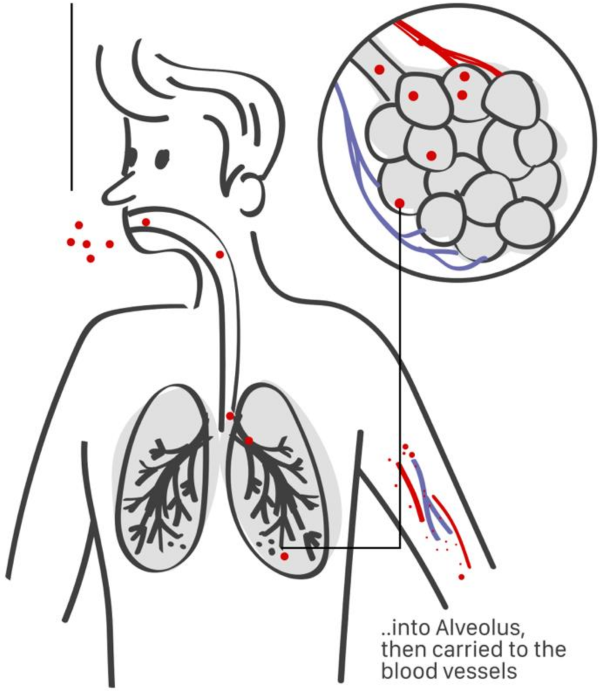
SPATIAL TRANSFER LEARNING FOR ESTIMATING PM2.5 IN DATA-POOR REGIONS

Shrey Gupta[1]*, Yongbee Park[5]*, Jianzhao Bi[2], Suyash Gupta[3], Andreas Züfle[1], Avani Wildani[1,5], Yang Liu[1]

[1]Emory University, US; [2]University of Washington, US; [3]University of California Berkeley, US; [4]Ingkle, Korea; [5]Cloudflare, US

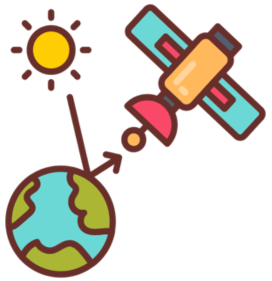
THE PM 2.5 PROBLEM

- Particulate Matter 2.5 ~ aerosols < 2.5 μm
- Poses significant public health concern. Small enough to:
  - Enter bloodstreams --> Heart diseases
  - Enter Lungs --> Pulmonary diseases
- Caused due to:
  - Vehicles
  - Wildfires
  - Industrial Processes



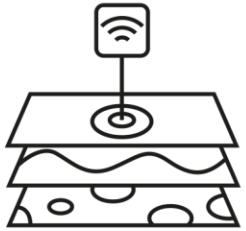
Simulation of PM2.5 entering and poisoning the body

NEED FOR TRANSFER LEARNING



Remote Sensing Data

Data collected is often inaccurate and compromised due to factors such as cloudy weather and high surface reflectance.

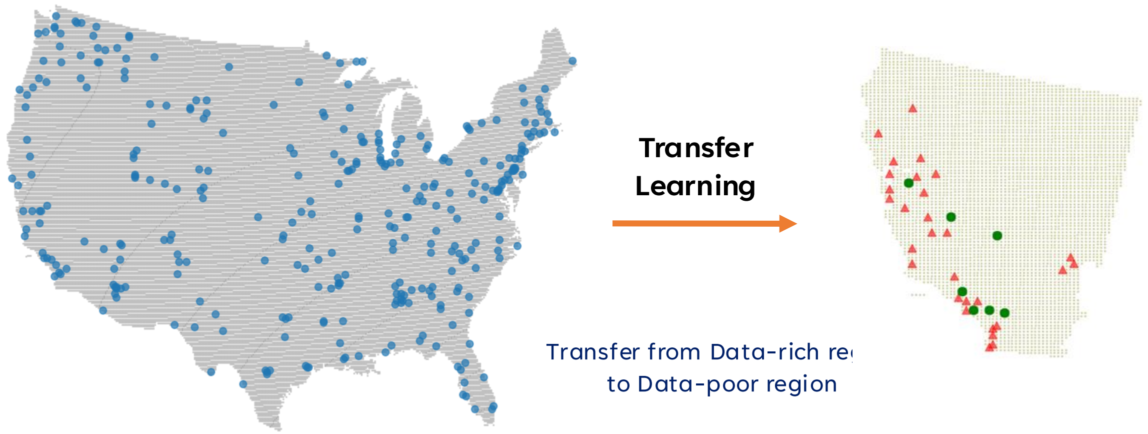


Installing Ground Sensors

Highly accurate data but installation, scaling and maintenance is costly for developing regions.

Transfer Learning to the Rescue!

Transfer knowledge from region with more data (data-rich) to region with less data (data-poor).



Transfer from Data-rich re
to Data-poor region

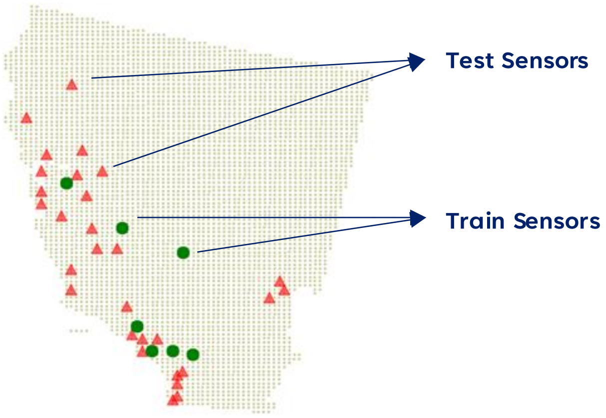
NEED FOR SPATIAL TRANSFER

Prior ($PM_{2.5}$) transfer studies focus on forecasting models.

- Models train on historical data for locations.
- Predict future values of same locations.

Limitations:

- [L1] Not suitable for missing temporal points.
- [L2.1] Not suitable for prediction on unknown locations.
- [L2.2] Not suitable for sparse train and test locations with low spatial autocorrelation.



California-Nevada w/ $PM_{2.5}$ sensors

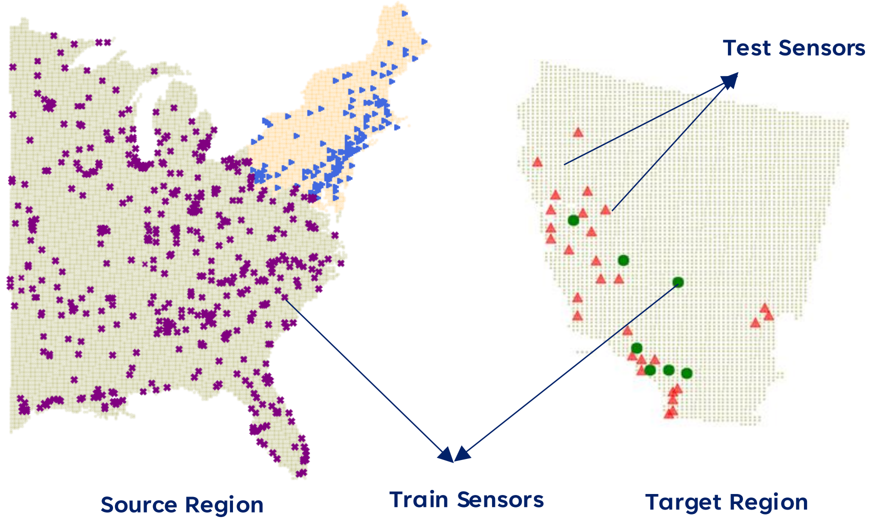
Solution:

- Instance Transfer Learning [L1]
- Capture spatial characteristics of the data [L2]

PROPOSED SOLUTION

Instance transfer learning (ITL)

- ITL models are unaffected by missing temporal data.
- These models combine source & target domains.



Combine source + target region data in ITL

Addition of a new feature that accounts for:

- **Spatial dependencies** – nearby locations have similar $PM_{2.5}$ levels
- **Semantic dependencies** – locations with similar meteorological and topographical conditions have similar $PM_{2.5}$ levels

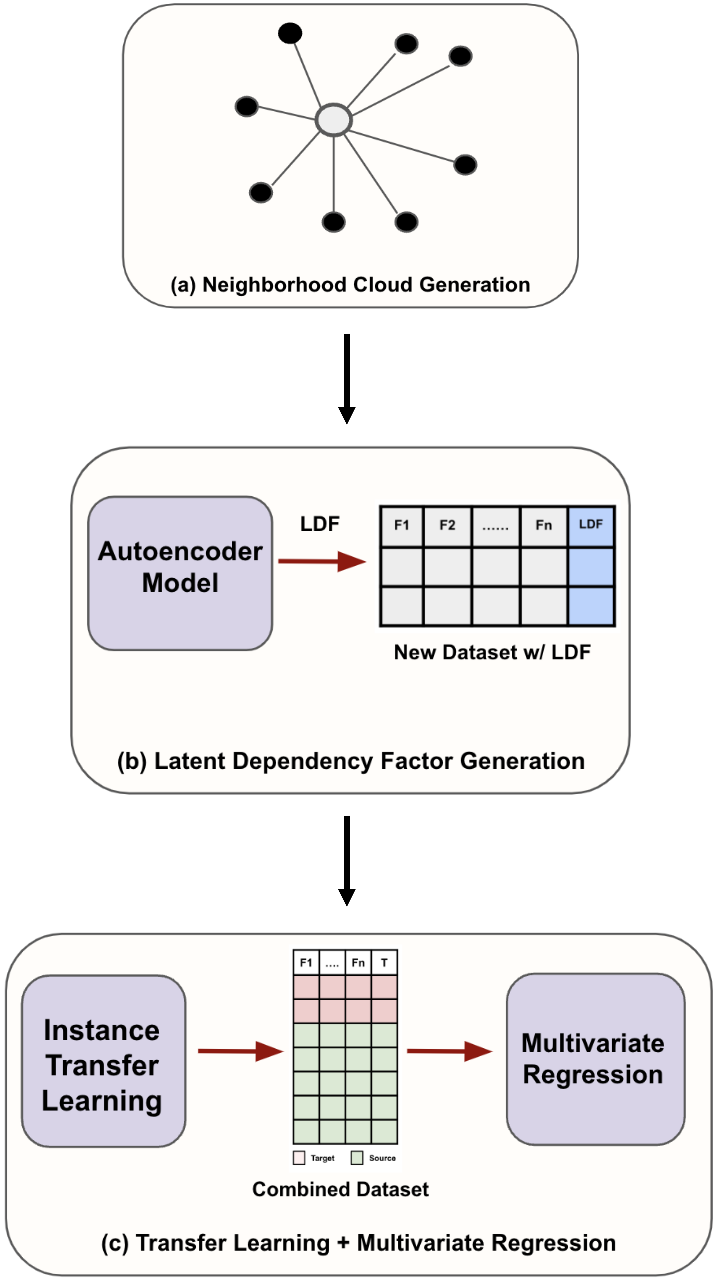
Meteorological Features		Topographical Features		Temporal Features	
F1	F2	F3	F4	F5	F6

+

LDF

CONTRIBUTIONS

- **Latent Dependency Factor (LDF):** We present a new feature (LDF) to represent spatial and semantic dependencies.
- **Two-stage Autoencoder Model:** We introduce a novel two-stage autoencoder model to generate LDF.
- **Spatial Transfer Learning:** We explore and design solution to the problem of spatial transfer learning.
- **Real-world Deployment:** We deploy our model on real-world data.



FRAMEWORK

STAGE I

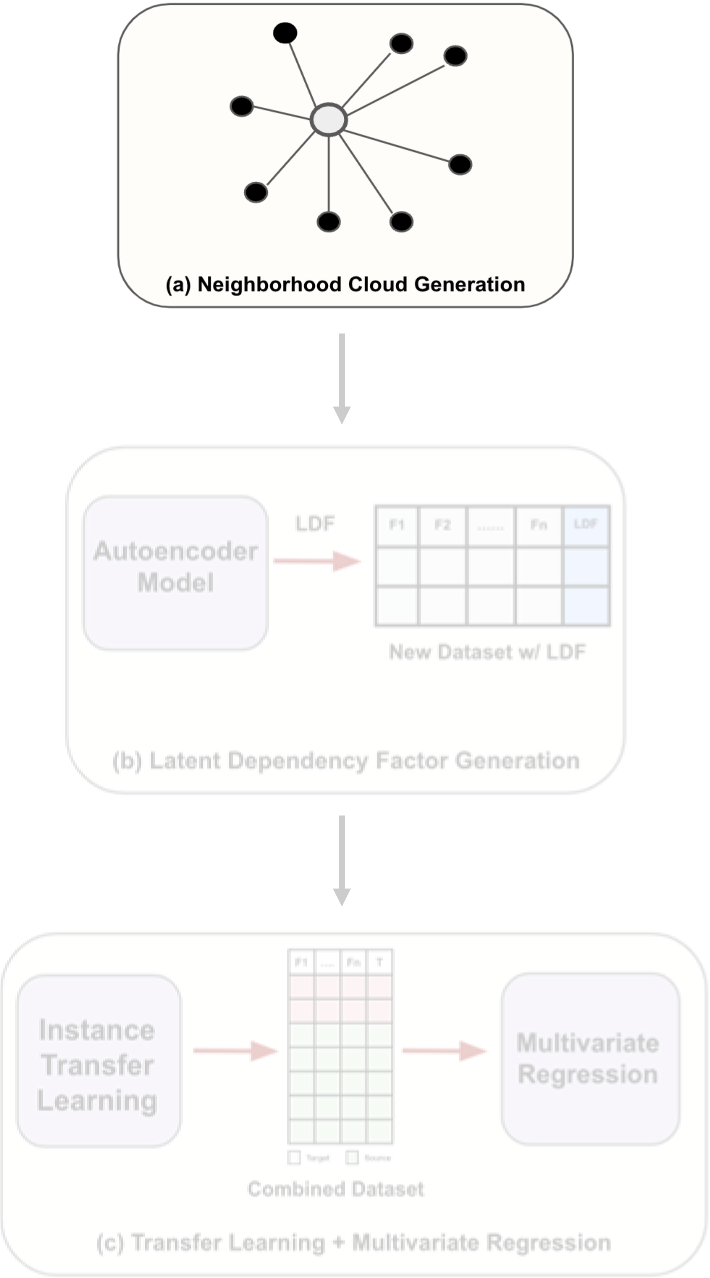
Neighborhood Cloud Generation

STAGE II

Latent Dependency Factor (LDF) Generation

STAGE III

Transfer Learning + Multivariate Regression

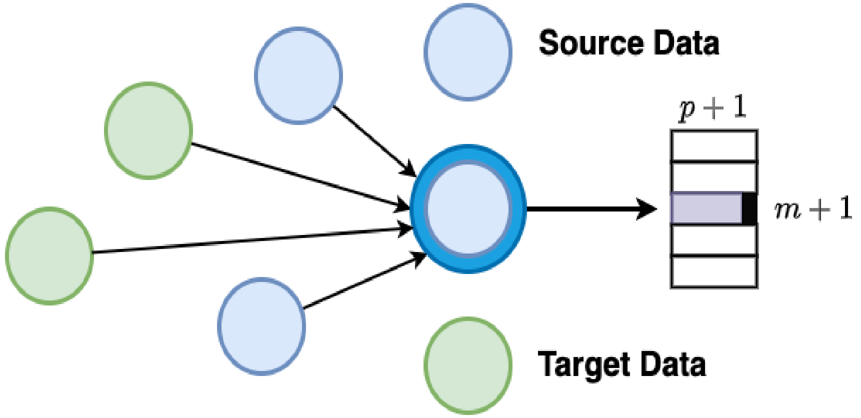


NEIGHBORHOOD CLOUD GENERATION

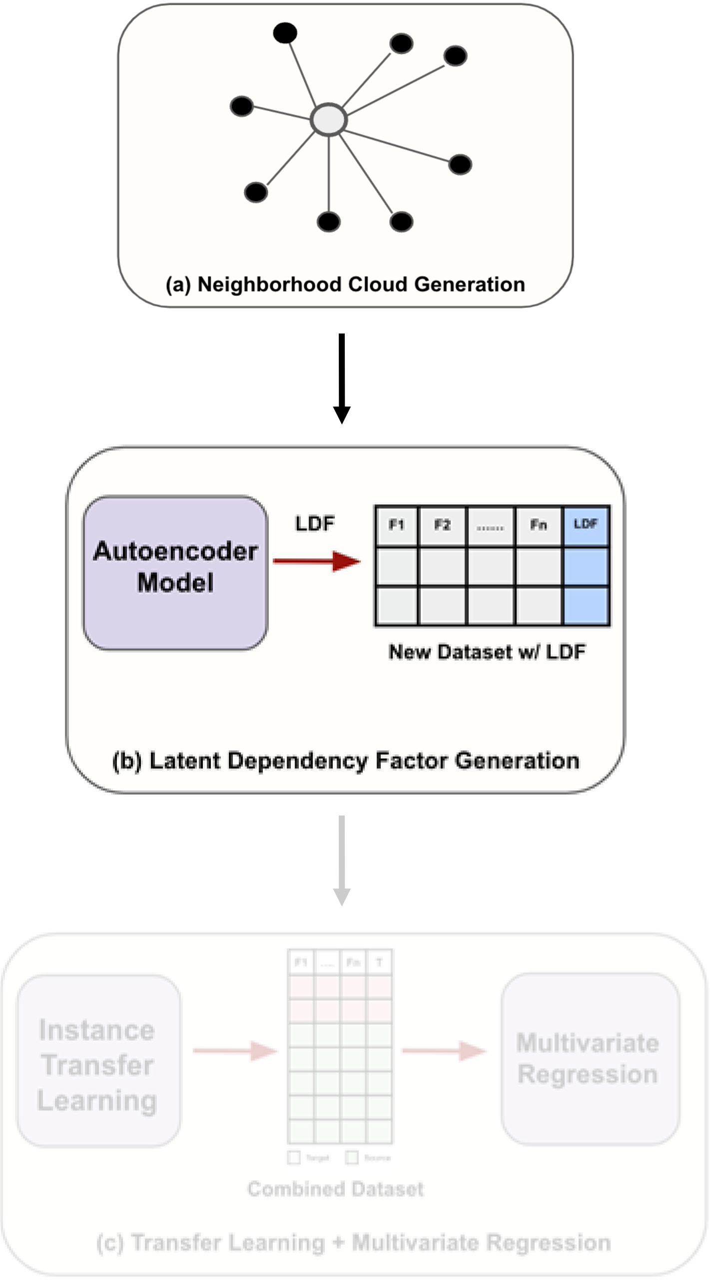
- Compute similarity between sensors (both target & source) and the objective location to find neighborhood cluster.

  Euclidean Distance (Similarity), $d(a,b) = \sqrt{(\Sigma (a_i - b_i)^2)}$

- Combine nearest m stations dataset (with p features) to generate cluster for each location.
- The data for each station is stacked to form a larger dataset – neighborhood cloud dataset.



Neighborhood cloud data generation

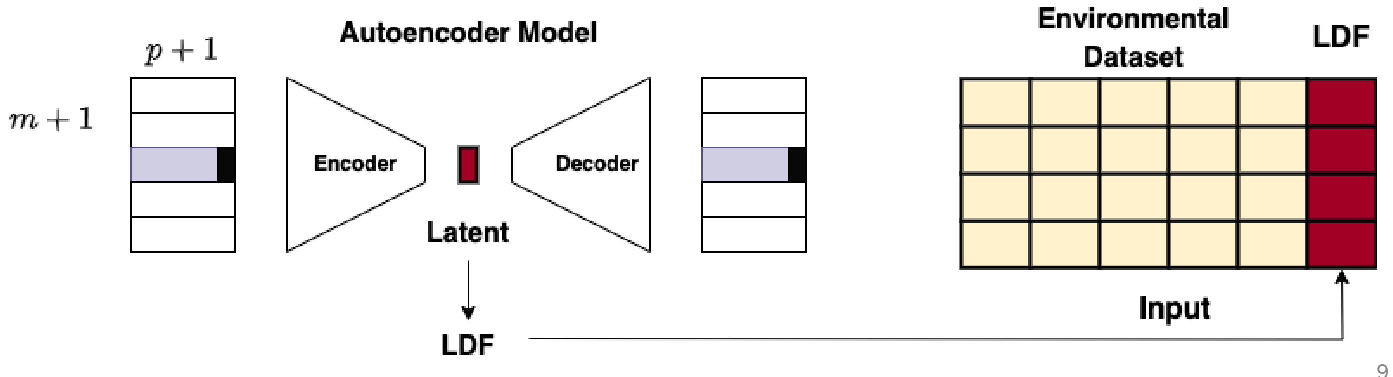


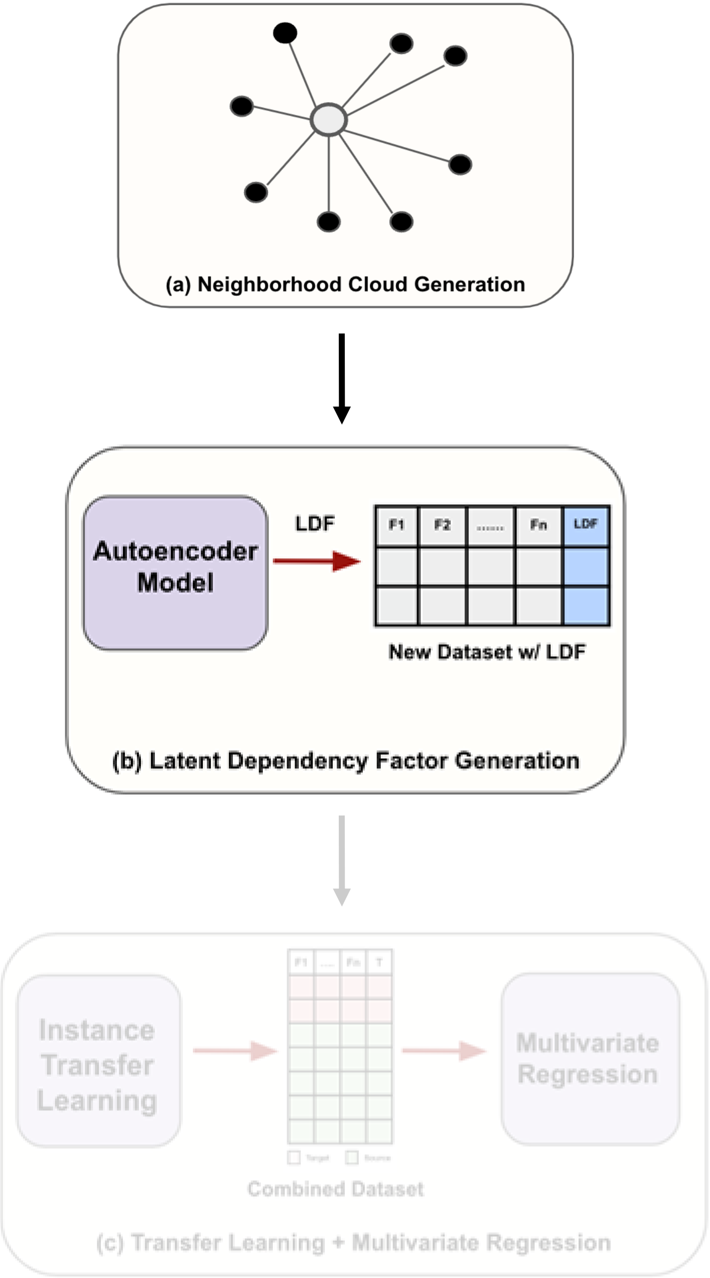
LATENT DEPENDENCY FACTOR (LDF) GENERATION

Stage I Autoencoder [Encoder-Decoder]:

- Generates the latent value using neighborhood cloud dataset.
- The encoder and decoder each have 3 1D CNN layers each.

  (The encoder-decoder model inbuilt with CNN allows to capture the spatial + semantic information across regions)
- The information from the 3 CNN layers is summed up using an FNN layer which outputs the LDF value.



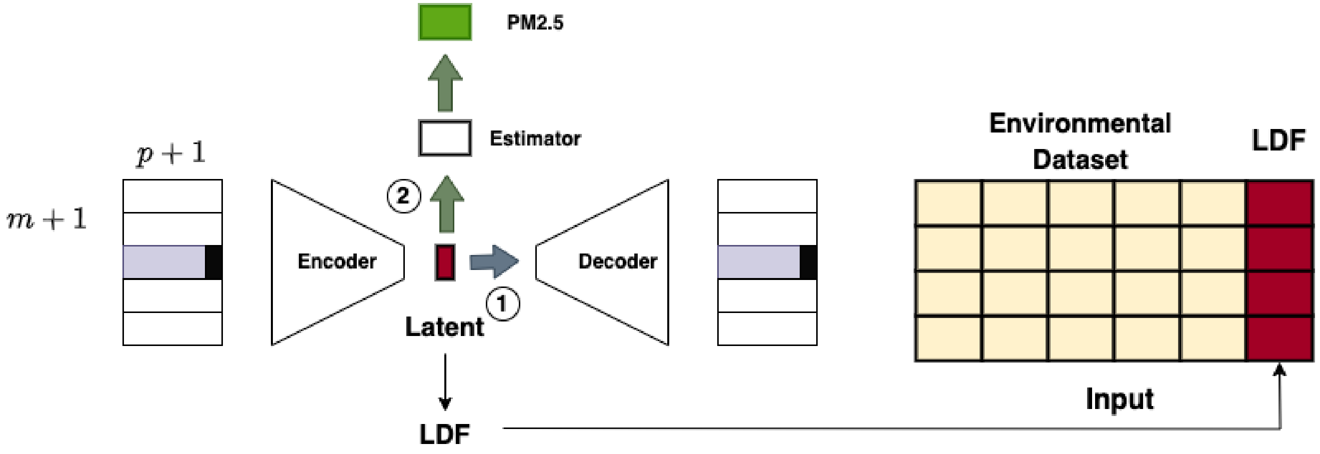


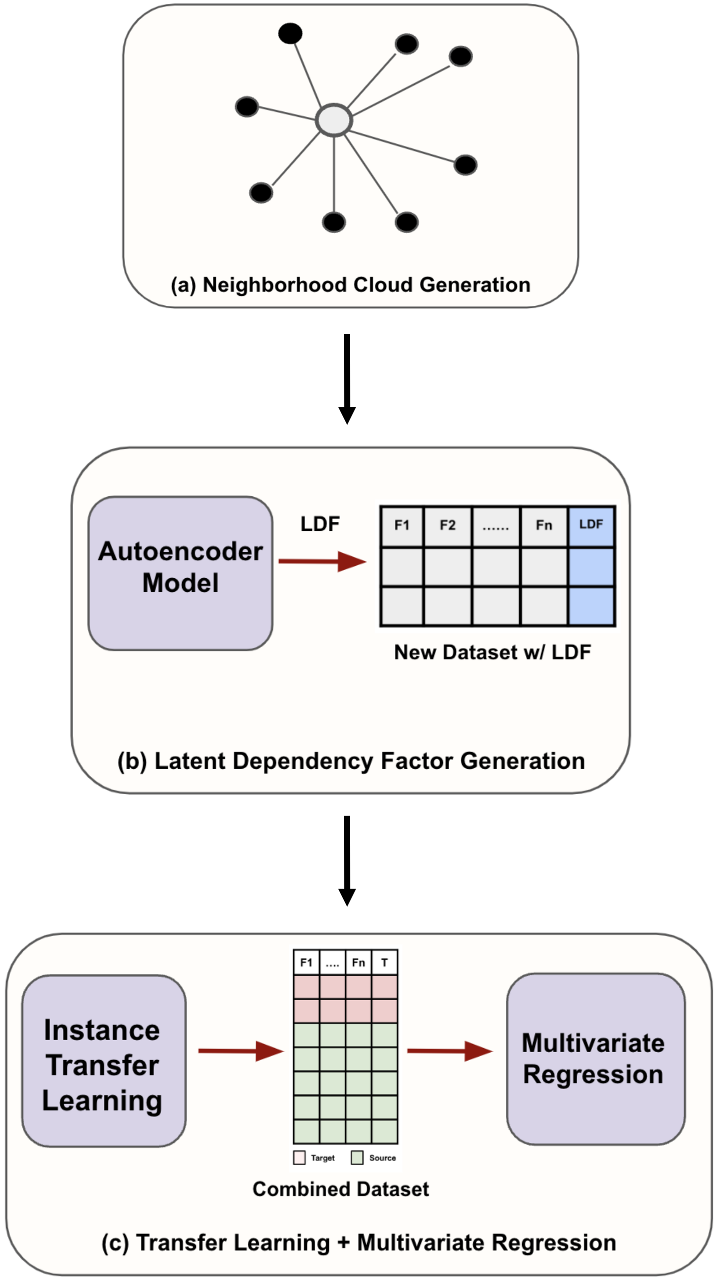
LATENT DEPENDENCY FACTOR (LDF) GENERATION

Stage II Autoencoder [Encoder-Estimator]:

- Increase attention on $PM_{2.5}$ value of objective location in the encoder-estimator stage to train an optimal LDF value.
- The estimator has single FNN layer.
- The autoencoder stages alternate between the two stages.

LDF-A: Consists of PM2.5 + Aerosol Optical Depth (AOD) in the encoder-estimator stage

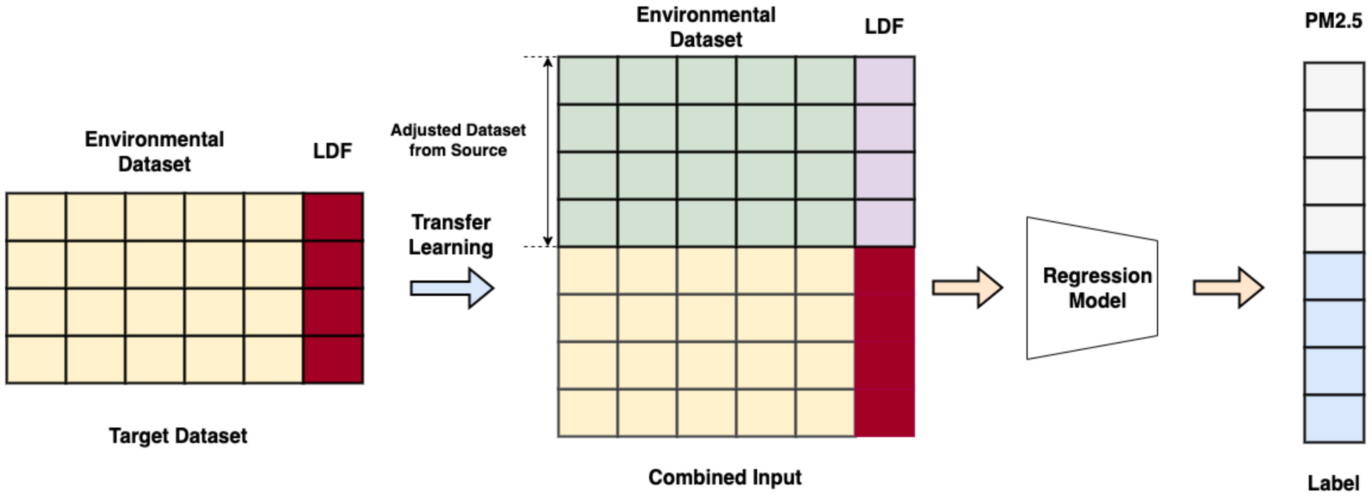




TRANSFER + REGRESSION

- Apply instance transfer learning on the LDF-combined dataset to generate source sample weights.

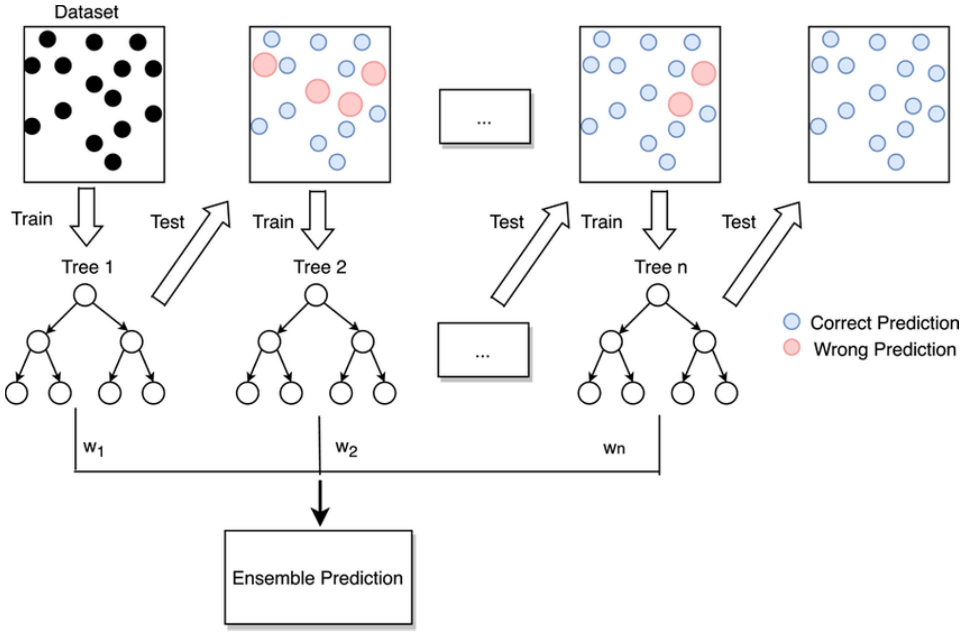
- Apply regression on the weighted source + target samples to predict $PM_{2.5}$ values.



ML MODELS

GRADIENT BOOSTING REGRESSION

- Ensemble model of Decision Tree to minimize pseudo-residuals (boosting algorithm).
- Applied on target region data.



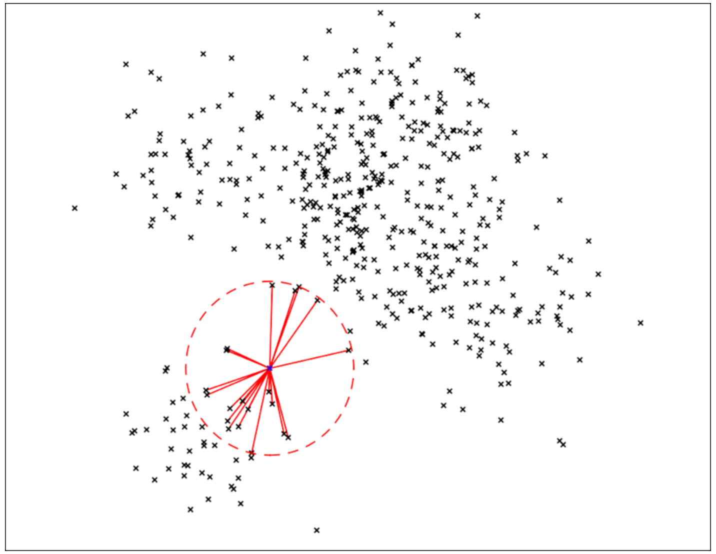
Gradient Boosting Regression

Image Courtesy: Zhang, Tao, et al. "Improving convection trigger functions in deep convective parameterization schemes using machine learning." *Journal of Advances in Modeling Earth Systems* (2021).

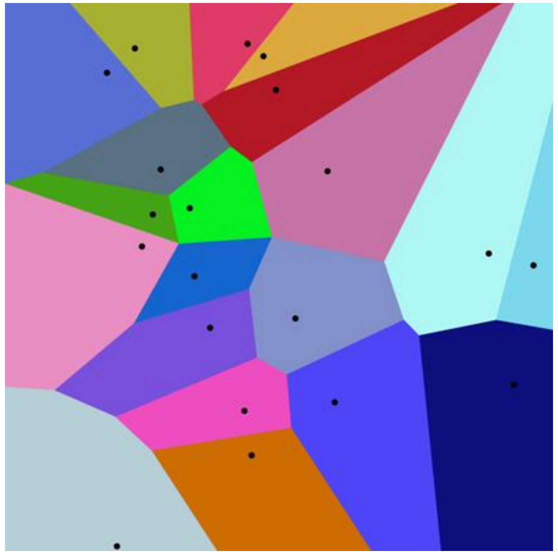
TRANSFER MODELS

NEAREST NEIGHBOR WEIGHING (NNW)

- Reweighs source samples by creating a Voronoi tessellation to calculate # target samples in it.
- Applied on source + target region data.



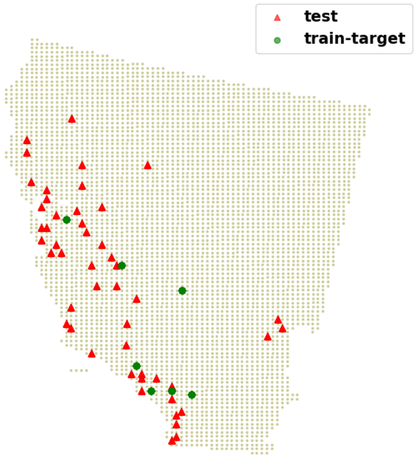
Nearest Neighbor Weighing (NNW)



Voronoi Tessellation

Image Courtesy [NNW]: erikbern.com/2015/09/24/nearest-neighbor-methods-vector-models-part-1.html
Image Courtesy [Voronoi]: https://en.wikipedia.org/wiki/Voronoi_diagram

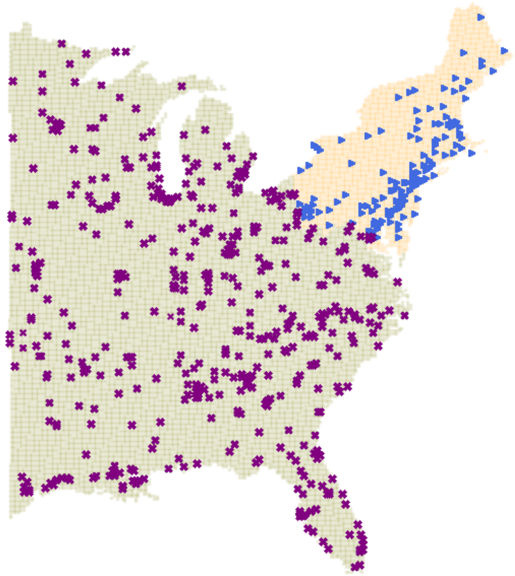
TARGET DATASETS (REGIONS)



California-Nevada

- **# $PM_{2.5}$ sensors:** 128
- **Dataset shape:** (249k, 27)
- **Features:** Meteorological, Topographical, and Geographical from year 2011.
- **Satellite samples (unlabeled) shape:** 19.5 M

SOURCE DATASETS (REGIONS)



Eastern and North-Eastern US

- **Eastern US** has **607** $PM_{2.5}$ sensors.
- **North-eastern US** has **147** $PM_{2.5}$ sensors.
- Dataset shape
  - **Eastern US:** (143k, 27)
  - **North-eastern US:** (37k, 27)
- **Features:** Meteorological, Topographical, and Geographical (Total Features = 77) from year 2011.
- **Common features with Cal-Nevada:** 27

EXPERIMENTAL SETUP

CALIFORNIA-NEVADA

Sampling:

- Sensors are grouped into sets of 5, 7, 9, 11 for CVs.
- Reported R^2 and RMSE values are averaged across 20 CVs.

Daily-data Matching:

Daily active sensors are matched across target & source to generate clusters.

RESULTS

Source: Eastern US

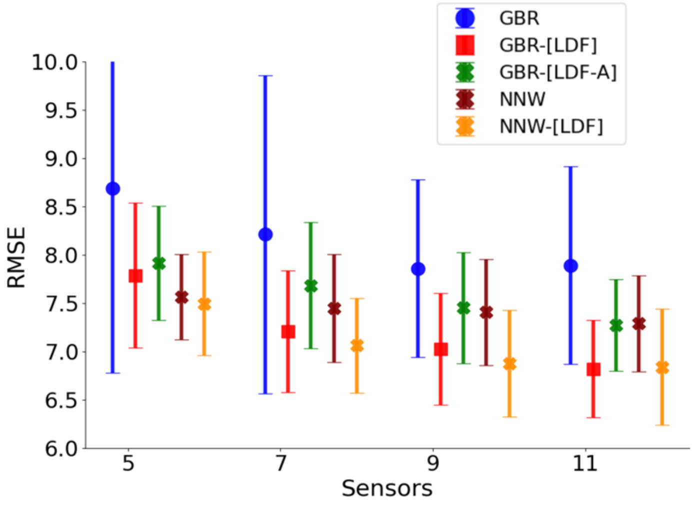
Models	5		7		9		11	
	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
GBR	-0.061	8.684	0.064	8.210	0.177	7.857	0.157	7.891
NNW	**0.236**	**7.563**	0.263	7.447	0.280	7.406	0.296	7.288
NNW [LDF]	**0.247**	**7.494**	**0.336**	**7.061**	**0.378**	**6.874**	**0.378**	**6.838**
NNW [LDF-A]	0.225	7.596	**0.298**	**7.230**	**0.359**	**6.973**	**0.359**	**6.924**

Source: North-Eastern US

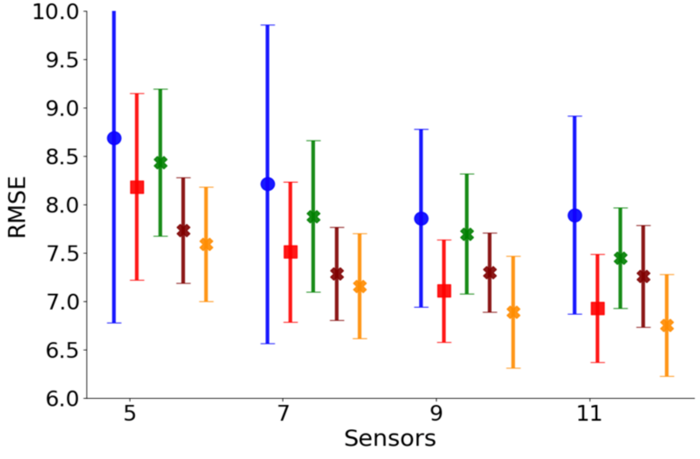
Models	5		7		9		11	
	R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
GBR	-0.061	8.684	0.064	8.210	0.177	7.857	0.157	7.891
NNW	0.199	7.732	0.294	7.286	0.301	7.297	0.298	7.257
NNW [LDF]	**0.225**	**7.592**	**0.317**	**7.157**	**0.376**	**6.886**	**0.392**	**6.751**
NNW [LDF-A]	**0.201**	**7.702**	**0.320**	**7.122**	**0.378**	**6.873**	**0.374**	**6.847**

ABLATION STUDY

- Ablation study compares GBR & transfer models using LDF-imputed data to validate performance.
- We observe that addition of the LDF feature improves the performance of GBR.
- GBR [LDF] performing as the second-best model.
- NNW [LDF] still outperforms GBR [LDF] indicating LDF is useful for transfer models.

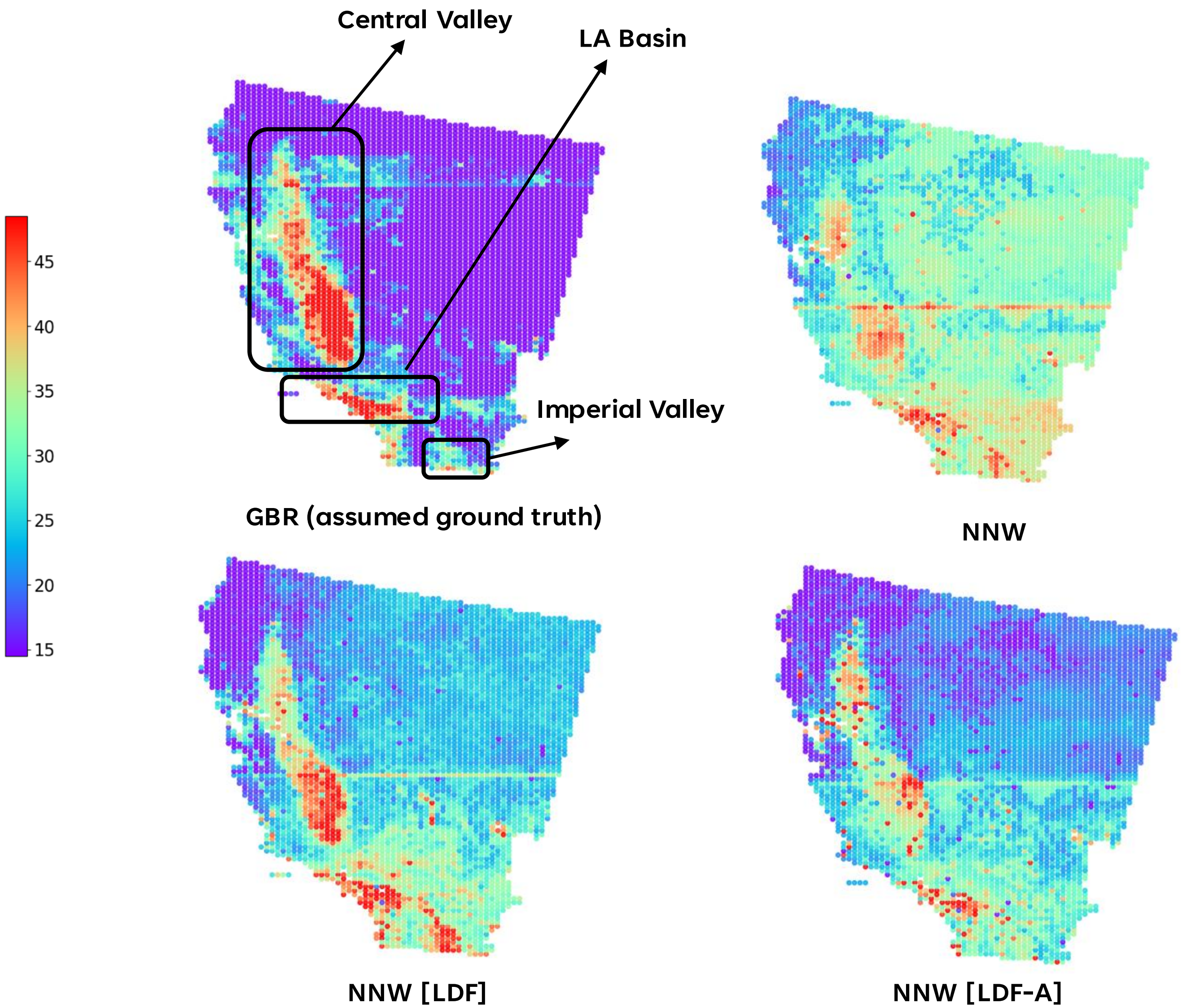


Source: Eastern US



Source: North-Eastern US

QUALITATIVE RESULTS [CAL-NEVADA]



- **NNW [LDF] model** provides most accurate $PM_{2.5}$ estimates in Central Valley and Los Angeles Basin but overestimates in the Imperial Valley.
- **NNW [LDF-A]** performs second-best; its estimates in the Central Valley are patchy.
- The **NNW model** shows obscure and patchy patterns; it underestimates in Central Valley and significantly overestimates in Imperial Valley.

FUTURE DIRECTIONS

Dataset Expansion:

Incorporate datasets lacking spatial and semantic dependencies to broaden the scope of $PM_{2.5}$ estimation.

Temporal Trend Integration:

Enhance the LDF feature to capture temporal trends, for time-series data.

Application to Other Domains:

Extending the LDF technique to domains like wildfire estimation and weather forecasting with similar spatial patterns.

CONCLUSION

- **Spatial transfer learning** is solved for the use case of transfer between highly sparse and distant source & target regions.
- **Latent Dependency Factor (LDF)** as a new 'spatial' feature is introduced.
- **Two-stage autoencoder model** is designed to generate LDF.
- **Quantitative results** show LDF shows a 19.34% improvement.
- **Qualitative results** show LDF captures varying concentration gradient accurately.

THANK YOU

Shrey Gupta*
Emory University

Yongbee Park*
Ingkle

Jianzhao Bi
University of Washington

Suyash Gupta
University of California, Berkeley

Andreas Züfle
Emory University

Avani Wildani
Cloudflare, Emory University

Yang Liu
Emory University

Checkout the Git repo: