Conditional Density Estimation with Histogram Trees

***Welcome to our poster** at poster session 1: Wed 11 Dec 11 a.m. PST — 2 p.m. PST*

NeurIPS 2024

Lincen Yang (presenting the slides) **& Matthijs van Leeuwen**
Leiden University, The Netherlands

Why conditional density estimation (CDE)?

- Get the full conditional distribution $P(Y|X)$, which provides more information than regression $E(Y|X)$.

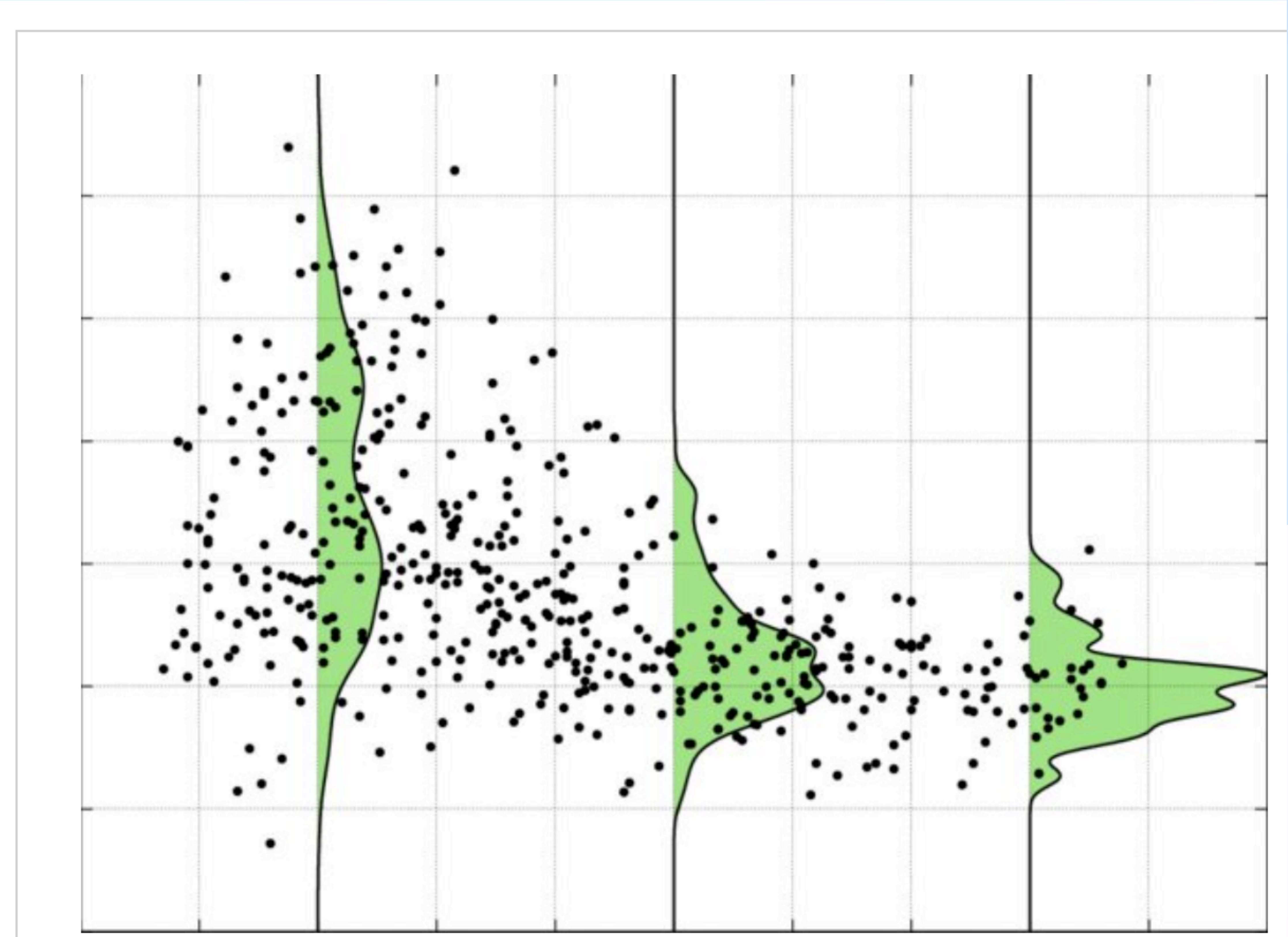


Figure from: Takeuchi, Ichiro, Kaname Nomura, and Takafumi Kanamori. "Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression." *Neural Computation* 21.2 (2009): 533-559.

Why conditional density estimation (CDE)?

- Get the full conditional distribution $P(Y|X)$, which provides more information than regression $E(Y|X)$.
- Useful for uncertainty quantification and knowledge discovery.

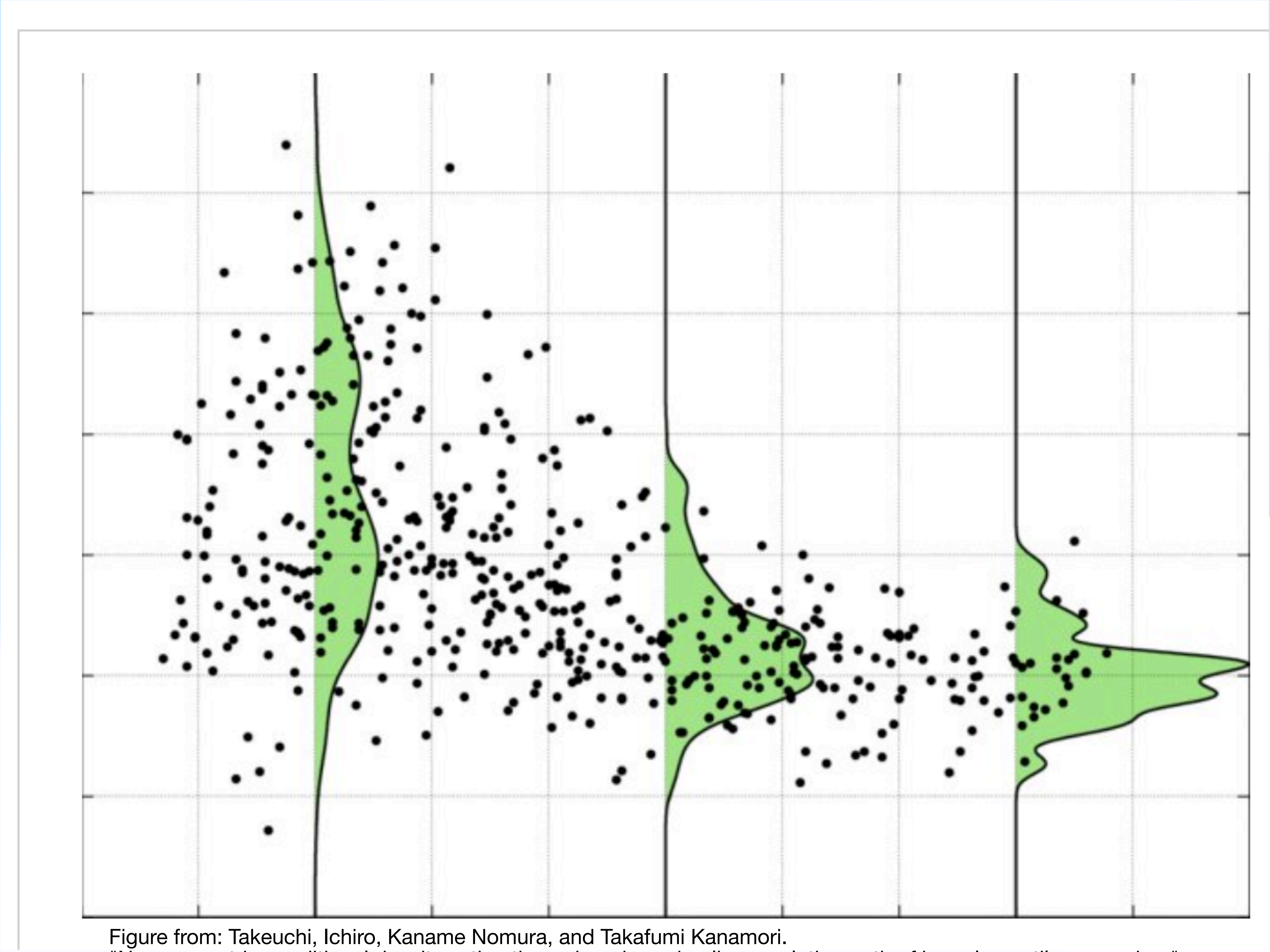


Figure from: Takeuchi, Ichiro, Kaname Nomura, and Takafumi Kanamori. "Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression." *Neural Computation* 21.2 (2009): 533-559.

Research Gap

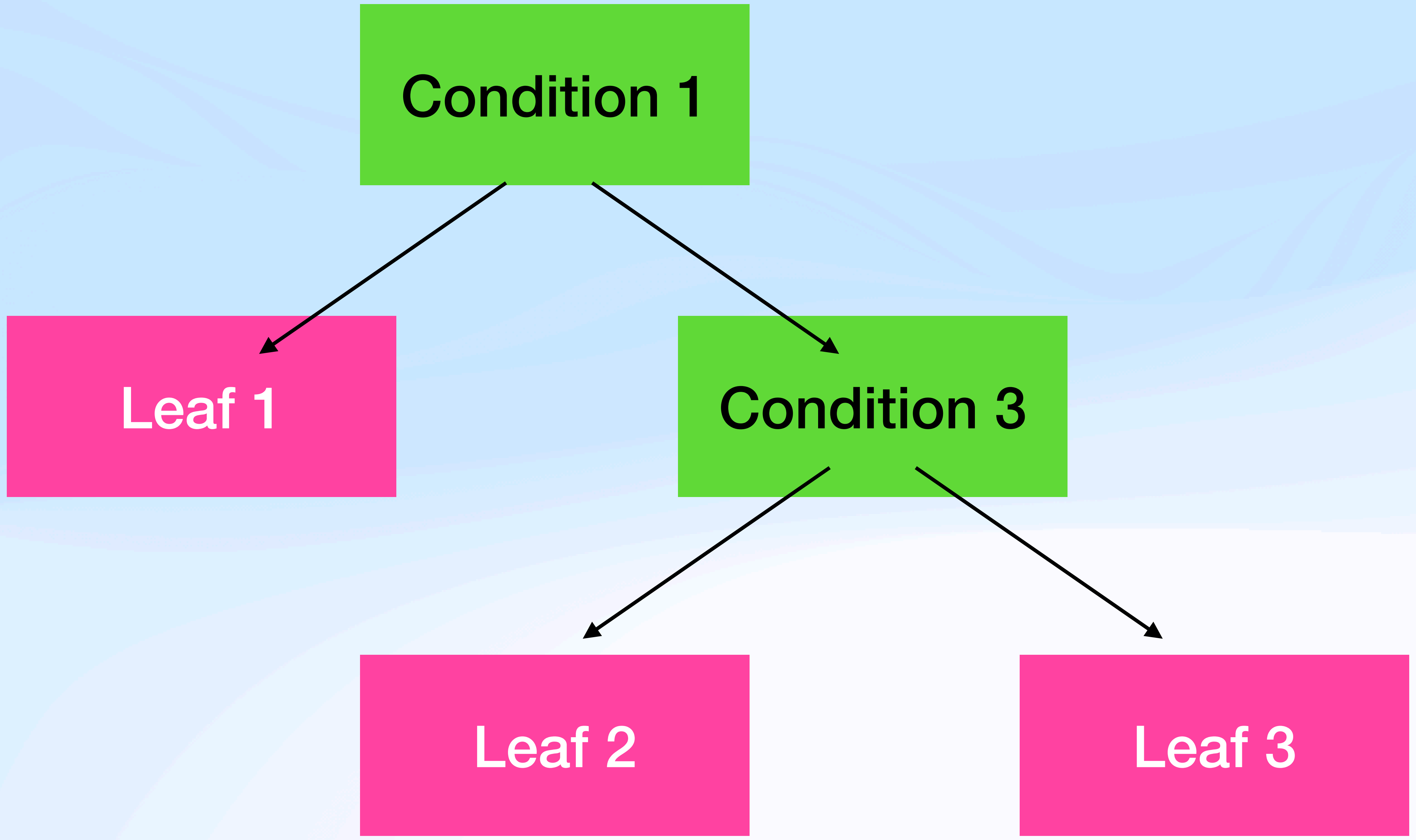
- Existing Methods for CDE
  - Kernel-based methods (the standard “shallow” methods for now)
  - Black-box methods (Normalizing Flows, Boosted trees, etc)

Research Gap

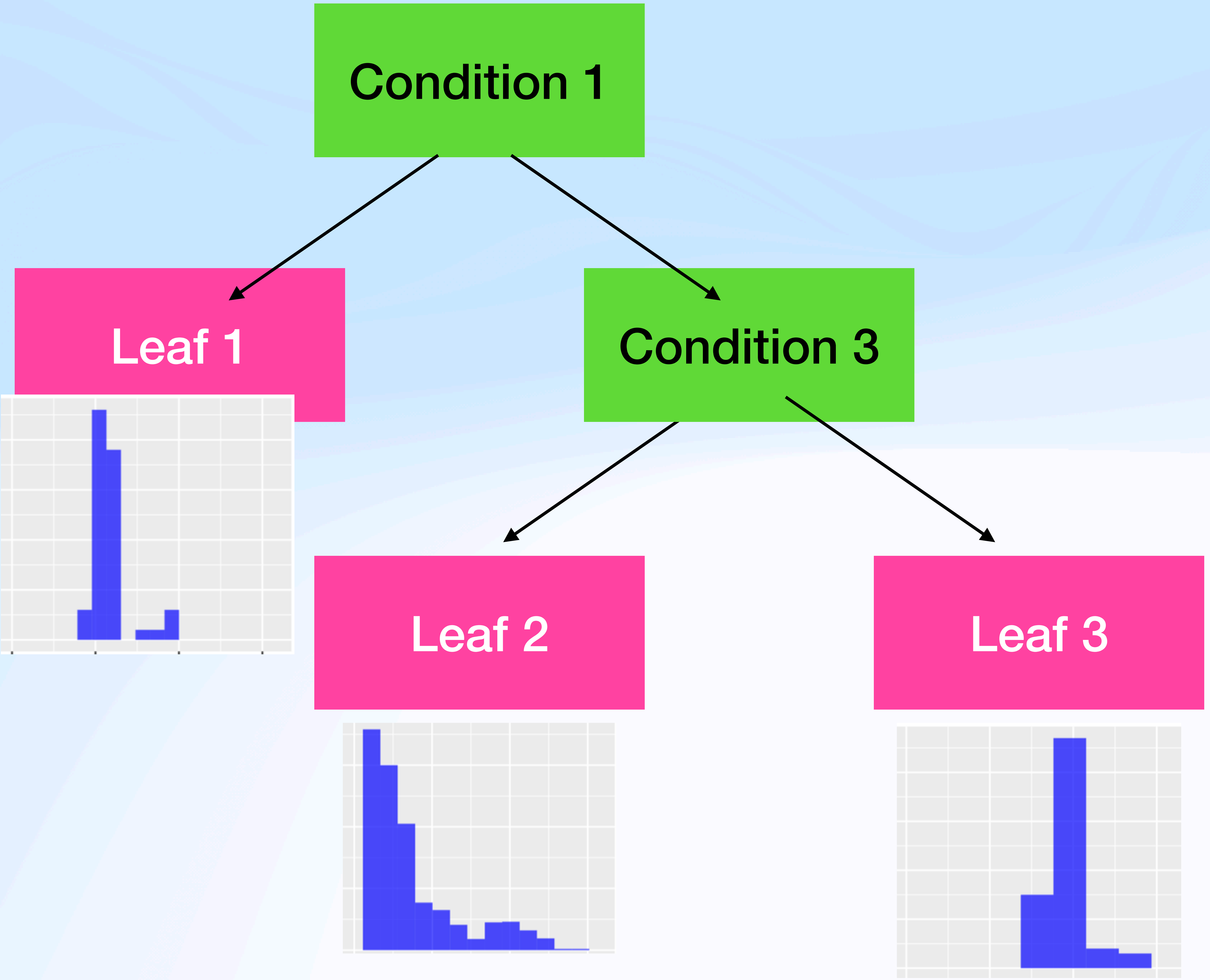
- Existing Methods for CDE
  - Kernel-based methods (the standard “shallow” methods for now)
  - Black-box methods (Normalizing Flows, Boosted trees, etc)

- **Intrinsically Interpretable models like decision trees have been understudied for conditional density estimation (CDE)!**
  - Arguably more interpretable than kernel-based methods

CDTree: Conditional Density Estimation Tree

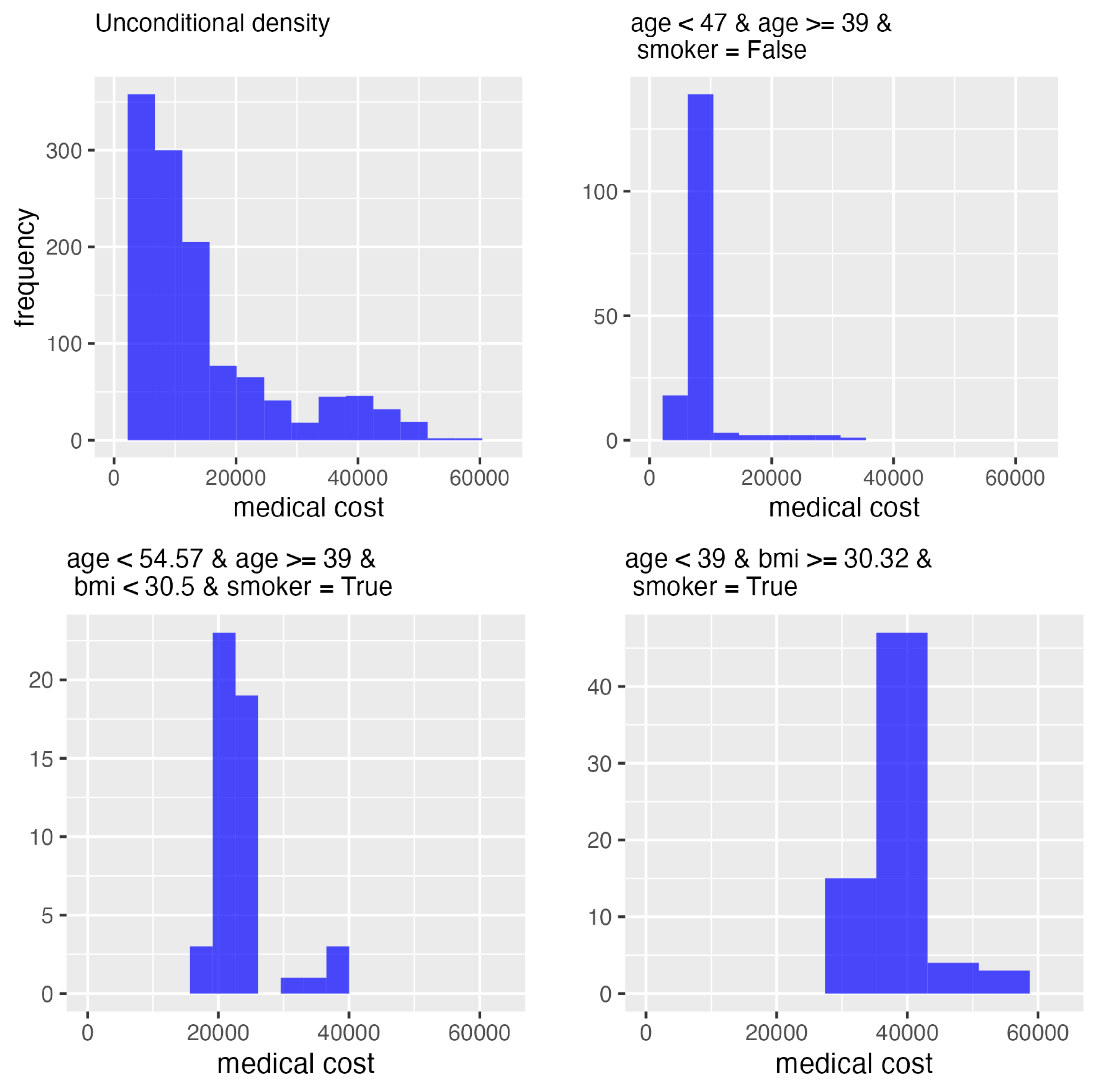


CDTree: Conditional Density Estimation Tree



CDTree: Conditional Density Estimation Tree

- Modeling the associations between medical costs and demographic & life style feature variables (e.g., smoker or not).



Learning CDTree, key features:

- Adopting the minimum description length (MDL) principle

$$M^* = \arg\min_{M \in \mathcal{M}} L(D \mid M) + L(M)$$

Learning CDTree, key features:

- Adopting the minimum description length (MDL) principle

$$M^* = \arg\min_{M \in \mathcal{M}} L(D \mid M) + L(M)$$

- $L(D \mid M)$: code length in bits needed to encode the data given model M

Learning CDTree, key features:

- Adopting the minimum description length (MDL) principle

$$M^* = \arg\min_{M \in \mathcal{M}} \boxed{L(D \mid M)} + \boxed{L(M)}$$

- $\boxed{L(D \mid M)}$: code length in bits needed to encode the data given model M
- $\boxed{L(M)}$: code length in bits needed to encode the model itself.

Learning CDTree, key features:

- Adopting the minimum description length (MDL) principle

$$M^* = \arg\min_{M \in \mathcal{M}} L(D \mid M) + L(M)$$

- $L(D \mid M)$: code length in bits needed to encode the data given model M
- $L(M)$: code length in bits needed to encode the model itself.
- In contrast, traditional optimization score often involves

$$M^* = \arg\min_{M \in \mathcal{M}} \boxed{\text{Loss function } \textit{(likelihood of data)}} + \alpha \boxed{\text{Tree Size}}$$

Learning CDTree, key features:

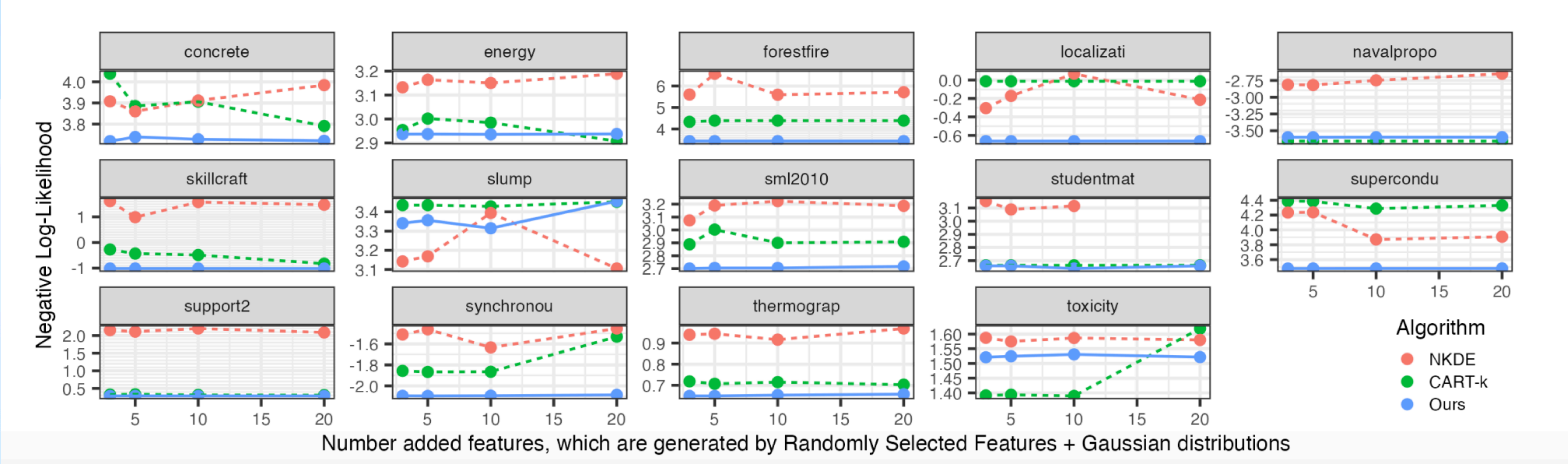
- No cross-validation for the hyper-parameter α to control overfitting
- Advantages:
  - Reduce runtime
  - Make the learned CDTree stable, *favoring interpretability*

Learning CDTree, key features:

- Iteratively grow the tree, **WITHOUT pruning**

Learning CDTree, key features:

- Iteratively grow the tree, **WITHOUT pruning**
- Advantages: Speed up the training & **Robust to “irrelevant" features**



Experiment Results

Predictive performance

Table 2: Negative log-likelihoods (smaller is better) on test sets. The best results among interpretable methods are shown in **bold**, and the best results among all interpretable and black-box models are marked by the underlines. The datasets are ordered by their numbers of columns (ascending).

	Interpretable models							Black-box models		
Datasets	CADET	CART-h	CART-k	CKDE	LSCDE	NKDE	*Ours*	LinCDE	MDN	NF
energy	3.55	3.09	3.06	**2.47**	3.38	3	2.93	2.93	2.78	2.86
synchrono	-2.93	-1.63	-1.86	**-3.59**	-1.25	-1.57	-2.11	-1.85	-2.94	-2.64
localizat	-0.23	-0.55	-0.01	-0.26	-0.61	-0.28	**-0.66**	-0.95	-0.68	-0.43
toxicity	1.8	1.5	1.38	**1.32**	1.34	1.55	1.53	1.29	1.24	1.23
concrete	4.17	3.75	3.93	**3.32**	3.66	3.91	3.72	3.47	2.97	3.18
slump	3.42	3.55	3.43	**2.35**	2.91	3.08	3.34	2.98	2.23	2.39
forestfir	134	3.96	4.39	4.85	4.68	5.55	**3.43**	4.35	3.26	3.23
navalprop	-3.53	-3.3	**-3.66**	-2.8	-2.88	-3.19	-3.6	-3.36	-4.12	-3.75
skillcraf	94.4	0.46	-0.42	1.54	1.61	1.56	**-1.02**	1.26	0.35	1.11
sml2010	6.52	2.85	2.89	**1.61**	3.14	3.12	2.7	2.97	2.15	2.61
thermogra	2.21	0.66	0.72	0.66	0.94	0.94	**0.64**	0.59	0.56	0.52
support2	97.3	0.51	0.32	2.09	2.46	2.13	**0.29**	1.48	1.53	1.24
studentma	3.83	**2.65**	2.66	2.89	4.19	3.11	2.66	2.59	3.85	3.54
supercond	9.6	3.84	4.36	4.55	4.17	4.19	**3.48**	3.87	3.33	3.5
rank (all)	8.79	5.68	6.04	5.11	7.46	7.68	4	4.46	**2.57**	3.21
rank (intp.)	6.07	3.43	3.86	3.14	4.57	4.86	**2.07**	—	—	—

Predictive performance

Table 2: Negative log-likelihoods (smaller is better) on test sets. The best results among interpretable methods are shown in **bold**, and the best results among all interpretable and black-box models are marked by the <u>underlines</u>. The datasets are ordered by their numbers of columns (ascending).

	Interpretable models							Black-box models		
Datasets	CADET	CART-h	CART-k	CKDE	LSCDE	NKDE	*Ours*	LinCDE	MDN	NF
energy	3.55	3.09	3.06	**<u>2.47</u>**	3.38	3	2.93	2.93	2.78	2.86
synchrono	-2.93	-1.63	-1.86	**<u>-3.59</u>**	-1.25	-1.57	-2.11	-1.85	-2.94	-2.64
localizat	-0.23	-0.55	-0.01	-0.26	-0.61	-0.28	**-0.66**	<u>-0.95</u>	-0.68	-0.43
toxicity	1.8	1.5	1.38	**1.32**	1.34	1.55	1.53	1.29	1.24	<u>1.23</u>
concrete	4.17	3.75	3.93	**3.32**	3.66	3.91	3.72	3.47	<u>2.97</u>	3.18
slump	3.42	3.55	3.43	**2.35**	2.91	3.08	3.34	2.98	<u>2.23</u>	2.39
forestfir	134	3.96	4.39	4.85	4.68	5.55	**3.43**	4.35	3.26	<u>3.23</u>
navalprop	-3.53	-3.3	**-3.66**	-2.8	-2.88	-3.19	-3.6	-3.36	<u>-4.12</u>	-3.75
skillcraf	94.4	0.46	-0.42	1.54	1.61	1.56	**<u>-1.02</u>**	1.26	0.35	1.11
sml2010	6.52	2.85	2.89	**<u>1.61</u>**	3.14	3.12	2.7	2.97	2.15	2.61
thermogra	2.21	0.66	0.72	0.66	0.94	0.94	**0.64**	0.59	0.56	<u>0.52</u>
support2	97.3	0.51	0.32	2.09	2.46	2.13	**<u>0.29</u>**	1.48	1.53	1.24
studentma	3.83	**2.65**	2.66	2.89	4.19	3.11	2.66	<u>2.59</u>	3.85	3.54
supercond	9.6	3.84	4.36	4.55	4.17	4.19	**3.48**	3.87	<u>3.33</u>	3.5
rank (all)	8.79	5.68	6.04	5.11	7.46	7.68	4	4.46	**2.57**	3.21
rank (intp.)	6.07	3.43	3.86	3.14	4.57	4.86	**2.07**	—	—	—

Predictive performance

Table 2: Negative log-likelihoods (smaller is better) on test sets. The best results among interpretable methods are shown in **bold**, and the best results among all interpretable and black-box models are marked by the underlines. The datasets are ordered by their numbers of columns (ascending).

	Interpretable models							Black-box models		
Datasets	CADET	CART-h	CART-k	CKDE	LSCDE	NKDE	*Ours*	LinCDE	MDN	NF
energy	3.55	3.09	3.06	**2.47**	3.38	3	2.93	2.93	2.78	2.86
synchrono	-2.93	-1.63	-1.86	**-3.59**	-1.25	-1.57	-2.11	-1.85	-2.94	-2.64
localizat	-0.23	-0.55	-0.01	-0.26	-0.61	-0.28	**-0.66**	-0.95	-0.68	-0.43
toxicity	1.8	1.5	1.38	**1.32**	1.34	1.55	1.53	1.29	1.24	1.23
concrete	4.17	3.75	3.93	**3.32**	3.66	3.91	3.72	3.47	2.97	3.18
slump	3.42	3.55	3.43	**2.35**	2.91	3.08	3.34	2.98	2.23	2.39
forestfir	134	3.96	4.39	4.85	4.68	5.55	**3.43**	4.35	3.26	3.23
navalprop	-3.53	-3.3	**-3.66**	-2.8	-2.88	-3.19	-3.6	-3.36	-4.12	-3.75
skillcraf	94.4	0.46	-0.42	1.54	1.61	1.56	**-1.02**	1.26	0.35	1.11
sml2010	6.52	2.85	2.89	**1.61**	3.14	3.12	2.7	2.97	2.15	2.61
thermogra	2.21	0.66	0.72	0.66	0.94	0.94	**0.64**	0.59	0.56	0.52
support2	97.3	0.51	0.32	2.09	2.46	2.13	**0.29**	1.48	1.53	1.24
studentma	3.83	**2.65**	2.66	2.89	4.19	3.11	2.66	2.59	3.85	3.54
supercond	9.6	3.84	4.36	4.55	4.17	4.19	**3.48**	3.87	3.33	3.5
rank (all)	8.79	5.68	6.04	5.11	7.46	7.[illegible]8	4	4.46	**2.57**	3.21
rank (intp.)	6.07	3.43	3.86	3.14	4.57	4.86	**2.07**	—	—	—

Model complexity: tree sizes

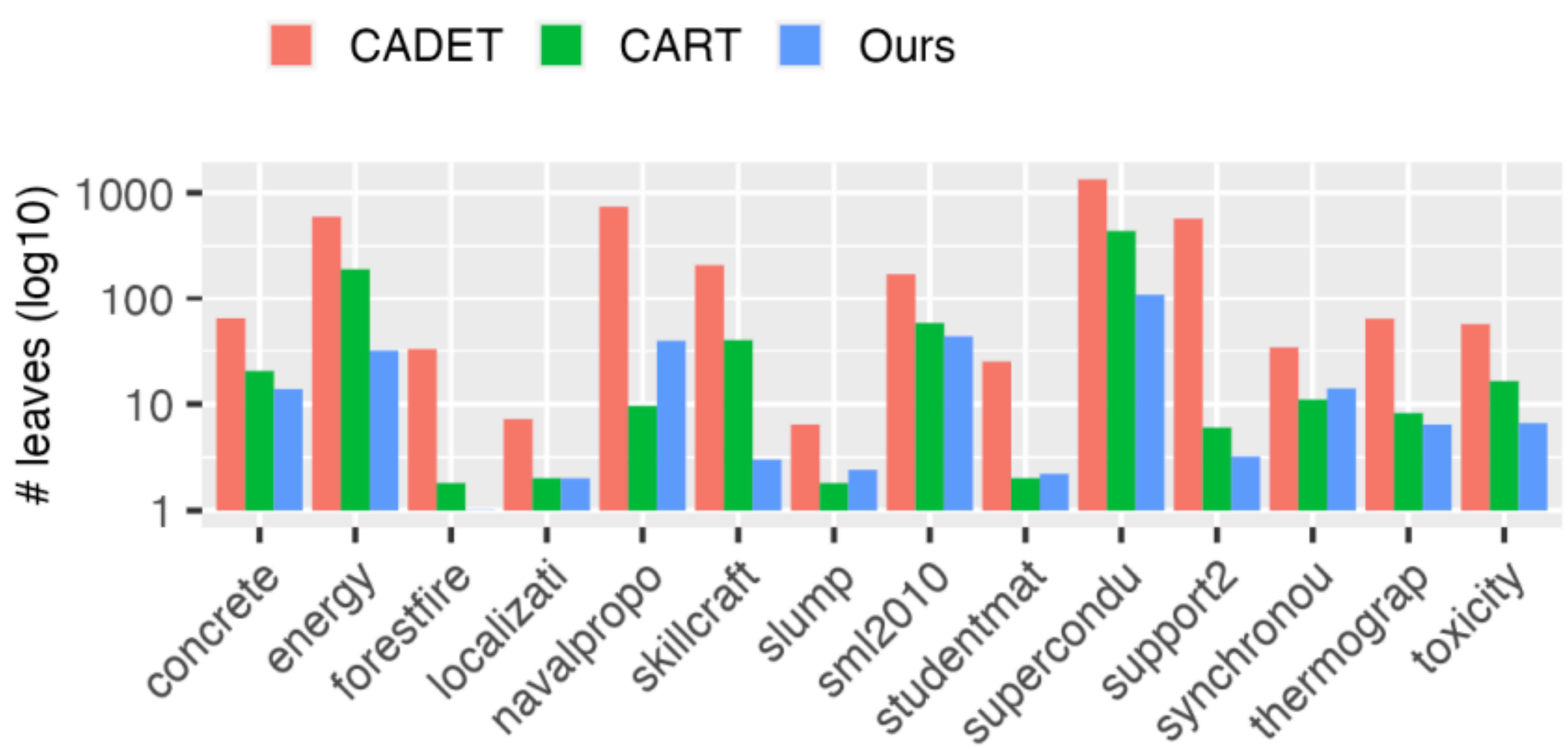


Figure: the number of leaves for tree-based methods

Conclusions

- We developed **CDTree**, the first dedicated decision tree method for non-parametric conditional density estimation.

Conclusions

- We developed **CDTree**, the first dedicated decision tree method for non-parametric conditional density estimation.
- We have demonstrated its competitive predictive performance and interpretability.

Conclusions

- We developed **CDTree**, the first dedicated decision tree method for non-parametric conditional density estimation.
- We have demonstrated its competitive predictive performance and interpretability.

- Github: https://github.com/ylincen/CDTree
- Paper: https://arxiv.org/pdf/2410.11449

Poster session 1: Wed 11 Dec 11 a.m. PST — 2 p.m. PST