AFRL-RI-RS-TR-2021-191

ZERO-KNOWLEDGE DISCOVERY USING DATA SMASHING

THE UNIVERSITY OF CHICAGO

NOVEMBER 2021

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-2021-191 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
MICHAEL J. MANNO
Work Unit Manager

/ S /
SCOTT PATRICK
Deputy Chief,
Intelligence Systems 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

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
NOVEMBER 2021	FINAL TECHNICAL REPORT	MAY 2017	MAY 2021

4. TITLE AND SUBTITLE

ZERO-KNOWLEDGE DISCOVERY USING DATA SMASHING

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-17-2-0124	N/A	62702E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
D3ME	00	09

6. AUTHOR(S)

The University of Chicago

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
The University of Chicago 5801 S Ellis Ave Chicago, IL 60637	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory/RIED 525 Brooks Road Rome NY 13441-4505	RI	AFRL-RI-RS-TR-2021-191

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

Project ZeD addresses zero knowledge inference in sequential data streams: the task of finding-models from raw data when we do not necessarily know the correct model structure a priori. Absence of such prior knowledge is becoming increasingly common for the complex questions we are now asking in biology, social systems, physics and engineering. We cannot reduce this exercise to one of simply tuning parameters and/or model calibration. Sparsity of reported de no-vo modeling paradigms that allow automated abduction of good models from raw data is a key bottleneck in automated problem solving. Our effort is designed to address this emerging gap. Leveraging fundamentally new insights into automated inference, such as the principle of data smashing and abductive learning of generative models of quantized stochastic processes, we en-vision transformative breakthroughs in automated problem solving.

15. SUBJECT TERMS

Project ZeD, D3M program, modeling paradigms, quantized stochastic processes

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	13

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
MICHAEL J. MANNO	**N/A**

TABLE OF CONTENTS

1.0 SUMMARY

Project ZeD addresses zero knowledge inference in sequential data streams: the task of finding-models from raw data when we do not necessarily know the correct model structure a priori. Absence of such prior knowledge is becoming increasingly common for the complex questions we are now asking in biology, social systems, physics and engineering. We cannot reduce this exercise to one of simply tuning parameters and/or model calibration. Sparsity of reported de novo modeling paradigms that allow automated abduction of good models from raw data is a key bottleneck in automated problem solving. Our effort is designed to address this emerging gap. Leveraging fundamentally new insights into automated inference, such as the principle of data smashing and abductive learning of generative models of quantized stochastic processes, we envision transformative breakthroughs in automated problem solving, contributing to Technical Area I of the D3M program. Within the scope of this project, we propose new learning primitives for sequential data streams (or data that may be transformed to sequential representation via some form of serialization) that focus on the issues described above.

2.0 INTRODUCTION

Within the scope of TA1, we propose a set of discoverable zero-knowledge modeling primitives that carry out diverse learning tasks applicable to broad set of data types including time series, sequential streams, and ultimately images and video. We call this the Z-primitive ecosystem, consisting of primitives that carry out zero-knowledge classification, infer low dimensional embedding of datasets directly from raw data, carry out task agnostic universal distance metric learning, execute zero-knowledge model validation, and finds good features with little or no human intervention – and we expect to achieve sample complexities significantly lower to competing approaches. Our modeling primitives are predicated on an algorithm for computing a universal similarity metric between data streams. This universal metric, used appropriately, addresses the featurization issue, often limiting or eliminating the need for expert intervention in the modeling process. Our modeling paradigm has specific mathematical properties that eschew over-fitting. The core data smashing algorithm can compute similarity in a rigorous sense between observed and model-predicted data streams, thereby providing a zero-knowledge domain-independent, and often task-agnostic, measure of how well the inferred model captures generalizable (and not over-fitted) patterns and/or dynamics. We will develop a generalized non-parametric indicator of data sufficiency (via estimation of the self-annihilation error where applicable), to indicate when, and where (datatype, modality or dimension), more data needs to be collected. Our new primitives will be discoverable by design, i.e., we will automatically determine when they are suitable for the answers/outcomes sought, given the kind and volume of data available. This will enable automatic composition within the goals of TA2 to distill accurate and succinct models of complex phenomena. Additionally, we will develop general algorithms to evaluate composition performance, and rigorously control for over-fitting and error rates.

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

At the center of the Z-primitive ecosystem design is the principle of data smashing. Data smashing is a zero-knowledge model-free feature-free approach for comparing finite sample paths from stochastic processes, that relies on the rigorously established notion of statistical anti-streams: Similarity between two streams is the degree to which one, when summed (in a group-theoretic sense) to the other's anti-stream, mutually annihilates all statistical structure to noise. Closer the smashed output is to flat white noise, closer are the hidden generators of the two processes. Thus, given two input streams, we can quantify statistical similarity in a rigorous universal sense. Additionally, given a corpora of sequential data, pairwise data smashing gives us an estimated distance matrix, which can then be used for classification, and finding manifold embeddings. Additionally, the data smashing algorithm yields a universal approach for model validation - smash the model generated data against the sample, and check if the output is close to noise. This is not limited to models we infer, and presents a fundamentally new approach to zero-knowledge model validation. In the sense that data smashing operates on streaming data this makes it reminiscent of sketching algorithms; however in sketching-based similarity detection approaches reported in the literature e.g. locality sensitive hashing (LSH), the metric of similarity is pre-defined. Here we estimate the metric with no prior knowledge and no features. In the sense that data smashing operates on streaming data makes it reminiscent of sketching algorithms; however in sketching-based similarity detection approaches reported in the literature e.g. locality sensitive hashing (LSH), the metric of similarity is pre-defined. Here we estimate the metric with no prior knowledge, and no features. The second important piece in our technical approach is zero-knowledge modeling of direction-specific dependence between the stochastic sources, which might be seen as a non-parametric non-linear computation of Granger causal influence. There do exist competing approaches to non-parametrically learning sequential data, e.g. infinite hidden
Markov models. However our formalism is more general, with explicit inference of non-trivial transition structures; we have no crafted priors or emission mechanisms, and we are guaranteed to learn good models with high probability (the primitives are PAC-efficient). Additionally, the state estimation problem in our framework is significantly simpler (no need to invoke expectation maximization routines), implying that we can use the inferred model easily and efficiently for prediction. In our case, patterns emergent in the data induce our model; more data does not lead to overtraining. Our probability of error falls exponentially with more observations – a consequence of error bounds achievable in learning empirical distributions (invoking Chernoff bounds and DKW inequality related results in nonparametric statistics). Since no structure is assume a priori, overfitting is much less of a problem, and we aim to offer demonstrable evidence to that effect.

Comparison with Current Technology:

Learning algorithms often presuppose structure, and calibrate parameters against observations.
This is problematic if our prior assumptions are incorrect. Even with model selection strategies we only "select" from a predetermined family of possibilities. We leverage our work on identifying generative models of quantized ergodic, quasi-stationary stochastic processes in the framework of finite state probabilistic automata, which is free from the above objections. Starting with a quantized data stream, we infer the number of causal states, their connectivity (how the states evolve), and the transition probabilities driving the stochastic dynamics. Patterns emergent in the data induce our model; more data does not lead to overtraining. In addition to the heuristic nature of

feature selection, machine learning algorithms typically necessitate the choice of a distance metric in the feature space. For example, the classic "nearest neighbor" k-NN classifier requires definition of proximity, and the k-means algorithm depends on pairwise distances in the feature space for clustering. To side-step the heuristic metric problem, recent approaches often learn appropriate metrics directly from data, attempting to "back out" a metric from side information or labeled constraints. Unsupervised approaches use dimensionality reduction and embedding strategies to uncover the geometric structure of geodesics in the feature space e.g. manifold learning. However, automatically inferred data geometry in the feature space is, again, strongly dependent on the initial choice of features. Since Euclidean distances between feature vectors are often misleading, heuristic features make it impossible to conceive of a task-independent universal metric. In contrast smashing is based on an application-independent notion of similarity between quantized sample paths observed from hidden stochastic processes. Our universal metric quantifies the degree to which the summation of the inverted copy of any one stream to the other annihilates the existing statistical dependencies, leaving behind flat white noise. We circumvent the need for features altogether and do not require training in the conventional sense. In the sense that data smashing operates on streaming data makes it reminiscent of sketching algorithms; however in sketching-based similarity detection approaches reported in the literature e.g. locality sensitive hashing (LSH), the metric of similarity is pre-defined. Here we estimate the metric with no prior knowledge, and no features. In the context of our causal inference algorithms, we expect to significantly advance the state of the art as well. Unfortunately, Granger causal inference has been mostly implemented with linearity assumptions; the original definition has no such restriction, and linear tests demonstrably fail to uncover non-linear influence. Non-linear tests exist; but often assume the class of allowed non-linearities; thus not quite removing presupposed structure. Non-parametric approaches, e.g. the Hiemstra-Jones test successfully dispense with pre-suppositions on structure, at the cost of failing to produce a generative model of inferred causation. We are left with a black box that answers questions without any insight on the dynamical structure of the system under inquiry. In contrast, our inference primitive makes no presupposition of linearity, is nonparametric, and yet produces direction-specific directional generative models of causal influence.

4.0 RESULTS AND DISCUSSION

Project ZeD is expected to have broad impact on how automated inference is applied to unravel data-intensive problems in science and engineering. Our approach is crucially applicable to scenarios and problem domains where human expertise is scarce, and model families are difficult to select a priori, or where the cost of failing to discover subtle emergent patterns is unacceptably high. For example, complex diseases seldom have known simple models. Clinical informatics enabled with zero-knowledge inference, will be potentially able to design new personalized and precision therapeutic interventions for complex diseases such as cancers. Such tools will be also vital in the social sciences, where we often have a priori unknown probabilistic principles of emergence. Same is true in complex engineering problems, and decision support system. We envision that our effort will significantly improve error rates beyond the state of art, enable automated problem solving pipelines, and accelerate scientific discovery in diverse fields. Thus, project ZeD is expected to emerge as a center piece to realizing the full potential of the D3M vision. Our initial focus is series learning, and particularly unsupervised series learning. As the project progresses, we plan to leverage our algorithms to produce novel tools applicable to more general data types, including text, images, video streams and graph corpora.

Deliverables of the project include validated software, along with source code. In addition to software, we provide detailed documentation, evaluation metrics, complexity and hardware requirements to execute our primitives and pipeline composers. Other deliverables include academic papers in high impact journals, presentations in relevant conferences, attending required project meetings, submit to project evaluations, and participate in integration exercises. The deliverable objectives met in this quarter include: 1) TA1 primitives update after
Summer 2018 Evaluation; 2) Addressing resource requirement issues such as runtime fluctuations in some of our primitives 2) Further improvements to our code base for easier integration in general with the rest of the D3M universe.

5.0 CONCLUSIONS

One concrete area we have been pursuing is the use of the technologies developed within this program to predict and prevent crime in the City of Chicago. We have begun discussion with some key players in this regard, including personnel at the Crime Lab at University of Chicago, and interested parties working with the Chicago Police Department. We are continuing our Workshop series on "Crimes of Prediction", where we present results from our spatio-temporal learning primitives in the context of the crime prediction problem to an audience of domain experts in sociology of crime. We hope to play a role in the design and implementation of predictive policing and general automation in the law enforcement apparatus in the City of Chicago. In addition, the ability of the Z-primitive ecosystem to distill validated generative models, which can then be used to predict complex system evolution, is highly monetizable, and we see significant potential here for commercialization and technology transfer. In particular, we anticipate interest from financial firms aiming to monetize volatility or offset risk by predicting big market moves, insurance firms that aim to mitigate risk via predictive analytics, pharmaceutical companies that wish to re-purpose drugs for therapeutic interventions in complex diseases, and medical device manufacturers interested in engineering the next generation smart wearables that leverage the model-free anomaly detection capability of data smashing to signal, log and mitigate medical events. In addition to commercial organizations, we are also considering technology transfer to government bodies interested in leveraging automated inference to address hard problems.

6.0 REFERENCES

N/A

APPENDIX A – PUBLICATIONS AND PRESENTATIONS

List the dates, times, title, event and speakers of any presentations made under this effort and the title author and publication information for any publication made under this effort.

Data Smashing 2.0: Sequence Likelihood (SL) Divergence For Fast Time Series Comparison
Yi Huang, Ishanu Chattopadhyay

https://arxiv.org/abs/1909.12243 arXiv is a free distribution service and an open-access archive for 1,975,103 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Materials on this site are not peer-reviewed by arXiv.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

ZeD	Zero-Knowledge Discovery
D3M	Data-Driven Discovery of Models
LSH	Locality Sensitive Hashing
NN	Nearest Neighbor
TA1	Technical Area 1
TA2	Technical Area 2