AFRL-RI-RS-TR-2018-048

MACHINE LEARNING ALGORITHMS FOR STATISTICAL PATTERNS IN LARGE DATA SETS

CARNEGIE MELLON UNIVERSITY

FEBRUARY 2018

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2018-048 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
JOHN SPINA
Work Unit Manager

/ S /
JON S. JONES
Technical Advisor, Information Intelligence
Systems and Analysis Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
FEBRUARY 2018	FINAL TECHNICAL REPORT	JAN 2012 – SEP 2017

4. TITLE AND SUBTITLE

MACHINE LEARNING ALGORITHMS FOR STATISTICAL PATTERNS IN LARGE DATA SETS

5a. CONTRACT NUMBER
FA8750-12-2-0324

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
602720E

6. AUTHOR(S)

Arthur Dubrawski

5d. PROJECT NUMBER
XDAT

5e. TASK NUMBER
A0

5f. WORK UNIT NUMBER
05

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

Carnegie Mellon University
500 Forbes Avenue
Pittsburgh PA 15213

8. PERFORMING ORGANIZATION REPORT NUMBER

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

Air Force Research Laboratory/RIEA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2018-048

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Modern data analysis operations are continuously flooded with streams of noisy, incomplete, and sometimes intentionally misleading data. Traditional analysis methods cannot scale to handle these issues. We developed a battery of new, efficient, parallel, statistical machine learning algorithms to push the boundaries of machine learning capabilities under these circumstances. We have made much of our mature algorithms available as open source tools and published in peer-reviewed academic journals and conferences. The algorithms cover a wide range of learning applications, but all rest on strong statistical foundations and in that sense that they all speak the same language. We have provided theoretical guarantees and proofs were possible and demonstrated the value of our algorithms on many interesting problems.

15. SUBJECT TERMS
Text Analysis, Text Exploitation, Situation Awareness of Text, Document Processing, Document Ingestion, Full Text Search, Information Extraction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**JOHN SPINA**
U	U	U	SAR	21	19b. TELEPHONE NUMBER (*Include area code*) **(315) 330-4032**

CONTENTS

LIST OF FIGURES

1 EXECUTIVE SUMMARY

Modern data analysis operations are continuously flooded with enormous streams of heterogeneous data and heterogeneous tasks. The data are notoriously noisy, incomplete, and sometimes intentionally misleading. Traditional analysis methods can not scale to handle the flood or the probabilistic data properties or both. We developed a battery of new, efficient, parallel, statistical machine learning algorithms to push the boundaries of machine learning capabilities under these circumstances. We have made much of our mature algorithms available as open source tools and published in peer-reviewed academic journals and conferences. The algorithms cover a wide range of learning applications, but all rest on strong statistical foundations and in that sense that they all speak the same language. We have provided theoretical guarantees and proofs were possible and demonstrated the value of our algorithms on many interesting problems. The methods described here have the ability to:

- Classify groups of data as indicating a particular phenomenon or class.
- Recognize change points, boundaries, and emerging trends.
- Perform fast learning and inference in a variety of graphical, Bayesian, and sparse models.
- Process streaming data and operate on time series.
- Make intelligent decisions about additional data to collect and analyses to perform.

2 INTRODUCTION AND KEY ACCOMPLISHMENTS

Modern data analysis operations are continuously flooded with enormous streams of heterogeneous data and heterogeneous tasks. We propose to develop and deliver machine learning algorithms and software to support such an operation. Here we list the primary issues caused by this heterogeneous flood and our approach to them:

Noisy Incomplete Data. Real data sets suffer from both missing and incorrect data. Incorrect data is a problem both due to noise in collection, intentional misinformation by adversaries, and the time-changing nature of data collection targets. We developed transfer learning methodologies to exploit data rich areas that are related to data starved problems. We have provided theoretical bounds on the associated approaches and explored a novel concept of active-transfer learning pairing two machine learning paradigms.

Large Data. Traditional algorithms and single processor “all-in-memory” executions are insufficient for large data sets. We created parallel and distributed frameworks for significantly speeding up common machine learning models and allowing the development of very large models and use of very large data sets. Our tools accelerate learning, make efficient use of limited communication channels, and do so on modest computational clusters.

Complex Data objects. We developed methods that can operate on groups of data as opposed to the traditional single-observation paradigm. In our paradigm the objects of learning are statistical distributions which are never observed directly, only sampled. Yet, we are able to characterize and

compare their attributes using nonparametric methods. Additionally, we have combined this concept with deep learning creating a deep distribution embedding model. These concepts push the boundaries of typical machine learning opening up the possibilities for how machine learning problems are constructed and defined.

3 METHODS

The majority of the algorithms below rely on non-parametric methods, using kernels to quantify the similarity between observations or abstract objects. In many cases they work directly in the kernel feature space using mean map embeddings and related techniques.

Internal evaluations were done during algorithm and software development using the wide variety of data sets accumulated and available in our labs at Carnegie Mellon and provided through the XDATA program. The local benchmark data sets include Wikipedia data, twitter data crawled from the web, numerous image data sets, telescope data from the Sloane Digital Sky Survey, and medical clinical data. These data sets have from millions up to a billion records and tested both computational scalability and modeling and prediction accuracy of the novel methods we developed.

4 ASSUMPTIONS AND PROCEDURES

By virtue of using non-parametric kernel based methods, much of our work does not make any assumptions on the underlying data generating distributions one wishes to learn. However, these methods typically require a bit more data than parametric algorithms to learn stable models. Additionally, our kernel and similarity based methods are best suited for numeric data. Categorical data present a difficulty in that one must come up with a suitable kernel and/or feature map transformation, and even though that is not unsurmountable, it was not our focus in this project.

5 RESULTS AND DISCUSSION

Below we describe a selection of project activities, results, and findings. In general, we developed many new algorithms and methods for addressing challenges in real-world machine learning problems and extending the utility of existing approaches. Section 7 enumerates publications funded or partially funded by this project.

5.1 Learning on Distributions

We developed methods to do machine learning on bags of samples as the fundamental unit of observation. The approach is to estimate the underlying probability distributions, computing a distance function and/or kernel on such pairs. Upon estimating the kernel values, common machine learning methods such as SVM or Gaussian Processes can be used to produce models for classification, regression, clustering, etc. The primary challenge in kernel based machine learning methods for distributions is to develop efficient estimates of kernel functions from samples.

We have explored estimators for linear and Radial Basis Function kernels and symmetric Kullback-Leibler divergence (SKL), Jensen-Shannon (JS), Squared Hellinger (H^2), and Total Variation (TV)

information-theoretic distances. We've also explored estimators for Maximum mean discrepancy (MMD) and Earth mover's distance (EMD). Some of these efforts were published in [21].

We demonstrated the effectiveness of these approaches on several machine learning tasks. [19] demonstrated distribution regression on estimating the number of mixture components in data generated from Gaussian mixtures and distribution classification on scene classification. [21] evaluated these approaches on dark matter halo mass prediction. In all of these applications raw data are complex and potentially irregular; however they can be treated as bags of samples admitting our approaches. Performance on all tasks is competitive with other cutting edge approaches.

In the course of this work, we proved several useful properties of kernel methods and mean maps [20]. We improved the uniform error bound of random Fourier Features, and developed a novel understandings of the embedding's variance, approximation error, and use in some machine learning methods. We also point out that surprisingly, of the two main variants of those features, the more widely used is strictly higher variance for the Gaussian kernel and has worse bounds [20].

5.1.1 Deep Mean Map Embeddings. The use of distributions and high-level features from deep architecture has become commonplace in modern computer vision. Both of these methodologies have separately achieved a great deal of success in many computer vision tasks. However, there has been little work attempting to leverage the power of these to methodologies jointly. To this end, we developed the idea of the Deep Mean Maps (DMMs) framework, a novel family of methods to non-parametrically represent distributions of features in convolutional neural network models. DMMs are able to both classify images using the distribution of top-level features, and to tune the top-level features for performing this task. In [21] we showed how to implement DMMs using a special mean map layer composed of typical CNN operations, making both forward and backward propagation simple. We illustrated the efficacy of DMMs at analyzing distributional patterns in image data in a synthetic data experiment. We also showed that we extending existing deep architectures with DMMs improve the performance of existing CNNs on several challenging real-world datasets.

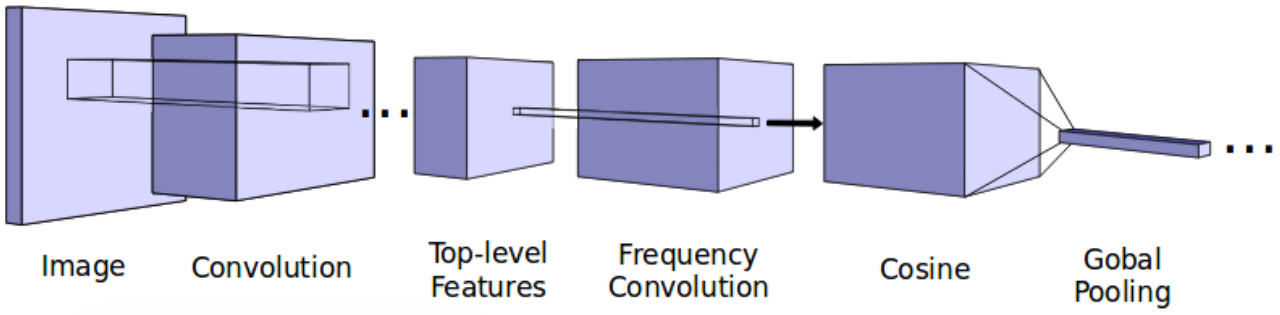

Figure 1: Sketch of deep mean map embedding architecture.

Suppose that the top convolution layer of a CNN (after any sub-sampling) produces a tensor $C_i \in R^{m \times h \times w}$ for the i^{th} input image fed forward through the network, where here m is the number of convolutional filters, and h, w are spatial dimensions of super pixels (see Figure 1). Then the layer's output can be viewed as a bag of vectors, with a different bag for each input. We defined a mean map

layer (MME layer) as a layer in the deep architecture that computes the mean map of these bags. This mean map layer can be represented in terms of standard CNN mechanisms, performing scaled 1x1 convolutions using random offsets combined with a cosine layer and global average pooling, depicted graphically in Figure 1. The scale parameter of the associated kernel becomes a free parameter in the network and can be learned using the usual back-propagation algorithm. After being fed through the MME layer (with the global-average pooling of random features), the i^{th} image in a batch will have the mean map embedding of its top-level convolution features computed. These vectors can then be fed through the network in a fully connected fashion to perform learning with respect to some loss.

We tested the approach on three image classification datasets. The deep mean map architecture nearly always outperformed baseline deep methods. This work demonstrates that the use of mean embeddings with random features allow the mean map layer to non-parametrically represent distributions of top-level features whilst still scaling to large image datasets. Also, inner products on the mean embeddings through the network are interpretable as RKHS inner products on the distributional embeddings, allowing one to build a strong theoretical foundation. Furthermore, we showed that the mean map layer may be implemented using typical CNN operations, making both forward and backward propagation simple to do with DMMs. Moreover, we illustrated the aptitude of the mean map layer at learning distributions of visual features for discrimination in a synthetic data experiment. We saw that even with very few instances the mean map layer allowed a network to quickly learn to distinguish visual distributional patterns in a sample efficient manner. Lastly, we showed that DMMs may be used to extend several existing state-of-the-art deep-architectures and improve their performance on various challenging real-world datasets. Indeed, the mean map layer prove flexible and capable of extending networks in a variety of ways; due to a propensity to generalize well, it is simple to choose an extension method with a straight-forward validation set approach. Thus, it is clear that the DMM framework can successfully build on the myriad of state-of-the-art deep architectures available.

5.2 Transfer Learning

Here we present work published in [24]. Transfer learning algorithms are used when one has sufficient training data for one supervised learning task (the source task) but only very limited training data for a second task (the target task) that is similar but not identical to the first. These algorithms use varying assumptions about the similarity between the tasks to carry information from the source to the target task. Common assumptions are that only certain specific marginal or conditional distributions have changed while all else remains the same. Alternatively, if one has only the target task, but also has the ability to choose a limited amount of additional training data to collect, then active learning algorithms are used to make choices which will most improve performance on the target task. These algorithms may be combined into active transfer learning, but previous efforts have had to apply the two methods in sequence or use restrictive transfer assumptions. We developed [24] two transfer learning algorithms that allow changes in all marginal and conditional distributions but assume the changes are smooth in order

to achieve transfer between the tasks. We also developed an active learning algorithm for the second method that yields a combined active transfer learning algorithm [24].

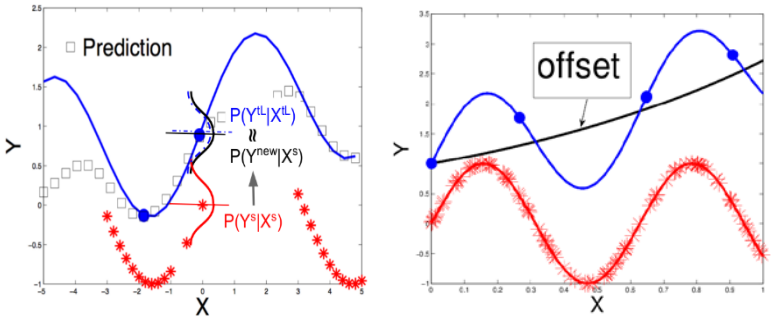

Figure 2: Depiction of two approaches to transfer learning: matching conditional distributions (left) and learning an offset function (right).

Our two approaches are to (i) match source and target domain kernel embedding of the conditional distributions (probability of label given feature vector) and (ii) learn an offset function that maps the source domain to the target, modeling the offset as a Gaussian process to exploit assumed smoothness. These approaches are depicted graphically in Figure 2. In [23], we demonstrated the algorithms on synthetic functions and a real-world task on estimating the yield of vineyards from images of the grapes, transferring models from one grape variety to another. Figure 3 depicts the results of the evaluation. Both of the proposed methods perform significantly better than baseline methods.

Previous work on covariate shift focuses on matching the marginal distributions on observations X across domains while assuming the conditional distribution $P(Y|X)$ stays the same. Relevant theory focusing on

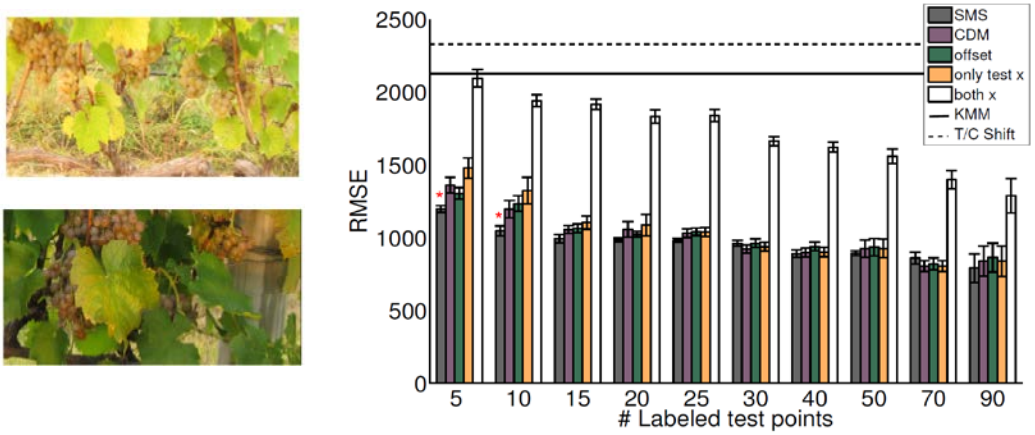

Figure 3: Example transfer learning problem: predict grape yield from images (left). The source domain is images of a different grape variety. The graph on the right shows regression performance for several approaches.

covariate shift has also been developed. Recent work on transfer learning under model shift deals with different conditional distributions $P(Y|X)$ across domains with a few target labels, while assuming the

changes are smooth. However, no analysis has been provided to say when these algorithms work. In [27], we analyzed transfer learning algorithms under the model shift assumption. Our analysis shows that when the conditional distribution changes, we are able to obtain a generalization error bound of that is $O(1/\lambda\sqrt{n})$ with respect to the labeled target sample size *n*, modified by the smoothness of the change (λ) across domains. Our analysis also sheds light on conditions when transfer learning works better than no-transfer learning (learning by labeled target data only). Furthermore, we extended the transfer learning algorithm from a single source to multiple sources.

5.3 Active Search

Active search is an important learning problem in which one uses a limited budget of label queries to discover as many members of a certain class as possible. Numerous real-world applications may be approached in this manner, including fraud detection, product recommendation, and drug discovery. Active search has model learning and exploration/exploitation features similar to those encountered in active learning and bandit problems, but algorithms for those problems do not fit active search. Previous work on the active search problem [3] showed that the optimal algorithm requires a look-ahead evaluation of expected utility that is exponential in the number of selections to be made and proposed a truncated look-ahead heuristic. Inspired by the success of myopic methods for active learning and bandit problems, we developed [25] a myopic method for active search on graphs. Our algorithm selects points by maximizing a score considering the potential impact of selecting a node, meant to emulate look-ahead while avoiding exponential search. The algorithm empirically outperforms popular approaches for active learning and bandit problems as well as truncated look-ahead of a few steps on real-world graphs.

The basic algorithm performs harmonic energy minimization (HEM) on a weighted undirected graph, however we augment the observed graph structure with pseudo-nodes (one for each graph node) to which observed labels or prior values are assigned. This augmentation guarantees a solution exists to the associated HEM problem. Look ahead is achieved by fast rank 1 matrix updates and a final score is computed as the product of the HEM node activation and the sum of unlabeled node activation upon one step look-ahead.

HEM is a common classifier for unlabeled nodes on undirected graphs (commonly referred to as label propagation), and is equivalent to the harmonic predictor on Gaussian random fields (GRFs). For active learning on GRFs, the commonly used *V*-optimality criterion queries nodes that reduce the L_2 (regression) loss. *V*-optimality satisfies a submodularity property showing that greedy reduction produces a (1-1/*e*) globally optimal solution. However, L_2 loss may not characterize the true nature of 0/1 loss in classification problems and thus may not be the best choice for active learning. We developed [12] a new criterion we call Σ-optimality, which queries the node that minimizes the sum of the elements in the predictive covariance. Σ-optimality directly optimizes the risk of the surveying problem, which is to determine the proportion of nodes belonging to one class. We extended submodularity guarantees from *V*-optimality to Σ-optimality using properties specific to GRFs. We further showed that GRFs satisfy the suppressor-free condition in addition to the conditional independence inherited from Markov random

fields. We tested Σ-optimality on coauthorship and citation graph problems. Our experiments showed that it outperforms V-optimality and other related methods on classification.

While the method for active search on graphs published in [25] performs well and can be quite flexible, initialization costs $O(n^3)$ operations and update costs $O(n^2)$ operations at each iteration. As a result the algorithm does not scale to large graphs. We developed two methods to scale the algorithm further. The first approach, published in [22], is to alter the similarity function to be a dot-product between feature vectors of data points, equivalent to having a linear kernel for the adjacency matrix. With this, we are able to scale tremendously: requiring only $O(nr + r^2)$ operations per iteration given r-dimensional features. We also show that a similarity function with low between class similarity is sufficient for active search on graphs to perform well [22]. The second approach is to use an approximate numerical solution for which the complexity is dependent only on edge density and damping parameter (walk termination probability), not on the number of observations. The approach is an adaptive implementation of the full algebraic multigrid solution for elliptical equations. We start at a trivially coarse grid approximation of the full adjacency graph, i.e. all unlabeled nodes are lumped together. When passing to the next level of refinement, we perform a heuristically constrained search of the nodes to find the one that will minimize the L_2 norm of the solution's residual. This node is then pulled from the aggregate group and promoted to singleton status. Because the solution decays along the graph according to the random walk termination probability, the number of nodes needed to achieve good approximation is dependent on that parameter and the edge density. In addition, this solution approach allows early termination when the solution vector has stabilized for the top few unlabeled nodes.

5.3.1 Generalized Queries. We developed light weight knowledge extraction capabilities using HEM over graphs, based on our work with active search. The approach computes the similarity of documents/observations as the weighted combination of the similarity of their constituent elements. We apply information extractors to obtain the constituent elements of a document

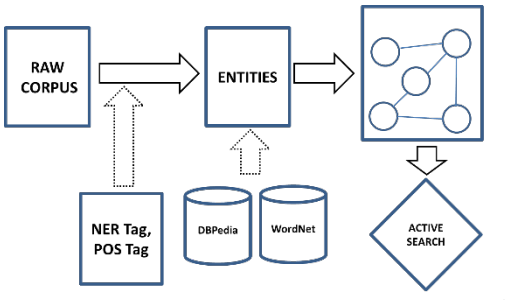

Figure 4: Flow diagram for augmenting document-mention graphs with additional information.

corpus. These include mentions of people, places, concepts, etc. We then create an incidence graph wherein the links indicate the presence of an element in a document. This is a bipartite graph. Additionally, we can add some external knowledge into the graph. One way to do that is to add a few layers of hypernyms from WordNet to the mentions extracted from the documents. Finally, we apply values to selected nodes which encode a query. Figure 3 shows a diagram of this process.

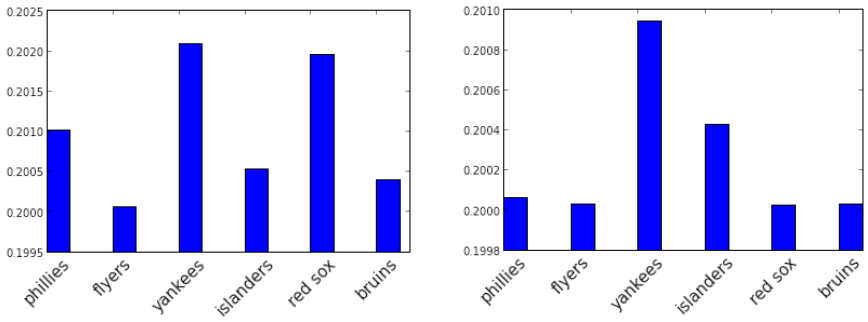

Figure 5: Node activations for sports team nodes under two query activations: "New York" (left) and "New York" + "Baseball" (right).

Anecdotal examples demonstrate successful knowledge extraction from a corpus. Building a graph using 20 news groups articles we were successfully able to answer questions such as: Which operating system are open source? Which operating system is most used by gamers? Which nations are the most democratic? Which hockey team is from New York? Such questions are answered by applying positive or negative values to node to encode the query. E.g. to identify baseball teams from New York one would place a positive value on the nodes for "Baseball" and "New York." Figure 4 shows node activation levels for sports teams with "New York" activated (left) and both "New York" and "Baseball" activated (right).

5.3.2 Region Sensing. The selection of data collection locations is a problem that has received significant research attention from classical design of experiments to various recent active learning algorithms. Typical objectives are to map an unknown function, optimize it, or find level sets in it. Each of these objectives focuses on an assessment of individual points. The introduction of set kernels has led to algorithms that instead consider labels assigned to sets of data points. In [9] we combined these two concepts and studied the problem of choosing data collection locations when the goal is to identify regions whose set of collected data would be labeled positively by a set classifier. We developed an algorithm for the case where the positive class is defined in terms of a region's average function value being above some threshold with high probability, a problem we call active area search. To this end, we model the latent function using a Gaussian process and use Bayesian quadrature to estimate its integral on predefined regions. Our method is the first which directly solves the active area search problem. In experiments it outperforms previous algorithms that were developed for other active search goals. Single-region search is also closely related to Σ-optimality described above.

We extended the ideas of active area search. In [10] we introduced the problem of active pointillistic pattern search (APPS) which generalizes active area search and seeks to discover regions of a domain exhibiting desired behavior with limited observations. Unusually, the patterns we consider are defined by large-scale properties of an underlying function that we can only observe at a limited number of points. Given a description of the desired patterns (in the form of a classifier taking functional inputs), we sequentially decide where to query function values to identify as many regions matching the pattern as possible, with high confidence. The expected reward of each unobserved point for a broad class of models in which the classifier salience can be written as a cumulative Gaussian of a linear function can be

computed analytically. We evaluated two instantiations of this class of algorithms on three difficult search problems: locating polluted regions in a lake via mobile sensors, forecasting winning electoral districts with minimal polling, and identifying vortices in a fluid flow simulation. Our algorithms compared very favorably against baseline methods.

5.3.3 Fast AS via Conjugate Sampling. In many applied methods of statistical inference, one wishes to sample from a multivariate normal Bayesian posterior distribution. We developed [11] a method for adapting conjugate gradient descent to efficiently sample from a high dimensional posterior. When the covariance matrix is sparse or structured this method dramatically reduces computational costs as compared to alternatives such as Thompson sampling.

5.4 User Specific Query Adaptation

We developed a first approach for modifying search algorithm behavior to suit user specific needs. In many cases search infrastructure is fixed, e.g. hash tables are precomputed and expensive to update. In these cases, one may still alter the behavior and performance of a search engine by modifying the query. We considered a region of interest (ROI) scenario in which a user is interested in searching for contents similar to a seed document, but only for a subset of seed. In this case, it is most reasonable to view the query not as a single seed document but rather a distribution over seed documents. I.e. we take the query to be the distribution of all documents that contain the region of interest. While this distribution may not known explicitly, in many cases it is possible to sample from it. In the case of images we sampled images, pasted the ROI on them, pushed these 'Frankenstein' images through the search engine's featurization and averaged the results. This average vector was the query vector. For many common choices of similarity function (e.g. inner product, cosine similarity, squared Mahalanobis distance, etc.) finding the document that maximizes expected similarity (over the implied query distribution) is equivalent to maximizing similarity with the average query vector. In these scenarios, one can submit the mean query vector to the search engine to provide user specific search capabilities. Components of this approach were developed for XDATA hackathons and made available at https://github.com/benbo/IQA_rest_api and https://github.com/benbo/IQA_ui**.**

5.5 Hyper(inter)active Learning

Machine learning requires substantial amounts of annotated data to train successful models. Yet, it is common place that annotated data is in short supply or completely unavailable. In some cases, human experts are capable of guiding, correcting, and/or adjudicating the understanding of intelligent problem solving agents. We aimed to reduce the perceived effort for training ML models by generalizing feedback or annotations beyond the traditional sample-label paradigm. The philosophy behind this concept is to increase the amount of useful information communicated to a computer in a form that is most palatable to human experts, while keeping the human effort manageable. Toward that end, we developed a nested PAC framework for classifiers of classifiers. This analysis suggested that it is possible to make significant gains in learning rates via reducing the size of the model hypothesis space in this fashion. We attempted to realize the idea, constructing a system that accepted feedback on clusters and clusters of clusters of

observations. Initial behavior was encouraging but the approach failed to mature. We shifted our approach and developed an algorithm that can accept label feedback as well as pair-wise comparisons [37]. This work suggested that adding the ability for a learning algorithm to consume comparisons can improve learning, but the improvement is only significant if direct labels are quite noisy. It has been shown to be effective in a range of benchmark cases and the work was submitted for publication at NIPS.

5.6 Explainable Analytics

We developed simple and fast analytics for explaining class distributions in data. Specifically, we developed an algorithm that searches for 2D axis-aligned boxes with high class purity. The resulting models are completely transparent and easy to understand. The low dimensional structures provide intuitive descriptors for new data sets, and domain experts can quite easily adjudicate any interesting findings. We also demonstrated that most observations in most data sets (between 20% and 80%, and ~60% on average, as measured over a large sample of real-world benchmark data sets) can be discriminated well with small sets of simple box models, while complex non-linear models are only needed for difficult and near-boundary cases. These methods were very effective in identifying gaps in training data for a gamma-ray source classification task [4], demonstrating as a side-effect the tendency of human data engineers to induce inadvertent biases that manifest in only a few dimensions of data at one time. This demonstrated the potential of our box-finding algorithm as a tool for quality management in data engineering.

5.7 Petuum – Framework for Large Scale Distributed Machine Learning

We developed Petuum [32], a general-purpose framework that systematically addresses data- and model-parallel challenges in large-scale ML, by leveraging several fundamental properties underlying ML programs that make them different from conventional operation-centric programs: error tolerance, dynamic structure, and non-uniform convergence; all stem from the optimization-centric nature shared in ML programs' mathematical definitions, and the iterative convergent behavior of their algorithmic solutions. These properties are taken advantage of, and distributed using bounded-latency network synchronization and dynamic load-balancing scheduling, which is efficient, programmable, and enjoys provable correctness guarantees. We have demonstrated how such a design in light of ML-first principles leads to significant performance improvements versus well-known implementations of several ML programs, allowing them to run in much less time and at considerably larger model sizes, on modestly-sized computer clusters.

Petuum has generated a spin-off company by the same name that "provides the next-generation omni-lingual (programmable in all languages), omni-mount (deployable on all hardware), and omni-source (compatible with all data formats) Operating System optimized for efficient and productive machine learning programming and computing. Built for the enterprise, Petuum's Operating System creates big data and artificial intelligence solutions that are high-quality, high-efficiency, high-availability, and low-maintenance" [www.petuum.com].

5.7.1 LDA and Efficient Sampling. When building large-scale machine learning (ML) programs, such as massive topics models or deep networks with up to trillions of parameters and training examples, one usually assumes that such massive tasks can only be attempted with industrial-sized clusters with thousands of nodes, which are out of reach for most practitioners or academic researchers. We considered this challenge in the context of topic modeling on web-scale corpora [38], and show that with a modest cluster of as few as 8 machines, we can train a topic model with 1 million topics and a 1-million-word vocabulary (for a total of 1 trillion parameters), on a document collection with 200 billion tokens— a scale not yet reported even with thousands of machines. Our major contributions include: (i) a new, highly efficient $O(1)$ Metropolis-Hastings sampling algorithm, whose running cost is (surprisingly) agnostic of model size, and empirically converges nearly an order of magnitude more quickly than current state-of-the-art Gibbs samplers; (ii) a structure-aware model-parallel scheme, which leverages dependencies within the topic model, yielding a sampling strategy that is frugal on machine memory and network communication; (iii) a differential data-structure for model storage, which uses separate data structures for high- and low-frequency words to allow extremely large models to fit in memory, while maintaining high inference speed; and (iv) a bounded asynchronous data-parallel scheme, which allows efficient distributed processing of massive data via a parameter server. Our distribution strategy is an instance of the model-and-data-parallel programming model underlying the Petuum framework for general distributed ML, and was implemented on top of the Petuum open-source system. We provide experimental evidence showing how this development puts massive models within reach on a small cluster while still enjoying proportional time cost reductions with increasing cluster size, in comparison with alternative options.

5.7.2 Poseidon. Deep learning models, which learn high-level feature representations from raw data, have become popular for machine learning and artificial intelligence tasks that involve images, audio, and other forms of complex data. A number of software "frameworks" have been developed to expedite the process of designing and training deep neural networks, such as Caffe, Torch, and Theano. Currently, these frameworks can harness multiple GPUs on the same machine, but are unable to use GPUs that are distributed across multiple machines; because even average-sized deep networks can take days to train on a single GPU when faced with 100s of GBs to TBs of data, distributed GPUs present a prime opportunity for scaling up deep learning. However, the limited inter-machine bandwidth available on commodity Ethernet networks presents a bottleneck to distributed GPU training, and prevents its trivial realization. To investigate how existing software frameworks can be adapted to efficiently support distributed GPUs, we developed Poseidon [39], a scalable system architecture for distributed inter-machine communication in existing deep learning frameworks. In order to assess Poseidon's effectiveness, we integrated Poseidon into the Caffe framework and evaluate its performance at training convolutional neural networks for object recognition in images. Poseidon features three key contributions that improve the training speed of deep neural networks on clusters: (i) a three-level hybrid architecture that allows Poseidon to support both CPU-only clusters as well as GPU-equipped clusters, (ii) a distributed wait-free backpropagation (DWBP) algorithm to improve

GPU utilization and to balance communication, and (iii) a dedicated structure-aware communication protocol (SACP) to minimize communication overheads. We empirically show that Poseidon converges to the same objective value as a single machine, and achieves state-of-the-art training speedup across multiple models and well-established datasets, using a commodity GPU cluster of 8. On the much larger ImageNet 22K dataset, Poseidon with 8 nodes achieves better speedup and competitive accuracy to recent CPU-based distributed deep learning systems which use 10s to 1000s of nodes.

5.7.3 Sufficient Factor Broadcasting. Matrix-parameterized models, including multiclass logistic regression and sparse coding, are used in machine learning (ML) applications ranging from computer vision to computational biology. When these models are applied to large-scale ML problems starting at millions of samples and tens of thousands of classes, their parameter matrix can grow at an unexpected rate, resulting in high parameter synchronization costs that greatly slow down distributed learning. To address this issue, we developed a Sufficient Factor Broadcasting (SFB) computation model [30] for efficient distributed learning of a large family of matrix-parameterized models, which share the following property: the parameter update computed on each data sample is a rank-1 matrix, i.e. the outer product of two "sufficient factors" (SFs). By broadcasting the SFs among worker machines and reconstructing the update matrices locally at each worker, SFB improves communication efficiency (communication costs are linear in the parameter matrix's dimensions, rather than quadratic) without effecting computational correctness. We present a theoretical convergence analysis of SFB, and empirically corroborate its efficiency on four different matrix-parameterized ML models.

Similarly, we developed an approach for adaptive network communication management for large-scale ML applications [29]. The system maximizes the network communication efficiency under a given inter-machine network bandwidth budget to minimize parallel error, while ensuring theoretical convergence guarantees for large-scale data-parallel ML applications and prioritizes messages most significant to algorithm convergence, further enhancing algorithm convergence. This system, called Bösen, is the first distributed implementation of the recently presented adaptive revision algorithm, which provides orders of magnitude improvement over a carefully tuned fixed schedule of step size refinements for some SGD algorithms. Experiments on two clusters with up to 1024 cores show that our mechanism significantly improves upon static communication schedules.

5.8 Dissemination

Our work has been distributed in many top machine learning conferences, including NIPS and AAAI, and peer-reviewed journals. Section 7 enumerates publications funded or partially funded by this project. This project has supported over 18 students and post-docs. Four doctoral dissertations [5][21][23][36] have been written under support of this project.

6 CONCLUSIONS

Taken as a whole, the work described above makes substantial progress against some of the common barriers to modern data analysis operations. The transfer learning methodologies we developed exploit data rich areas that are related to data starved problems. The theoretical bounds give insight as to how well transfer learning can be expected to work. In particular, our bounds show that if the offset (difference) between the source and target domains is smooth, then one can gain significantly from transfer learning. This formalizes the notion of 'relatedness' often used a motivation for transfer learning methodologies. We further extended these capabilities when target data is in short supply by developing active transfer learning. Another common barrier is the 'size' of modern data sets. We created parallel and distributed frameworks for significantly speeding up common machine learning models and allowing the development of very large models and use of very large data sets. Our tools accelerate learning, make efficient use of limited communication channels, and do so on modest computational clusters. Finally, we pushed the boundaries of what is considered the typical unit of analysis: the observation. Extending the observational concept to a bag of measurements drawn from a common distribution enables learning on heterogeneous irregular data, something that traditional methods cannot do. This extends standard machine learning, which uses as unit of analysis an individual, perhaps multivariate data point, towards recognizing and accepting more abstract objects as units of analysis. To that end, we have also initiated extensions of machine learning to enable a range of levels of abstraction of user-provided labels. We envision a new hyper-paradigm in which the data as well as its annotations can jointly exist in a spectrum of abstractions, and that future algorithms will optimally choose subsets of representational as well as used dialogue abstracts, to make the process the most natural and convenient to users, while boosting the results of learning.

Our efforts are exploratory in many ways, chipping away at the common barriers and challenges in modern data science. Even though open problems still remain, our work has demonstrated that e.g. kernel based methods work well to characterize distributions, but how to create of kernels for arbitrary complex objects remains a question of interest. Similarly, our results in transfer learning, active search, and hyper-active learning warrant further investigation. Questions remain such as: Can transfer learning be efficiently applied to large libraries of historical data? How best to pick a few data sets for transfer? How can high-level abstract information be used for learning models in a general and flexible manner?

7 PUBLICATIONS ORIGINATING FROM THIS WORK

[1] Dai, W., Kumar, A., Wei, J., Ho, Q., Gibson, G.A. and Xing, E.P., "High-Performance Distributed ML at Scale through Parameter Server Consistency Models." *AAAI*, 2015. pp. 79-87.

[2] De-Arteaga, M., Dubrawski, A., and Huggins, P., "Canonical Autocorrelation Analysis." *arXiv preprint arXiv:1511.06419*, 2015.

[3] Garnett, R., Krishnamurthy, Y., Xiong, X., Schneider, J. and Mann, R., "Bayesian optimal active search and surveying." *International Conference on Machine Learning (ICML)*, 2012.

[4] Gisolfi, N., Fiterau, M., and Dubrawski, A., "Finding Meaningful Gaps to Guide Data Acquisition for a Radiation Adjudication System." *AAAI*, 2015.

[5] Huang, Tzu-Kuo, "Exploiting Non-Sequence Data in Dynamic Model Learning." *Dissertations*. 561. http://repository.cmu.edu/dissertations/561, 2013.

[6] Huang, T.K. and Schneider, J., "February. Spectral learning of hidden Markov models from dynamic and static data." *International Conference on Machine Learning*, 2013. pp. 630-638.

[7] Huang, T.K. and Schneider, J., "Learning hidden Markov models from non-sequence data via tensor decomposition." *Advances in Neural Information Processing Systems*, 2013. pp. 333-341.

[8] Kandasamy, K., Al-Shedivat, M. and Xing, E.P., "Learning HMMs with Nonparametric Emissions via Spectral Decompositions of Continuous Matrices." *Advances in Neural Information Processing Systems, 2016*. pp. 2865-2873.

[9] Ma, Y., Garnett, R. and Schneider, J., "Active area search via Bayesian quadrature." *Artificial Intelligence and Statistics*, 2014. pp. 595-603.

[10] Ma, Y., Sutherland, D., Garnett, R. and Schneider, J., "Active pointillistic pattern search." *Artificial Intelligence and Statistics*, 2015. pp. 672-680.

[11] Ma, Y., Garnett, R., Schneider, J. and Wilson, A.G., "Fast Bayesian Optimization via Conjugate Sampling." *NIPS 2016 Workshop on Practical Bayesian Nonparametric*, 6, 2017.

[12] Ma, Y., Garnett, R. and Schneider, J., "Σ-optimality for active learning on Gaussian random fields." *Advances in Neural Information Processing Systems*, 2013. pp. 2751-2759.

[13] Neiswanger, W., Wang, C. and Xing, E., "Asymptotically exact, embarrassingly parallel MCMC." Proceedings of the 30th International Conference on Conference on Uncertainty in Artificial Intelligence, 2014.

[14] Oliva, J.B., Sutherland, D.J., Póczos, B. and Schneider, J., "Deep mean maps." *arXiv preprint arXiv:1511.04150*, 2015.

[15] Oliva, J., Neiswanger, W., Póczos, B., Schneider, J. and Xing, E., "Fast distribution to real regression." *Artificial Intelligence and Statistics*, 2014. pp. 706-714.

[16] Rabbany, R., Eswaran, D., Dubrawski, A.W. and Faloutsos, C., "Beyond Assortativity: Proclivity Index for Attributed Networks (PRONE)." Pacific-*Asia Conference on Knowledge Discovery and Data Mining*, 2017. pp. 225-237.

[17] Sutherland, D.J., Póczos, B., and Schneider, J., "Active learning and search on low-rank matrices." *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013. pp. 212-220.

[18] Sutherland, D.J., Xiong, L., Póczos, B., and Schneider, J., "Kernels on Sample Sets via Nonparametric Divergence Estimates." *arXiv preprint arXiv: 1202.0302*, 2012.

[19] Sutherland, D.J., Oliva, J.B., Póczos, B. and Schneider, J.G., "Linear-Time Learning on Distributions with Approximate Kernel Embeddings." *AAAI*, 2106. pp. 2073-2079.

[20] Sutherland, D.J. and Schneider, J., "On the error of random Fourier features." *arXiv preprint arXiv:1506.02785*, 2015.

[21] Sutherland, D.J., "Scalable, Flexible and Active Learning on Distributions." *Doctoral dissertation, Carnegie Mellon University*, 2016.

[22] Venkatesan, S., Miller, J.K., Schneider, J., and Dubrawski, A., "Scaling Active Search using Linear Similarity Functions." *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017. pp. 2878–2884.

[23] Wang, X., "Active Transfer Learning." Doctoral dissertation, Carnegie Mellon University, 2016.

[24] Wang, X., Huang, T.K. and Schneider, J., "Active transfer learning under model shift." *International Conference on Machine Learning*, 2014. pp. 1305-1313.

[25] Wang, X., Garnett, R. and Schneider, J., "Active search on graphs." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013. pp. 731-738.

[26] Wang, X., Garnett, R. and Schneider, J., "An Impact Criterion for Active Graph Search." *NIPS workshop on Bayesian Optimization and Decision Making*, 2012.

[27] Wang, X. and Schneider, J., "Flexible transfer learning under support and model shift." *Advances in Neural Information Processing Systems*, 2014. pp. 1898-1906.

[28] Wang, Y.X., Sadhanala, V., Dai, W., Neiswanger, W., Sra, S. and Xing, E., "Parallel and distributed block-coordinate Frank-Wolfe algorithms." *International Conference on Machine Learning*, 2016. pp. 1548-1557.

[29] Wei, J., Dai, W., Qiao, A., Ho, Q., Cui, H., Ganger, G.R., Gibbons, P.B., Gibson, G.A. and Xing, E.P., "Managed communication and consistency for fast data-parallel iterative analytics." *Proceedings of the Sixth ACM Symposium on Cloud Computing*, 2015. pp. 381-394.

[30] Xie, P., Kim, J.K., Zhou, Y., Ho, Q., Kumar, A., Yu, Y. and Xing, E., "Distributed machine learning via sufficient factor broadcasting." *arXiv preprint arXiv:1511.08486*, 2015.

[31] Xie, P., Deng, Y. and Xing, E., "Diversifying restricted boltzmann machine for document modeling." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015. pp. 1315-1324.

[32] Xing, E.P., Ho, Q., Dai, W., Kim, J.K., Wei, J., Lee, S., Zheng, X., Xie, P., Kumar, A. and Yu, Y., "Petuum: A new platform for distributed machine learning on big data." *IEEE Transactions on Big Data*, Vol. **1**, No. 2, 2015. pp. 49-67.

[33] Xing, E.P., Ho, Q., Xie, P. and Wei, D., "Strategies and principles of distributed machine learning on big data." *Engineering*, Vol. **2**, No. 2, 2016. pp. 179-195.

[34] Xiong, L., Póczos, B., and Schneider, J., "Efficient Learning on Point Sets." *IEEE International Conference on Data Mining (ICDM)*, 2013.

[35] Xiong, L. and Schneider, J.G., "Learning from Point Sets with Observational Bias." *UAI*, 2014. pp. 898-906.

[36] Xiong, Liang, "On Learning from Collective Data." *Dissertations*. 560. http://repository.cmu.edu/dissertations/560, 2013.

[37] Xu, Y., Zhang, H., Miller, K., Singh, A. and Dubrawski, A., "Noise-Tolerant Interactive Learning from Pairwise Comparisons with Near-Minimal Label Complexity." *arXiv preprint arXiv:1704.05820*. 2017.

[38] Yuan, J., Gao, F., Ho, Q., Dai, W., Wei, J., Zheng, X., Xing, E.P., Liu, T.Y. and Ma, W.Y., "LightLDA: Big topic models on modest computer clusters." *Proceedings of the 24th International Conference on World Wide Web*, 2015. pp. 1351-1361.

[39] Zhang, H., Zheng, Z., Xu, S., Dai, W., Ho, Q., Liang, X., Hu, Z., Wei, J., Xie, P. and Xing, E.P., "Poseidon: An Efficient Communication Architecture for Distributed Deep Learning on GPU Clusters." *USENIX Annual Technical Conference*, 2017.

8 LIST OF ACRONYMS

APPS	Active pointillistic pattern search
CNN	Convolutional neural network
CPU	Central processing unit
DMM	Deep mean map
DWBP	Distributed wait-free backpropagation
EMD	Earth mover's distance
GB	Gigabyte
GPU	Graphical processing unit
GRF	Gaussian random field
H^2	Squared Hellinger
HEM	Harmonic energy minimization
JS	Jensen-Shannon
ML	Machine learning
MMD	Maximum mean discrepancy
MME	Mean map layer (of a CNN)
PAC	Probably approximately correct
RKHS	Reproducing kernel Hilbert space
ROI	Region of interest
SACP	Structure-aware communication protocol
SF	Sufficient factors
SFB	Sufficient factor broadcasting
SGD	Stochastic gradient descent
SKL	Symmetric Kullback-Leibler divergence
SVM	Support vector machine
TB	Terabyte
TV	Total variation