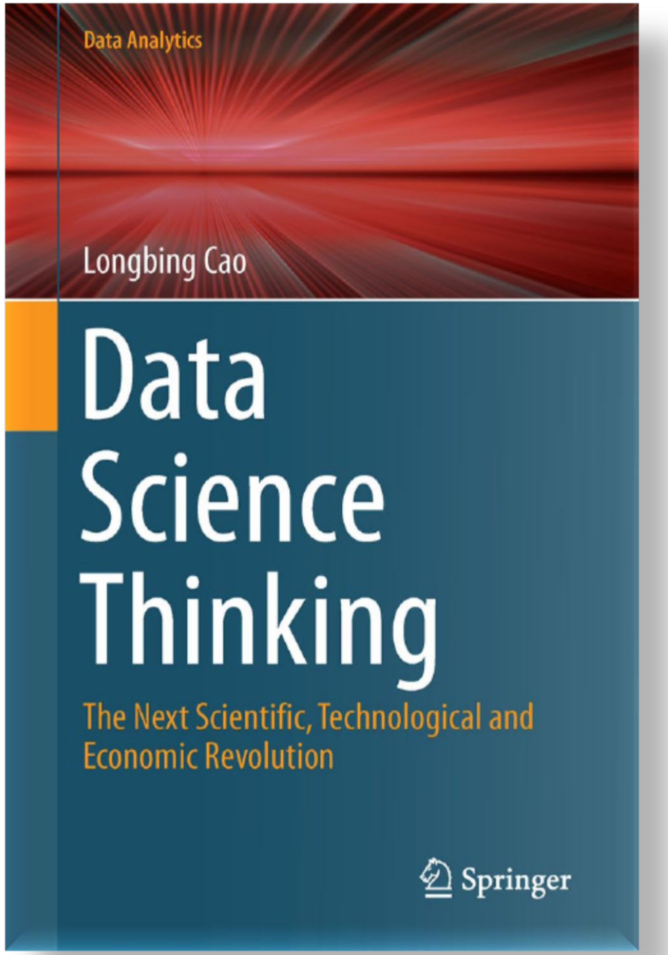
Some of Critical Challenges and Opportunities in Data science

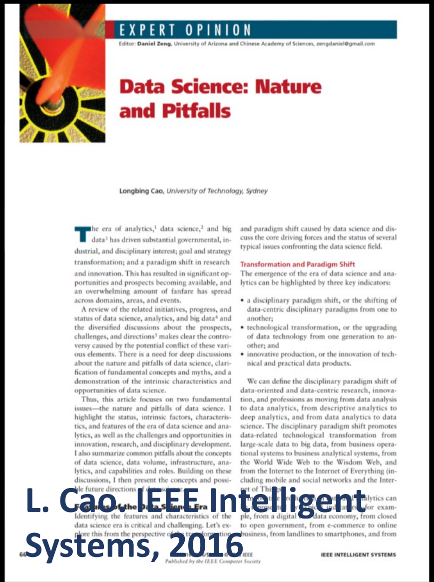
Longbing Cao | University of Technology Sydney

The Data Science Lab
www.datasciences.org



The World Has been Fundamentally Transformed by Data Science and Data-driven Intelligence

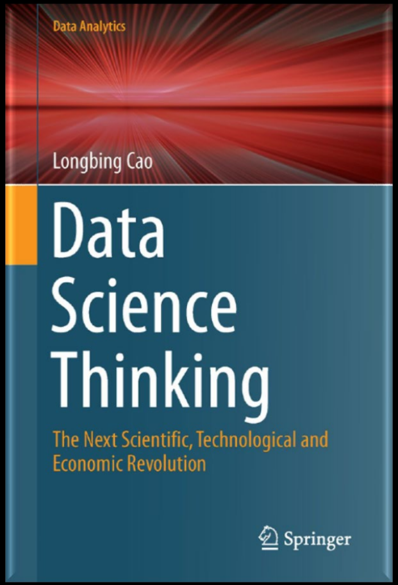
Trends vs. Controversies



L. Cao. IEEE Intelligent Systems, 2016

L. Cao. Communications of the ACM, 2017

L. Cao. IEEE Intelligent Systems, 2019

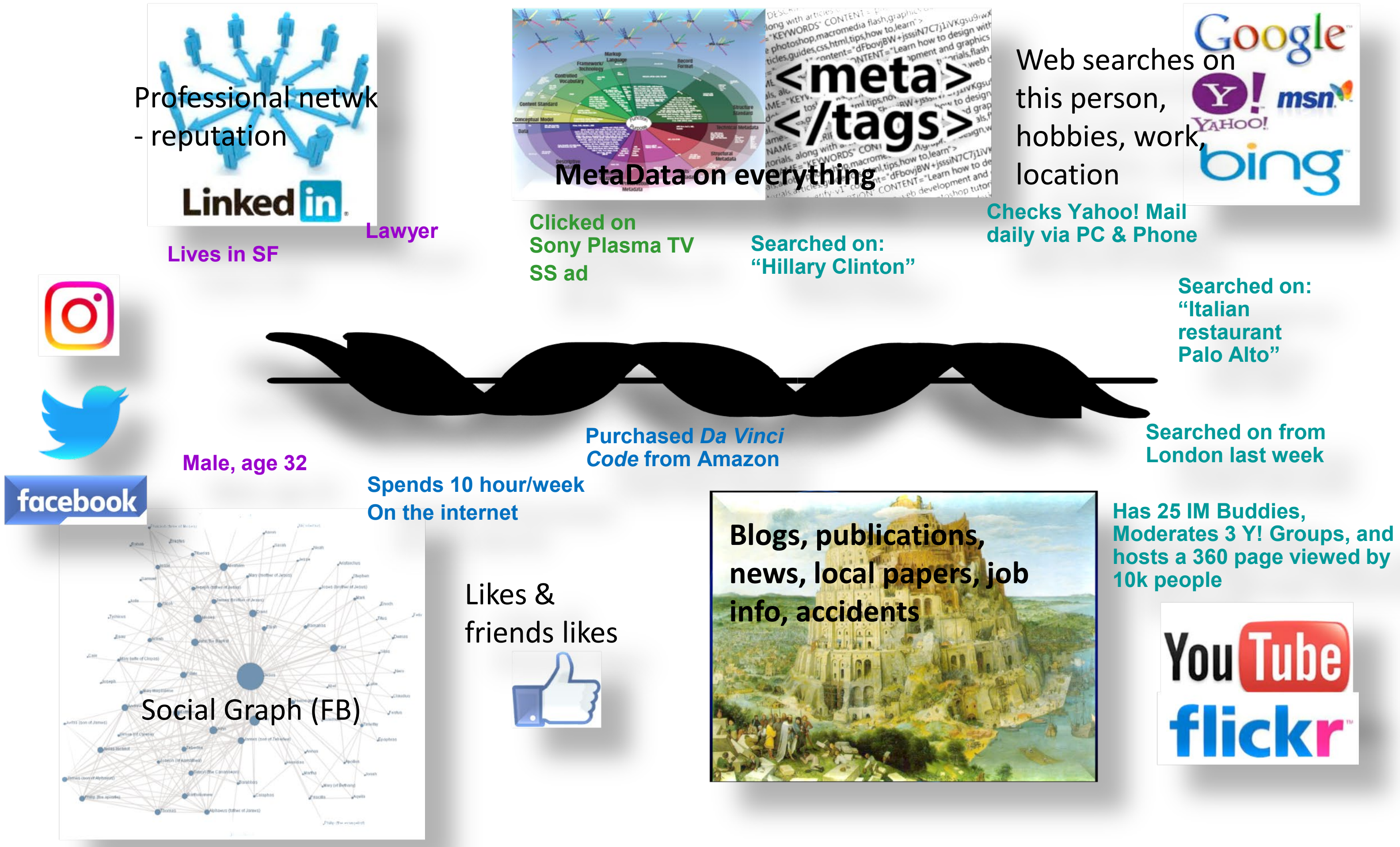


L. Cao. ACM Computing Survey, 2017

50 Years of data science
vs. immature data science discipline

D. Donoho, "50 Years of Data Science," 2015; http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Ubiquitous data silos
vs. Incomplete data DNA and data genomics

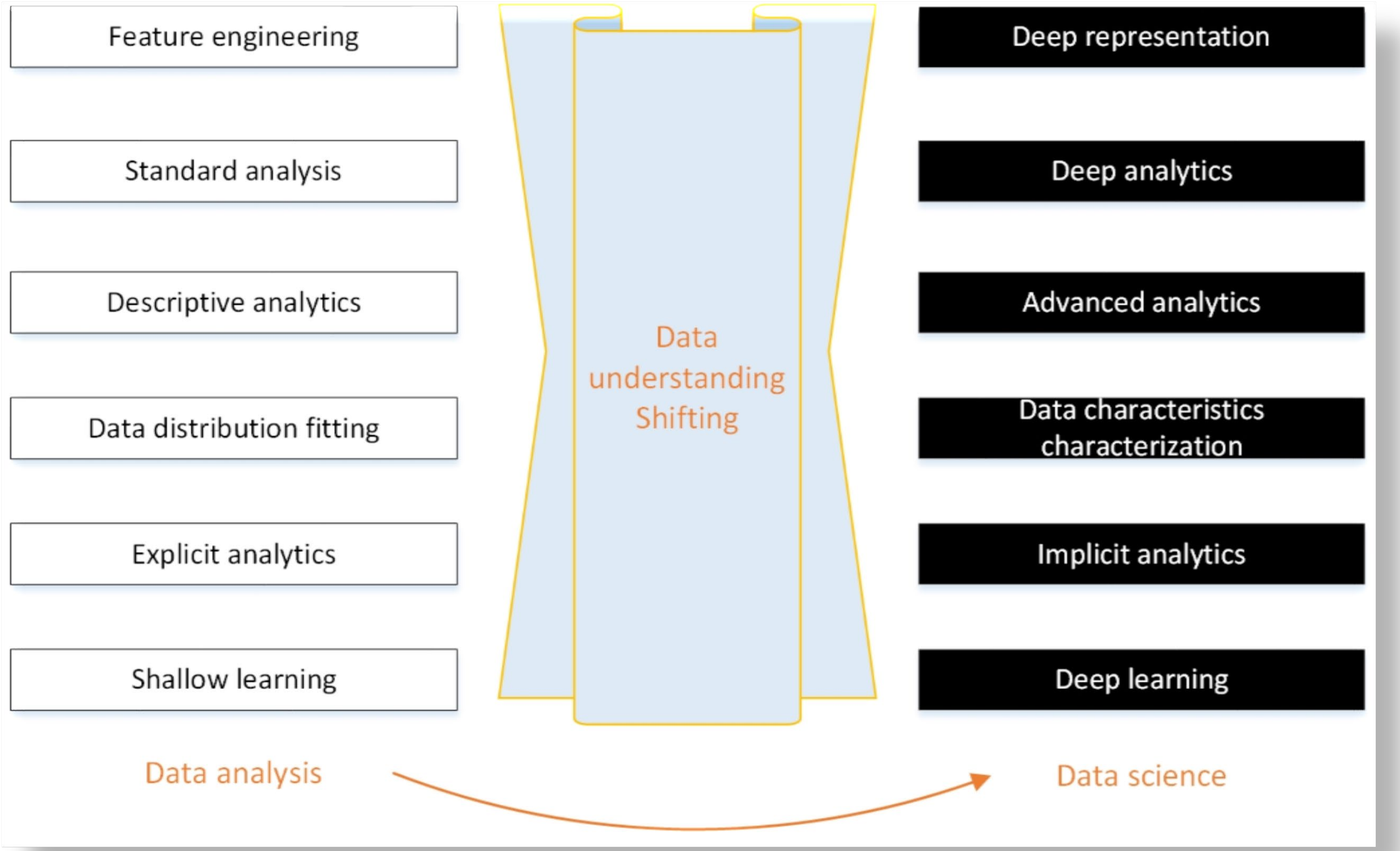


L. Cao. Data Science: A Comprehensive Overview, ACM Computing Survey, 2017

We have NOT built human and organizational data DNA/genomics

Data silos: every body, every organization, every where, every thing, every time, every behavior

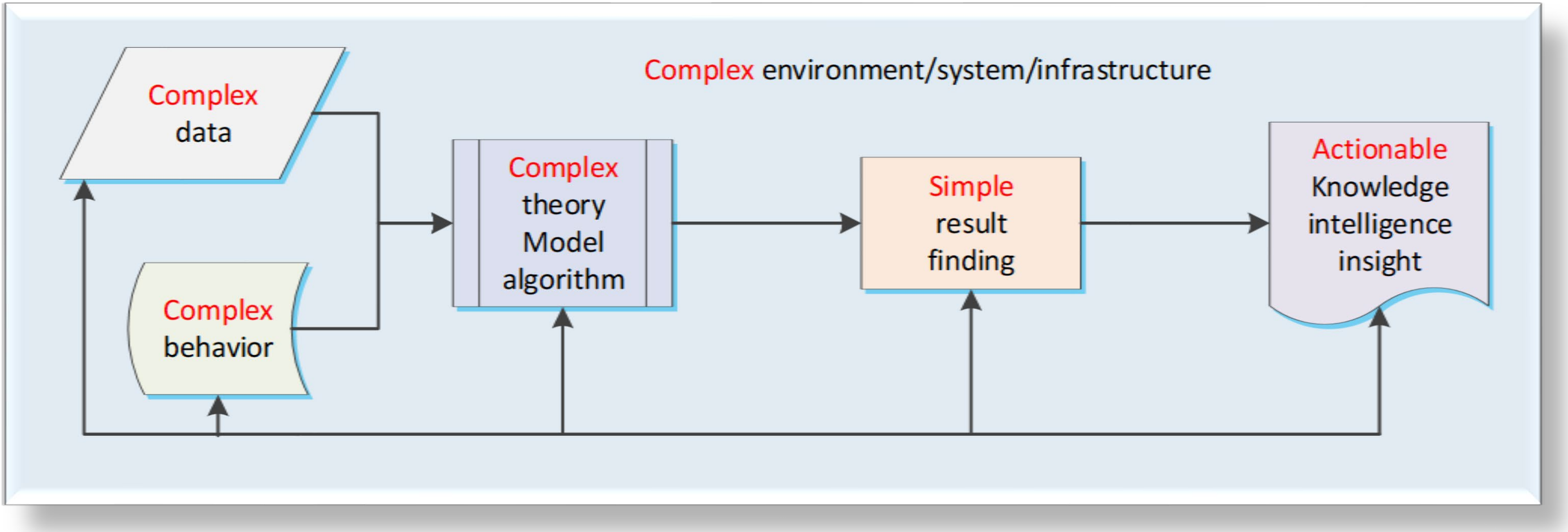
Paradigm shift: Well-developed data analysis → Immature data science



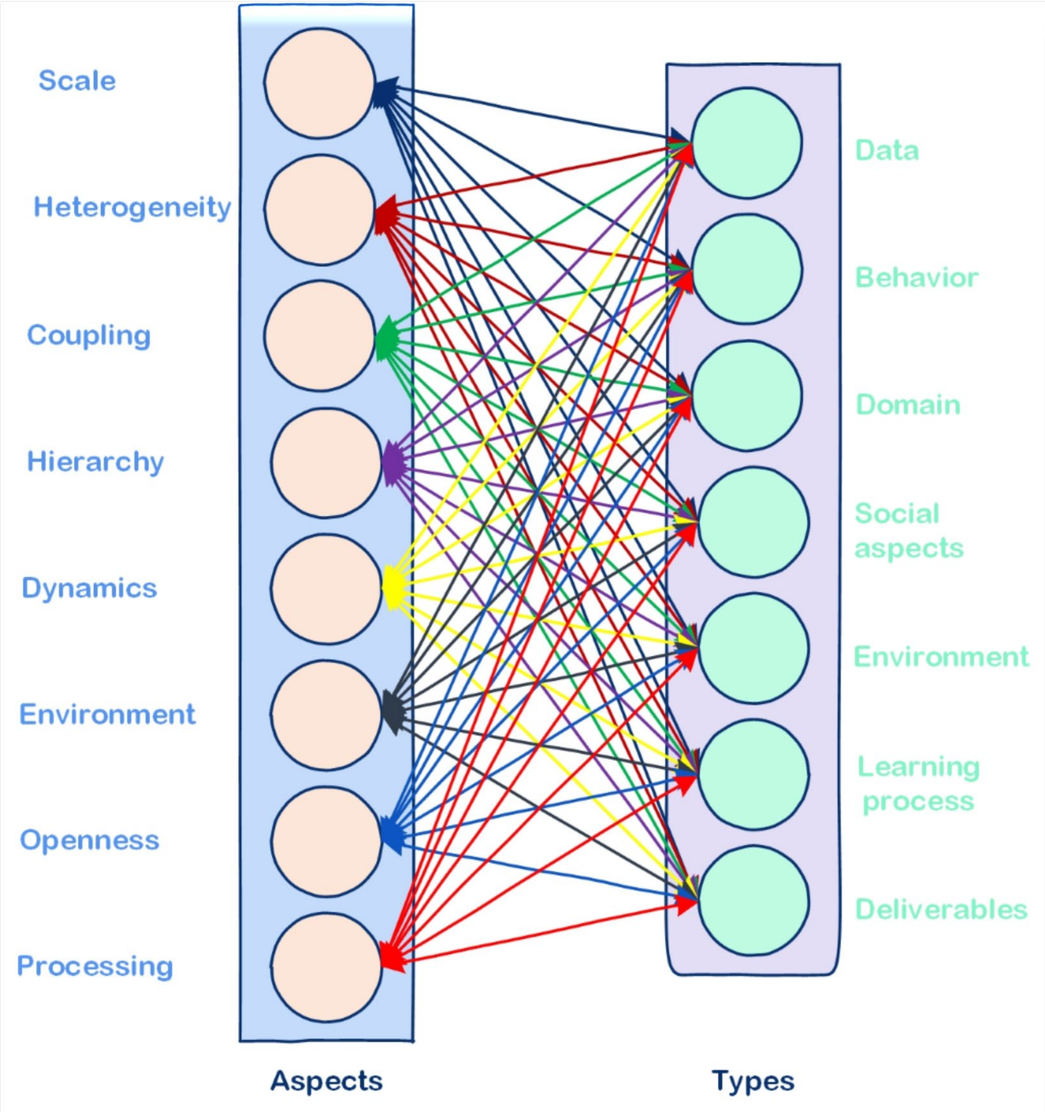
L. Cao. Data science thinking, Springer, 2018

Complex real world

vs. often simple, specific solutions and results



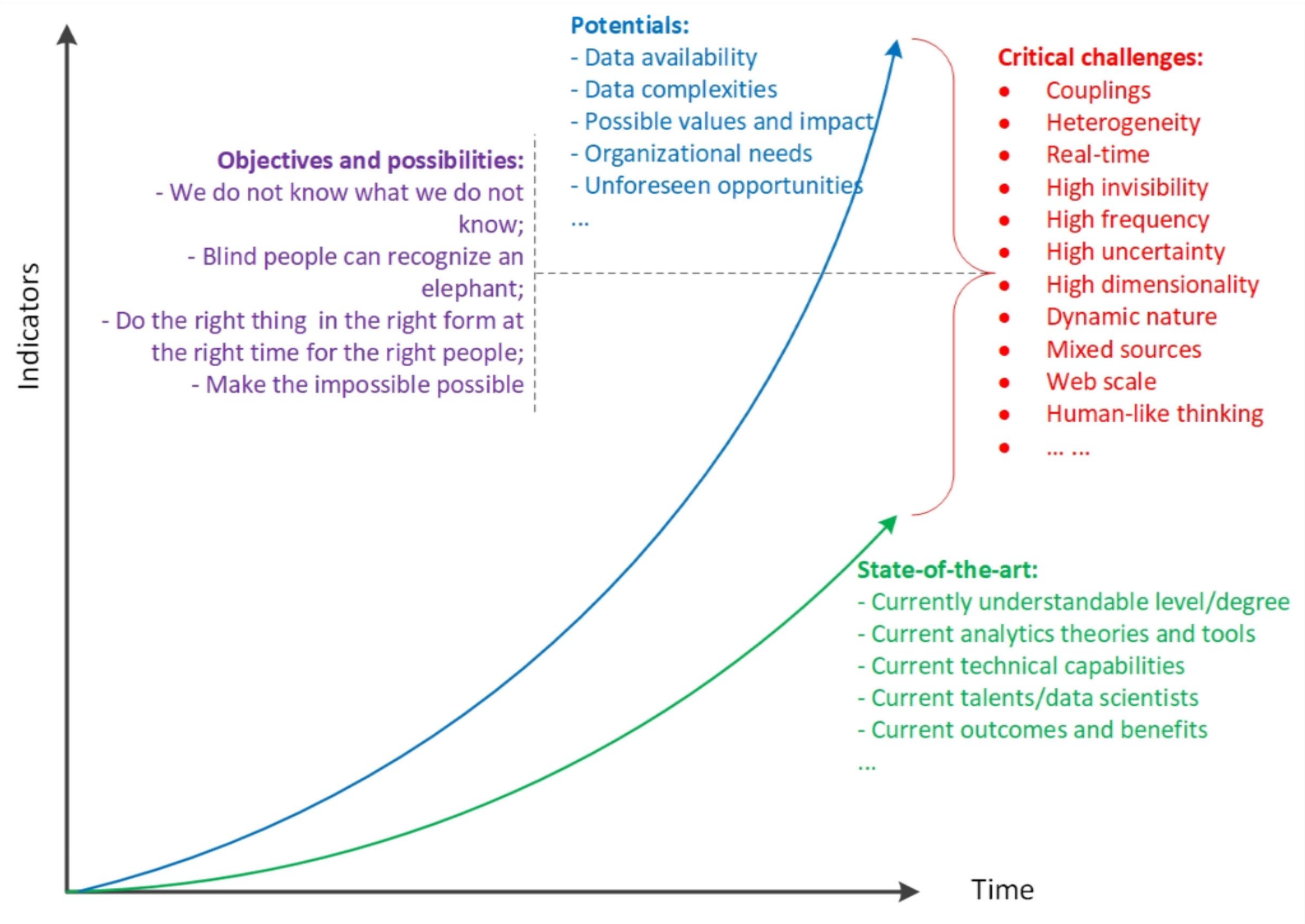
X-complexities and X-intelligences vs. Highly simplified assumptions



L. Cao, C. Zhang, R. Dai. [Intelligence Metasynthesis in Building Business Intelligence Systems], LNCS4845, 2007

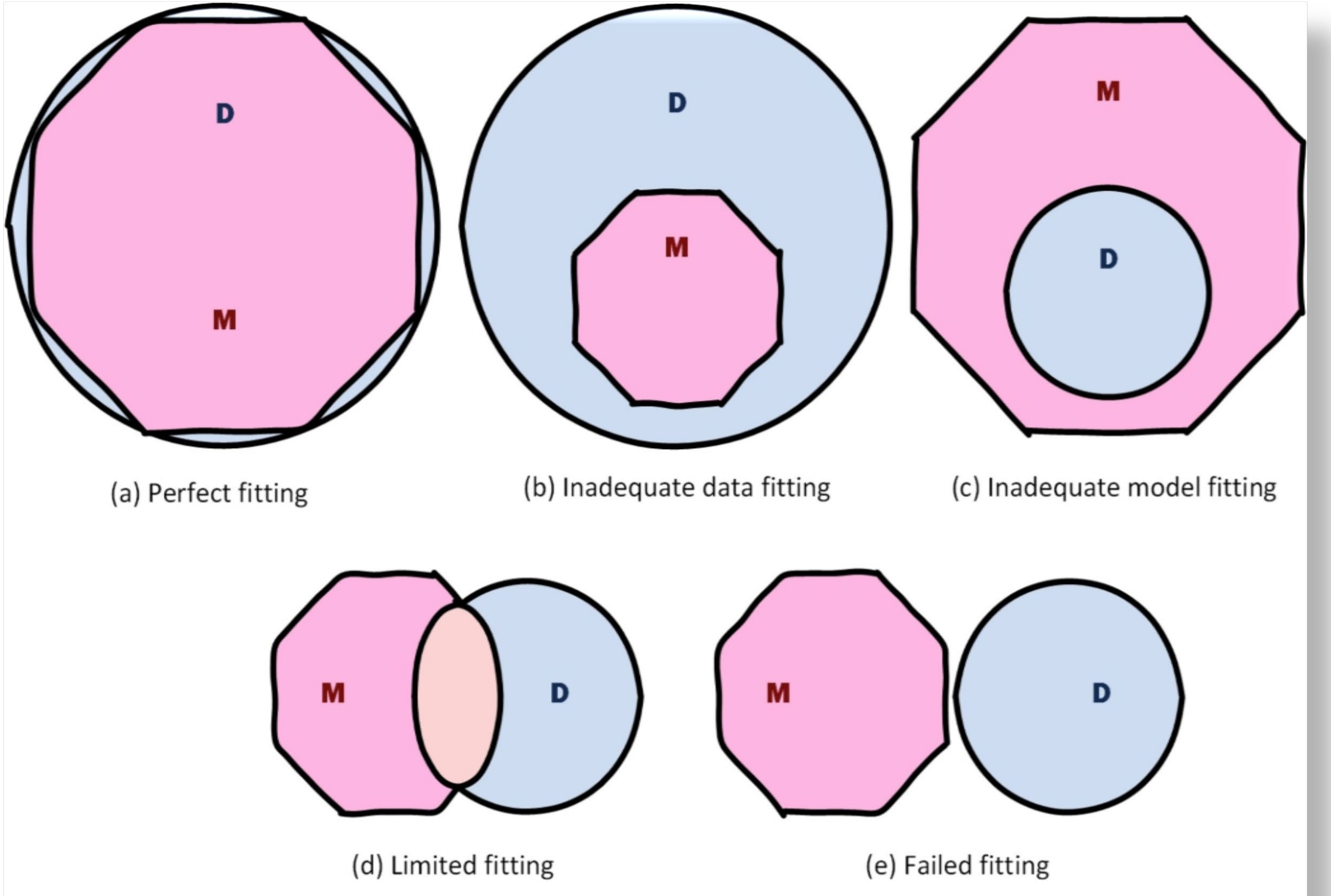
L. Cao. Data science: Challenges and directions, Communications of the ACM, 2017

Massive data potential vs. Significant capability/capacity gaps



L. Cao. Data science thinking, Springer, 2018

Fantastic theories and models
vs. Tailored data fitting and low actionability



L. Cao. Data science thinking, Springer, 2018

Manuscript

Click here to download Manuscript Berrar_ JDSA_Friedman_test_submitted.tex

Noname manuscript No.
(will be inserted by the editor)

Statistical comparison of machine learning algorithms: paradoxes, dilemmas, and open problems

Daniel Berrar

Int. J. Data Science and Analytics

Received: date / Accepted: date

Abstract The experimental comparison of machine learning algorithms is routinely underpinned by null hypothesis significance tests. When multiple classifiers are compared on multiple data sets, global null hypothesis tests are nowadays widely applied. The Friedman test has established itself as the method of choice for this purpose. Here, we analyze paradoxes, dilemmas, and open problems that this common practice entails. Our conclusion is that the Friedman test is not suitable for the statistical comparison of multiple classifiers over multiple data sets. Alternative methods for multiple testing are no solution, however, because the problem is a deeper one: the p-value is a recondite measure, and benchmark studies in machine learning would benefit from abandoning statistical significance.

Keywords Friedman test; p-value; significance test; paradoxes

1 Introduction

Significance tests have become firmly embedded in the minds and habits of machine learning researchers. Specifically, such tests nowadays routinely accompany comparative studies and are even sometimes stipulated in guidelines for reviewers. In arguably one of the most common experimental designs, several classifiers are compared based on their performance over multiple benchmark data sets. Here, the Friedman test has established itself as the method of choice to test the global null hypothesis that there is no difference in performance [17].

D. Berrar
Data Science Laboratory
Department of Information and Communications Engineering
Tokyo Institute of Technology, Japan
E-mail: daniel.berrar@ict.e.titech.ac.jp

We believe that the widespread popularity of such tests is due to a genuine desire of researchers to underpin the interpretation of their experimental studies with an objective, rigorous method as a safeguard against chance findings. However, there are a number of underrated paradoxes, dilemmas, and open problems that are due to this practice. Our most important results is that the widely used Friedman test is not suitable for the comparison of learning algorithms. We also argue that alternative omnibus tests are no solution, either, because the problem is a deeper one: the p-value is of very limited use for model evaluation and selection.

Arguments against the p-value have been made for decades, notably in psychology [44, 12, 46, 47, 23] and biomedicine [26, 43, 52]. The problem is not only that significance tests are frequently misused and p-values misinterpreted [27], but also that they are an impediment to cumulative scientific knowledge [46]. In 2016, the American Statistical Association (ASA) addressed the p-value problem, concluding with a set of guidelines for the proper use of p-values and significance tests [54]. The special issue "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$", published in The American Statistician in 2019, contains 43 papers on the p-value problem, but without converging on a consensus on the role of p-values in statistical inference [55]. Decades of criticisms of the p-value have had virtually no impact on the statistical practice in empirical research [11], and it is questionable whether the ASA statement will be able to improve the status quo [33]. The decision rule $p < 0.05$ is still almost always the decisive factor in the decision process of whether a study will or will not be accepted for publication [37].

Like many other sciences, the field of machine learning embraced the p-value in order to make statistical inferences [45, 18, 17]. Recently, however, the use of sig-

International Journal of Data Science and Analytics (2019) 7:247–257
https://doi.org/10.1007/s41060-018-0148-4

REGULAR PAPER

CrossMark

Should significance testing be abandoned in machine learning?

Daniel Berrar[1] · Werner Dubitzky[2]

Received: 12 April 2018 / Accepted: 26 July 2018 / Published online: 3 August 2018

Abstract
Significance testing has become a mainstay in machine learning, with the p value being firmly embedded in the current research practice. Significance tests are widely believed to lend scientific rigor to the interpretation of empirical findings; however, their problems have received only scant attention in the machine learning literature so far. Here, we investigate one particular problem, the *Jeffreys–Lindley paradox*. This paradox describes a statistical conundrum: the p value can be close to zero, convincing us that there is overwhelming evidence against the null hypothesis. At the same time, however, the posterior probability of the null hypothesis being true can be close to 1, convincing us of the exact opposite. In experiments with synthetic data sets and a subsequent thought experiment, we demonstrate that this paradox can have severe repercussions for the comparison of multiple classifiers over multiple benchmark data sets. Our main result suggests that significance tests should not be used in such comparative studies. We caution that the reliance on significance tests might lead to a situation that is similar to the reproducibility crisis in other fields of science. We offer for debate four avenues that might alleviate the looming crisis.

Keywords Jeffreys–Lindley paradox · p Value · Significance test · Bayesian test · Classification

1 Introduction

Significance testing is increasingly used in machine learning and data science, particularly in the context of comparative classification studies [9]. For example, the Friedman test has been widely used for comparing multiple classifiers over multiple data sets [18]. Suppose that we wish to compare a new classifier with three other classifiers. Let us assume that we compare their performance over 50 benchmark data sets. We use the Friedman test to test the global null hypothesis of equal performance between the four classifiers. Suppose that

This paper is an extended version of the DSAA2017 Research Track paper titled "On the Jeffreys–Lindley paradox and the looming reproducibility crisis in machine learning" [11].

✉ Daniel Berrar
daniel.berrar@ict.e.titech.ac.jp

Werner Dubitzky
werner.dubitzky@helmholtz-muenchen.de

1 Data Science Laboratory, Department of Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, Japan

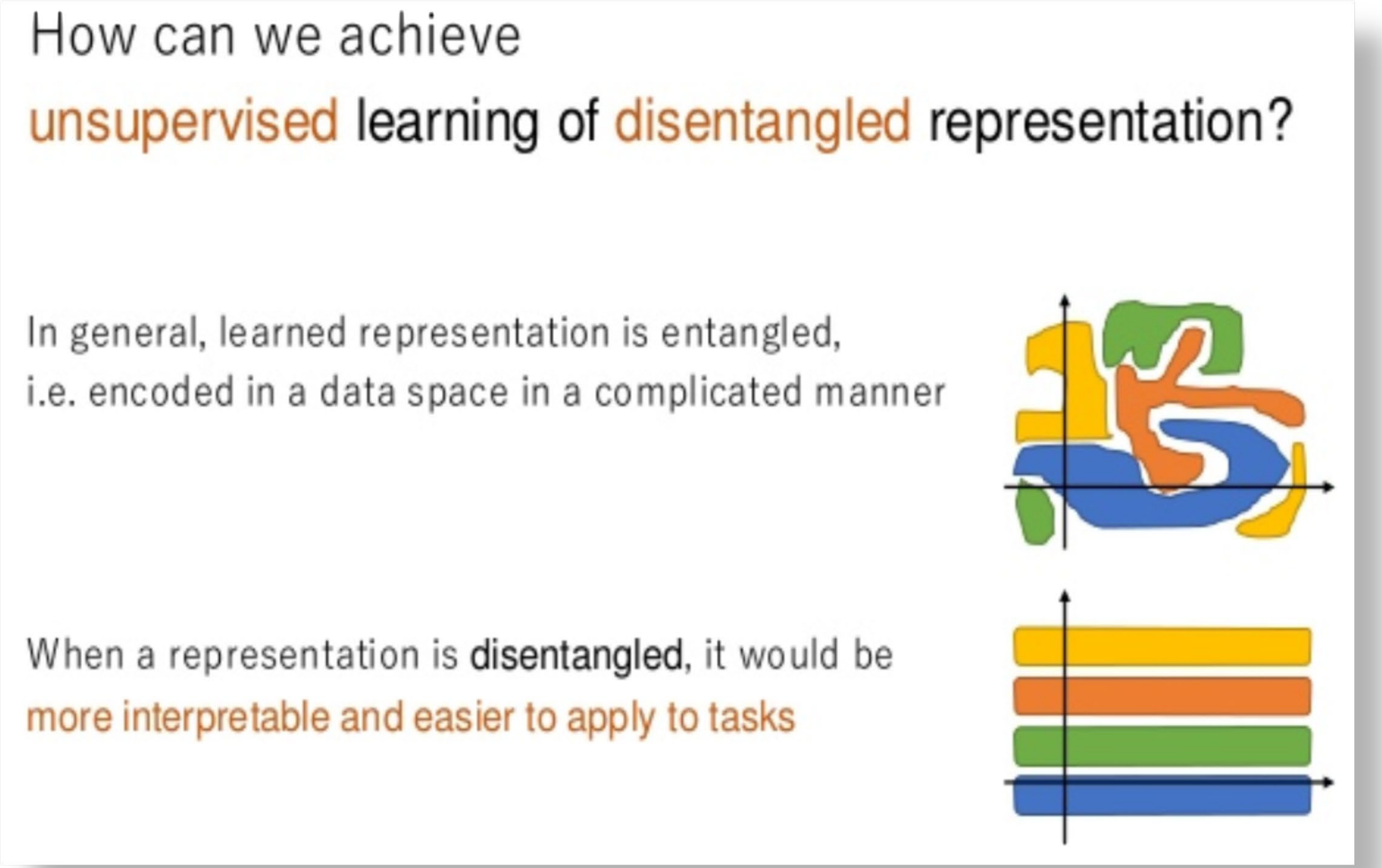
2 Research Unit Scientific Computing, German Research Center for Environmental Health, Helmholtz Zentrum München, Munich, Germany

we obtain a p value of 0.001. How should we interpret this result? We would like to invite the reader to briefly ponder over this question.

The question might seem silly, as the answer seems all too obvious: "Reject the null hypothesis of equal performance." But is this the correct interpretation? As we will discuss, the answer to this question is far more complicated than it seems. Paradoxically, the p value can be close to 0, yet the posterior probability in favor of the null hypothesis can be close to 1. In other words, it is possible to obtain a very small p value, but the evidence after the experiment can convince us that the null hypothesis is almost cer was first observed by Jeffrey conundrum in his seminal pap it has since become widely k or *Jeffreys–Lindley paradox*.

The statistical literature co dox; however, there is no c scientific communication [1 problem in the context of co ies [11]. Here, we report the The goal of the present study of the paradox for machine le tistical evaluation of learning

Coupled/entangled nature/realities vs. decoupled and disentangled representations

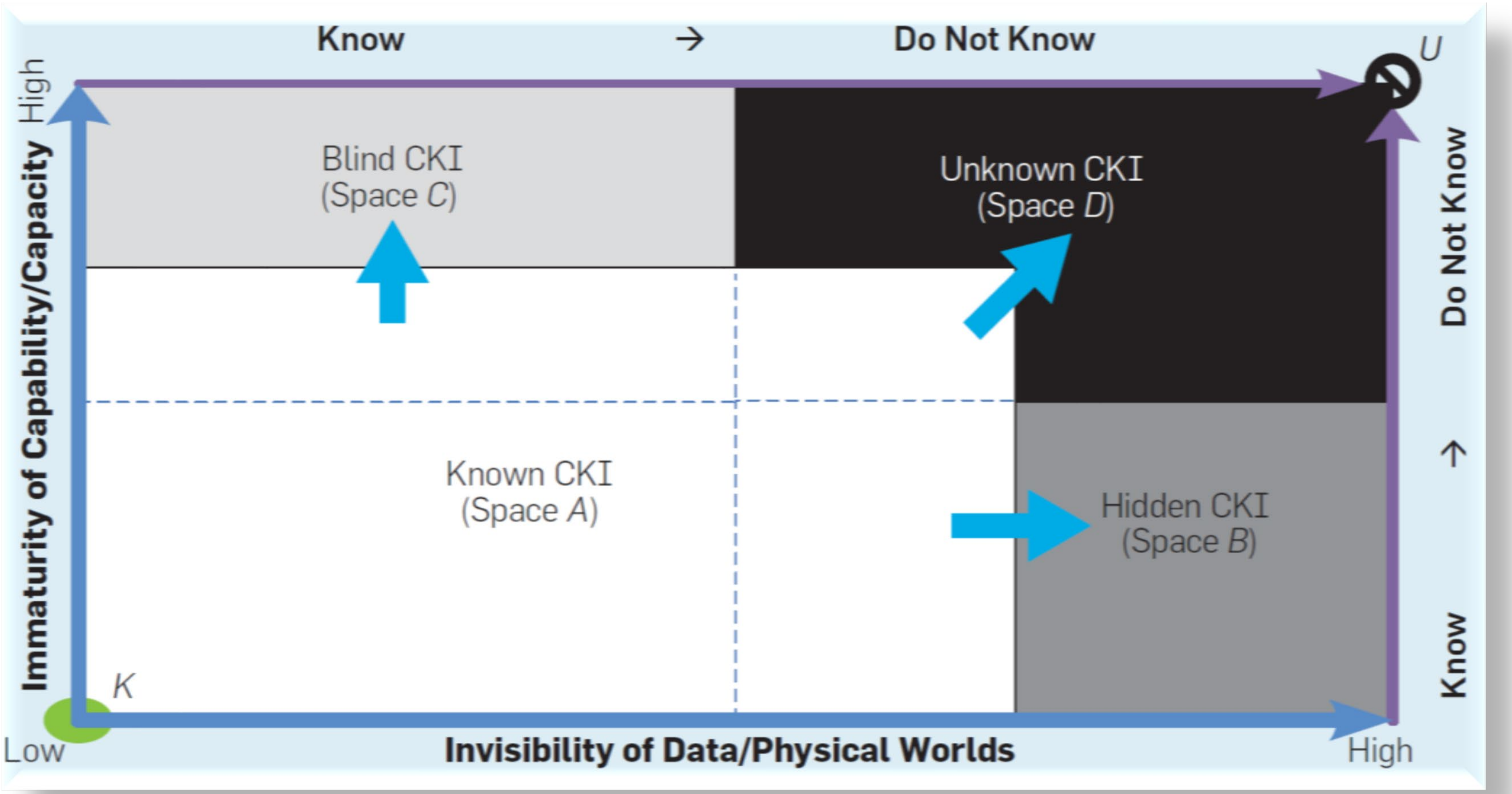


Generative Adversarial Networks (GANs) and Disentangled Representations, NeurIPS2018

Couplings in real-life data, behaviors and systems:

- Value couplings
- Feature couplings
- Relation couplings
- Structure couplings
- Distribution couplings
- Object couplings
- Ensembled model couplings
- Objective couplings
- Result couplings

The status has not been fundamentally changed: We do not know what we do not know



L. Cao. Data science: Challenges and directions, Communications of the ACM, 2017

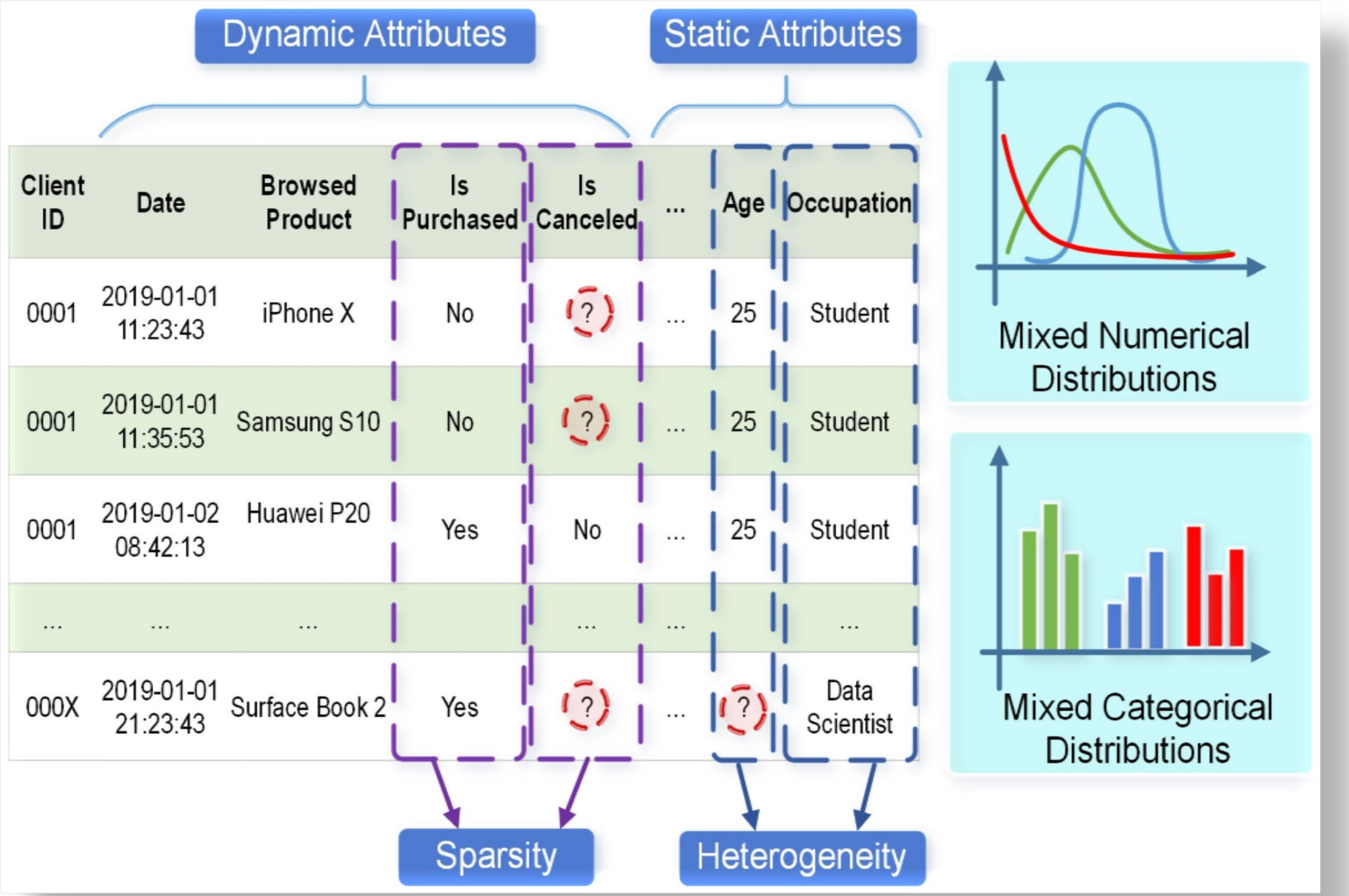
Data science and New-generation AI: The unknown world

L. Cao. Data science: Challenges and directions, Communications of the ACM, 2017

One Specific Challenge

Non-IID Data, Behaviors and Systems

Data/behavior/system non-IIDness
vs. IID assumptions and learning systems



Real-life data/behavior/systems:

- Low quality:
  - Sparsity
  - Imbalanced
  - Noisy
  - Redundant
- Interactive and coupled:
  - Interactive vs. relational
  - Coupled vs. disentangled
  - M*couplings
- Heterogeneous and mixed:
  - Distributions
  - Structures
  - Interactions/couplings
  - Static and dynamic

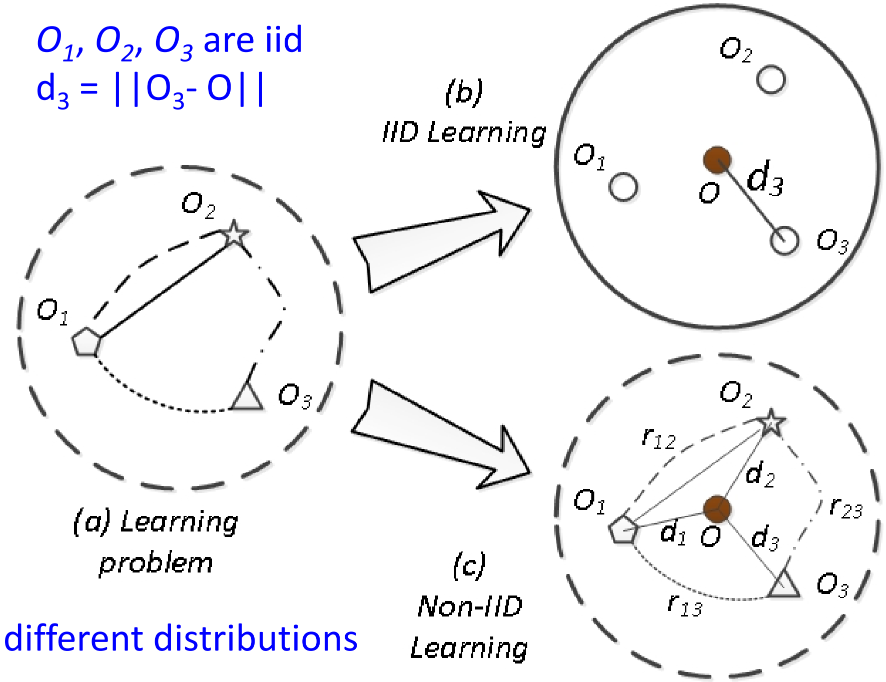
Non-IID Learning

Tutorials: CIKM/KDD/IJCAI tutorials

Website: noniid.datasciences.org

Non-IID Learning: fundamental yet challenging

O_1, O_2, O_3 are iid
$d_3 = ||O_3 - O||$



O_1, O_2, O_3 share different distributions
$d_3 = ||O_3 - O||$
$= ||\ O_3(r_{13}, r_{23}) - O(d_1, d_2)\ ||$

IIDness:

Independence +

Identical Distribution

Non-IIDness:

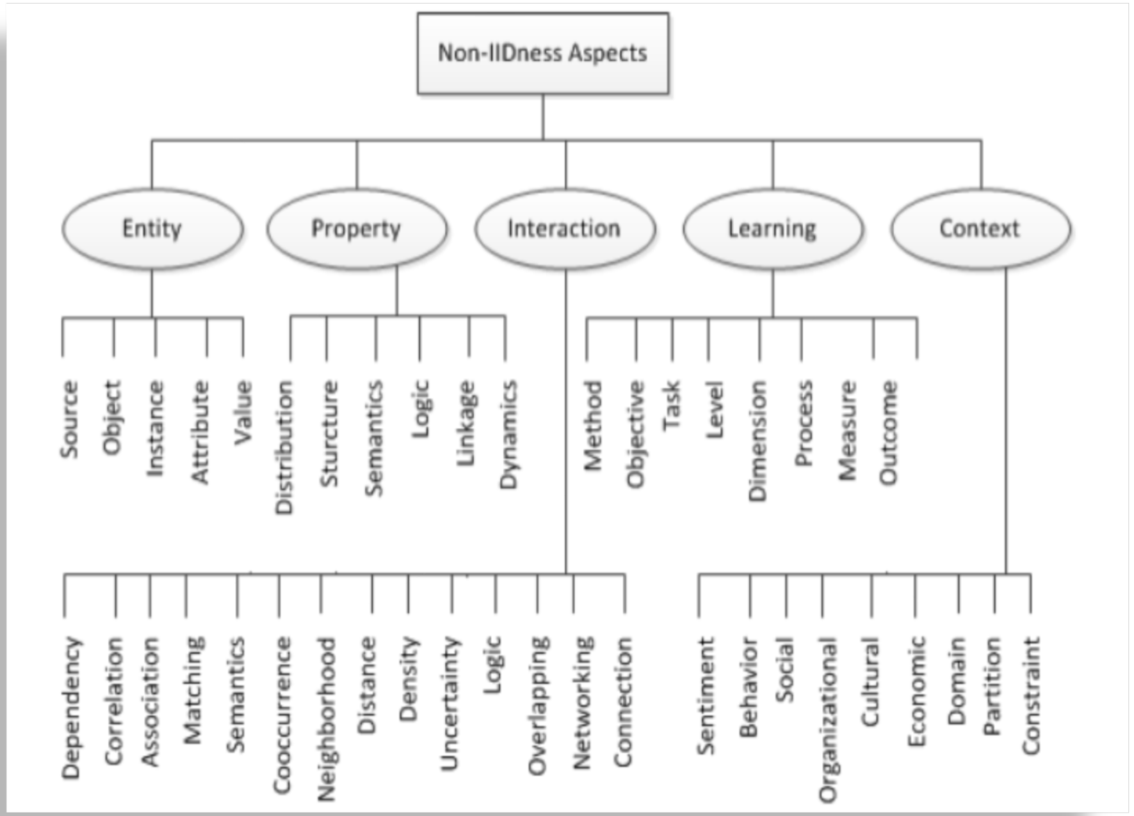
Couplings +

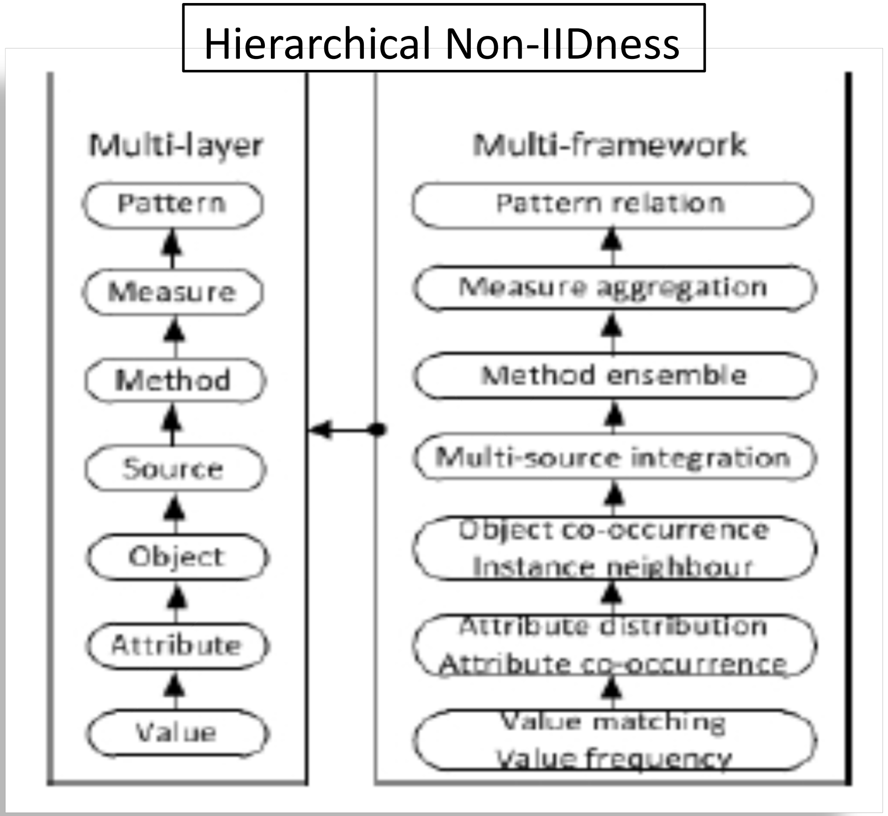
Heterogeneities

IID learning dominates classic analytics and learning in AI, KDD, ML, CVPR, and statistics research and methods

Non-IID power: Rich aspects of non-IIDness

Non-IIDness does not limit itself to statistical dependency and non-identical distributions

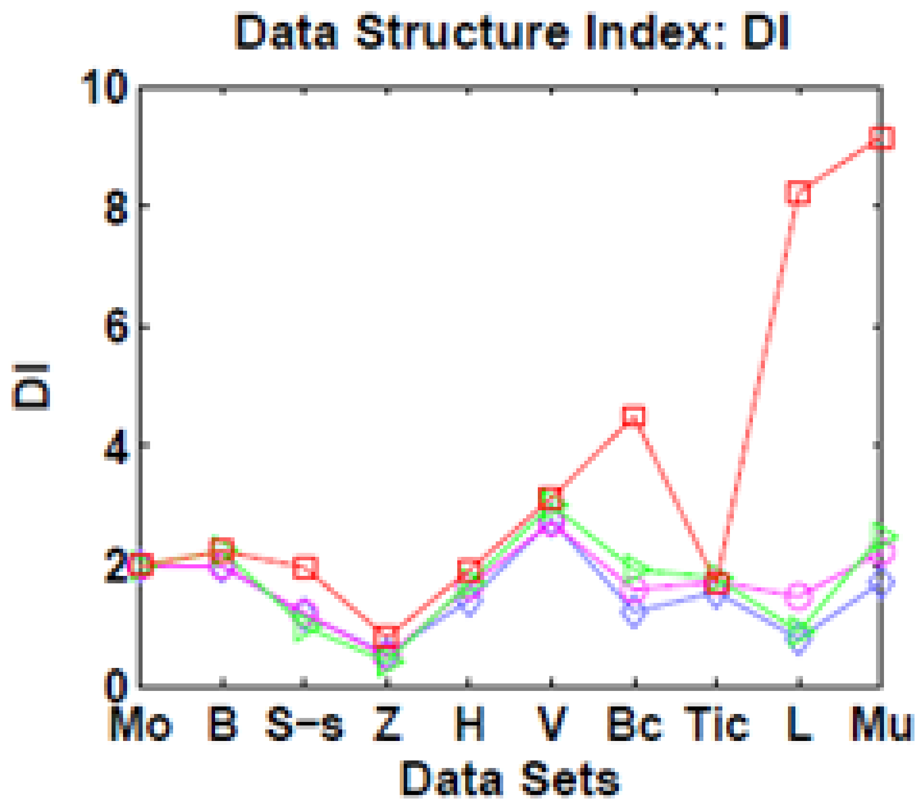




Cao, Longbing. *Coupling Learning of Complex Interactions*, IP&M, 51(2): 167-186 (2015)

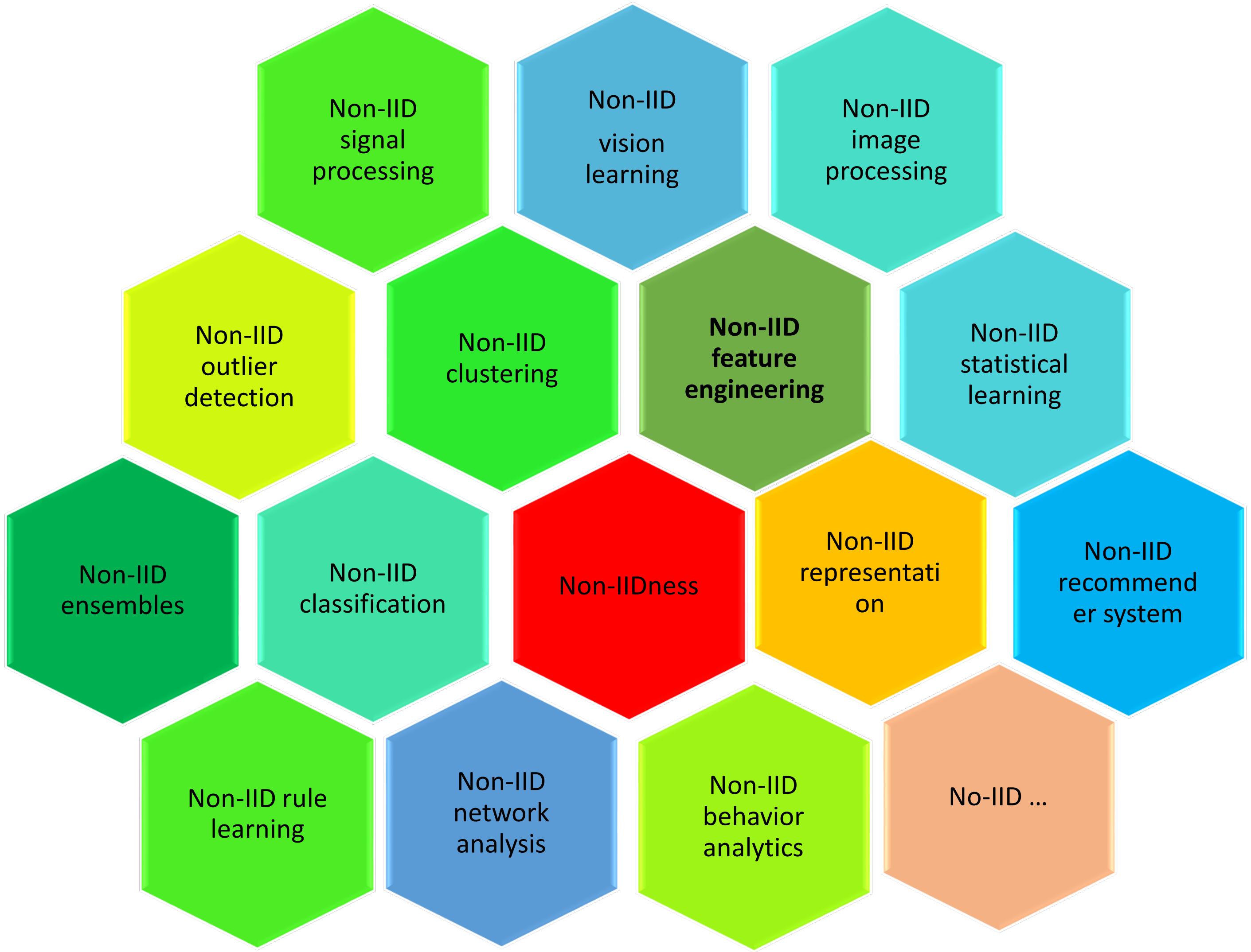
IID Risk: Problems of IID learning and results

- Results learned by IID analytical/learning methods and algorithms on non-IID data could be:
  - suboptimal
  - incomplete
  - biased,
  - misleading
  - incorrect



C. Wang, et al. [Coupled Attribute Similarity Learning on Categorical Data](), IEEE Transactions on Neural Networks and Learning Systems, 26(4): 781-797 (2015)

Non-IID Learning: A Significant Area



Non-IID paradigm

Real-world data, behavior and systems are non-IID, requiring a non-IID paradigm to understand:

- Data/behavior/system non-IIDness
- Non-IID similarity/dissimilarity metrics/measures
- Non-IID representations
- Non-IID learning systems
- Non-IID objective functions
- Non-IID optimization theory
- Non-IID inference theory
- New perspectives …

Non-IID Metric Learning

C. Zhu, L. Cao, Q. Liu, J. Yin and V. Kumar. Heterogeneous Metric Learning of Categorical Data with Hierarchical Couplings. TKDE, 2018.

Motivation

Name	Gender	Performance	Commitment	Class
John	M	A	H	c1
Mary	F	B	H	c1
Sarah	F	B	I	c1
David	M	C	L	c1
Alice	F	C	I	c2
Edward	M	D	L	c2

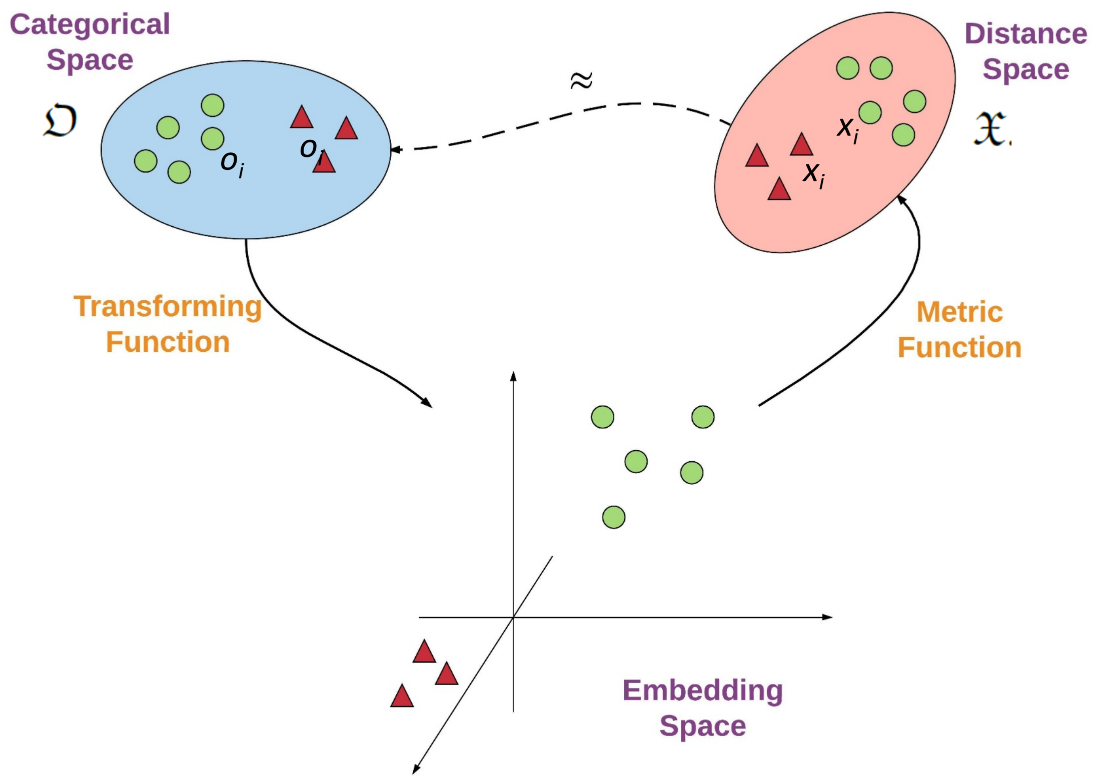
Hamming distance: Dis(H,I) = Dis(H,L) = 1

High (H) level commitment is closer to intermediate (I) instead of low (L) level.

Frequency-based distance: Dis(H, I) = 0

H commitment is different from I.

Problem statement

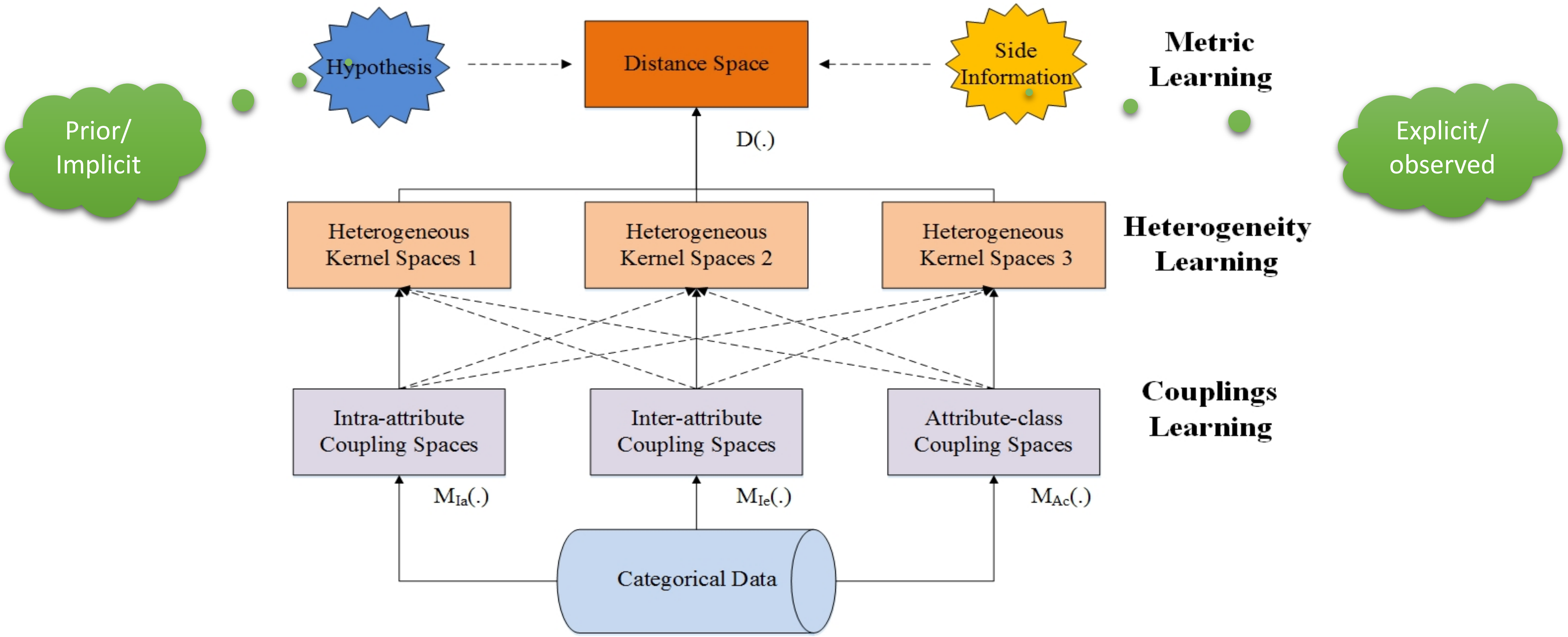


$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \widetilde{Div}(\mathfrak{O}||\mathfrak{X}) \\ \text{subject to} \quad & \mathsf{o} \sim \mathfrak{O} \\ & \mathbf{x} \sim \mathfrak{X} \\ & d(\mathsf{o}_i, \mathsf{o}_j) = \mathbf{x}_i \odot \mathbf{x}_j. \end{aligned}$$

Distance metric d(., .) satisfies:

1) $d(\mathsf{o}_i, \mathsf{o}_j) + d(\mathsf{o}_j, \mathsf{o}_k) \geq d(\mathsf{o}_i, \mathsf{o}_k)$,
2) $d(\mathsf{o}_i, \mathsf{o}_j) \geq 0$,
3) $d(\mathsf{o}_i, \mathsf{o}_j) = d(\mathsf{o}_j, \mathsf{o}_i)$.

The HELIC framework: A multikernel approach



HELIC: Heterogeneous Metric Learning with hIerarchical Couplings

Coupling learning: Value-to-class couplings

Learning Intra-attribute Couplings

$$m_{Ia}^{(j)}(\mathsf{v}_i^{(j)}) = \frac{|g^{(j)}(\mathsf{v}_i^{(j)})|}{n_o}.$$

Capture value frequency

Learning Inter-attribute Couplings

Capture value co-occurrence

$$m_{Ie}^{(j)}(\mathsf{v}_i^{(j)}) = \left[\; p(\mathsf{v}_i^{(j)}|\mathsf{v}_{*1}), \quad \cdots, \quad p(\mathsf{v}_i^{(j)}|\mathsf{v}_{*|V_*|}) \;\right]^\top$$

Learning Attribute-class Couplings

Capture value distribution in each class

$$m_{Ac}^{(j)}(\mathsf{v}_i^{(j)}) = \left[\; p(\mathsf{v}_i^{(j)}|c_1) \quad \cdots \quad p(\mathsf{v}_i^{(j)}|c_{n_c}) \;\right]^\top$$

Heterogeneity learning: Distributions, structures, couplings, etc.

Construct Kernel Spaces:

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{m}_1, \mathbf{m}_1) & k(\mathbf{m}_1, \mathbf{m}_2) & \cdots & k(\mathbf{m}_1, \mathbf{m}_{n_v^{(j)}}) \\ k(\mathbf{m}_2, \mathbf{m}_1) & k(\mathbf{m}_2, \mathbf{m}_2) & \cdots & k(\mathbf{m}_2, \mathbf{m}_{n_v^{(j)}}) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{m}_{n_v^{(j)}}, \mathbf{m}_1) & k(\mathbf{m}_{n_v^{(j)}}, \mathbf{m}_2) & \cdots & k(\mathbf{m}_{n_v^{(j)}}, \mathbf{m}_{n_v^{(j)}}) \end{bmatrix}$$

Using various kernel functions for the value-to-class coupling spaces, a set of kernel matrices $\{\mathbf{K_1}, \cdots, \mathbf{K_{n_k}}\}$ can be obtained. Further, a set of transformation matrices $\{\mathbf{T_1}, \cdots, \mathbf{T_{n_k}}\}$ can be learned to guarantee that the space of the p-th transformed kernel $\mathbf{K}'_p$ only contains the p-th kernel sensitive information, where the $\mathbf{K}'_p$ is defined as:

$$\mathbf{K}'_p = \mathbf{T}_p \cdot \mathbf{K}_p$$

Metric learning

With a positive semi-definite matrix $\omega_p = \alpha_p \mathbf{T}_p^\top \mathbf{T}_p$, the metric d_{ij} is calculated as :

$$d_{ij} = \sum_{p=1}^{n_k} \mathbf{k}_{p,ij}^\top \boldsymbol{\omega}_p \mathbf{k}_{p,ij}$$

where $\mathbf{k}_{p,ij} = \mathbf{K}_{p,i\cdot} - \mathbf{K}_{p,j\cdot}$

The distance can be represented as

$$\boldsymbol{\omega} = \begin{bmatrix} \boldsymbol{\omega}_1^{\mathrm{diag}} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\omega}_2^{\mathrm{diag}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\omega}_{n_k}^{\mathrm{diag}} \end{bmatrix}$$

$$d_{ij} = \sum_{p=1}^{n_k} \mathbf{k}_{p,ij}^\top \boldsymbol{\omega}_p \mathbf{k}_{p,ij}$$

$$\mathbf{k}_{ij} = \begin{bmatrix} \mathbf{k}_{1,ij}^\top & \mathbf{k}_{2,ij}^\top & \cdots & \mathbf{k}_{n_k,ij}^\top \end{bmatrix}^\top$$

Metric learning: Objective function

Objective function:

$$\begin{aligned}\underset{\boldsymbol{\omega}, b}{\text{minimize}} \quad & \frac{1}{n_o^2} \sum_{i,j \in N_o} \xi_{ij} + \underline{\lambda \|\boldsymbol{\omega}\|_1} \\ \text{subject to} \quad & \boldsymbol{\omega} \succcurlyeq 0, \\ & \boldsymbol{\omega}_{kl} = 0 \quad for \quad k \neq l, \\ & \underline{1 + r_{ij}(\mathbf{k}_{ij}^\top \boldsymbol{\omega} \mathbf{k}_{ij} - b) \leqslant \xi_{ij}} \\ & \xi_{ij} \geqslant 0, \forall i, j \in N_o.\end{aligned}$$

$$r_{ij} = \begin{cases} 1, & c(\mathrm{o}_i) = c(\mathrm{o}_j) \\ -1, & c(\mathrm{o}_i) \neq c(\mathrm{o}_j) \end{cases}$$

Selecting the kernels for their sensitive data distribution

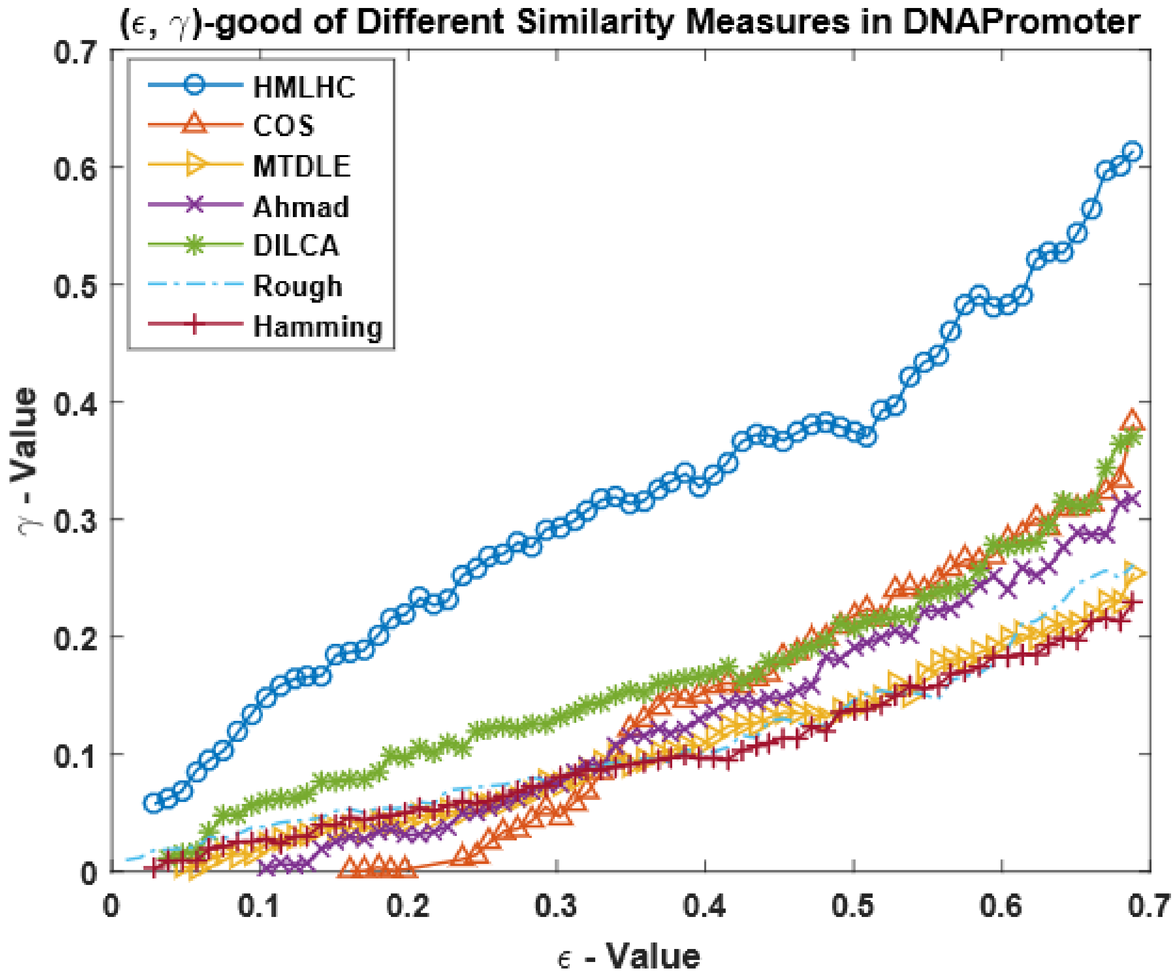
Force the distance between objects from different classes larger than a margin

Representation performance of HELIC

KNN Classification F-score (%) with Different Distance Measures

Data	HELIC	COS	MTDLE	Ahmad	DILCA	Rough	Hamming	Δ%
Zoo	**100***	**100***	**100***	**100***	**100***	97.75±11.11	**100***	0.00%
DNAPromoter	**92.90±5.85***	75.89±13.35	81.67±10.19	79.98±9.14	90.33±10.31	81.16±10.30	78.05±12.00	2.85%
Hayesroth	**90.85±5.07***	79.64±9.71	68.54±10.55	52.26±10.20	54.60±12.58	81.50±8.59	61.73±12.40	11.47%
Audiology	**75.44±7.60***	41.51±7.20	36.70±7.50	54.29±8.96	64.83±8.04	36.37±7.60	58.55±10.30	16.36%
Housevotes	**96.65 ± 3.40**	94.28 ± 4.95	91.09 ± 5.55	95.81 ± 4.15	94.90 ± 4.14	91.59 ± 5.14	93.77 ± 5.30	0.88%
Spect	**53.09 ±10.35***	51.31±9.16*	52.94±9.48*	52.70±9.69*	51.11±8.97*	51.18±7.90*	51.98±8.85*	0.28%
Mofn3710	**94.39 ±5.86***	79.35±9.07	68.74±10.58	79.35±9.07	71.21±8.42	77.70±11.44	74.82±8.08	18.95%
Monks3	**100***	34.85±0.00	99.88±0.52*	34.85±0.00	34.85±0.00	**100***	92.06±5.24	0.00%
ThreeOf9	**91.01 ±2.93***	32.00±0.00	75.88±8.41	32.00±0.00	32.00±0.00	78.84±5.09	78.84±5.09	15.44%
Balance	**58.91 ±1.31***	21.25±0.00	41.80±5.82	21.25±0.00	21.25±0.00	39.32±4.25	39.32±4.25	40.93%
Crx	**83.26±5.68***	78.58±4.74	77.54±5.68	82.79 ±3.86*	81.02±4.08	77.63±5.12	78.28±4.87	0.57%
Mammographic	**79.61 ±4.59***	70.22±7.12*	70.14±7.10*	70.20±7.02*	70.22±7.81*	69.79±7.11 *	69.95±7.29*	13.37%
Flare	**59.88 ± 3.36***	57.01 ± 4.38*	57.11 ± 3.09	54.41 ± 3.39	55.61 ± 3.13	55.88 ± 4.38	54.98 ± 4.00	4.85%
Titanic	**23.33 ± 2.48***	10.54 ± 1.76	10.06 ± 0.62	10.06 ± 0.99	10.54 ± 1.76	10.54 ± 1.76	10.54 ± 1.76	32.48 %
DNAnominal	**93.12 ± 1.05***	77.52 ± 1.21	52.22 ± 0.00	80.33 ± 1.48	91.65 ± 1.39	81.46 ± 1.75	69.11 ± 1.45	1.60 %
Splice	**93.69 ± 1.11***	77.25 ± 2.19	24.45 ± 0.00	79.85 ± 2.07	84.96 ± 2.21	81.05 ± 1.81	69.29 ± 2.24	10.28 %
Krvskp	**96.98 ± 1.06***	91.77 ± 1.66	90.04 ± 1.65	92.46 ± 1.74	91.39 ± 2.05	89.00 ± 1.43	91.48 ± 1.68	4.89%
Led24	**63.37 ± 1.94***	62.11 ± 1.85*	41.35 ± 2.74	61.81 ± 1.98*	62.58 ± 1.85*	47.89 ± 2.37	41.57 ± 2.19	1.26 %
Mushroom	**100 ± 0.00***	99.98 ± 0.06*	**100 ± 0.00***	**100 ± 0.00 ***	**100 ± 0.00***	**100 ± 0.00 ***	**100 ± 0.00***	0.00%
Krkopt	**53.62 ± 1.71***	52.66 ± 0.78*	NA	52.50 ± 0.96*	52.57 ± 1.02*	39.05 ± 0.70	10.42 ± 0.10	1.82%
Adult	**84.91 ± 0.86***	68.13 ± 1.12	NA	68.20 ± 1.07	68.16 ± 1.14	67.76 ± 1.04	68.01 ± 1.04	24.50%
Connect4	**56.33 ± 0.78***	48.23 ± 0.73	NA	46.95 ± 0.49	46.65 ± 0.55	53.22 ± 0.73	45.81 ± 0.72	5.84%
Census	**68.93 ± 0.55***	66.88 ± 0.40	NA	67.47 ± 0.43	66.66 ± 0.42	66.96 ± 0.55	67.16 ± 0.37	2.64%
Mean	**78.71***	63.95	65.27	63.89	65.09	68.51	65.47	14.89%

Representation quality of HELIC



Classification performance of HELIC

KNN Classification F-score (%) with Couplings

Dataset	HELIC-KNN	HC-KNN	Δ%
Zoo	100	100	0%
DNAPromoter	92.90±5.85	94.93±7.00	0%
Hayesroth	90.85±5.07	85.89±6.39	5.77%
Audiology	75.44±7.60	54.94±11.85	37.31%
Housevotes	96.65 ± 3.40	95.43 ± 4.46	1.28%
Spect	53.09±10.35	51.40±9.51	3.28%
Mofn3710	94.39±5.86	94.92±3.36	0%
Monks3	100	100	0%
ThreeOf9	91.01±2.93	89.96±2.92	1.17%
Balance	58.91±1.31	59.64±1.46	0%
Crx	83.26±5.68	82.43±4.39	1.01%
Mammographic	79.61±4.59	70.31±7.00	13.23%
Flare	59.88 ± 3.36	55.40 ± 3.93	8.09%
Titanic	23.33 ± 2.48	12.15 ± 1.65	92.02%
DNAnominal	93.12 ± 1.05	91.83 ± 1.64	1.40%
Splice	93.69 ± 1.11	75.88 ± 2.03	23.47%
Krvskp	96.98 ± 1.06	92.49 ± 0.92	4.85%
Led24	63.37 ± 1.94	57.71 ± 2.46	9.81%
Mushroom	100 ± 0.00	100 ± 0.00	0.00%
Krkopt	53.62 ± 1.71	52.44 ± 1.58	2.25%
Adult	84.91 ± 0.86	84.32 ± 0.80	0.70%
Connect4	56.33 ± 0.78	43.07± 0.50	30.79%
Census	68.93 ± 0.55	64.23 ± 0.49	7.32%
Mean	78.71	74.32	5.91%

- HC: only learn the hierarchical couplings.
- HELIC: learn both hierarchical couplings and heterogeneity.

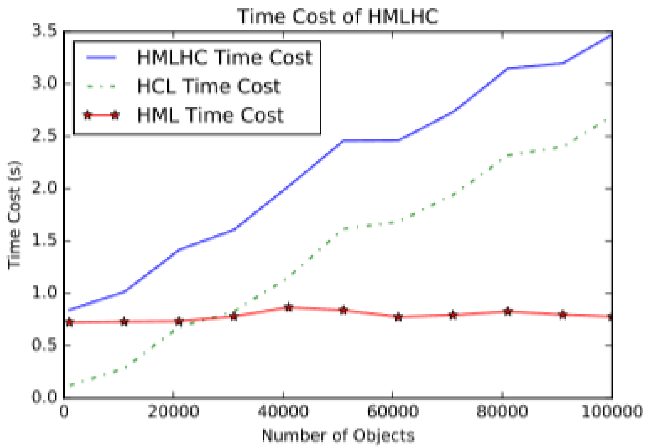
Flexibility of HELIC

LR, RF and SVM Classification F-score (%) with HELIC and MTDLE

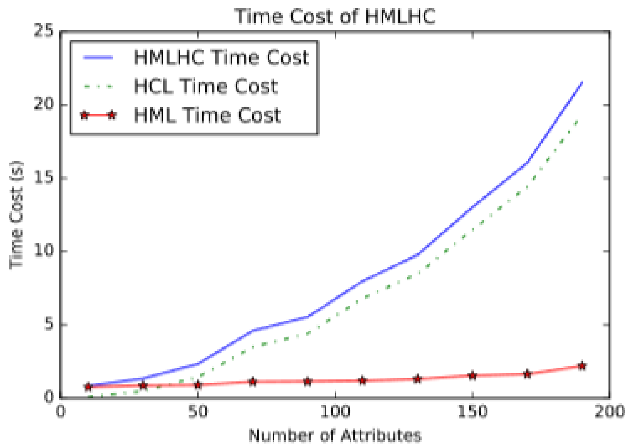
Data	HELIC-LR	MTDLE-LR	Δ%	HELIC-RF	MTDLE-RF	Δ%	HELIC-SVM	MTDLE-SVM	Δ%
Zoo	100	92.50 ± 11.75	8.11%	100	99.64 ± 1.63	0.36%	100	100	0%
DNAPromoter	98.48 ± 3.70	89.84 ± 10.89	9.62%	93.88 ± 9.02	74.87 ± 11.89	25.39%	97.98 ± 4.15	89.88±10.35	9.01%
Hayesroth	83.56 ± 6.53	83.23 ± 8.16	0.40%	82.51±7.85	79.80± 10.66	3.40%	84.44 ± 8.62	81.64 ± 8.76	3.43%
Audiology	73.63 ± 6.33	49.88 ± 10.26	47.61%	73.04 ± 7.30	39.23 ± 13.19	86.18%	73.47 ± 6.07	62.15±10.70	18.21%
Spect	69.10±12.68	51.31 ± 8.79	34.67%	69.38±11.94	69.17 ±15.11	3.04%	69.65±12.22	69.33 ± 12.33	0.46%
Mofn3710	100	83.13 ± 16.47	20.29%	81.62±9.03	67.97± 9.94	20.08%	100	100	0%
Monks3	97.21 ± 1.79	100	0%	100	99.88 ± 0.52	0.12%	100	100	0%
ThreeOf9	80.54 ± 5.05	79.52 ± 5.20	1.29%	99.71±0.96	97.14 ± 2.60	2.65%	79.37±5.61	79.46 ± 5.48	0%
Balance	91.24 ± 7.00	63.94 ± 0.06	42.70%	58.52±1.86	58.17 ± 2.24	0.60%	97.45±2.49	98.09 ± 2.44	0%
Crx	85.76 ± 4.86	83.96 ± 4.82	2.14%	85.15±3.72	84.21 ± 4.00	1.12%	84.98±4.79	76.10 ± 5.99	11.67%
Mammographic	82.62 ± 5.13	82.36 ± 4.53	0.32%	82.75±5.36	80.61 ± 4.78	2.65%	82.59±4.32	80.91 ± 5.45	2.08%
Mean	87.96	78.51	12.04%	84.99	77.84	9.19%	88.61	85.91	3.14%

The HELIC framework can be incorporated into different classifiers

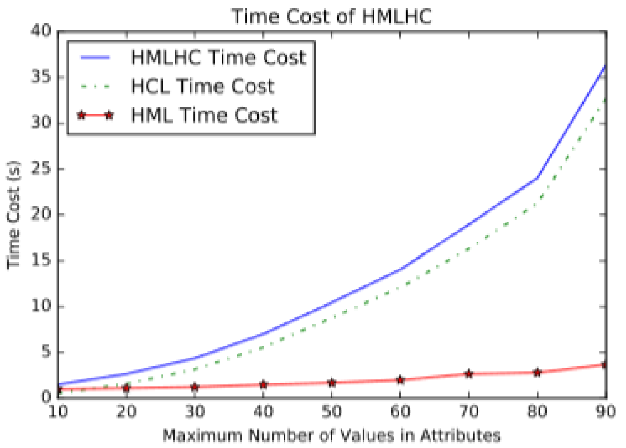
Scalability of HELIC



(a) Time Cost v.s. Number of Objects.



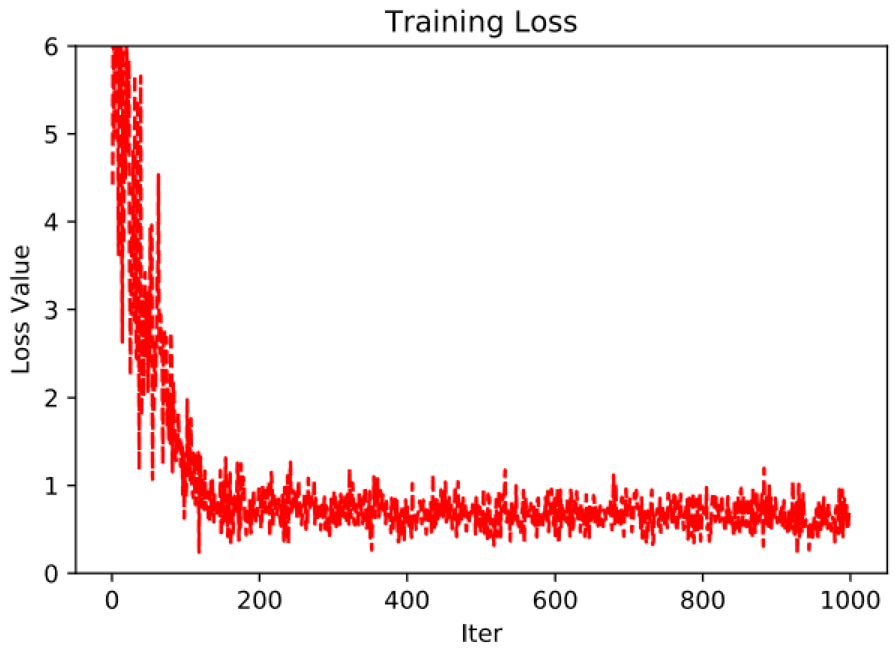
(b) Time Cost v.s. Number of Attributes.



(c) Time Cost v.s. Number of Attribute Values.

The Time Cost of HELIC w.r.t. Data Factors: Object Number n_o, Attribute Number n_a, and Maximum Number of Attribute Values n_{mv}. The solid line refers to the total time cost of HELIC. The dotted line refers to the time cost of the hierarchical coupling learning parts. The star line refers to the time cost of the heterogeneous metric learning parts.

Scalability of HELIC



Comments

What if different categorical attributes have different non-IIDness?

What if the input are mixed with non-IID numerical data and non-IID categorical data?

Change kernel representations to other representations e.g., deep representations, probabilistic representations?

How to address the curse of non-IIDness?

Statistical Learning of Large, Sparse, Dynamic and Multisource data

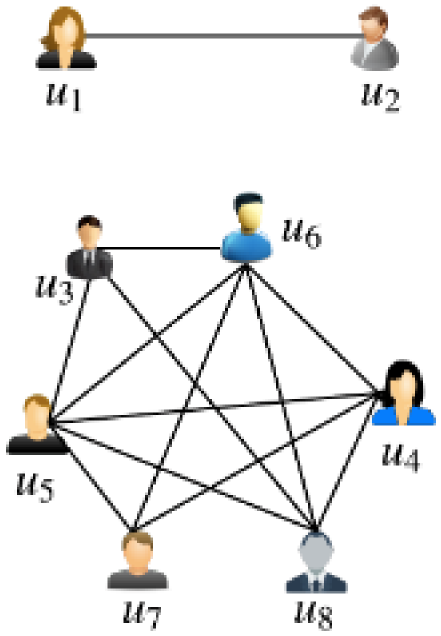
Tutorials: PAKDD19/AAAI20 tutorials

T. Do and L. Cao. Gamma-Poisson Dynamic Matrix Factorization Embedded with Metadata Influence, NIPS2018.

Large, sparse, dynamic and multi-source data

	The Godfather	The Dark Knight	Goodfellas	Toy Story 3	Alien
u_1	5	3	5	4	?
u_2	5	?	5	?	?
u_3	1	3	?	?	?
u_4	1	?	?	?	?
u_5	1	3	?	4	?
u_6	1	3	?	4	?
u_7	?	3	?	5	?
u_8	?	?	?	?	?

(a) Rating table



(b) User friendship

	Age	Location	Occupation	Education
u_1	28	NY	Developer	Bac
u_2	27	NY	Nurse	Bac
u_3	42	HI	Prof.	PhD
u_4	40	HI	Prof.	PhD
u_5	43	HI	Prof.	PhD
u_6	41	HI	Prof.	PhD
u_7	42	HI	Prof.	PhD
u_8	45	HI	Prof.	PhD

(c) User metadata

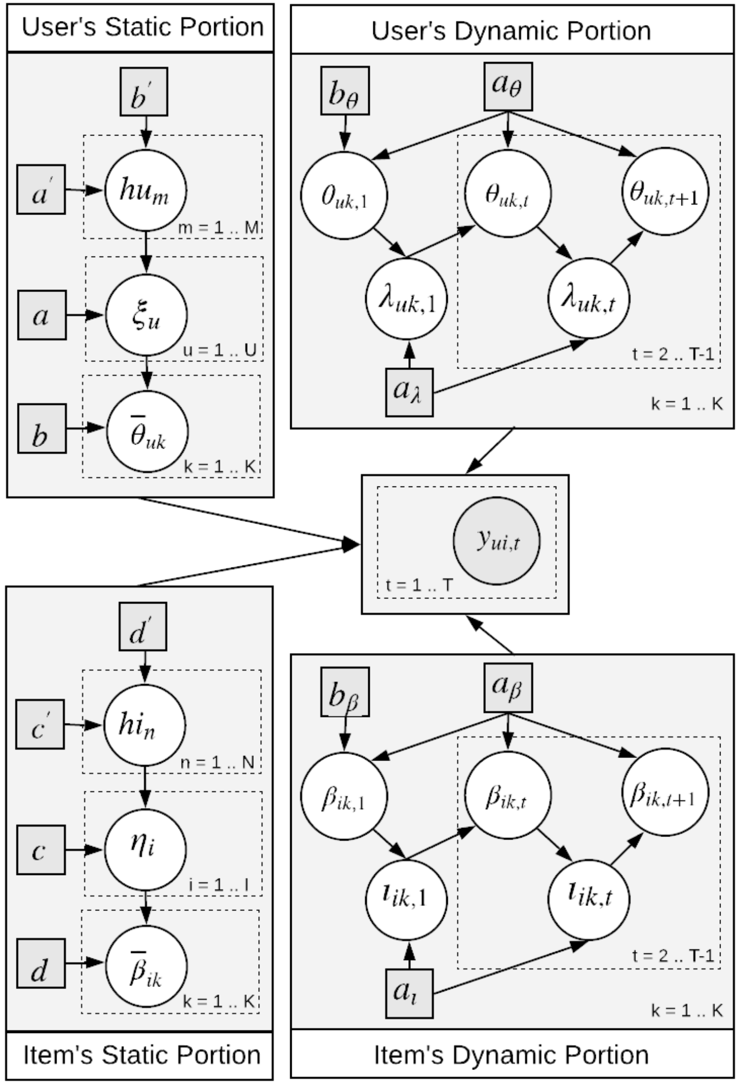
Challenges to statistical learning

- Latent feature learning
- Latent relation learning
- Matrix factorization
- Dynamic learning
- Incorporating multisource data
- Inference
- Sampling

Gamma-Poisson dynamic matrix factorization model incorporated with metadata influence (mGDMF)

Enrich prior of user's static portion using metadata

Enrich prior of item's static portion using metadata



Model dynamics of a user

Model dynamics of item

mGDMF: Generative process

1. **Metadata Integration:**
   (a) For each user:
      i. Draw the weight of m^{th} attribute in user metadata $hu_m \sim Gamma(a', b')$
      ii. Draw latent user preference $\xi_u \sim Gamma(a, \prod_{m=1}^{M} hu_m^{fu_{u,m}})$
      iii. Draw global static factor $\overline{\theta}_{uk} \sim Gamma(b, \xi_u)$
   (b) For each item:
      i. Draw the weight of n^{th} attribute in item metadata $hi_n \sim Gamma(c', d')$
      ii. Draw latent item attractiveness $\eta_i \sim Gamma(c, \prod_{n=1}^{N} hi_n^{fi_{i,n}})$
      iii. Draw global static factor $\overline{\beta}_{ik} \sim Gamma(d, \eta_i)$
2. **Dynamic Modeling:**
   (a) For each user:
      i. Draw initialized state of local dynamic factor $\theta_{uk,1} \sim Gamma(a_\theta, a_\theta b_\theta)$
      ii. For each time slice $t > 1$:
         A. Draw auxiliary variable $\lambda_{uk,t-1} \sim Gamma(a_\lambda, a_\lambda \theta_{uk,t-1})$
         B. Draw local dynamic factor $\theta_{uk,t} \sim Gamma(a_\theta, a_\theta \lambda_{uk,t-1})$
   (b) For each item:
      i. Draw initialized state of local dynamic factor $\beta_{ik,1} \sim Gamma(a_\beta, a_\beta b_\beta)$
      ii. For each time slice $t > 1$:
         A. Draw auxiliary variable $\iota_{ik,t-1} \sim Gamma(a_\iota, a_\iota \beta_{ik,t-1})$
         B. Draw local dynamic factor $\beta_{ik,t} \sim Gamma(a_\beta, a_\beta \iota_{ik,t-1})$
3. **For each rating:**
   (a) Draw $y_{ui,t} \sim Poisson(\sum_k (\theta_{uk,t} + \overline{\theta}_{uk})(\beta_{ik,t} + \overline{\beta}_{ik}))$

Inference

- Variational Inference for mGDMF (still statistically i.i.d. though):
  - The mean-field family assumes each distribution is independent of the others.

$$
\begin{aligned}
q(hu, hi, \xi, \eta, \overline{\theta}, \overline{\beta}, \lambda, \iota, \theta, \beta, z) = & \prod_{m} q(hu_m|\zeta_m) \prod_{n} q(hi_n|\rho_n) \prod_{u} q(\xi_u|\kappa_u) \prod_{i} q(\eta_i|\tau_i) \\
& \prod_{u,k} q(\overline{\theta}_{uk}|\overline{\nu}_{uk}) \prod_{i,k} q(\overline{\beta}_{ik}|\overline{\mu}_{ik}) \prod_{u,k,t} q(\theta_{uk,t}|\nu_{uk,t}) \prod_{i,k,t} q(\beta_{ik,t}|\mu_{ik,t}) \\
& \prod_{u,k,t} q(\lambda_{uk,t}|\gamma_{uk,t}) \prod_{i,k,t} q(\iota_{ik,t}|\omega_{ik,t}) \prod_{u,i,t,k} q(z_{ui,t,k}|\phi_{ui,t,k})
\end{aligned} \tag{3}
$$

We use the class of conditionally conjugate priors for hu_m, hi_n, ξ_u, η_i, $\overline{\theta}_{uk}$, $\overline{\beta}_{ik}$, θ_{uk}, $\lambda_{uk,t}$, β_{ik}, $\iota_{ik,t}$ and $z_{ui,t,k}$ to update the variational parameters $\{\zeta, \rho, \kappa, \tau, \overline{\nu}, \overline{\mu}, \nu, \gamma, \mu, \omega, \phi\}$. For the Gamma distribution, we update both hyper-parameters: *shape* and *rate*.

Inference

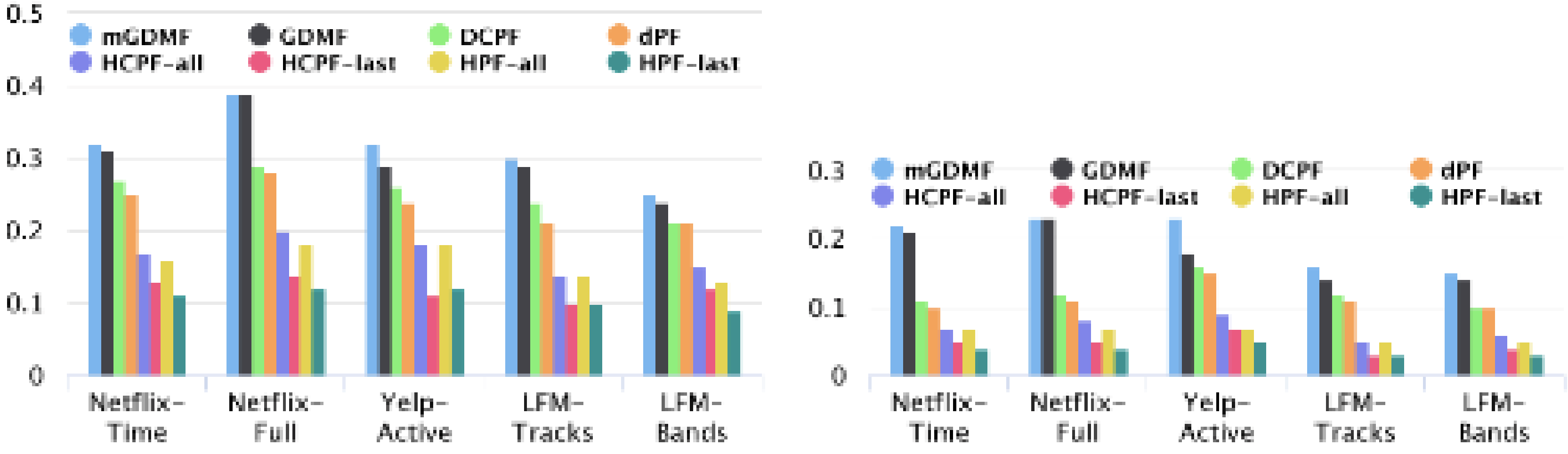
Table 1: Latent Variables, Type, Variational Variables and Variational Update for Users. Similar variables for items (i.e., hi_n, η_i, $\overline{\beta}_{ik}$, β_{ik}, $\iota_{ik,t}$) can be found in the supplementary. $\aleph_m$ is the number of users having the m^{th} attribute, K is the number of latent components, and $\Psi(.)$ is the *digamma* function. The Gamma distribution is parameterized by *shape* (shp) and *rate* (rte).

Latent Variable	Type	Variational Variable	Variational Update
hu_m	Gamma	$\zeta_m^{shp}, \zeta_m^{rte}$	$a' + \aleph_m a,\ b' + \sum_u \frac{\kappa_u^{shp}}{\kappa_u^{rte}}$
ξ_u	Gamma	$\kappa_u^{shp}, \kappa_u^{rte}$	$a + Kb,\ \prod_{m=1}^{M} \left(\frac{\zeta_m^{shp}}{\zeta_m^{rte}}\right)^{fu_{u,m}} + \sum_k \frac{\overline{\nu}_{uk}^{shp}}{\overline{\nu}_{uk}^{rte}}$
$z_{ui,t,k}$	Mult	$\phi_{ui,t,k}$	$(exp\{\Psi(\nu_{uk,t}^{shp}) - log(\nu_{uk,t}^{rte})\} + exp\{\Psi(\overline{\nu}_{uk}^{shp}) - log(\overline{\nu}_{uk}^{rte})\})$ $*(exp\{\Psi(\mu_{ik,t}^{shp}) - log(\mu_{ik,t}^{rte}\} + exp\{\Psi(\overline{\mu}_{ik}^{shp}) - log(\overline{\mu}_{ik}^{rte}))\})$
$\overline{\theta}_{uk}$	Gamma	$\overline{\nu}_{uk}^{shp}, \overline{\nu}_{uk}^{rte}$	$b + \sum_{i,t} y_{ui,t}\phi_{ui,t,k},\ \frac{\kappa_u^{shp}}{\kappa_u^{rte}} + \sum_i \left(\frac{\overline{\mu}_{ik}^{shp}}{\overline{\mu}_{ik}^{rte}} + \sum_t \frac{\mu_{ik,t}^{shp}}{\mu_{ik,t}^{rte}}\right)$
$\theta_{uk,t}$	Gamma	$\nu_{uk,t}^{shp}$	$a_\theta + a_\lambda + \sum_i y_{ui,t}\phi_{ui,t,k}$
		$\nu_{uk,1}^{rte}$	$a_\theta b_\theta + a_\lambda \frac{\gamma_{uk,1}^{shp}}{\gamma_{uk,1}^{rte}} + \sum_i \left(\frac{\overline{\mu}_{ik}^{shp}}{\overline{\mu}_{ik}^{rte}} + \frac{\mu_{ik,1}^{shp}}{\mu_{ik,1}^{rte}}\right)$
		$\nu_{uk,t,(t>1)}^{rte}$	$a_\theta \frac{\gamma_{uk,t-1}^{shp}}{\gamma_{uk,t-1}^{rte}} + a_\lambda \frac{\gamma_{uk,t}^{shp}}{\gamma_{uk,t}^{rte}} + \sum_i \left(\frac{\overline{\mu}_{ik}^{shp}}{\overline{\mu}_{ik}^{rte}} + \frac{\mu_{ik,t}^{shp}}{\mu_{ik,t}^{rte}}\right)$
$\lambda_{uk,t}$	Gamma	$\gamma_{uk,t}^{shp}, \gamma_{uk,t}^{rte}$	$a_\lambda + a_\theta,\ a_\lambda \frac{\nu_{uk,t}^{shp}}{\nu_{uk,t}^{rte}} + a_\theta \frac{\nu_{uk,t+1}^{shp}}{\nu_{uk,t+1}^{rte}}$

Experiments

- Datasets:
  - (1) Netflix-Time, Netflix-Full [Li et al., 2011].
  - (2) Yelp-Active [Jerfel et al., 2017].
  - (3) LFM-Tracks, LFM-Bands [Ò. Celma Herrada, 2009].
- Baseline methods:
  - Static:
    - HPF [Gopalan et al., 2015], HCPF [Basbug and Engelhard, 2016] as it outperforms many baselines in MF including NMP, LDA and PMF.
    - PF-last and HCPF-last are trained by using the last time slice in the training set as the observations.
    - HPF-all and HCPF-all are trained on all training ratings.
  - Dynamic:
    - dPF [Charlin et al., 2016] and DCPF [Jerfel et al., 2017].
    - dPF was shown to outperform state-of-the-art dynamic collaborative filtering algorithms, specifically, BPTF and TimeSVD++.

Effect of metadata and dynamic data modeling



(a) *Precision*@50 (b) *Recall*@50

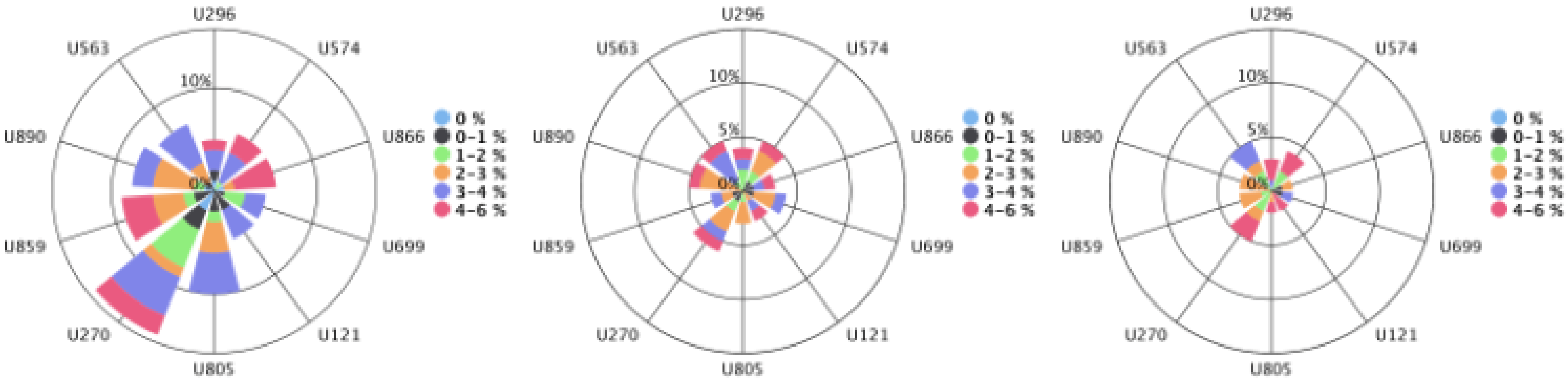
Figure 1: Top-50 Recommendations Compared with Baselines.

Effect of metadata and dynamic data modeling

Table 2: Predictive Performance on Five Datasets w.r.t. NDCG and AUC.

	Netflix-Time		Netflix-Full		Yelp-Active		LFM-Tracks		LFM-Bands	
	NDCG	AUC	NDCG	AUC	NDCG	AUC	NDCG	AUC	NDCG	AUC
mGDMF	**0.389**	**0.9145**	**0.403**	**0.9321**	**0.494**	**0.8650**	**0.310**	**0.8245**	**0.367**	**0.8217**
GDMF	0.367	0.9121	0.398	0.9320	0.416	0.8512	0.275	0.8101	0.354	0.8139
DCPF	0.293	0.9023	0.315	0.8991	0.357	0.8418	0.231	0.8098	0.275	0.8011
dPF	0.257	0.9012	0.301	0.8901	0.332	0.8321	0.210	0.8019	0.298	0.8122
HCPF-all	0.241	0.8012	0.245	0.8370	0.243	0.8032	0.209	0.7010	0.213	0.7121
HCPF-last	0.183	0.7423	0.201	0.7600	0.172	0.7312	0.132	0.5893	0.160	0.6101
HPF-all	0.231	0.8035	0.250	0.8124	0.248	0.8130	0.179	0.7084	0.184	0.7013
HPF-last	0.162	0.7213	0.198	0.7540	0.145	0.6810	0.143	0.6050	0.141	0.5982
$\delta_{min}(\%)$	32.76	1.35	27.94	3.67	38.38	2.76	34.20	1.82	23.15	1.70
$\delta_{max}(\%)$	140.12	26.78	103.54	23.62	240.69	27.12	134.85	44.83	160.28	37.36

Effect of sparse users/items and ‘cold-start’



(a) mGDMF (b) GDMF (c) DCPF

Figure 2: Percentage (%) of Sparse Items Recommended Precisely for 10 Users by mGDMF, GDMF and DCPF.

Case study of mGDMF-based recommendation

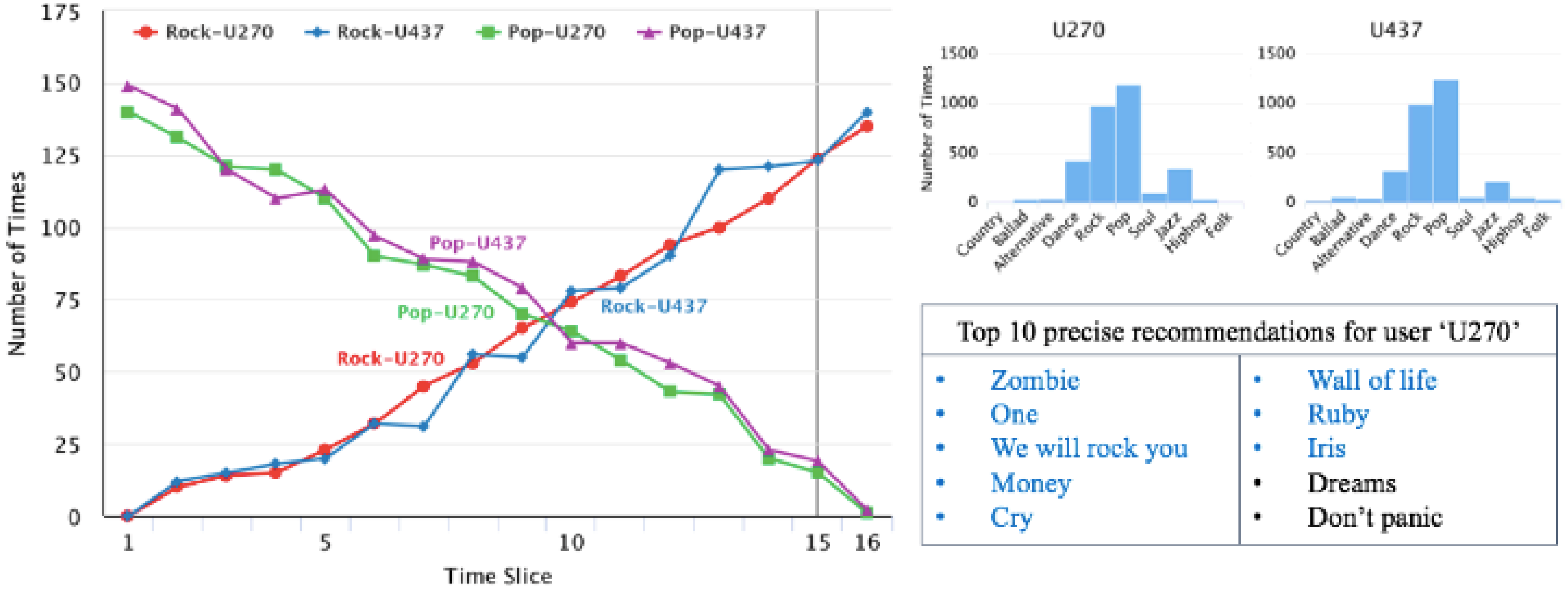


Figure 3: Analysis on two users '$U270$' and '$U437$' with the same metadata in Last.fm. The number of times that users listened to two 'rock' and 'pop' tracks with 16 time slices is shown on the left. The distribution of the number of times that $U270$ and $U437$ listened to top 10 'rock' and 'pop' tracks and the top10 precise recommendations by mGDMF are shown on the right.

Comments

How to cope with observable variables with different distributions?

When latent variables are non-IID, how to conduct the sampling and inference?

When multiple distinct distributions are coupled, how to statistically learn them in one model?

How can deep Bayesian learning capture various non-IIDness in complex data?

Learning from low quality, ultrahigh-dimensional data

Learning Representations of Ultra... for Random Distance-based Outlier Detection, KDD2018

Sparse Modeling-based Sequential Ensemble Learning for Effective Outlier Detection in High-dimensional Numeric Data. AAAI2018.

Learning Homophily Couplings from Non-IID Data for Joint Feature Selection and Noise-Resilient Outlier Detection. IJCAI2017

Selective Value Coupling Learning for Detecting Outliers in High-Dimensional Categorical Data. CIKM2017.

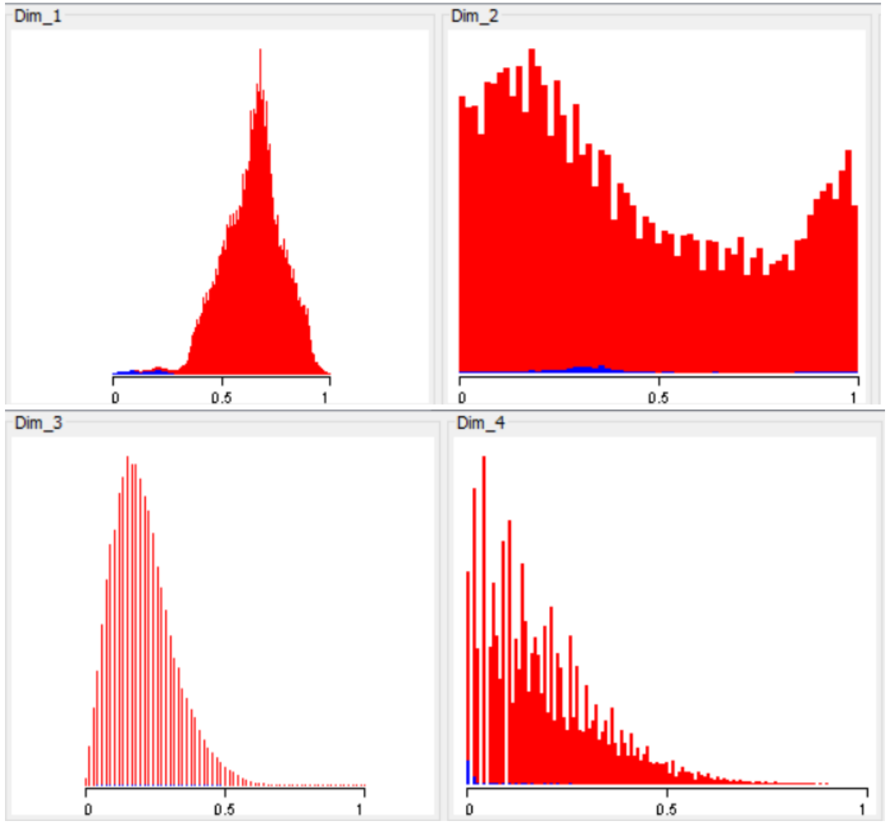
Unsupervised Feature Selection for Outlier Detection by Modelling Hierarchical Value-Feature Couplings. ICDM2016.

Non-IID Real-life Data

Couplings

Source: http://www.diabeticrockstar.com

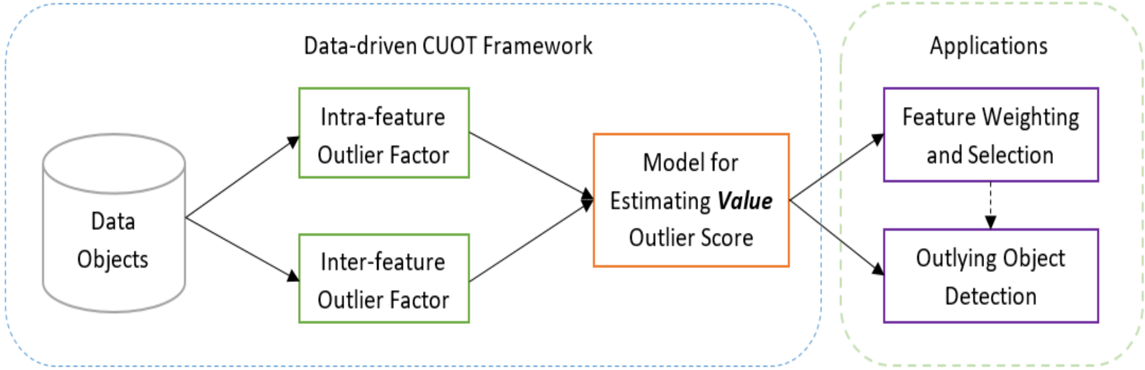
Heterogeneity



Four features from the *CoverType* data set

Non-IID value-based approach

Learning value outlierness from data with non-IID values



Data	CBRW	CBRWie	CBRWia	MarP$^+$	MarP	FPOF	COMP	FORE
BM	0.6287	**0.6566**	0.5999	0.5778	0.5584	0.5466	0.6267	0.5762
Census	0.6678	0.6579	**0.6832**	0.6033	0.5899	0.6148	0.6352	0.5378
AID362	**0.6640**	0.6324	0.6034	0.6152	0.6270	○	0.6480	0.6485
w7a	0.6484	**0.7338**	0.4453	0.4565	0.4723	○	0.5683	0.4053
CMC	**0.6339**	0.6323	0.6179	0.5623	0.5417	0.5614	0.5669	0.5746
APAS	0.8190	0.8624	**0.8739**	0.6208	0.6193	○	0.6554	0.4792
CelebA	0.8462	**0.9108**	0.7135	0.7352	0.7358	0.7380	0.7572	0.6797
Chess	**0.7897**	0.4058	0.7766	0.6854	0.6447	0.6160	0.6387	0.6124
AD	0.7348	**0.8270**	0.7250	0.7033	0.7033	○	•	0.7084
SF	0.8812	0.8833	**0.8867**	0.8469	0.8446	0.8556	0.8526	0.7865
Probe	0.9906	**0.9907**	0.9434	0.9795	0.9800	0.9867	0.9790	0.9762
U2R	0.9651	0.9640	0.8817	0.8848	0.8848	0.9156	**0.9893**	0.9781
LINK	0.9976	0.9976	0.9976	0.9977	0.9977	**0.9978**	0.9973	0.9917
R10	**0.9905**	0.9903	0.9823	0.9866	0.9866	○	0.9866	0.9796
CT	0.9703	0.9703	0.9388	0.9770	**0.9773**	0.9772	0.9772	0.9364
Avg.(Top-10)	0.7314	0.7202	0.6925	0.6407	0.6337	0.6554	0.6610	0.6009
Avg.(All)	0.8152	0.8077	0.7779	0.7488	0.7442	0.7810	0.7770	0.7247
	CBRW vs.	0.7959	0.0392	0.0012	0.0008	0.0115	0.0147	0.0040
p-value		CBRWie vs.	0.4225	0.0969	0.0592	0.4316	0.3167	0.0446
			CBRWia vs.	0.1460	0.1223	0.2886	0.8490	0.0979

Intra-feature couplings:

$$\sigma(v) = \frac{1}{2}[base(m) + dev(v)]$$

$$base(m) = 1 - freq(m)$$

$$dev(v) = \frac{freq(m) - freq(v)}{freq(m)}$$

Inter-feature couplings:

$$\boldsymbol{q}_v = [\eta(u,v), \dots, \eta(w,v)]^\intercal$$
$$= [\frac{freq(u,v)}{freq(v)}, \dots, \frac{freq(w,v)}{freq(v)}]^\intercal,$$

Objective function:

$$object\_score(x) = \sum_{f \in F} w_f \times value\_score(g_f(x)) \quad (9)$$

where $w_f = \frac{rel(f)}{\sum_{f \in F} rel(f)}$ is a feature weighting component.

✓ CBRW obtains more than 12%, 12%, 13%, 7% and 17% improvement on these 10 data sets

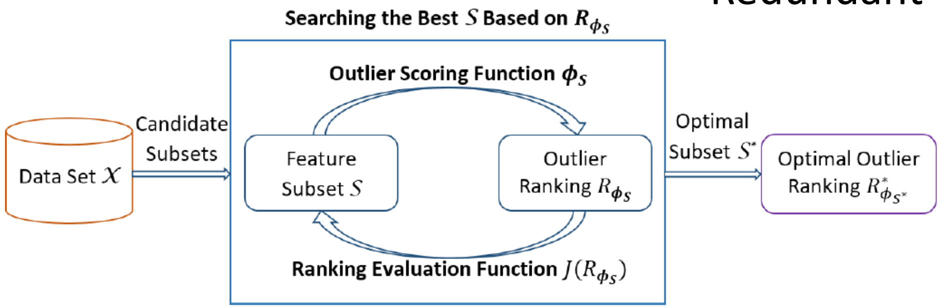
Guansong Pang, Longbing Cao, Ling Chen. Identifying Outliers in Complex Categorical Data by Modeling Feature Value Couplings. IJCAI16.

End-to-end learning from low-quality complex data

- Highly imbalanced
- Highly sparse

- High to ultrahigh-dimensional
- Noisy
- Redundant

- AUC: 7% and 21% improvement over COMP and FPOF
- P@n: 37% and 90% over COMP and FPOF



$$J(R_{\phi_S}, k) = \frac{\Delta_S}{|\mathcal{S}|} = \frac{1}{k|\mathcal{S}|}\sum_{\boldsymbol{x}\in\mathcal{O}}[\phi_S(\boldsymbol{x}) - \phi_S(\boldsymbol{x}')]$$

				POP	CBRW			ZERO			iForest		
Data	$\vert\mathcal{F}\vert$	$\vert\mathcal{F}'\vert$	$\vert\mathcal{F}''\vert$	-	POFS	CBFS	DSFS	POFS	CBFS	DSFS	POFS	CBFS	DSFS
w7a	300	180	26	**0.8673**	0.8220	0.7738	0.5155	0.7701	0.7885	0.5155	0.5893	0.7674	0.5155
wap.wc	4229	2537	3570	**1.0000**	0.9026	0.8739	0.6387	0.7339	0.7429	0.5395	0.5902	0.6816	0.5121
R8	9467	5680	2006	**0.9479**	NA	NA	0.9249	0.8902	NA	0.8758	0.8370	NA	0.8426
CAL16	253	151	194	0.9928	0.9930	0.9928	**0.9931**	0.9910	0.9900	0.9903	0.9828	0.9824	0.9811
AD	1555	933	49	**0.9290**	0.7845	0.7456	0.7432	0.7547	0.7587	0.7428	0.7345	0.7723	0.7435
CAL28	727	436	564	**0.9608**	0.9603	0.9604	0.9599	0.9566	0.9584	0.9540	0.9488	0.9524	0.9421
CelebA	39	23	34	**0.8968**	0.8901	0.8818	0.8502	0.8519	0.8511	0.7722	0.8038	0.8213	0.6973
PCMAC	3039	1823	1256	**0.6935**	0.6759	0.6678	0.6413	0.5952	0.5793	0.4959	0.5509	0.5425	0.4745
BASE	4320	2592	1895	0.6521	0.6294	0.6558	0.5760	0.5396	0.5897	0.4375	0.5096	0.5417	0.4233
WebKB	6601	3960	3487	0.7306	0.7449	NA	0.7251	0.7377	NA	0.6995	0.7292	NA	0.6891
RELA	4080	2448	2101	**0.7449**	0.7256	0.7352	0.6984	0.6580	0.6793	0.5987	0.6268	0.6459	0.5844
Arrhy	64	38	13	**0.6762**	0.6095	0.6527	0.5625	0.6074	0.6540	0.5626	0.6065	0.6543	0.5624
			Average	**0.8410**	0.7943	0.7940	0.7357	0.7572	0.7592	0.6820	0.7091	0.7362	0.6640
			P-value	-	0.0098	0.0117	0.0010	0.0024	0.0020	0.0005	0.0005	0.0020	0.0005

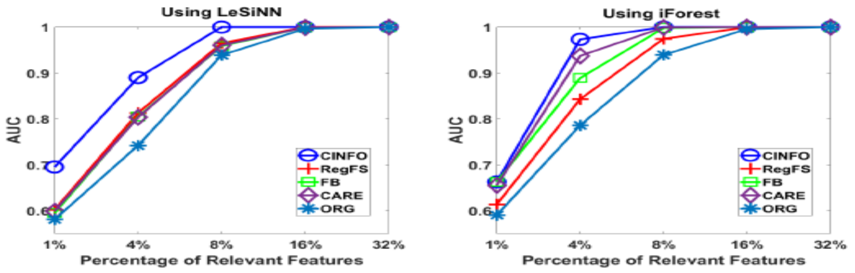


Figure 2: AUC Performance on Data with Different Levels of Noisy Features. 'ORG' denotes the bare LeSiNN/iForest. All methods obtain AUC of one with more than 32% relevant features.

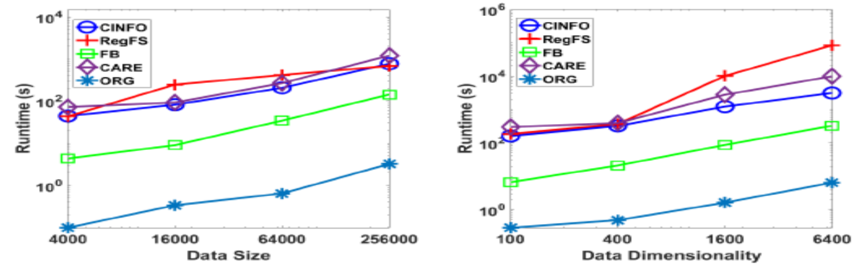


Figure 3: Runtime of CINFO and Its Competitors Using LeSiNN. 'ORG' denotes the bare LeSiNN. Logarithmic scales are used. Similar trends can be expected for using iForest as the outlier detector, since LeSiNN and iForest have similar time complexities.

Comments

Real-life data is often highly complex, while quality may not be good

Enterprise data is often of low quality but with ultrahigh-dimensionality

Existing models on such data for risk analysis often either do not deliver actionable results or do not work at all

Concluding remarks

We are lucky in the era of data science and new-generation AI, however

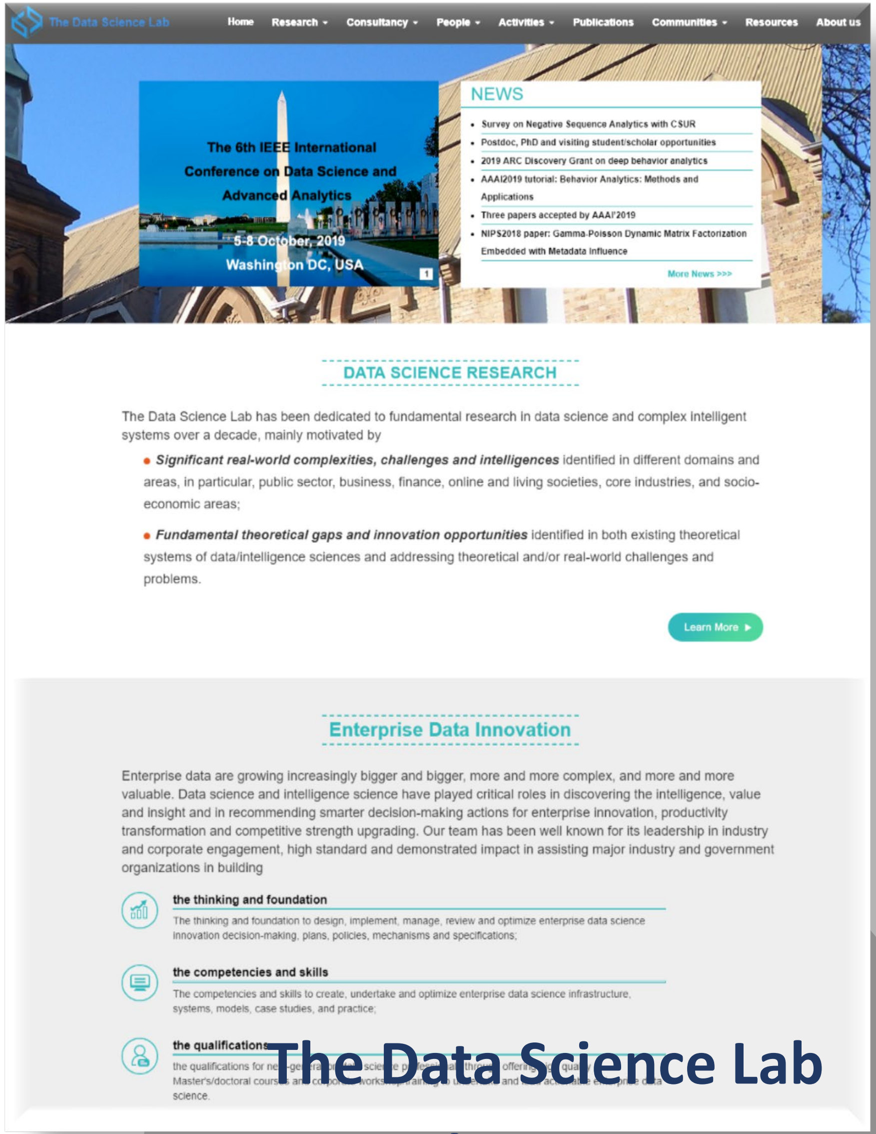
Many intrinsic working mechanisms and challenges in complex data, behaviors and systems may be still unclear, invisible, and unrepresentable

Today's data science is at its early stage, machine learning and AI are highly tailored for particular circumstances, assumptions and purposes

Today's capabilities and capacities for understanding, representing, recognizing and learning data complexities and intelligences are still limited and far from fully capturing their intricate nature

While recent community interest has shifted to topics including data science/AI ethics, interpretability, reproducibility, and autoML, many fundamental issues in building actionable analytics and learning theories and systems are still open

Thank You Very Much


The Data Science Lab | Home | Research | Consultancy | People | Activities | Publications | Communities | Resources | About us

The 6th IEEE International Conference on Data Science and Advanced Analytics
5-8 October, 2019
Washington DC, USA

NEWS
- Survey on Negative Sequence Analytics with CSUR
- Postdoc, PhD and visiting student/scholar opportunities
- 2019 ARC Discovery Grant on deep behavior analytics
- AAAI2019 tutorial: Behavior Analytics: Methods and Applications
- Three papers accepted by AAAI'2019
- NIPS2018 paper: Gamma-Poisson Dynamic Matrix Factorization Embedded with Metadata Influence

More News >>>

DATA SCIENCE RESEARCH

The Data Science Lab has been dedicated to fundamental research in data science and complex intelligent systems over a decade, mainly motivated by

- ***Significant real-world complexities, challenges and intelligences*** identified in different domains and areas, in particular, public sector, business, finance, online and living societies, core industries, and socio-economic areas;
- ***Fundamental theoretical gaps and innovation opportunities*** identified in both existing theoretical systems of data/intelligence sciences and addressing theoretical and/or real-world challenges and problems.

Learn More

Enterprise Data Innovation

Enterprise data are growing increasingly bigger and bigger, more and more complex, and more and more valuable. Data science and intelligence science have played critical roles in discovering the intelligence, value and insight and in recommending smarter decision-making actions for enterprise innovation, productivity transformation and competitive strength upgrading. Our team has been well known for its leadership in industry and corporate engagement, high standard and demonstrated impact in assisting major industry and government organizations in building

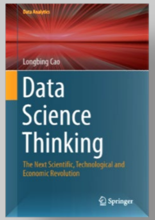
the thinking and foundation
The thinking and foundation to design, implement, manage, review and optimize enterprise data science innovation decision-making, plans, policies, mechanisms and specifications;

the competencies and skills
The competencies and skills to create, undertake and optimize enterprise data science infrastructure, systems, models, case studies, and practice;

the qualifications

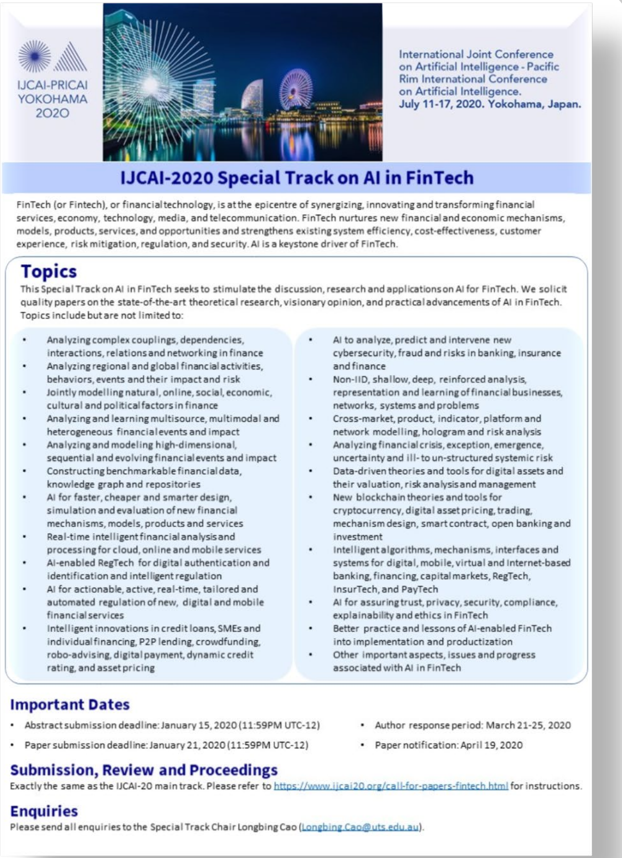
The Data Science Lab

www.datasciences.org



IEEE/ACM/ASA DSAA2020

- Postdoc fellowship
- PhD scholarships


IJCAI-PRICAI YOKOHAMA 2020

International Joint Conference on Artificial Intelligence - Pacific Rim International Conference on Artificial Intelligence.
July 11-17, 2020. Yokohama, Japan.

IJCAI-2020 Special Track on AI in FinTech

FinTech (or Fintech), or financial technology, is at the epicentre of synergizing, innovating and transforming financial services, economy, technology, media, and telecommunication. FinTech nurtures new financial and economic mechanisms, models, products, services, and opportunities and strengthens existing system efficiency, cost-effectiveness, customer experience, risk mitigation, regulation, and security. AI is a keystone driver of FinTech.

Topics

This Special Track on AI in FinTech seeks to stimulate the discussion, research and applications on AI for FinTech. We solicit quality papers on the state-of-the-art theoretical research, visionary opinion, and practical advancements of AI in FinTech. Topics include but are not limited to:

- Analyzing complex couplings, dependencies, interactions, relations and networking in finance
- Analyzing regional and global financial activities, behaviors, events and their impact and risk
- Jointly modelling natural, online, social, economic, cultural and political factors in finance
- Analyzing and learning multisource, multimodal and heterogeneous financial events and impact
- Analyzing and modeling high-dimensional, sequential and evolving financial events and impact
- Constructing benchmarkable financial data, knowledge graph and repositories
- AI for faster, cheaper and smarter design, simulation and evaluation of new financial mechanisms, models, products and services
- Real-time intelligent financial analysis and processing for cloud, online and mobile services
- AI-enabled RegTech for digital authentication and identification and intelligent regulation
- AI for actionable, active, real-time, tailored and automated regulation of new, digital and mobile financial services
- Intelligent innovations in credit loans, SMEs and individual financing, P2P lending, crowdfunding, robo-advising, digital payment, dynamic credit rating, and asset pricing
- AI to analyze, predict and intervene new cybersecurity, fraud and risks in banking, insurance and finance
- Non-IID, shallow, deep, reinforced analysis, representation and learning of financial businesses, networks, systems and problems
- Cross-market, product, indicator, platform and network modelling, hologram and risk analysis
- Analyzing financial crisis, exception, emergence, uncertainty and ill- to un-structured systemic risk
- Data-driven theories and tools for digital assets and their valuation, risk analysis and management
- New blockchain theories and tools for cryptocurrency, digital asset pricing, trading, mechanism design, smart contract, open banking and investment
- Intelligent algorithms, mechanisms, interfaces and systems for digital, mobile, virtual and Internet-based banking, financing, capital markets, RegTech, InsurTech, and PayTech
- AI for assuring trust, privacy, security, compliance, explainability and ethics in FinTech
- Better practice and lessons of AI-enabled FinTech into implementation and productization
- Other important aspects, issues and progress associated with AI in FinTech

Important Dates

- Abstract submission deadline: January 15, 2020 (11:59PM UTC-12)
- Paper submission deadline: January 21, 2020 (11:59PM UTC-12)
- Author response period: March 21-25, 2020
- Paper notification: April 19, 2020

Submission, Review and Proceedings

Exactly the same as the IJCAI-20 main track. Please refer to https://www.ijcai20.org/call-for-papers-fintech.html for instructions.

Enquiries

Please send all enquiries to the Special Track Chair Longbing Cao (Longbing.Cao@uts.edu.au).