Online Semi-Supervised Learning with Mix-Typed Streaming Features

Di Wu[1], Shengda Zhuo[2], Yu Wang[2], Zhong Chen[3], Yi He[4]

[1]College of Computer and Information Science, Southwest University

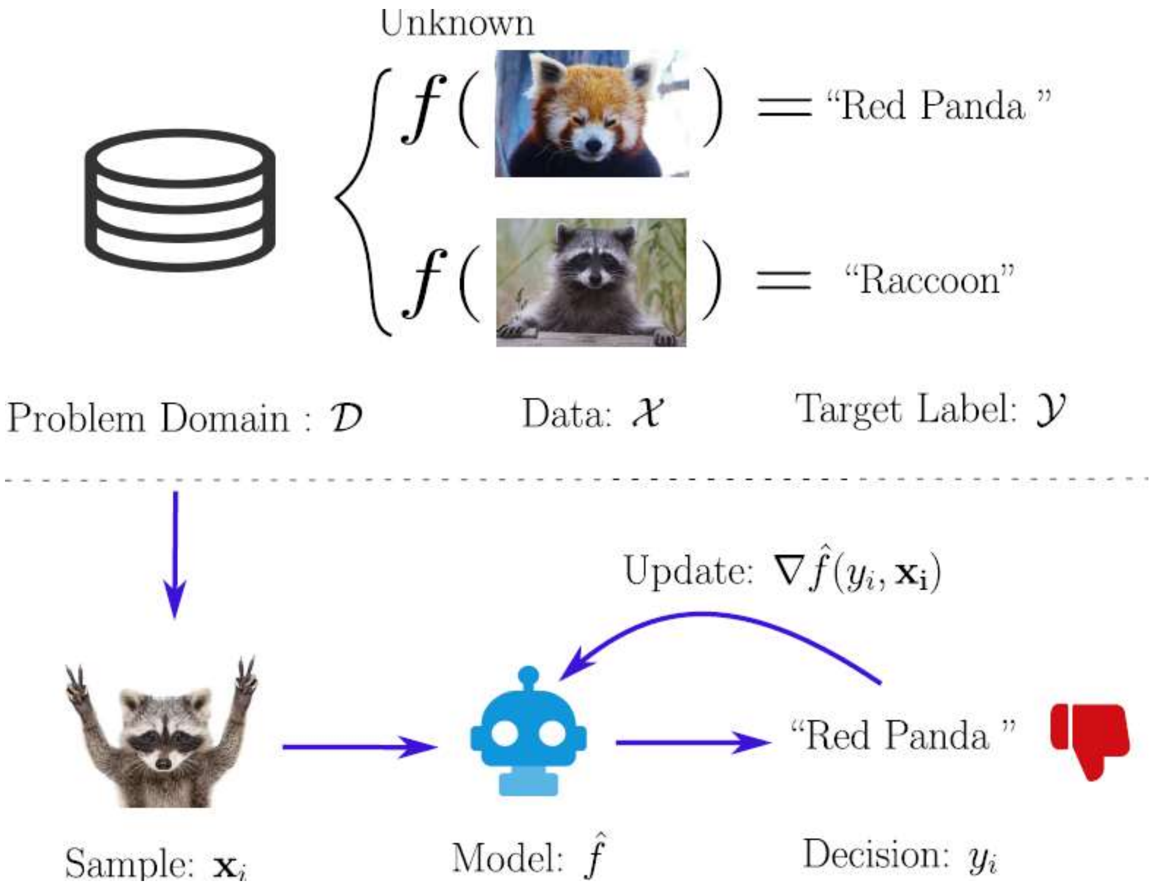
[2]Institute of Artificial Intelligence and Blockchain, Guangzhou University

[3]Department of Computer Science, Xavier University of Louisiana

[4]Department of Computer Science, Old Dominion University

Corresponds to: yihe@cs.odu.edu

Learning from data: Statistical or Sequential? It is a question



Statistical Learning:

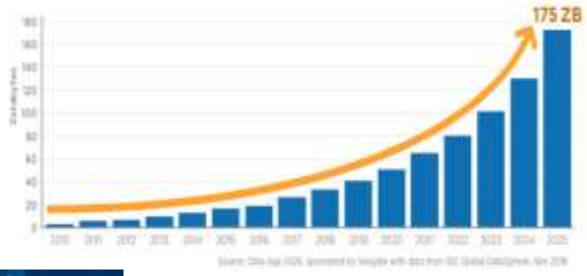
$$\exists h \in \mathcal{H}, \quad h: \sim\mathcal{X} \mapsto \mathcal{Y}$$

$$\min_{h} E_{(x,y)\in\mathcal{X}\times\mathcal{Y}} [\mathcal{L}(h(x), y)]$$

Sequential (Online) Learning:

$$\min_{h_1,\dots,h_T} \frac{1}{N}\sum_{t=1}^{N} \mathcal{L}\left(h_t(x_{t+1}), y_{t+1}\right)$$

Why online Learning



In 2020's cyberspace, 20 ZB (=2.2 x 10^{13} GB) data generated

High-speed streams call for real-time analytics

what organizations can do with that collected data

Volume

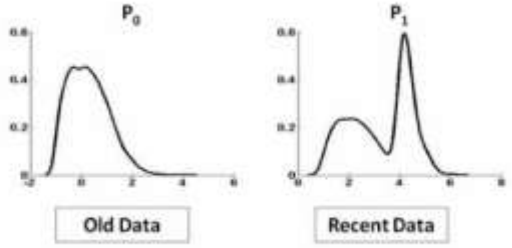
Velocity

Value

Big Data

Variety

Veracity



Patterns like societal interests evolve over time

Gathered data could have missing pieces, may be inaccurate.

How much we lose by learning online?

- By "Lose" we mean "Regret":

$$Regret_T(\mathcal{A}) = \overbrace{\sum_{t=1}^{T} \ell(f_{t-1}; (\mathbf{x}_t, y_t))}^{\text{Cumulative Risk}} - \overbrace{\inf_{f^\star \in \mathcal{F}} \sum_{t=1}^{T} \ell(f^\star, (\mathbf{x}_t, y_t)))}^{\text{Hindsight Optimum}} \quad \in \mathcal{O}(\sqrt{T})$$

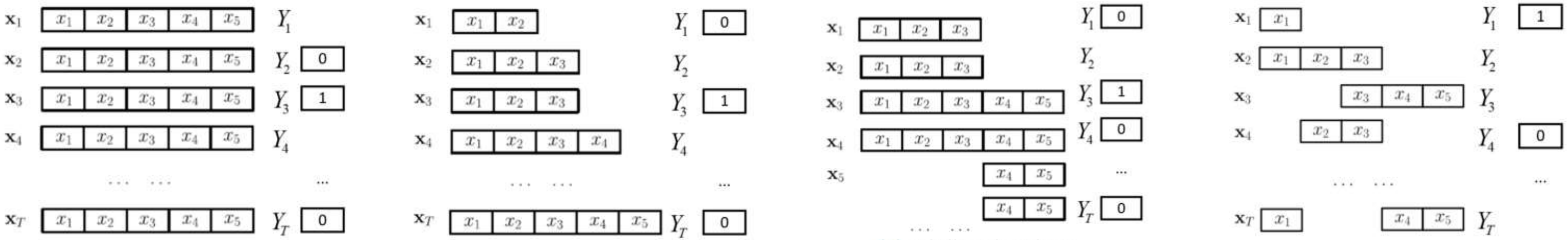
[Cesa-Bianchi and Lugosi, 2006; Zinkevich, ICML'03]

"A good online learner suffers asymptotically no regret"

Goes beyond time horizon: Streaming Features

As new data points and label arrive:

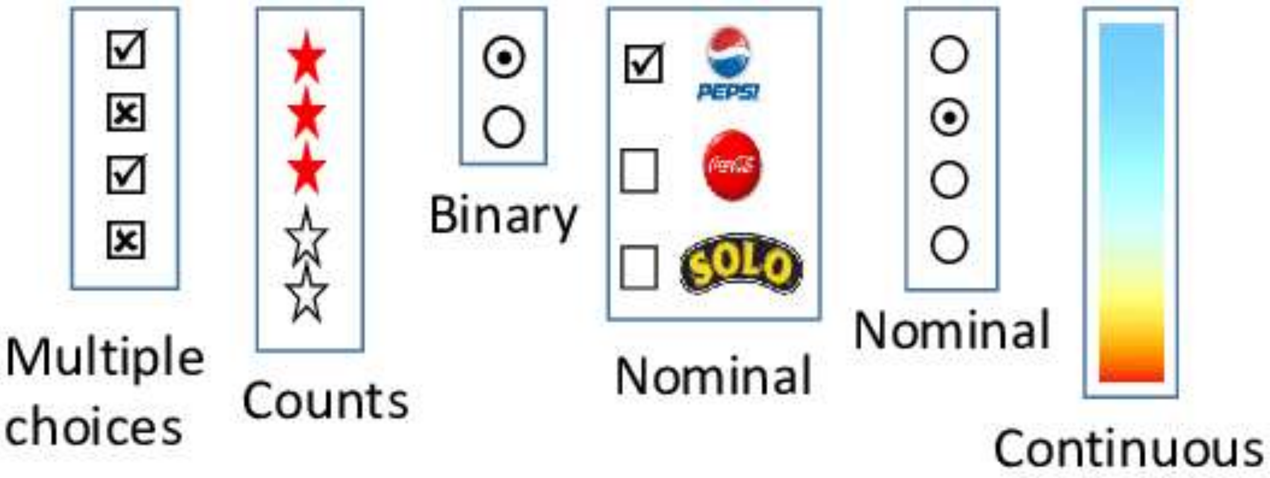
- Increasing more features – trapezoidal DS [Zhang et al., ICDM'15, T.KDE'16]
- Shifts batch to batch – feature-evolvable DS [Hou et al., NeurIPS'17, AAAI'20; Zhang et al., ICML'20,]
- Variable Feature Space(VFS) [Beyazit et al., AAAI'19; He et al., IJCAI'19, AAAI'20]
- Label Scarcity [Hou et al., NeurIPS'17; He et al.,AAAI' 21]



(a) Conventional OL (b) Trapezoidal (c) Minibatch Evlove (d) VFS

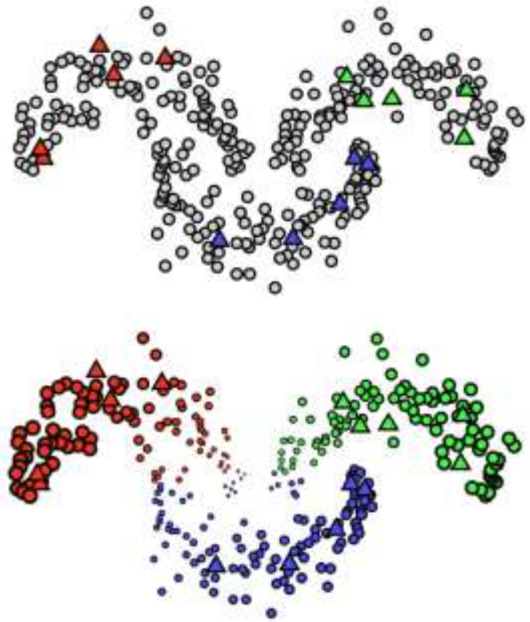
Limitation of Current Doubly-Streaming Data Learners

- Do not support mixed data types:



(Credit to Kien Do et al.)

- Do not support label propagation:



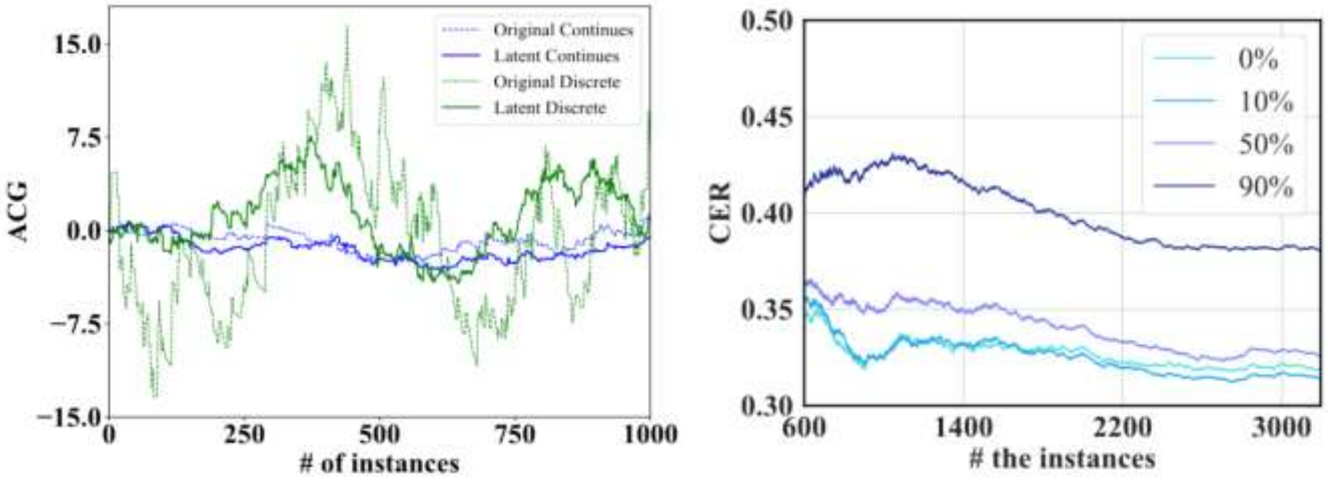
(Credit to Ahmet Iscen et al.)

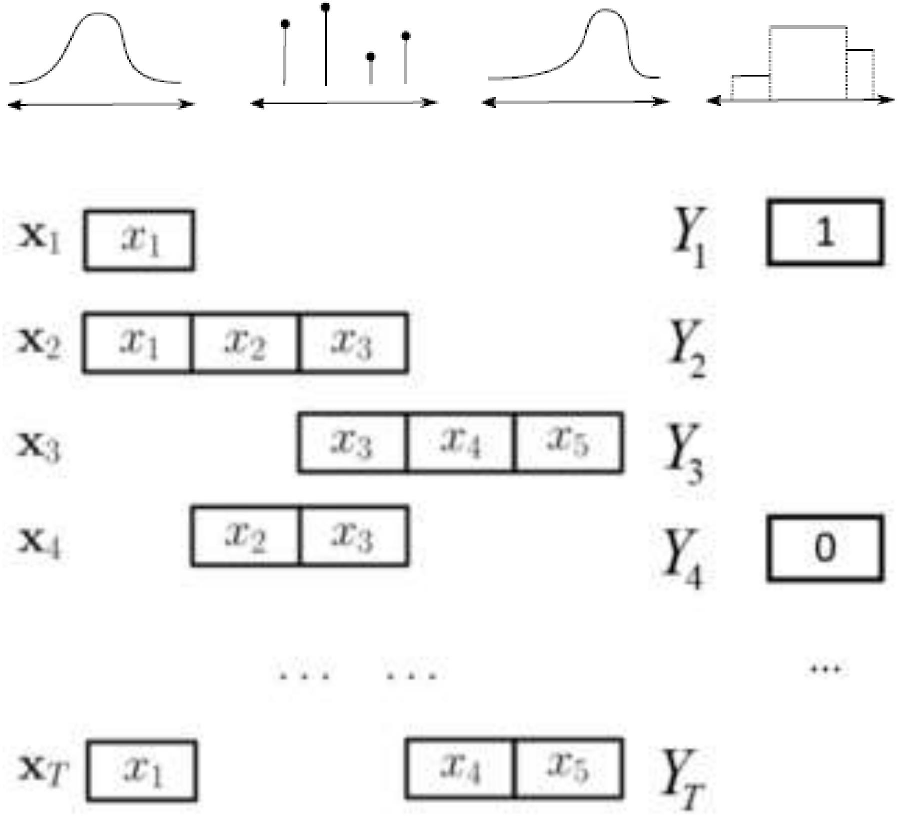
Which is ubiquitous

“Online Mixed Data” face “Streaming Features”

- Preprocessing, e.g., normalization?

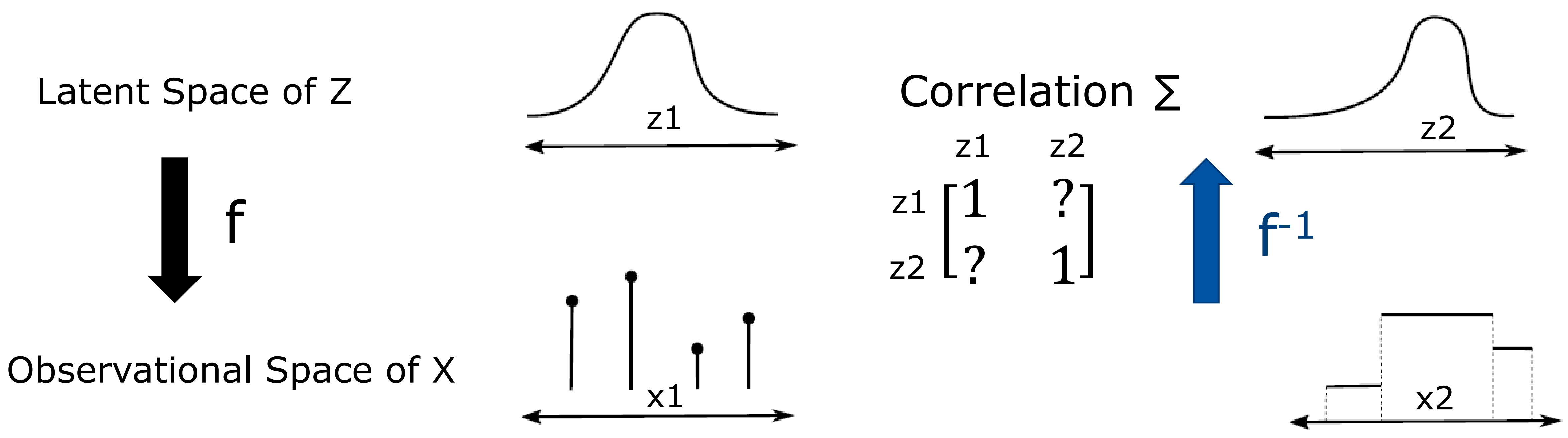
  1. Unknow data volume
  2. Missing feature entries exacerbate bias
  3. Missing labels cause the model not to update





Our Method: Space Mapping and Geometric structure building

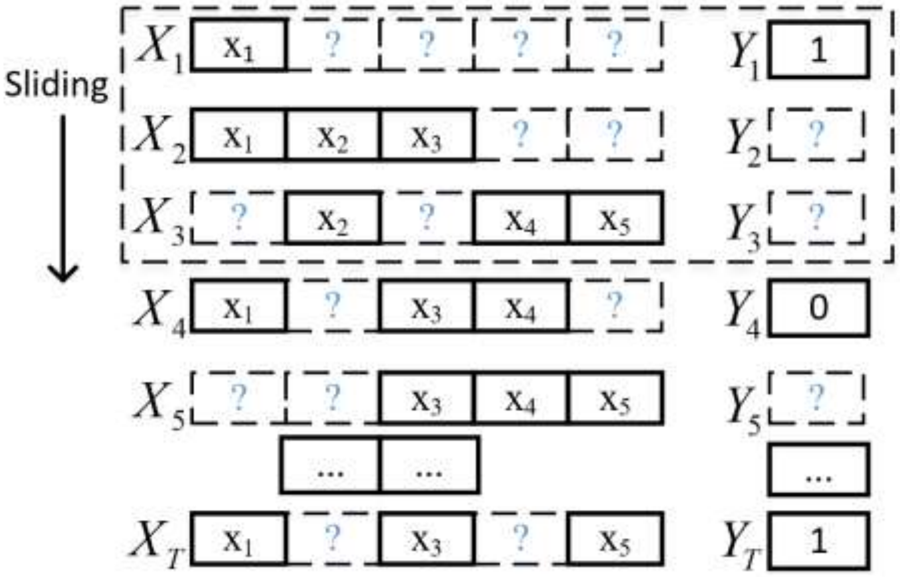
Definition 1 (GC (Masarotto and Varin 2012)). *For* $\forall\, \mathbf{x} \in \mathbb{R}^d$ *that follows the* GC *is a random vector, there is a correlation matrix* Σ *and an element-wise monotone function* $\mathrm{g} : \mathbb{R}^d \mapsto \mathbb{R}^d$ *to make that* $\mathbf{x} = \mathrm{g}(\mathbf{z})$ for $\mathbf{z} \sim N_d(\mathbf{0}, \Sigma)$.

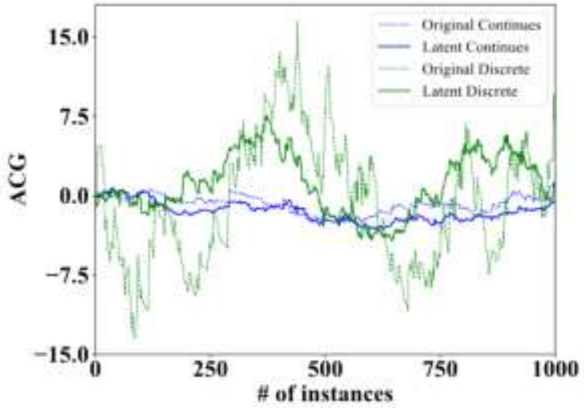


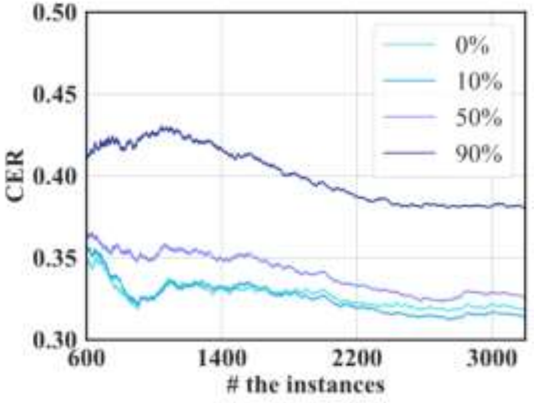
Our Idea

- Missing entries completion
  1. Establishing relationship among features
  2. Learner enjoys a complete observation
- Construction of missing labels
  1. Establishing geometric structure
  2. Learner update the model with pseudo-labels
- Discrete features relaxation

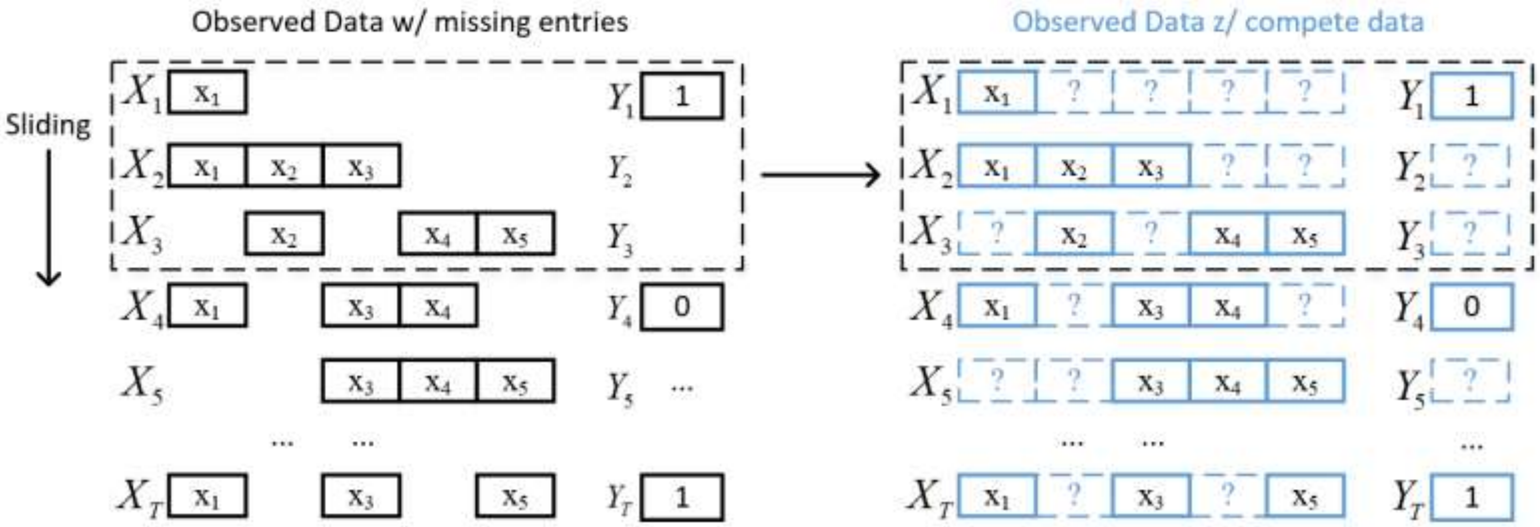
  Expedite convergence -> lower regret





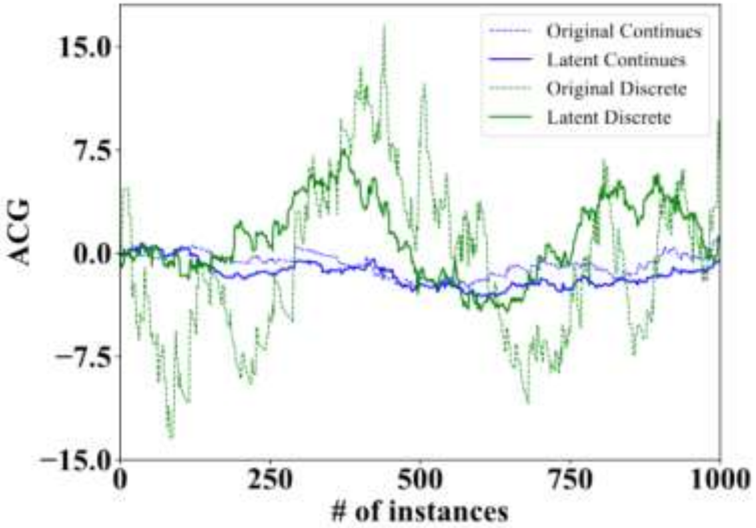


Our Method: Space Mapping and Geometric structure building

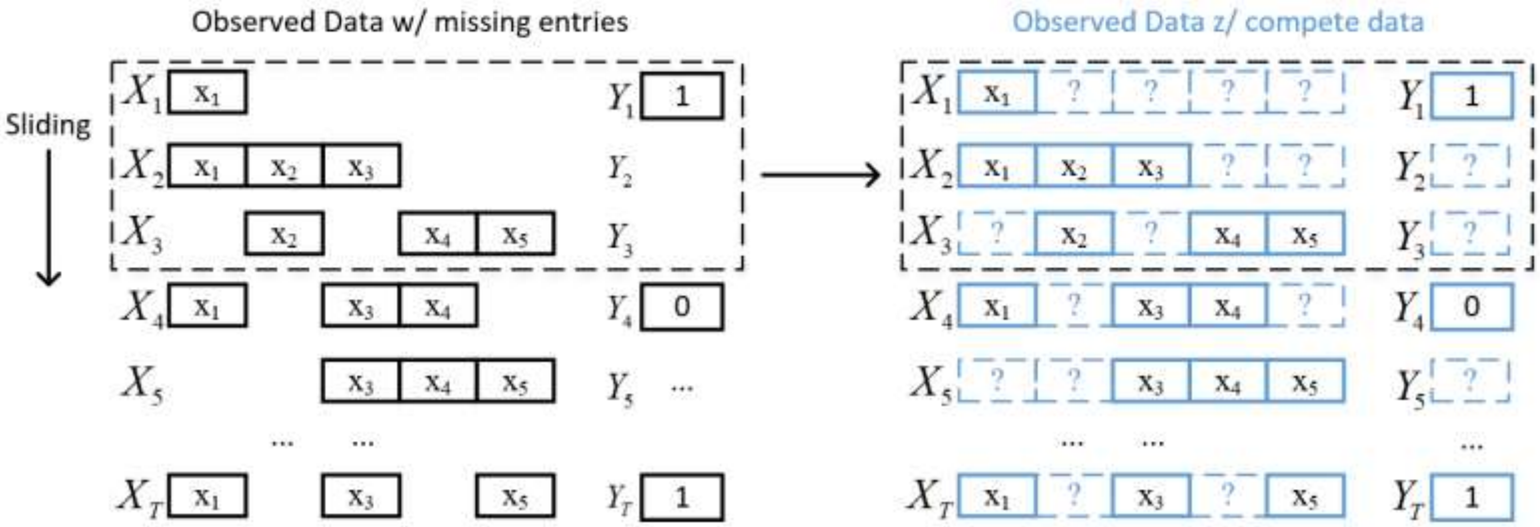


Online Correlation Estimation [Zhao and Udell, KDD'20]

- Latent representations (continuous)
- stabilize the oscillating gradients (discrete)

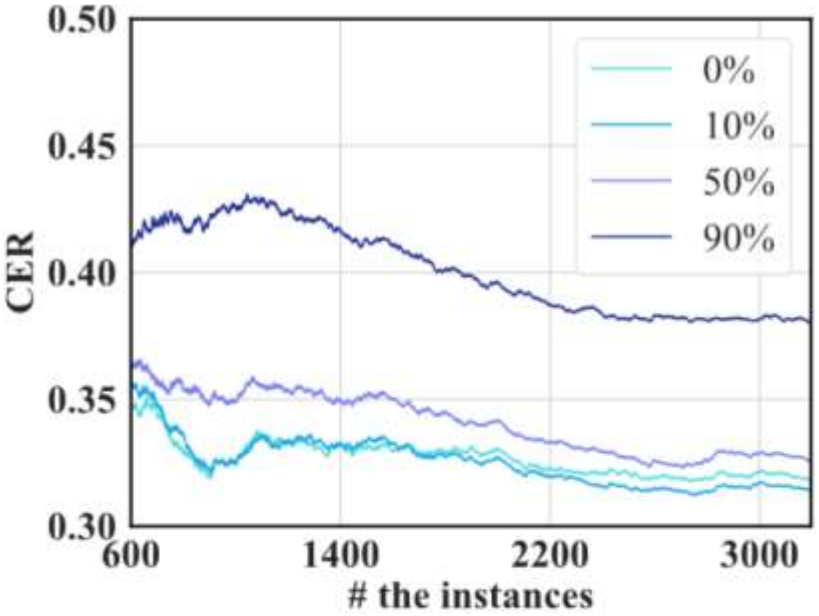


Our Method:
Space Mapping and Geometric structure building



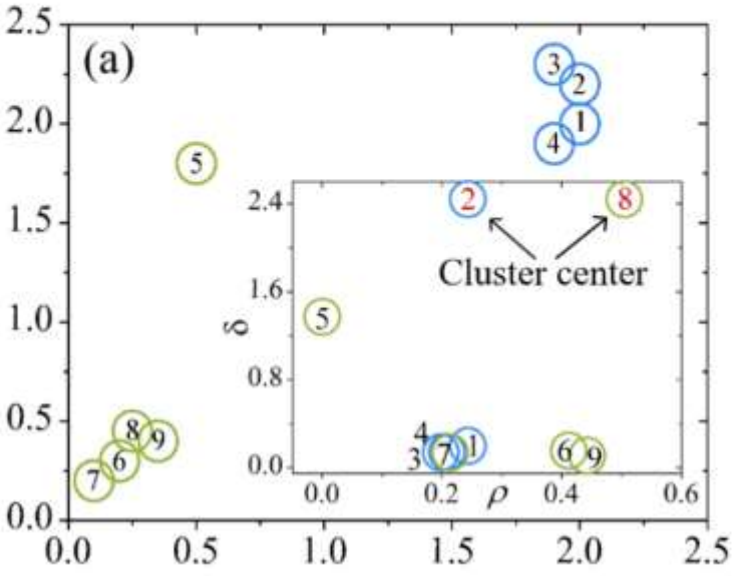
Online Correlation Estimation [Zhao and Udell, KDD'20]

- Accuracy of different labels missing

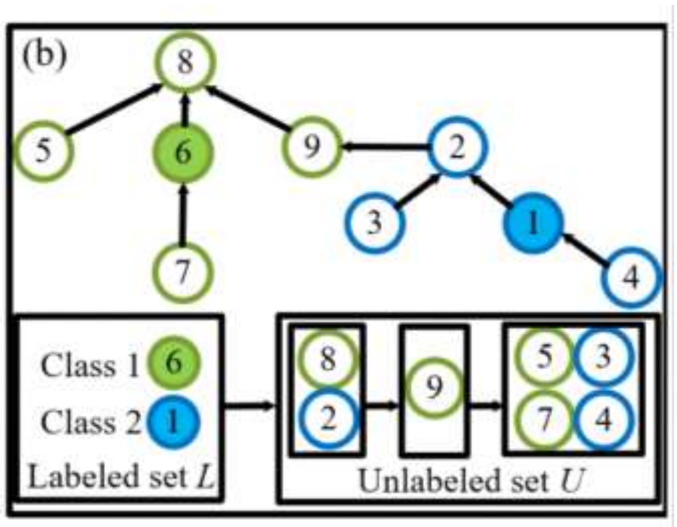


Our Method: Space Mapping and Geometric structure building

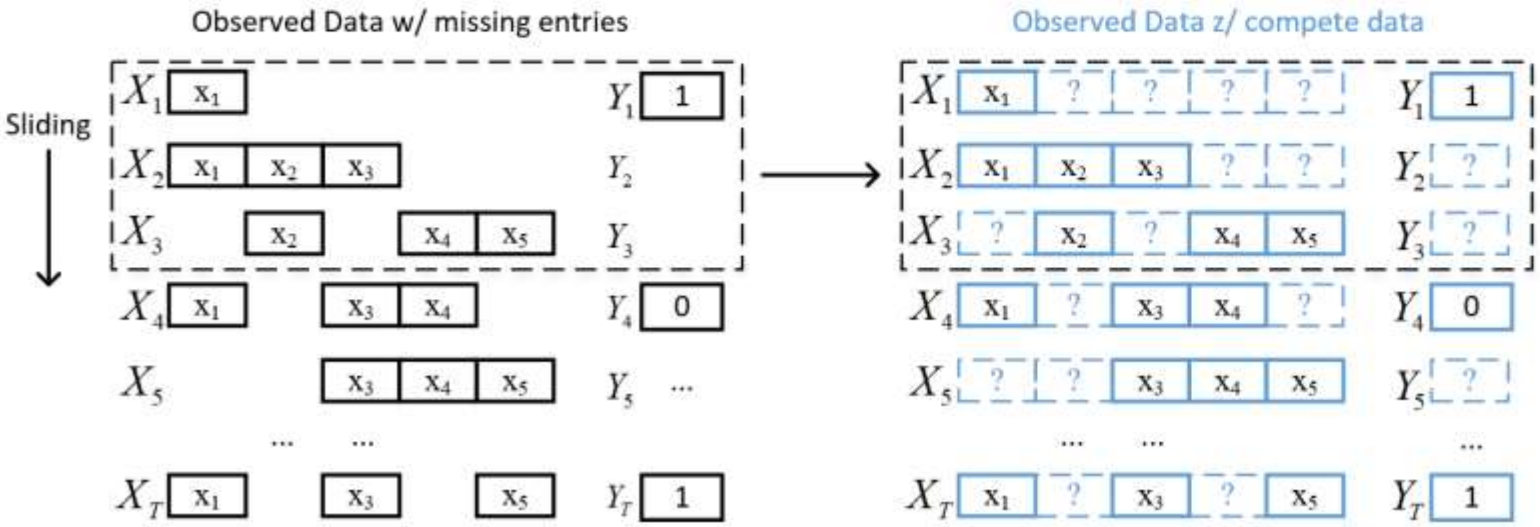
- Learning Cluster center



- Constructing geometric spaces



Ensemble Learning to further improve



Learner 1

Learner 2

$$\hat{y_t} = \alpha_1 \cdot y_O + \alpha_2 \cdot y_Z$$

$$\alpha_1 + \alpha_2 = 1$$

Too few observable entries

1. Imputation noises
2. Weak predictions

Hedge Reweighing: $\mathcal{O}(\sqrt{T})$ Regret

$$\alpha_1 = e^{-\mu R_O(T)} / \left(e^{-\mu R_O(T)} + e^{-\mu R_Z(T)}\right)$$

Empirical Results

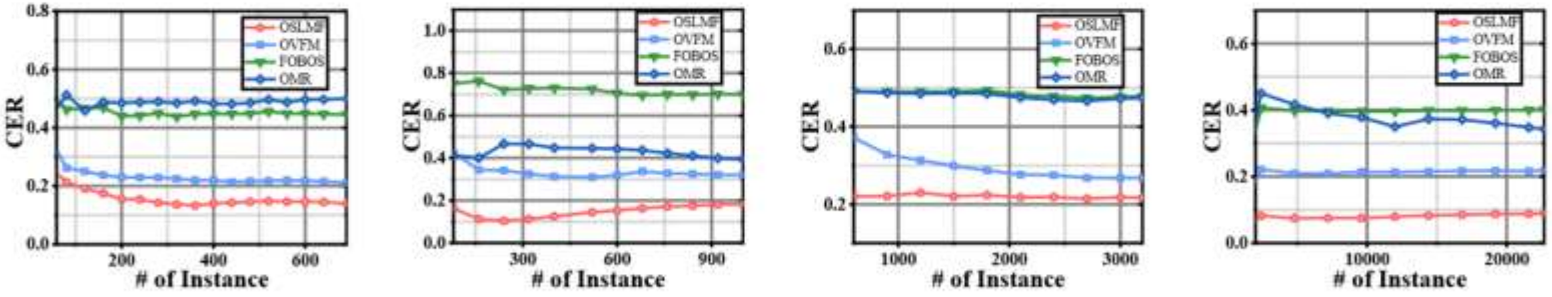
- Superiority over trapezoidal and VPS competitors
- In mixed data, more advanced methods tend to lose to simple baseline

	Trapezoidal Data Streams				Capricious Data Streams			
Dataset	FOBOS	OMR	OLSF	OSLMF	FOBOS	OMR	OVFM	OSLMF
wpbc	.237 ± .000	.345 ± .000	.366 ± .001	**.235 ± .003**	.248 ± .000•	.320 ± .000•	.309 ± .000•	.567 ± .001
ionosphere	.342 ± .000	.443 ± .000	.230 ± .000	**.225 ± .000**	.479 ± .000	.418 ± .000•	.269 ± .000•	.466 ± .000
wdbc	.577 ± .000	.460 ± .000	.347 ± .000	**.187 ± .000**	.628 ± .000	.399 ± .000	.113 ± .000	**.110 ± .000**
australian	.497 ± .000	.491 ± .000	.486 ± .000	**.356 ± .000**	.455 ± .000	.492 ± .001	.255 ± .000	**.194 ± .000**
credit-a	.445 ± .000	.415 ± .000	.312 ± .000	**.186 ± .000**	.445 ± .000	.484 ± .000	.484 ± .000	**.416 ± .000**
wbc	.345 ± .000	.394 ± .000	.455 ± .000	**.219 ± .000**	.162 ± .000	.461 ± .000	.072 ± .000	**.059 ± .000**
diabetes	.349 ± .000	.376 ± .000	.331 ± .000	**.170 ± .000**	.349 ± .000	.426 ± .000	.399 ± .000	**.331 ± .004**
dna	.518 ± .000	.496 ± .000	.499 ± .000	**.462 ± .000**	.511 ± .000	.496 ± .000	.282 ± .000	**.229 ± .000**
german	.300 ± .000	.381 ± .000	.407 ± .000	**.227 ± .000**	.700 ± .000	.372 ± .000	.321 ± .000	**.227 ± .000**
splice	.500 ± .000	.493 ± .000	.375 ± .000	**.311 ± .000**	.519 ± .000	.400 ± .001	.498 ± .000	**.424 ± .000**
kr-vs-kp	.482 ± .000	.523 ± .000	.239 ± .000	**.221 ± .000**	.478 ± .000	.242 ± .000	.280 ± .000	**.241 ± .000**
magic04	.665 ± .000	.529 ± .000	.374 ± .000	**.348 ± .000**	.689 ± .000	.438 ± .000	.317 ± .000	**.091 ± .000**
a8a	.375 ± .000	.482 ± .003	.273 ± .004	**.179 ± .001**	.401 ± .003	.368 ± .001	.191 ± .001	**.086 ± .001**
stream	.615 ± .000	.472 ± .000	.233 ± .000	**.230 ± .000**	.621 ± .000	.471 ± .000	.231 ± .000	**.224 ± .000**
loss/win	0/14	0/14	0/14	0/42	1/13	2/12	2/12	5/37
p-value	.0005	.0005	.0005	−−	.0008	.0015	.0071	−−
F-rank	3.286	3.124	2.500	1.000	3.357	3.071	2.286	1.285

Table 2: The comparison results on cumulative error rates. We repeated the experiment 10 times for each dataset, averaged the cumulative error rate (CER), and calculated the variance of the 10 times values. Experimental results (CER ± Variance) for 14 data sets in the case of trapezoidal and capricious data streams. • indicates the cases that OSLMF loses the comparison.

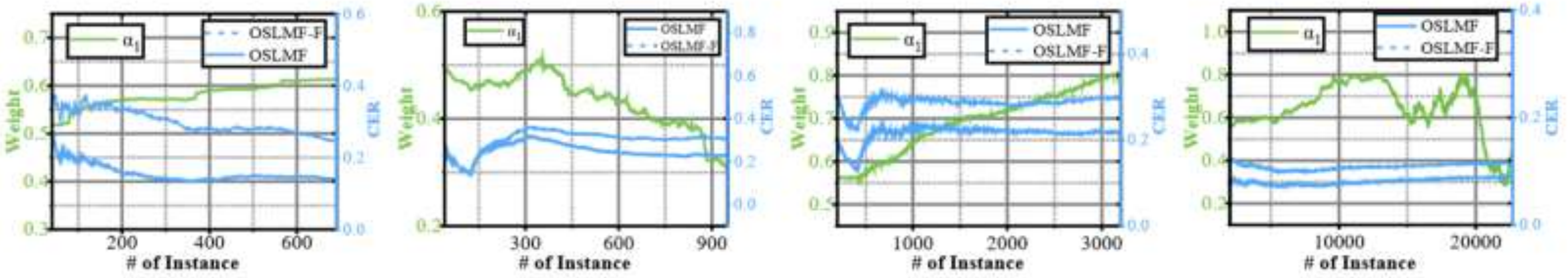
Empirical Results

- Superiority over trapezoidal and VPS competitors
- In mixed data, more advanced methods tend to lose to simple baseline



(a) australian (b) german (c) kr-vs-kp (d) a8a

Figure 3: CER trends of four methods in capricious data streams.



(a) australian (b) dna (c) kr-vs-kp (d) a8a

Figure 4: Temporal variation of ensemble weight α_1 and CERs of OSLMF and its ablation variant OSLMF-F.

Thanks

Q & A