Synthetic data for learning-based knowledge discovery

William Shiao
University of California, Riverside
Riverside, CA, USA
wshia002@ucr.edu

Evangelos E. Papalexakis
University of California, Riverside
Riverside, CA, USA
epapalex@cs.ucr.edu

ABSTRACT

Recent advances in deep learning have demonstrated the ability of learning-based methods to tackle very hard downstream tasks. Historically, this has been demonstrated in predictive tasks, while tasks more akin to the traditional KDD (Knowledge Discovery in Databases) pipeline have enjoyed proportionally fewer advances. Can learning-based approaches help with inherently hard problems within the KDD pipeline, such as *"how many patterns are in the data"*, *"what are different structures in the data"*, and *"how can we robustly extract those structures?"* In this vision paper, we argue for the need for synthetic data generators to empower cheaply-supervised learning-based solutions for knowledge discovery. We describe the general idea, early proof-of-concept results which speak to the viability of the paradigm, and we outline a number of exciting challenges that await, and a set of milestones for measuring success.

1. INTRODUCTION

Supervised and self-supervised learning has made and continues to be making tremendous strides. Numerous examples include (but are not limited to) language models [7; 14], vision models [12; 5], graph neural networks [26; 11], and even some "general purpose" models that can work for multiple data types and tasks [3]. The superiority of these modern deep learning models is primarily shown in downstream tasks that are predictive in nature, e.g., image classification, speech recognition, General Language Understanding Evaluation (GLUE) [21] tasks, graph node classification or link prediction.

In stark contrast to traditional downstream tasks, tasks that relate to what we collectively call Knowledge Discovery in Databases (KDD) or "the KDD process" [9] have enjoyed considerably less attention and, as a result, significantly fewer advances. This disparity, at first glance, is rather understandable since tasks that pertain to the KDD process are much more open-ended than prediction or classification-based downstream tasks and are inherently unsupervised in nature.

However, when we look at the state of the art of the KDD process overall, there has been steady and significant progress made in introducing new mining algorithms, new pre or post-processing techniques, and new evaluation techniques, but for the most part, the "glue" of any practical such pipeline is by-and-large human-based. Many design choices and algorithmic hyperparameters in that pipeline are typically chosen by an experienced data scientist and are a result of the application of a number of heuristics and copious amounts of trial-and-error experimentation.

A natural question that arises is whether recent advances in deep (self-)supervised learning can transform the way that practitioners perform the KDD process in similar ways that they have transformed the way in which we approach classification and prediction problems in real life. For instance, can we use cutting-edge deep learning methods to solve inherently hard problems which lie at the heart of the KDD process, such as "Are there any interesting patterns in my data? If so, how many, and what kinds of structure(s) do they follow?" Furthermore, can we do so while having no real supervision—without real data with annotations that directly answer those questions? In this vision paper, we propose a "Blue Sky" idea, borrowing the terminology from the initiative set forth by the Computing Research Association (CRA) [6], to tackle the above question, towards transforming the process of knowledge discovery.

2. PROPOSED VISION

The Blue Sky idea: The key to transforming data discovery is the design of *high-quality realistic synthetic data* used in conjunction with cutting-edge deep (self-)supervised machine learning models. An overview of the proposed idea is shown in Figure 1.

Unlike "traditional" supervised approaches, this paradigm introduces "cheap" supervision where human involvement is ideally zero (or close to zero), thus remaining essentially unsupervised. Furthermore, this eliminates the current need for running the analytical pipeline (or parts thereof) multiple times, in trial-and-error mode, in order to manually or heuristically determine the best result out of the myriad executions. Because of the quick response/inference time of modern deep models, this idea has the potential to decrease the KDD process execution time by orders of magnitude.

In addition to practicality and scalability, this idea, extending existing efforts for uncertainty quantification in "traditional" supervised scenarios, can allow for robust hypothesis testing and provide uncertainty bounds on the presence of certain types of structure in real data.

Finally, this idea has the potential to allow us to solve problems for which we currently have no widely accepted solution by generalizing from examples and problems that are "easier" and for which we have acceptable solutions, by leveraging the problem structure (see Section 3 for an example).

Why is it a Blue Sky idea? The proposed idea has the potential to transform the traditional KDD process, which

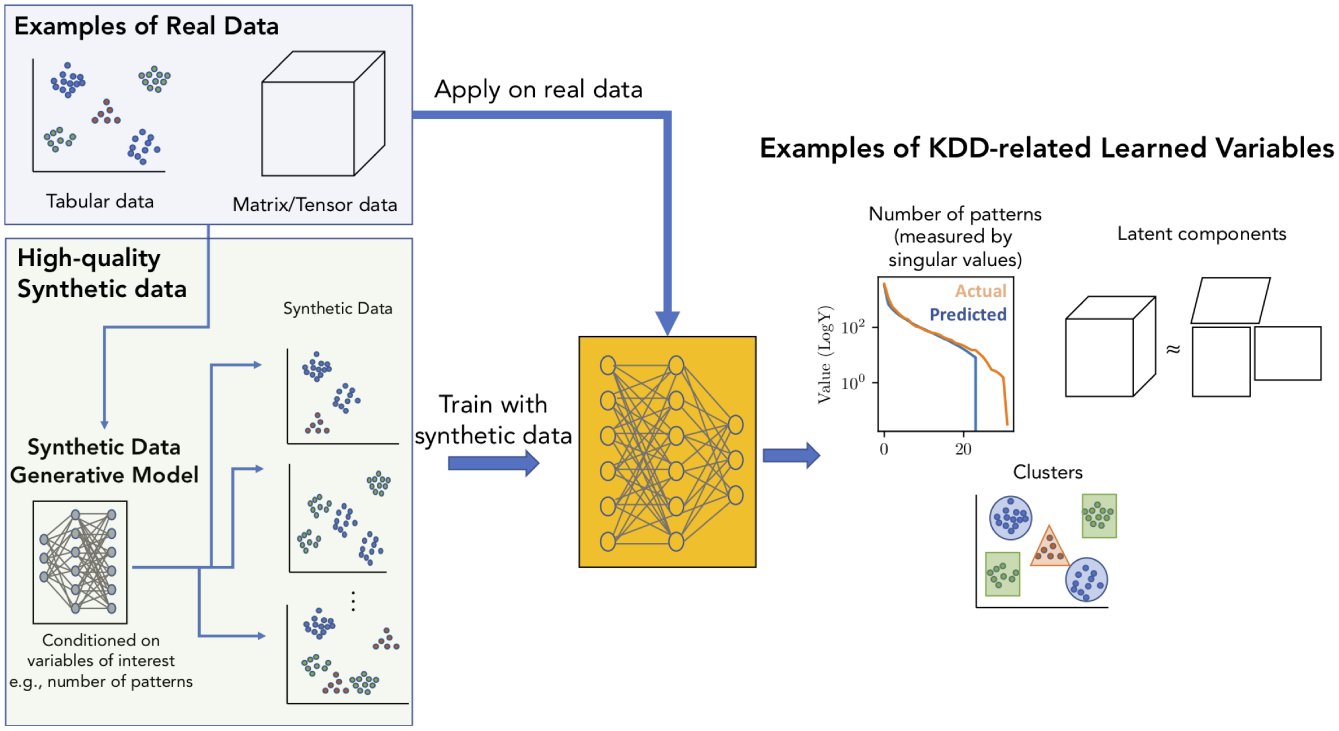

Figure 1: Overview of the proposed vision.

is especially useful and relevant in emerging domains where insights and structure in the data are the desired outputs. Furthermore, problems that this proposed idea is promising to tackle are extremely hard and usually left to be solved manually by the end user of a given algorithm/pipeline employing heuristics.

At the heart of it, the proposed idea aims towards a unified and generalizable framework for a very heterogeneous and multi-faceted process and can ultimately push the frontier of data mining in the design of automated and personalized KDD pipelines.

Why should the community ponder over it? The data mining community has the collective expertise and domain knowledge necessary for this kind of endeavor and can inject it into the data generation process.

Moreover, the proposed idea is a treasure trove of interesting and hard research problems: It poses a number of fascinating and unique challenges that can not only advance the state of the art in the KDD process, but are also poised to advance generative models and (self-)supervised learning since both the kinds of data to be generated and learn generalizable representations from are novel in that context.

Why now? Currently, the data mining and machine learning communities have a mature understanding of a number of crucial components that are necessary for executing this research agenda, ranging from recent advances in generative models (from adversarial generation [10] to diffusion models [15]) and deep (self-)supervised models. The twist, of course, is that this understanding pertains to the current use of the above methods, which is not necessarily aligned with the proposed use, which in itself poses interesting challenges.

It is important to acknowledge here the broad and profound impact that synthetic data have already had in data mining and machine learning, starting from classical and powerful oversampling techniques, such as SMOTE [4], to the advances of Generative Adversarial Networks (GANs) [10] and the remarkable results produced by Diffusion [15] and Transformer-based Large Language Models [1; 20]

The key novelty here is that our proposed synthetic data generation is not focused on data augmentation or mere generation of realistic data (with the term "mere" meant here strictly as a qualifier of single-purpose and by no means implies that such generation is trivial) but should rather focus on hidden patterns in the data and the inclusion of "knobs" such as the "number of patterns", which render the synthetic data more suitable for exploring different aspects of the KDD process.

3. PROOF OF CONCEPT

We would like to offer two particular data points of reference which provide preliminary results for the viability of our general idea. The particular hard problem at hand that we have been focusing on is the identification of the low-rank in matrix or tensor data, from which one can draw parallels to problems such as identifying the number of clusters in data [8].

In recent work [23], we demonstrated that we can successfully learn the singular value profile of a given matrix, which is essentially what is needed in order to identify the full and the low rank of that matrix. Given that this has been successful in matrix data, can we generalize it to tensor data, where this problem is extremely hard and wide open, by leveraging the algebraic structure of the two different prob-

lems? In concurrent work [18], we show that by using simple but carefully-designed synthetic tensor data, where the low rank is known, we can accurately learn the low rank.
We, by no means, claim that we have solved this problem, however, those two instances provide strong evidence for the viability of our proposed paradigm.

4. RESEARCH CHALLENGES

A number of exciting research challenges need to be addressed for this paradigm shift to take effect.

4.1 Designing data generators

The design of synthetic data generators is of paramount importance. Generators ought to obey the following properties:
P1: Generate *realistic* data which closely mimic the distribution of real data.
P2: Offer *control over parameters* of importance to the KDD process (e.g., number of clusters in the data).
P3: Offer substantial *diversity* in the generated data points such that they can be used to successfully train a model that learns generalizable features.
For example, we recently introduced generation of graph adjacency matrices [17] and tensors [16], where the rank is a controllable parameter of the generator.

4.2 Evaluating realism

When we are generating synthetic data, even though our goal is not the generation of novel-looking data (e.g., images), we still have to make sure that the generated data are realistic, in that the closely follow the distribution of the real data.
When measuring realism, we may need to take modality-specific approaches (e.g., treat images differently from graphs), and when generating synthetic data for novel and emerging applications, we have to carefully decide upon "realism" criteria that which we can use to hold our data to this important test. We can derive such an example from our recent work [16] where the goal is to generate multiplex graphs. Even though there exist established realism criteria for single graphs, applying them on each individual view of the multiplex graph is not enough, since a major consideration for the output data is that each generated graph view is not independent from the rest. Thus, in order to capture this relation across graph views, and how close it is to real data, we would have to define novel tests. In the particular case at hand, we opted for viewing the generated multiplex graph as a tensor, and measure how "compressible" it is for different decomposition ranks, and subsequently compared this behavior against the one observed for real-world multiplex graphs when treated as such.
Finally, beyond realism in the raw feature dimensions of the data (such as realism in produced images), in this case we should be able to measure realism in the hidden pattern dimensions of the data as well. For example, in the application of community detection in graphs, earlier work has demonstrated that in many real-world graphs communities have hyperbolic shapes [2]. In this case, a "community" is essentially a hidden pattern in the generated graph data, and ensuring that its generation adheres to this real-world observation, when supported by the data and application of interest, can enhance the realism of the latent patterns in our synthetic data.

4.3 Limited real data & knowledge-guided generation

Modern generative models assume that we have some seed real data available from which we learn their distribution and successfully generate new data points. What if we have no real data available, or the amount of data is rather insufficient for generating a diverse-enough synthetic dataset?
In such data-scarce scenarios, we may resort to model and knowledge-guided design of synthetic data, a process which would essentially bring knowledge-guided machine learning approaches [13] to our paradigm, and where we would infuse model-based knowledge to the data generation to compensate for the lack of real data.

4.4 Representation learning

How can we learn effective representations from structured or unstructured data which work for KDD-process downstream tasks?
It may be tempting to immediately endeavor to learn those representations fully automatically using deep learning modeling. As in most scenarios, doing so without having a firm grip over the different kinds of bias that are introduced in the generated process may yield suboptimal representations. Thus, in conjunction with fully-automated representation learning, domain-expertise-guided feature generation may be a reasonable first step which would allow us to understand what features work and what features fail (such as in our proof of concept work, where we define a set of descriptive features for tensor data, based on years of expertise [18]), and progressively "graduate" to fully-automated representations.

4.5 Generalization and transfer across tasks

This challenge is highly related to the previous one of representation learning. However, it underscores an important requirement for our approach to be generalizable and transferable when there exist structural similarities across tasks and when solutions exist for simpler tasks, and we wish to generalize to harder instances.

4.6 Designing end-to-end KDD pipelines

When we integrate all the different advances together into an analytical pipeline, this may look vastly different from existing pipelines. For example, we may be able to tailor entire pipelines to a specific problem and accordingly build multiple personalized KDD pipelines. Alternatively, we may opt for a generalist solution where a single powerful pipeline can handle most cases.
In addition to building the pipeline, under this approach, we may be able to offer more robust uncertainty quantification while reducing the execution time of a single pipeline by orders of magnitude, which may invite us to rethink the overall design, especially as it may integrate with domain experts in the loop.

4.7 Robust evaluation

Given that the nature of most problems that our proposed idea is poised to tackle is extremely hard, evaluation poses

a unique challenge in itself. As mentioned above, this new paradigm may allow us to revisit the design of the analytical pipeline, where interaction and potential evaluation by a domain expert may be much more scalable than ever before. We anticipate that evaluation should heavily rely on the help of domain experts, either directly or indirectly. For instance, when evaluating the accuracy of tensor rank learning in [18], we rely on chemometrics expertise which links rank to the number of chemicals in a mix.

5. MEASURING SUCCESS

In order to measure success of the proposed approach, the following milestones have to be progressively met, ideally for a number of different KDD-related problems. **M1**: Solve problems that we can already solve exactly (e.g., matrix rank and singular value profile): Success is measured by how far we are from the exact solution

M2: Solve problems for which we have widely acceptable and easy-to-use heuristics (e.g., matrix *low* rank or finding the number of clusters in K-means using the "elbow method"): Success is measured by how far we are from solutions produced by data scientist experts using heuristics afforded to them and their best judgement.

M3: Apply to problems for which there is no widely accepted heuristic solution (e.g., tensor low rank): Success will be measured by focusing on real-world application domains in collaboration with domain experts.

(a) Direct measures for M3: Do the results agree with what domain experts know to be true (after translating KDD terms such as "cluster" to the domain language, such as "phenotype")? Do domain experts evaluate results favorably to existing methods? **(b) Indirect measures for M3**: Did the application of the learning-based KDD process in a particular domain lead to a significant discovery in that domain?

6. DISCUSSION AND CONCLUSION

Our idea has parallels to another emerging direction which involves the use of Reinforcement Learning (RL) in solving hard data mining problems, such as fine-tuning the popular DBSCAN clustering algorithm [24]. We believe that the two approaches are synergistic and we are interested in exploring their interplay.

As this paper is meant to start a discussion around this topic and explore the opportunities and limitations of the proposed direction, we envision that there is a set of problems that where the proposed direction can have more immediate impact:

- Learning-based solutions developed as part of this vision can serve as:
  - **Auxiliary parts of a KDD pipeline**, such as replacing or augmenting existing heuristics that guide the discovery (such as Cluster Validation Indices [19])
  - **More optimistic: Main parts of a KDD pipeline**, where the learning-based solution will be able to learn either elements of the desired solution (e.g., cluster membership between two different points) or the entire desired solution (e.g., cluster assignments, alignment between data points [25; 22], etc)
- **Generalizing from simpler to harder problems**, where we can develop models in cases where there exist exact analytical descriptions for the sought-after patterns or latent variables, and work to extend them to cases where such analytical solutions no longer exist.

In closing, in this vision paper, we propose the transformation of the KDD process through the use of synthetic data which can train powerful deep learning models tailored to tackling the hardest problems in knowledge discovery from data.

7. ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under CAREER grant no. IIS 2046086 and CREST Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (MECIS) grant no. 2112650, and by the Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation here on.

8. REFERENCES

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] M. Araujo, S. Günnemann, G. Mateos, and C. Faloutsos. Beyond blocks: Hyperbolic community detection. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pages 50–65. Springer, 2014.

[3] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] C. R. A. (CRA). Blue sky ideas. `https://cra.org/ccc/visioning/blue-sky/`.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 606–610. SIAM, 2005.

[9] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, et al. Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88, 1996.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. 6 2014.

[11] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[13] A. Karpatne, R. Kannan, and V. Kumar. *Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data*. CRC Press, 2022.

[14] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[16] W. Shiao, B. A. Miller, K. Chan, P. Yu, T. Eliassi-Rad, and E. E. Papalexakis. Tengan: adversarially generating multiplex tensor graphs. *Data Mining and Knowledge Discovery*, 38(1):1–21, 2024.

[17] W. Shiao and E. E. Papalexakis. Adversarially generating rank-constrained graphs. In *2021 IEEE DSAA*, pages 1–8. IEEE, 2021.

[18] W. Shiao and E. E. Papalexakis. Frappe: Fast rank approximation with explainable features for tensors. *arXiv preprint arXiv:2206.09316*, 2022.

[19] W. Shiao, U. S. Saini, Y. Liu, T. Zhao, N. Shah, and E. E. Papalexakis. Carl-g: Clustering-accelerated representation learning on graphs. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2036–2048, 2023.

[20] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[21] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[22] Y. Wu, U. S. Saini, J. Chen, and E. E. Papalexakis. Tenalign: Joint tensor alignment and coupled factorization. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 568–577. IEEE, 2022.

[23] D. Xu, W. Shiao, J. Chen, and E. E. Papalexakis. Sv-learn: Learning matrix singular values with neural networks. In *2022 IEEE ICDM Workshops*. IEEE, 2022.

[24] R. Zhang, H. Peng, Y. Dou, J. Wu, Q. Sun, Y. Li, J. Zhang, and P. S. Yu. Automating dbscan via deep reinforcement learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2620–2630, 2022.

[25] S. Zhang, H. Tong, Y. Xia, L. Xiong, and J. Xu. Nettrans: Neural cross-network transformation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 986–996, 2020.

[26] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, and L. Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.