fully exploit the diverse and rich feature-sets of modern client environments while retaining functionality in the absence of these features. For instance, we describe how an application may use a modern JavaScript library to enhance interactivity and end-user experience while also maintaining usability when the library is unavailable to the client (perhaps due to incompatible software). These methods form a framework that we have developed as part of our work on a Web application for presenting large volumes of scientific datasets to nonspecialists.

- ***A low-cost scalable Web mapping service for climate data***. Erik Albert and Sudarshan S. Chawathe. In Proceedings of the 19th annual Harold W. Borns, Jr. Symposium, Orono, Maine, May 2011.

  > We describe the design and implementation of a method to serve hundreds of terabytes of image data (tiles) for a Web-based mapping service. The method allows the service to scale gracefully from a few dozen to thousands of concurrent connections. Map tiles are stored in implicit form in a database and the corresponding bit-mapped images are computed as needed using an efficient stored-procedure implementation. The implementation is also particularly well suited to deployment in the cloud computing environment.

- ***P301dx: Interactive data analysis***. Mark Royer, Sudarshan S. Chawathe, Andrei V. Kurbatov, and Paul A. Mayewski. In Proceedings of the 19th annual Harold W. Borns, Jr. Symposium, Orono, Maine, May 2011.

  > The Project 301 Data Explorer, P301dx, is a software workbench for climate-change data. It aids scientists with the tasks of storing, integrating, sharing, analyzing, and visualizing such data. The primary goal of Project 301 is improving the efficiency and effectiveness of the process of transforming raw data into easily interpretable scientific results.

- ***Information Systems for Passenger Guidance in Transit Systems***, Sudarshan S. Chawathe. Invited presentation at the Symposium on Engineering and Technologies for the Metro Bogota (Metrosimposio), May 2010

- ***Low-Latency Indoor Localization Using Bluetooth Beacons***. Sudarshan S. Chawathe. In Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems (ITSC), St. Louis, Missouri, October 2009

- ***Effective Whitelisting for Filesystem Forensics***. Sudarshan S. Chawathe. In Proceedings of the 7th IEEE Intelligence and Security Informatics Conference (ISI), Richardson, Texas, June 2009

- ***Beacon Placement for Indoor Localization using Bluetooth***. Sudarshan S. Chawathe. In Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems (ITSC), pages 980–985, Beijing, China, October 2008

- ***Using Dead Drops to Improve Data Dissemination in Very Sparse Equipped Traffic***. Sudarshan S. Chawathe. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), pages 962–967, Eindhoven, Netherlands, June 2008

- ***Protecting Transportation Infrastructure***. Daniel Zeng, Sudarshan S. Chawathe, Hua Huang, and Fei-Yue Wang. *IEEE Intelligent Systems*, 22(5):8–11, September/October 2007

- ***Marker-Based Localizing for Indoor Navigation***. Sudarshan S. Chawathe. In Proceedings of the 10th International IEEE Conference on Intelligent Transportation Systems (ITSC), pages 885–890, Seattle, Washington, October 2007

- ***Interimistic Data Dissemination***. Sudarshan S. Chawathe and Abheek Anand. *Information Systems and e-Business Management (ISeB)*, 5(3):229–253, June 2007

- ***Segment-Based Map Matching***. Sudarshan S. Chawathe. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), pages 1190–1197, Istanbul, Turkey, June 2007

- ***Organizing Hot-Spot Police Patrol Routes***. Sudarshan S. Chawathe. In Proceedings of the 5th IEEE Intelligence and Security Informatics Conference (ISI), pages 78–85, New Brunswick, New Jersey, May 2007

- ***Protecting Transportation Infrastructure***. Daniel Zeng, Sudarshan S. Chawathe, and Fei-Yue Wang. *IEEE Intelligent Transportation Systems Society Newsletter*, 2007. Republished as a selected paper from the IEEE Intelligent Systems

- ***Inter-Vehicle Data Dissemination in Sparse Equipped Traffic***. Sudarshan S. Chawathe. In Proceedings of the 9th International IEEE Conference on Intelligent Transportation Systems (ITSC), pages 273–280, Toronto, Canada, September 2006

- ***Strategic Web-Service Agreements***. Sudarshan S. Chawathe. In Proceedings of the 4th IEEE International Conference on Web Services (ICWS), pages 119–126, Chicago, Illinois, September 2006

- ***Tracking Changes in Healthcare Documents***. Sudarshan S. Chawathe. In Proceedings of the 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS), pages 137–142, Salt Lake City, Utah, June 2006

- ***Distributing the Cost of Securing a Transportation Infrastructure***. Sudarshan S. Chawathe. *IEEE Intelligent Transportation Systems Society Newsletter*, 8(2):17–21, June 2006. Republished as one of two selected papers from the ISI-2006 conference

- ***Distributing the Cost of Securing a Transportation Infrastructure***. Sudarshan S. Chawathe. In Proceedings of the 4th IEEE Intelligence and Security Informatics Conference (ISI), pages 596–601, San Diego, California, May 2006

- ***Fair Policies for Travel on Neighborhood Streets***. Sudarshan S. Chawathe. In Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems (ITSC), pages 1027–1032, Vienna, Austria, September 2005

- ***Book Review: Perspectives on Intelligent Transportation Systems***. Sudarshan S. Chawathe. *IEEE Intelligent Transportation Systems Society Newsletter*, 7(3):14–15, September 2005

- ***Differencing Data Streams***. Sudarshan S. Chawathe. In Proceedings of the 9th International Database Engineering and Applications Symposium (IDEAS), pages 273–284, Montreal, Canada, July 2005

- ***XSQ: A Streaming XPath Engine***. Feng Peng and Sudarshan S. Chawathe. *ACM Transactions on Database Systems (TODS)*, 30(2):577–623, June 2005

- ***Data Management in Interimistic Environments***. Abheek Anand and Sudarshan S. Chawathe. In Proceedings of the Third Workshop on E-Business (WeB), Washington, D.C., December 2004

- ***Real-Time Traffic-Data Analysis***. Sudarshan S. Chawathe. In Proceedings of the 7th IEEE International Conference on Intelligent Transportation Systems (ITSC), pages 112–117, Washington, D.C., October 2004

- ***Control of Personal Location Data***. Sudarshan S. Chawathe. In Proceedings of the Location Privacy Workshop, Schoodic Peninsula, Acadia National Park, Maine, August 2004

- ***Managing RFID Data***. Sudarshan S. Chawathe, Venkat Krishnamurthy, Sridhar Ramachandran, and Sanjay Sarma. In Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), pages 1189–1195, Toronto, Canada, August 2004

- ***Privacy-Preserving Inter-Database Operations***. Gang Liang and Sudarshan S. Chawathe. In Proceedings of the Symposium on Intelligence and Security Informatics (ISI), volume 3073 of *Lecture Notes in Computer Science (LNCS)*, pages 66–82, Tucson, Arizona, June 2004
- ***Skipping Streams with XHints***. Akhil Gupta and Sudarshan S. Chawathe. Technical Report CS-TR-4566, Computer Science Department, University of Maryland, College Park, Maryland, February 2004
- ***Privacy-Preserving Inter-Database Operations***. Gang Liang and Sudarshan S. Chawathe. Technical Report CS-TR-4564 (UMIACS-TR-2004-09), University of Maryland, College Park, February 2004
- ***Efficient Peer-to-Peer Namespace Searches***. Vijay Gopalakrishnan, Bobby Bhattacharjee, Sudarshan S. Chawathe, and Pete Keleher. Technical Report CS-TR-4568, University of Maryland, College Park, Maryland, February 2004
- ***Cooperative Data Dissemination in a Serverless Environment***. Abheek Anand and Sudarshan S. Chawathe. Technical Report CS-TR-4562, Computer Science Department, University of Maryland, College Park, Maryland, January 2004
- ***XPaSS: A Multiple-Query Streaming XPath Query Engine***. Feng Peng and Sudarshan S. Chawathe. Technical Report CS-TR-4565, Computer Science Department, University of Maryland, College Park, Maryland, January 2004
- ***Streaming XPath Subquery Evaluation***. Feng Peng and Sudarshan S. Chawathe. Technical Report CS-TR-4560, Computer Science Department, University of Maryland, College Park, Maryland, January 2004
- ***Optimal Buffering for Streaming XPath Evaluation***. Feng Peng and Sudarshan S. Chawathe. Technical Report CS-TR-4561, Computer Science Department, University of Maryland, College Park, Maryland, January 2004
- ***Semistructured Data in Relational Databases***. Sudarshan S. Chawathe, chapter 25, pages 1–19. Practical Handbook of Internet Computing. CRC Press, 2004
- ***XSQ: A Streaming XPath Engine***. Feng Peng and Sudarshan S. Chawathe. Technical Report CS-TR-4493, Department of Computer Science, University of Maryland, May 2003
- ***XPath Queries on Streaming Data***. Feng Peng and Sudarshan S. Chawathe. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), pages 431–442, San Diego, California, June 2003
- ***Tracking Hidden Groups Using Communications***. Sudarshan S. Chawathe. In Proceedings of the NSF/NIJ Symposium on Intelligence and Security Informatics (ISI), volume 2665 of *Lecture Notes in Computer Science (LNCS)*, pages 195–208, Tucson, Arizona, June 2003
- ***XSQ: Streaming XPath Queries***. Feng Peng and Sudarshan S. Chawathe. In Proceedings of the 19th International Conference on Data Engineering (ICDE), pages 780–782, Bangalore, India, March 2003. Demonstration description
- ***Efficient Peer-to-Peer Searches Using Result-Caching***. Bobby Bhattacharjee, Sudarshan S. Chawathe, Vijay Gopalkrishnan, Pete Keleher, and Bujor Silaghi. In Proceedings of the International Workshop on Peer-to-Peer Systems (IPTPS), pages 225–236, Berkeley, California, February 2003
- ***Managing Historical XML Data***. Sudarshan S. Chawathe, volume 57 of *Advances in Computers*, chapter 3, pages 109–169. Elsevier Science, 2003
- ***SEuS: Structure Extraction using Summaries***. Shayan Ghazizadeh and Sudarshan S. Chawathe. In Steffen Lange, Ken Satoh, and Carl H. Smith, editors, Proceedings of the 5th International Conference on Discovery Science, volume 2534 of *Lecture Notes in Computer Science (LNCS)*, pages 71–85, Lubeck, Germany, November 2002

- ***Tracking Moving Clutches in Streaming Graphs***. Sudarshan S. Chawathe. Technical Report CS-TR-4376, Computer Science Department, University of Maryland, College Park, Maryland, October 2002

- ***XSQ: Streaming XPath Queries***. Feng Peng and Sudarshan S. Chawathe. Technical Report CS-TR-4401 (UMIACS-TR-2002-81), Computer Science Department, University of Maryland, College Park, Maryland, September 2002

- ***Discovering Freuqent Structures using Summaries***. Shayan Ghazizadeh and Sudarshan Chawathe. Technical report, University of Maryland, Computer Science Department, 2002

- ***Discovering Frequent Structures using Summaries***. Shayan Ghazizadeh and Sudarshan S. Chawathe. Technical Report CS-TR-4364, Computer Science Department, University of Maryland, College Park, Maryland, November 2001

- ***VQBD: Exploring Semistructured Data***. Sudarshan S. Chawathe, Thomas Baby, and Jihwang Yeo. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), page 603, Santa Barbara, California, May 2001. Demonstration description

- ***VQBD: Visualizing, Querying, and Browsing Semistructured Data***, Sudarshan S. Chawathe, Thomas Baby, and Jihwang Yeo, November 2000. Extended version of demonstration description. `http://cs.umaine.edu/~chaw/`

- ***Comparing Hierarchical Data in External Memory***. Sudarshan S. Chawathe. In Proceedings of the International Conference on Very Large Data Bases (VLDB), pages 90–101, Edinburgh, Scotland, September 1999

- ***Describing and Manipulating XML Data***. Sudarshan S. Chawathe. *Bulletin of the IEEE Technical Committee on Data Engineering*, 22(3):3–9, September 1999

- ***Managing Historical Semistructured Data***. Sudarshan S. Chawathe, Serge Abiteboul, and Jennifer Widom. *Theory and Practice of Object Systems*, 5(3):143–162, August 1999

- ***Managing Change in Heterogeneous Autonomous Databases***. Sudarshan S. Chawathe. PhD thesis, Stanford University, 1999

- ***Representing and Querying Changes in Semistructured Data***. Sudarshan S. Chawathe, Serge Abiteboul, and Jennifer Widom. In Proceedings of the International Conference on Data Engineering (ICDE), pages 4–13, Orlando, Florida, February 1998

- ***An Expressive Model for Comparing Tree-Structured Data***. Sudarshan S. Chawathe and Hector Garcia-Molina. Technical report, Stanford University Database Group, November 1997

- ***Representing and Querying Changes in Heterogeneous Semistructured Databases (Demonstration Description)***. S. Chawathe, V. Gossain, X. Liu, J Widom, and S. Abiteboul. Technical report, Stanford University Database Group, November 1997. Available at `http://www-db.stanford.edu`

- ***Meaningful Change Detection in Structured Data***. Sudarshan S. Chawathe and Hector Garcia-Molina. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), pages 26–37, Tuscon, Arizona, May 1997

- ***Representing and Querying Changes in Semistructured Data (Extended Version)***, S. Chawathe, S. Abiteboul, and J. Widom. Available at `http://www-db.stanford.edu`, 1997

- ***Meaningful Change Detection in Structured Data***, S. Chawathe and H. Garcia-Molina. Available at `http://www-db.stanford.edu/`, 1997. Extended version

- ***Representative Objects: Concise Representations of Semistructured, Hierarchial Data***. Svetlozar Nestorov, Jeffrey D. Ullman, Janet Wiener, and Sudarshan S. Chawathe. In Proceedings of the International Conference on Data Engineering (ICDE), pages 79–90, Birmingham, U.K., 1997

- ***Change Detection in Hierarchically Structured Information***. Sudarshan S. Chawathe, Anand Rajaraman, Hector Garcia-Molina, and Jennifer Widom. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), pages 493–504, Montréal, Québec, June 1996

- ***A standard textual interchange format for the Object Exchange Model (OEM)***. Roy Goldman, Sudarshan S. Chawathe, Arturo Crespo, and Jason McHugh. Technical report, Stanford University Database Group, June 1996

- ***A Toolkit for Constraint Management in Heterogeneous Information Systems***. Sudarshan S. Chawathe, Hector Garcia-Molina, and Jennifer Widom. In Proceedings of the International Conference on Data Engineering (ICDE), pages 56–65, New Orleans, Louisiana, 1996

- ***Change Detection in Hierarchically Structured Information***. S. Chawathe, A. Rajaraman, H. Garcia-Molina, and J. Widom. Technical report, Dept. of Computer Science, Stanford University, 1995. Available at `http://www-.stanford.edu`

- ***Change Detection in Hierarchically Structured Information***. S. Chawathe, A. Rajaraman, H. Garcia-Molina, and J. Widom. Technical report, Stanford University Database Group, 1995. Available at `http://www-db.stanford.edu`

- ***The Tsimmis Project: Integration of Heterogeneous Information Sources***. S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. In Proceedings of 100th Anniversary Meeting of the Information Processing Society of Japan, pages 7–18, Tokyo, Japan, October 1994

- ***The Tsimmis Project: Integration of Heterogeneous Information Sources***. Sudarshan S. Chawathe, Hector Garcia-Molina, Joachim Hammer, Kelly Ireland, Yannis Papakonstantinou, Jeffrey D. Ullman, and Jennifer Widom. In Proceedings of 100th Anniversary Meeting of the Information Processing Society of Japan, pages 7–18, Tokyo, Japan, October 1994

- ***Flexible Constraint Management for Autonomous Distributed Databases.*** Sudarshan S. Chawathe, Hector Garcia-Molina, and Jennifer Widom. *Data Engineering Bulletin*, 17(2):23–27, June 1994

- ***Constraint Management in Loosely Coupled Distributed Databases***. Sudarshan S. Chawathe, Hector Garcia-Molina, and Jennifer Widom. Technical report, Computer Science Department, Stanford University, 1994. Available at `http://www-db.stanford.edu`

- ***On Index Selection Schemes for Nested Object Hierarchies***. Sudarshan S. Chawathe, Ming-Syan Chen, and Philip S. Yu. In Proceedings of the International Conference on Very Large Data Bases (VLDB), pages 331–341, Santiago de Chile, 1994

- ***Constraint Management in Loosely Coupled Distributed Databases***. Sudarshan S. Chawathe, Hector Garcia-Molina, and Jennifer Widom. Technical report, Computer Science Department, Stanford University, 1993. Available at `http://www-db.stanford.edu`