Attention-based Learning for Missing Data Imputation in HoloClean

Richard Wu[1], Aoqian Zhang[1], Ihab F. Ilyas[1]
Theodoros Rekatsinas[2]

UNIVERSITY OF WATERLOO

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

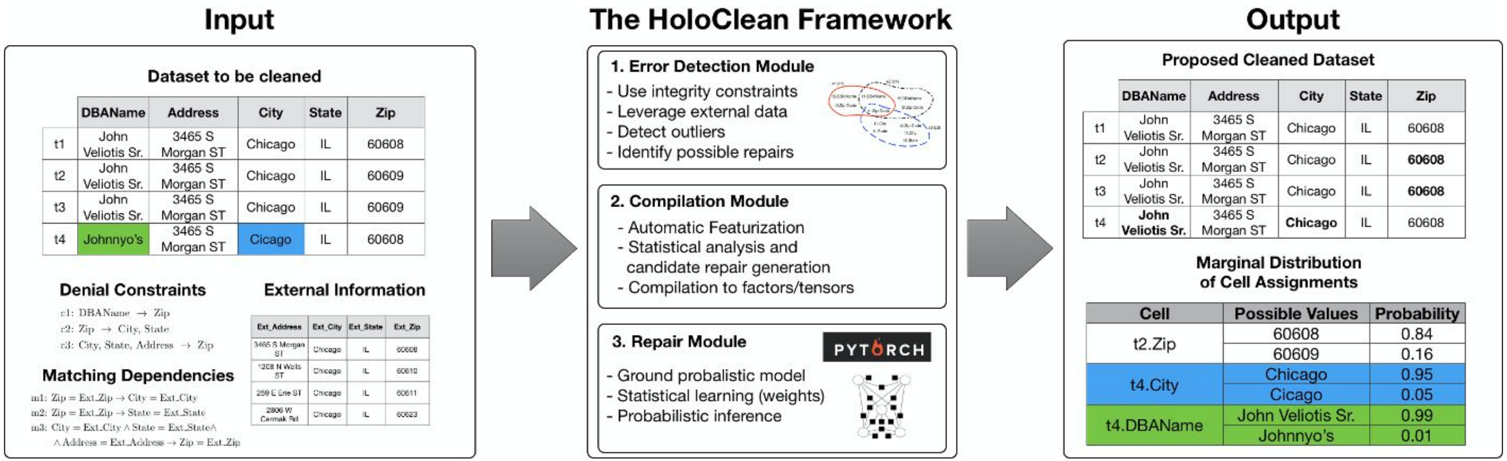
Problem

- Missing data is a **persistent** problem in many fields
  - Sciences
  - Data mining
  - Finance
- Missing data can reduce downstream statistical power
- Most models require complete data

Modern ML for Data Cleaning: HoloClean

- Framework for holistic data repairing driven by probabilistic inference
- Unifies **qualitative** (integrity constraints and external sources) with **quantitative** data repairing methods (statistical inference)

Available at www.holoclean.io



Missing Values in *Real* Data sets

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
0	douglas	Male	8/6/1993	12:42 PM	97308	6.945	True	Marketing
1	thomas	Male	3/31/1996	6:53 AM	61933	4.170	True	NaN
2	maria	Female	4/23/1993	11:17 AM	130590	11.858	False	Finance
3	jerry	Male	3/4/2005	1:00 PM	138705	9.340	True	Finance
4	larry	Male	1/24/1998	4:47 PM	101004	1.389	True	Client Services
5	dennis	Male	4/18/1987	1:35 AM	115163	10.125	False	Legal
6	ruby	Female	8/17/1987	4:20 PM	65476	10.012	True	Product
7	NaN	Female	7/20/2015	10:43 AM	45906	11.598	NaN	Finance
8	angela	Female	11/22/2005	6:29 AM	95570	18.523	True	Engineering
9	frances	Female	8/8/2002	6:51 AM	139852	7.524	True	Business Development
10	louise	Female	8/12/1980	9:01 AM	63241	15.132	True	NaN
11	julie	Female	10/26/1997	3:19 PM	102508	12.637	True	Legal
12	brandon	Male	12/1/1980	1:08 AM	112807	17.492	True	Human Resources

Challenges

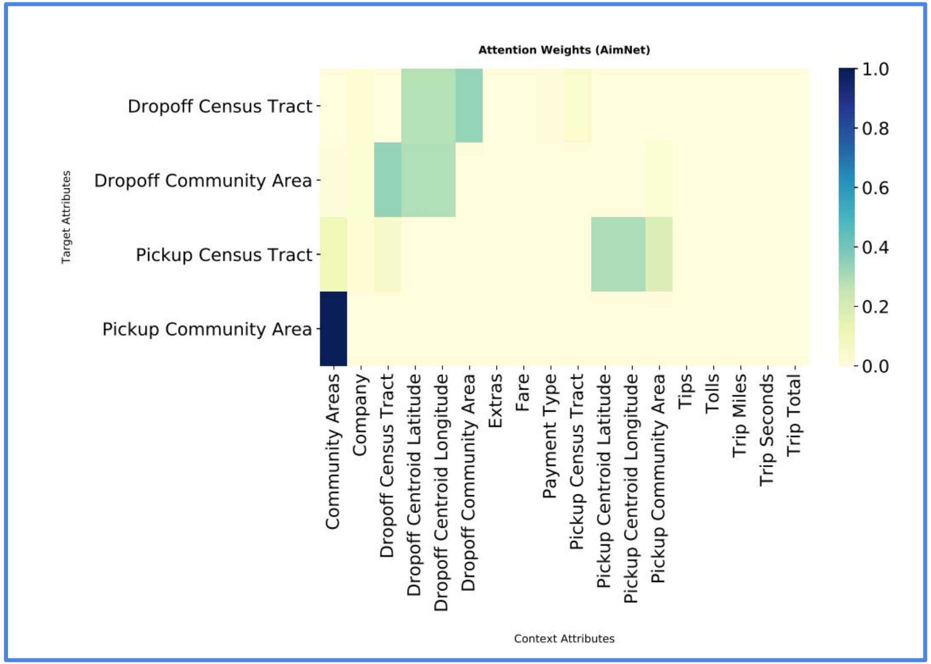
- Values may not be missing completely at random (MCAR/i.i.d.) but **systematically**
- Mixed types (discrete and continuous) introduce **mixed distributions**
- **Drawbacks** of current methods:
  - Heuristic-based (impute **mean/mode**)
  - Requires predefined **rules**
  - **Complex** ML models that are **difficult to train**, **slow**, **hard to interpret**

Contribution

*A **simple** attention architecture that exploits **structure** across attributes*

Our results:

- >**54%** lower run time than baselines
- *Missing at random (MCAR)* : **3%** higher accuracy and **26.7%** reduction in normalized-RMS
- *Systematic*: **43%** higher accuracy and **7.4%** reduction in normalized-RMS



How does AimNet improve on the MVI problem?

Key idea:
Exploit the structure in data

model that learns **schema-level relationships** between attributes

dot product attention

Architecture overview

(1) Model mixed data

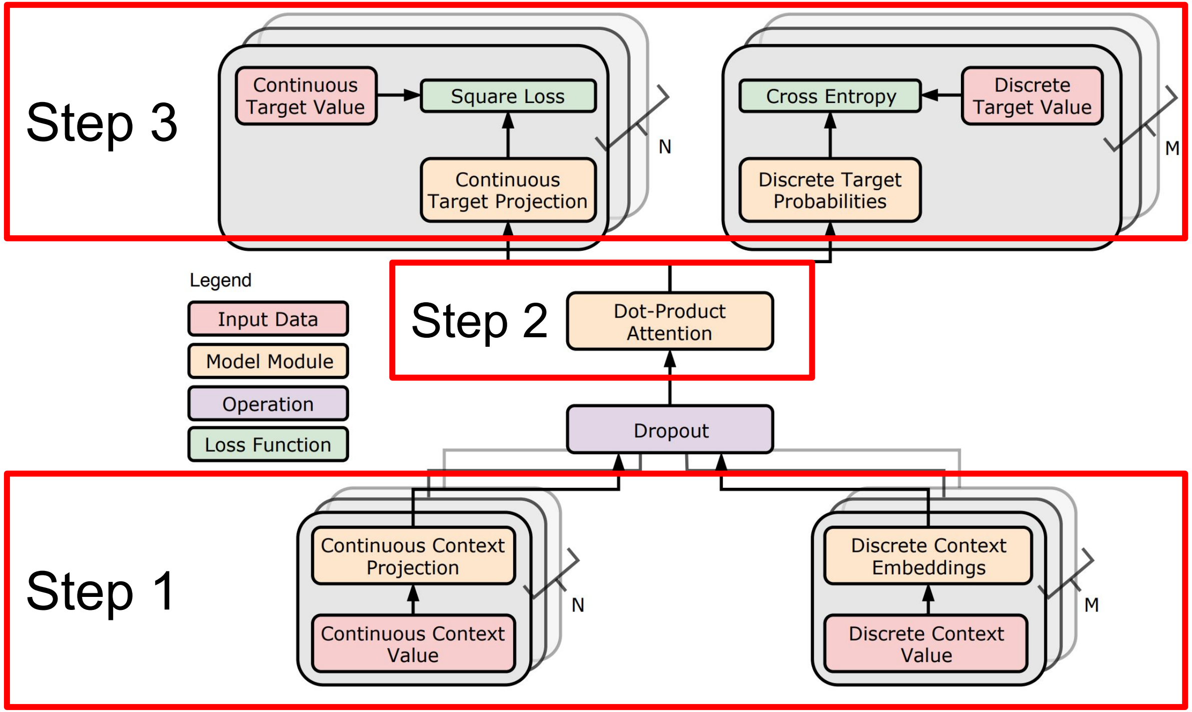
- Encode w/ non-linear layers (continuous)
- Embedding lookup (discrete)

(2) Identify relevant context

- Attention helps identify schema-level importance

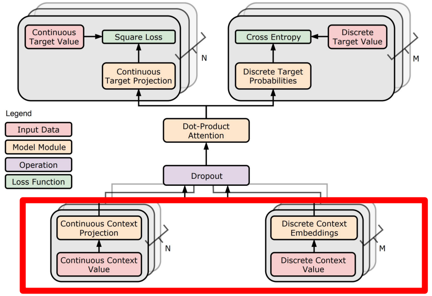
(3) Prediction

- Inverse of encoding (continuous)
- Softmax over possible values (discrete)



Learned via self-supervision: mask and predict observed values

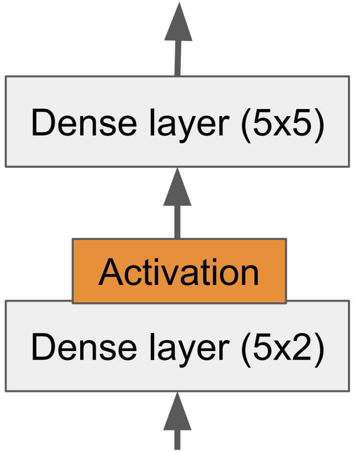
How do we encode mixed types?



Convert **context values** to **vector embeddings.**

Output: embeddings

[0.1, 1.2, -5, 2, 15]



[-12, 3.5]

Continuous values

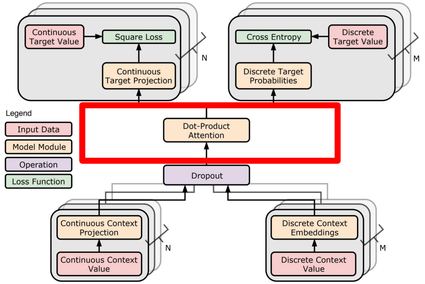
[1, 0, -1.3, 5, -7]

(Name, Joe)	[0, 2, -1, 2.5, 1]
(City, Chicago)	**[1, 0, -1.3, 5, -7]**
(Zip Code, 10010)	...

(City, Chicago)

Discrete values

Input: raw data



Attention layer

Attention where Q/K are derived from **attributes** rather than values

[-1, 5, 0.5, 1.2, -2]

Output: context vector

Target: County (K)

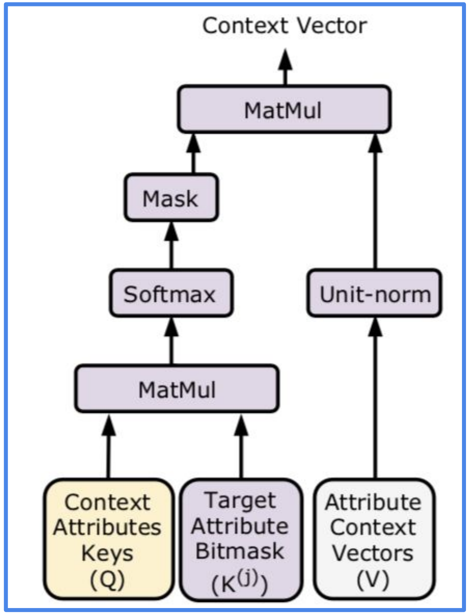
[0.09, 0.90, 0.01]

$\mathrm{softmax}(QK_{County}^{T})$

(City, Chicago) [1, 0, -1.3, 5, -7] (V_{City})

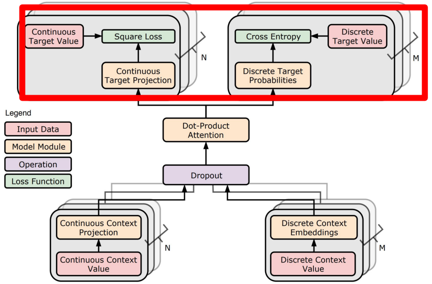
(Zip code, 60603) [1.2, 0.5, -2, 3, 5] ($V_{Zip\ code}$)

(Age, 35) [0, 1, 2, 3, -1.5] (V_{age})



Prediction

Input: context vector [-1, 5, 0.5, 1.2, -2]



Dense layer (5x5)

Activation

Dense layer (1x5)

Output: 100600

Salary (continuous)

matmul

County A: $[0, 100, 0, 0, 0]^T$

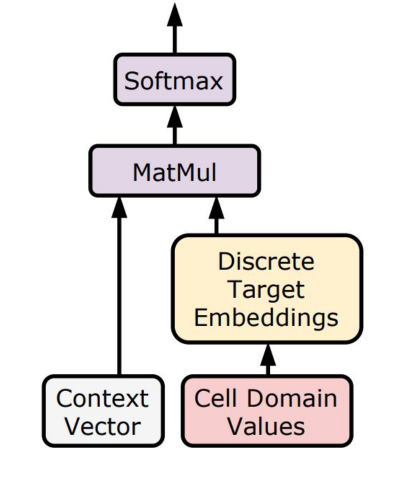
County B: $[0, 0, 0, 0, 50]^T$

softmax

[0.99, 0.01]

Output: County A

County (discrete)



(c) Discrete Target Probabilities

Questions

- Can AimNet impute missing completely at random (**MCAR/i.i.d.**) values?
- Does AimNet's emphasis on structure help it with **systematic bias** in missing values?
- Can we **interpret** the structure that AimNet learns in the data?

Experimental setup

- 14 real data sets
- Missing types
  - **MCAR/i.i.d.**
  - **Systematic**
- Evaluation
  - Accuracy (discrete)
  - normalized-RMS (continuous)

Mostly discrete

Data Set	$\|r\|$	# Continuous Attributes	# Discrete Attributes
Tic-Tac-Toe	958	0	10
Hospital	1000	2	14
Mammogram	831	1	5
Thoracic	470	3	14
Contraceptive	1473	2	8
Solar Flare	1066	3	10
NYPD	32399	4	13
Credit	653	6	10
Australian	691	6	9
Chicago	400k	11	7
Balance	625	4	1
Eye EEG	14976	14	1
Phase	9628	4	0
CASP	45730	10	0

Mostly continuous

$$NRMS_j = \left(\sum_i^{n_j} ((y_i - \hat{y}_i)^2) / (n_j \cdot \sigma(\vec{y})^2)\right)^{-1/2}$$

- Training: self-supervised learning where targets = observable values

Experiment results

- >**54%** lower run time than baselines
- *Missing at random (MCAR)* : **3%** higher accuracy and **26.7%** reduction in normalized-RMS
- *Systematic*: **43%** higher accuracy and **7.4%** reduction in normalized-RMS

Attention identifies **structure between attributes** that helps it deal with systematic bias in missing values

MCAR (20%)

AimNet outperforms on both **discrete** and **continuous** attributes on almost all data sets

- 3% in accuracy
- 26.7% in NRMS

HCQ	XGB	MIDAS	GAIN	MF	MICE
HoloClean with quantization	XGBoost	Denoising Autoencoder	GAN	Random Forest	Linear regression with multiple iterations

data set	Accuracy on discrete attributes (ACC ± std)						
	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
Tic-Tac-Toe	**0.61 ± 0.01**	0.53 ± 0.01	0.57 ± 0.02	0.46 ± 0.01	0.32 ± 0.01	0.52 ± 0.01	0.58 ± 0.02
Hospital	**0.99 ± 0.0**	**0.99 ± 0.0**	0.97 ± 0.01	0.24 ± 0.0	0.13 ± 0.01	**0.99 ± 0.0**	0.82 ± 0.01
Mammogram	**0.75 ± 0.01**	**0.74 ± 0.02**	**0.74 ± 0.02**	**0.74 ± 0.01**	0.35 ± 0.02	0.68 ± 0.02	0.64 ± 0.02
Thoracic	**0.86 ± 0.01**	0.84 ± 0.01	**0.85 ± 0.01**	0.84 ± 0.01	0.59 ± 0.09	**0.86 ± 0.01**	0.38 ± 0.4
Contraceptive	**0.65 ± 0.01**	**0.64 ± 0.01**	0.63 ± 0.01	0.63 ± 0.01	0.42 ± 0.01	0.63 ± 0.02	0.57 ± 0.01
Solar Flare	**0.78 ± 0.02**	**0.77 ± 0.02**	0.76 ± 0.01	0.65 ± 0.01	0.48 ± 0.02	0.76 ± 0.02	0.67 ± 0.02
NYPD	0.92 ± 0.0	0.89 ± 0.0	**0.93 ± 0.0**	0.79 ± 0.01	0.14 ± 0.01	0.92 ± 0.0	0.72 ± 0.0
Credit	**0.76 ± 0.01**	0.73 ± 0.02	0.75 ± 0.01	0.61 ± 0.02	0.4 ± 0.01	**0.76 ± 0.01**	0.68 ± 0.01
Australian	**0.72 ± 0.02**	0.69 ± 0.02	0.71 ± 0.02	0.61 ± 0.03	0.45 ± 0.01	**0.73 ± 0.01**	0.63 ± 0.02
Balance	**0.79 ± 0.04**	**0.78 ± 0.04**	0.75 ± 0.05	0.68 ± 0.08	0.5 ± 0.05	0.69 ± 0.05	0.72 ± 0.05
Eye EEG	0.71 ± 0.01	0.63 ± 0.01	0.81 ± 0.01	0.55 ± 0.01	0.54 ± 0.01	**0.87 ± 0.01**	0.54 ± 0.01

data set	NRMS on continuous attributes (NRMS ± std)						
	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
Hospital	**0.72 ± 0.06**	1.4 ± 0.36	0.87 ± 0.07	611.12 ± 129.15	1.19 ± 0.02	0.86 ± 0.07	1.13 ± 0.13
Mammogram	**0.91 ± 0.04**	1.03 ± 0.08	0.96 ± 0.06	1.12 ± 0.05	1.0 ± 0.11	0.99 ± 0.05	1.27 ± 0.11
Thoracic	**1.1 ± 0.41**	1.78 ± 2.33	1.76 ± 2.45	1.71 ± 1.17	**1.5 ± 0.95**	1.66 ± 1.61	3.62 ± 4.82
Contraceptive	**0.84 ± 0.02**	1.06 ± 0.05	0.87 ± 0.04	1.09 ± 0.03	1.13 ± 0.05	0.88 ± 0.04	1.14 ± 0.06
Solar Flare	**0.94 ± 0.15**	**0.94 ± 0.13**	1.21 ± 0.53	17698.23 ± 8959.74	**0.96 ± 0.16**	**0.96 ± 0.2**	1.22 ± 0.34
NYPD	0.15 ± 0.01	1.28 ± 0.53	0.14 ± 0.0	0.62 ± 0.04	3.19 ± 0.41	**0.1 ± 0.0**	0.37 ± 0.01
Credit	**0.94 ± 0.03**	1.84 ± 0.83	1.09 ± 0.15	1.29 ± 0.24	1.18 ± 0.09	1.04 ± 0.13	1.84 ± 0.83
Australian	**0.94 ± 0.03**	2.47 ± 2.0	1.09 ± 0.12	1.22 ± 0.13	1.24 ± 0.18	1.07 ± 0.24	1.58 ± 0.28
Balance	**0.92 ± 0.02**	1.37 ± 0.08	1.0 ± 0.03	1.02 ± 0.02	1.03 ± 0.03	1.08 ± 0.05	1.26 ± 0.07
Eye EEG	0.4 ± 0.0	0.94 ± 0.34	0.39 ± 0.0	0.84 ± 0.01	0.65 ± 0.04	**0.35 ± 0.0**	0.62 ± 0.0
Phase	**0.45 ± 0.01**	0.54 ± 0.03	**0.45 ± 0.01**	0.95 ± 0.01	0.76 ± 0.14	0.5 ± 0.01	0.63 ± 0.01
CASP	0.45 ± 0.02	1.45 ± 0.14	0.43 ± 0.02	0.82 ± 0.01	0.72 ± 0.04	**0.41 ± 0.02**	0.64 ± 0.03

Chicago taxi data set

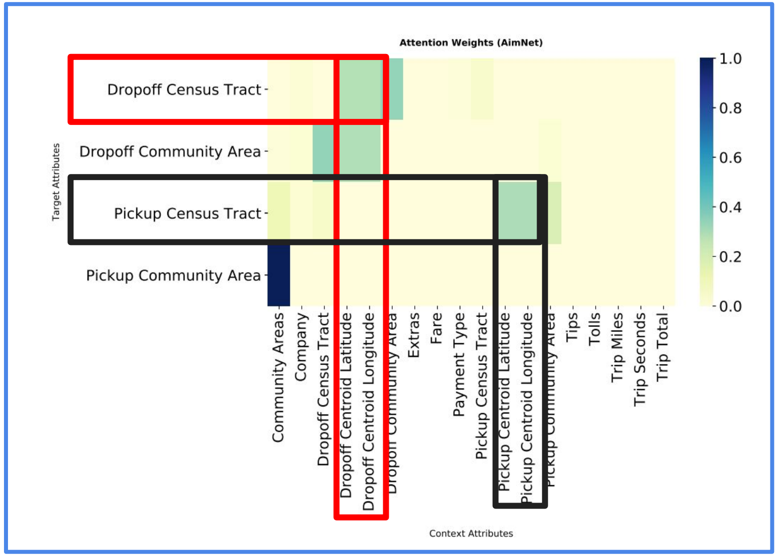
- Benchmark in TFX data validation pipeline
- Pickup/dropoff info, fare, company
- Naturally-occurring missing values w/ ground truth
- Systematic bias between companies

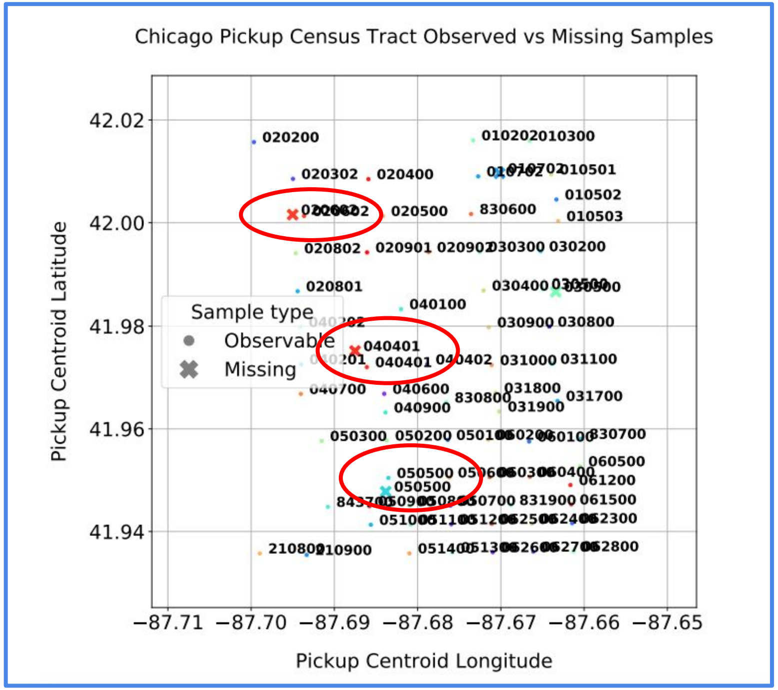
	Company	Pickup Census Tract	Pickup Centroid Latitude	Pickup Centroid Longitude
2366	Chicago Medallion Leasing INC	nan	41.975171	-87.687516
78445	Dispatch Taxi Affiliation	nan	41.975171	-87.687516
57109	Taxi Affiliation Services	17031040401	41.972036	-87.686100

All within "17031040401" census tract

Chicago taxi: naturally-occurring missing data

- Values are missing systematically (not i.i.d.)
- **Attention** learns relationship between **Census Tract** and **Latitude/Longitude**





Chicago taxi results

AimNet outperforms baselines by a **huge margin**

- Accuracy: **73%** vs 27% (XGB)
- Run time: **53 mins.** vs 124 mins (HoloClean w/ Quantization)

Accuracy on discrete attributes for the Chicago data set						
AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
$\mathbf{0.73 \pm 0.01}$	0.07 ± 0.0	0.27 ± 0.0	0.09 ± 0.01	0.01 ± 0.01	0.3 ± 0.0	—
Run time (minutes) for the Chicago data set						
AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
53	124	5350	176	186	7439	—

What if we inject systematic errors into other real data sets?

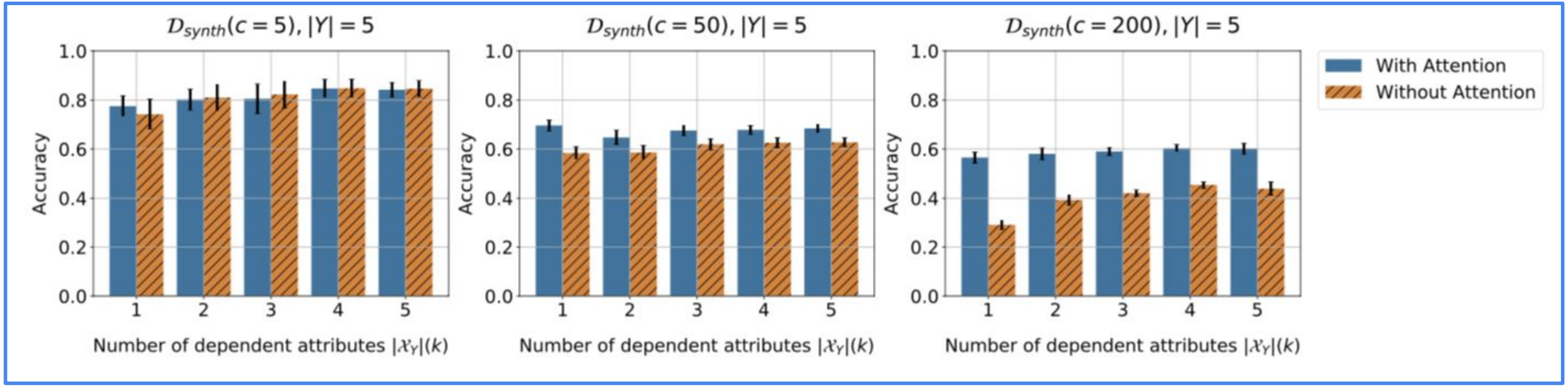
AimNet still outperforms baselines in almost all cases

data set	Attribute	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
		Accuracy on discrete attributes (ACC ± std)						
Balance	class	**0.83 ± 0.07**	0.4 ± 0.37	0.48 ± 0.32	0.7 ± 0.16	0.5 ± 0.12	0.46 ± 0.34	**0.78 ± 0.15**
NYPD	ADDR_PCT_CD	**0.67 ± 0.03**	0.19 ± 0.05	0.41 ± 0.05	0.13 ± 0.01	0.04 ± 0.04	0.59 ± 0.03	0.23 ± 0.01
	BORO_NM	**0.92 ± 0.07**	**0.85 ± 0.09**	0.58 ± 0.18	0.78 ± 0.04	0.23 ± 0.03	0.84 ± 0.09	0.58 ± 0.09
	PATROL_BORO	**0.83 ± 0.07**	0.69 ± 0.03	0.57 ± 0.17	0.6 ± 0.06	0.13 ± 0.02	0.72 ± 0.1	0.58 ± 0.07
data set	Attribute	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
		NRMS on continuous attributes (NRMS ± std)						
Phase	A	**0.75 ± 0.13**	1.26 ± 0.46	**0.81 ± 0.18**	1.45 ± 0.5	2.04 ± 1.73	0.94 ± 0.18	1.41 ± 0.66
	B	**0.77 ± 0.07**	0.99 ± 0.27	**0.83 ± 0.16**	1.18 ± 0.33	1.54 ± 0.54	0.98 ± 0.13	1.32 ± 0.4
	C	**0.79 ± 0.11**	1.25 ± 0.26	**0.81 ± 0.13**	1.44 ± 0.25	1.33 ± 0.34	1.09 ± 0.23	1.29 ± 0.1
	D	**0.62 ± 0.09**	1.1 ± 0.24	**0.65 ± 0.09**	1.47 ± 0.56	1.57 ± 0.63	0.85 ± 0.17	1.03 ± 0.13
Chicago	Trip Total	0.82 ± 0.21	2.38 ± 1.16	**0.48 ± 0.32**	3.84 ± 0.7	1.87 ± 0.7	**0.71 ± 0.57**	—
	Fare	**1.35 ± 0.76**	5.18 ± 1.94	**1.16 ± 0.73**	12.73 ± 7.93	58.09 ± 36.03	2.78 ± 3.0	—
	Tips	**0.52 ± 0.01**	1.29 ± 0.09	**0.5 ± 0.12**	1.59 ± 0.49	11.64 ± 6.76	0.8 ± 0.28	—

Does the attention layer actually help?

As the **domain size increases**, attention leads to better performance

- Learns schema-level dependencies



5 classes 50 classes 200 classes

Architecture summary

- **Encode**: learns projections for **continuous** and embeddings for **discrete** data
- **Structure**: new variation of attention to learn structural dependencies between **attributes**
- **Prediction**: mixed-type prediction using projections (continuous) and softmax classification (discrete)

Conclusion

- A **simple attention-based architecture** modestly outperforms existing methods on i.i.d. missing values
- AimNet outperforms state of the art in the presence of **systematically** missing values by **a large margin**
- **Attention** mechanism learns **structural properties of the data** which improves MVI with systematic bias

Appendix

Hyperparameter Sensitivity

Accuracy on discrete attributes for the NYPD data set				
	dropout	max domain size	embedding size	AimNet
base	0.25	50	64	0.921
(dropout rate)	0.0			0.918
	0.5			0.920
(max domain size)		10		0.917
		100		0.921
(embedding size)			16	0.920
			32	0.921
			128	**0.922**
			256	0.920

NRMS on continuous attributes for the NYPD data set				
	dropout rate	max domain size	embedding size	AimNet
base	0.0	50	64	**0.150**
(dropout rate)	0.25			0.281
	0.5			0.509
(embedding size)			16	0.159
			32	**0.150**
			128	0.153
			256	**0.144**

Multi-task and Single-task

Data Set	Accuracy on discrete attributes (ACC ± std) Single	Multi-Task
Tic-Tac-Toc	$\mathbf{0.61 \pm 0.01}$	$\mathbf{0.61 \pm 0.01}$
Hospital	$\mathbf{0.99 \pm 0.0}$	$\mathbf{0.99 \pm 0.0}$
Mammogram	$\mathbf{0.75 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$
Thoracic	$\mathbf{0.86 \pm 0.01}$	$\mathbf{0.86 \pm 0.01}$
Contraceptive	$\mathbf{0.65 \pm 0.01}$	$\mathbf{0.65 \pm 0.01}$
Solar Flare	$\mathbf{0.78 \pm 0.02}$	$\mathbf{0.78 \pm 0.01}$
NYPD	$\mathbf{0.92 \pm 0.0}$	$\mathbf{0.92 \pm 0.0}$
Credit	$\mathbf{0.76 \pm 0.01}$	$\mathbf{0.76 \pm 0.01}$
Australian	$\mathbf{0.72 \pm 0.02}$	$\mathbf{0.72 \pm 0.01}$
Chicago	$\mathbf{0.73 \pm 0.01}$	0.7 ± 0.03
Balance	$\mathbf{0.79 \pm 0.04}$	$\mathbf{0.79 \pm 0.04}$
Eye EEG	$\mathbf{0.71 \pm 0.01}$	0.69 ± 0.01
Data Set	**NRMS on continuous attributes (NRMS ± std) Single**	**Multi-Task**
Hospital	$\mathbf{0.72 \pm 0.06}$	$\mathbf{0.77 \pm 0.06}$
Mammogram	$\mathbf{0.91 \pm 0.04}$	$\mathbf{0.93 \pm 0.03}$
Thoracic	$\mathbf{1.1 \pm 0.41}$	$\mathbf{1.3 \pm 0.83}$
Contraceptive	$\mathbf{0.84 \pm 0.02}$	$\mathbf{0.84 \pm 0.02}$
Solar Flare	$\mathbf{0.94 \pm 0.15}$	$\mathbf{0.87 \pm 0.16}$
NYPD	$\mathbf{0.15 \pm 0.01}$	$\mathbf{0.15 \pm 0.01}$
Credit	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.94 \pm 0.03}$
Australian	$\mathbf{0.94 \pm 0.03}$	$\mathbf{0.93 \pm 0.02}$
Eye EEG	$\mathbf{0.4 \pm 0.0}$	0.44 ± 0.0
Phase	$\mathbf{0.45 \pm 0.01}$	$\mathbf{0.45 \pm 0.01}$
CASP	$\mathbf{0.45 \pm 0.02}$	0.48 ± 0.01

Data Set	Run time (seconds) Single	Multi-Task
Tic-Tac-Toc	**9**	38
Hospital	**18**	148
Mammogram	**9**	39
Thoracic	**11**	69
Contraceptive	**15**	102
Solar Flare	**15**	131
NYPD	**378**	6320
Credit	**15**	198
Australian	**16**	145
Chicago	**3180**	462756
Balance	**11**	12
Eye EEG	**188**	5306
Phase	**66**	250
CASP	**648**	5800

MCAR (40% missing) results

Data Set	Accuracy on discrete attributes (ACC ± std)						
	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
Tic-Tac-Toc	$\mathbf{0.53 \pm 0.01}$	0.5 ± 0.01	0.52 ± 0.01	0.44 ± 0.02	0.35 ± 0.01	0.5 ± 0.01	0.46 ± 0.01
Hospital	$\mathbf{0.95 \pm 0.0}$	$\mathbf{0.95 \pm 0.0}$	0.91 ± 0.01	0.24 ± 0.01	0.14 ± 0.02	0.94 ± 0.01	0.7 ± 0.01
Mammogram	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.72 \pm 0.01}$	$\mathbf{0.72 \pm 0.02}$	0.71 ± 0.02	0.35 ± 0.01	0.66 ± 0.02	0.63 ± 0.02
Thoracic	$\mathbf{0.85 \pm 0.01}$	0.84 ± 0.01	0.84 ± 0.01	0.83 ± 0.02	0.52 ± 0.15	$\mathbf{0.85 \pm 0.01}$	0.75 ± 0.03
Contraceptive	$\mathbf{0.63 \pm 0.01}$	$\mathbf{0.63 \pm 0.01}$	0.62 ± 0.01	0.62 ± 0.01	0.43 ± 0.01	0.62 ± 0.01	0.55 ± 0.01
Solar Flare	$\mathbf{0.76 \pm 0.01}$	0.75 ± 0.01	0.75 ± 0.01	0.66 ± 0.01	0.46 ± 0.02	0.74 ± 0.01	0.65 ± 0.01
NYPD	0.87 ± 0.0	0.85 ± 0.0	$\mathbf{0.88 \pm 0.0}$	0.75 ± 0.0	0.15 ± 0.01	$\mathbf{0.88 \pm 0.0}$	0.58 ± 0.0
Credit	$\mathbf{0.73 \pm 0.01}$	0.7 ± 0.01	$\mathbf{0.73 \pm 0.01}$	0.6 ± 0.01	0.39 ± 0.01	$\mathbf{0.73 \pm 0.01}$	0.63 ± 0.01
Australian	$\mathbf{0.7 \pm 0.01}$	0.66 ± 0.01	0.68 ± 0.01	0.6 ± 0.01	0.46 ± 0.01	0.69 ± 0.01	0.59 ± 0.01
Balance	$\mathbf{0.73 \pm 0.03}$	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.71 \pm 0.03}$	0.64 ± 0.03	0.45 ± 0.05	0.63 ± 0.05	0.64 ± 0.04
Eye EEG	0.67 ± 0.01	0.62 ± 0.01	0.73 ± 0.01	0.55 ± 0.01	0.52 ± 0.03	$\mathbf{0.78 \pm 0.01}$	0.53 ± 0.01

Data Set	NRMS on continuous attributes (NRMS ± std)						
	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
Hospital	$\mathbf{0.81 \pm 0.04}$	1.1 ± 0.08	0.92 ± 0.07	440.63 ± 61.35	2.26 ± 1.18	0.89 ± 0.04	1.23 ± 0.09
Mammogram	$\mathbf{0.92 \pm 0.02}$	1.02 ± 0.04	0.98 ± 0.05	1.12 ± 0.08	1.05 ± 0.06	1.01 ± 0.03	1.25 ± 0.07
Thoracic	$\mathbf{0.94 \pm 0.01}$	1.09 ± 0.05	1.03 ± 0.11	5.64 ± 7.16	1.23 ± 0.22	0.99 ± 0.06	1.32 ± 0.12
Contraceptive	$\mathbf{0.9 \pm 0.02}$	1.12 ± 0.04	0.94 ± 0.02	1.11 ± 0.02	1.17 ± 0.05	0.99 ± 0.02	1.23 ± 0.06
Solar Flare	$\mathbf{0.93 \pm 0.09}$	$\mathbf{0.98 \pm 0.09}$	$\mathbf{1.0 \pm 0.1}$	10772.7 ± 4057.37	$\mathbf{1.0 \pm 0.09}$	1.04 ± 0.11	1.16 ± 0.07
NYPD	0.32 ± 0.01	0.44 ± 0.13	0.28 ± 0.0	0.69 ± 0.03	3.63 ± 0.17	$\mathbf{0.22 \pm 0.01}$	0.62 ± 0.01
Credit	$\mathbf{0.97 \pm 0.03}$	1.24 ± 0.03	1.26 ± 0.41	1.15 ± 0.07	1.2 ± 0.08	1.12 ± 0.18	1.34 ± 0.11
Australian	$\mathbf{0.96 \pm 0.02}$	1.23 ± 0.03	1.19 ± 0.2	1.14 ± 0.12	1.27 ± 0.16	1.07 ± 0.13	1.6 ± 0.7
Eye EEG	0.48 ± 0.0	0.71 ± 0.03	0.47 ± 0.0	0.91 ± 0.01	1.0 ± 0.28	$\mathbf{0.44 \pm 0.0}$	0.67 ± 0.01
Phase	$\mathbf{0.52 \pm 0.01}$	0.58 ± 0.0	0.53 ± 0.01	0.97 ± 0.01	1.14 ± 0.26	0.58 ± 0.01	0.73 ± 0.01
CASP	0.5 ± 0.01	1.5 ± 0.26	$\mathbf{0.49 \pm 0.01}$	0.88 ± 0.01	0.83 ± 0.09	$\mathbf{0.48 \pm 0.01}$	0.73 ± 0.03

MCAR (60% missing) results

Data Set	Accuracy on discrete attributes (ACC ± std)						
	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
Tic-Tac-Toc	**0.48 ± 0.01**	**0.48 ± 0.0**	0.47 ± 0.01	0.43 ± 0.01	0.39 ± 0.01	0.44 ± 0.01	0.4 ± 0.01
Hospital	**0.86 ± 0.01**	**0.86 ± 0.01**	0.68 ± 0.01	0.24 ± 0.0	0.11 ± 0.01	0.79 ± 0.01	0.37 ± 0.01
Mammogram	**0.69 ± 0.01**	**0.69 ± 0.01**	**0.68 ± 0.01**	**0.68 ± 0.01**	0.34 ± 0.02	0.62 ± 0.02	0.58 ± 0.02
Thoracic	**0.85 ± 0.01**	0.84 ± 0.01	0.83 ± 0.0	0.84 ± 0.01	0.51 ± 0.13	0.84 ± 0.01	0.72 ± 0.04
Contraceptive	**0.62 ± 0.01**	**0.62 ± 0.01**	0.61 ± 0.01	0.61 ± 0.01	0.42 ± 0.02	0.6 ± 0.01	0.53 ± 0.01
Solar Flare	**0.72 ± 0.01**	**0.72 ± 0.01**	0.71 ± 0.01	0.66 ± 0.01	0.45 ± 0.03	0.7 ± 0.01	0.61 ± 0.01
NYPD	0.77 ± 0.0	0.76 ± 0.0	**0.79 ± 0.0**	0.67 ± 0.0	0.15 ± 0.0	0.78 ± 0.0	0.45 ± 0.0
Credit	**0.69 ± 0.01**	0.66 ± 0.01	0.68 ± 0.01	0.6 ± 0.01	0.38 ± 0.02	0.68 ± 0.01	0.57 ± 0.01
Australian	**0.67 ± 0.01**	0.65 ± 0.01	0.65 ± 0.01	0.59 ± 0.01	0.45 ± 0.02	0.65 ± 0.01	0.54 ± 0.02
Balance	**0.67 ± 0.03**	**0.64 ± 0.04**	0.63 ± 0.03	0.51 ± 0.06	0.46 ± 0.04	0.54 ± 0.03	0.55 ± 0.04
Eye EEG	0.63 ± 0.01	0.6 ± 0.01	0.66 ± 0.01	0.54 ± 0.01	0.52 ± 0.03	**0.67 ± 0.01**	0.52 ± 0.01

Data Set	NRMS on continuous attributes (NRMS ± std)						
	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
Hospital	**0.9 ± 0.04**	1.16 ± 0.12	1.01 ± 0.08	139.97 ± 24.15	3.79 ± 0.36	0.95 ± 0.05	1.27 ± 0.09
Mammogram	**0.94 ± 0.02**	1.05 ± 0.06	1.0 ± 0.05	1.13 ± 0.06	1.1 ± 0.11	1.01 ± 0.04	1.28 ± 0.08
Thoracic	**0.99 ± 0.02**	1.13 ± 0.05	1.17 ± 0.11	3.54 ± 5.28	1.18 ± 0.09	1.07 ± 0.04	1.43 ± 0.2
Contraceptive	**0.94 ± 0.01**	1.15 ± 0.04	1.01 ± 0.02	1.12 ± 0.02	1.31 ± 0.14	1.14 ± 0.04	1.29 ± 0.05
Solar Flare	**0.98 ± 0.06**	**1.01 ± 0.06**	1.05 ± 0.09	4454.64 ± 1610.59	1.07 ± 0.17	1.13 ± 0.14	1.24 ± 0.19
NYPD	0.56 ± 0.0	0.58 ± 0.04	0.45 ± 0.0	0.8 ± 0.01	3.51 ± 0.14	**0.42 ± 0.01**	0.98 ± 0.0
Credit	**0.99 ± 0.01**	1.33 ± 0.19	1.23 ± 0.25	1.14 ± 0.11	1.24 ± 0.11	1.13 ± 0.13	1.43 ± 0.26
Australian	**0.98 ± 0.01**	1.37 ± 0.26	1.29 ± 0.42	1.1 ± 0.04	1.25 ± 0.12	1.13 ± 0.19	1.38 ± 0.09
Eye EEG	0.59 ± 0.0	0.82 ± 0.04	**0.57 ± 0.0**	0.97 ± 0.01	1.61 ± 0.24	**0.57 ± 0.0**	0.79 ± 0.0
Phase	**0.64 ± 0.0**	0.71 ± 0.01	0.65 ± 0.0	1.0 ± 0.0	1.45 ± 0.51	0.71 ± 0.01	0.91 ± 0.01
CASP	**0.58 ± 0.01**	2.06 ± 0.51	0.59 ± 0.01	0.94 ± 0.01	1.2 ± 0.22	0.62 ± 0.0	0.88 ± 0.03

Census Tracts form Voronoi-like cells