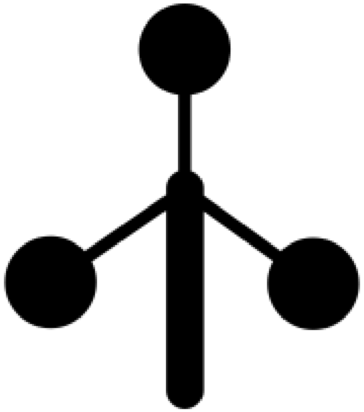
Learning with Previously Unseen Features

Yuan Shi
Computer Science Department

Craig A. Knoblock
Information Sciences Institute

University of Southern California

Motivating Example



weather station

temperature, pressure sensor → humidity

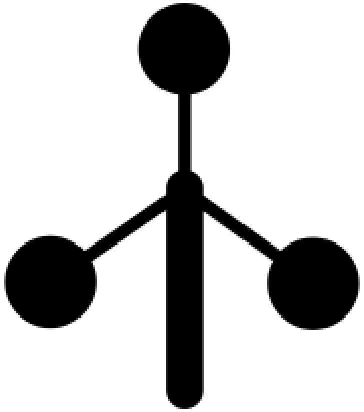
Training data:

Feature | Label

Temperature (°F)	Pressure (in)	Humidity (%)
73	29.9	65
68	29.3	72
71	29.4	73
68	29.1	77

Motivating Example

Training data:



Feature		Label
Temperature (°F)	**Pressure (in)**	**Humidity (%)**
73	29.9	65
68	29.3	72
71	29.4	73
68	29.1	77

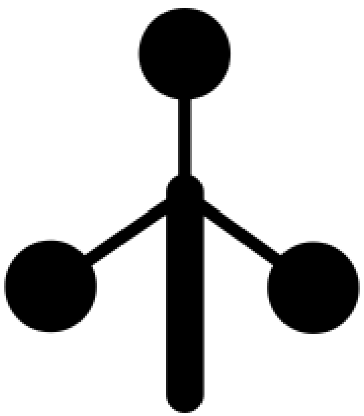
Test data:

Dew point (°F)	Temperature (°F)	Pressure (in)
60	69	29.9
62	68	29.3
61	72	29.4
65	68	29.1

Humidity?

Motivating Example

Training data:



Feature		Label
Temperature (°F)	**Pressure (in)**	**Humidity (%)**
73	29.9	65
68	29.3	72
71	29.4	73
68	29.1	77

Test data:

Dew point (°F)	Temperature (°F)	Pressure (in)
60	69	29.9
62	68	29.3
61	72	29.4
65	68	29.1

Dew point can help predict humidity!

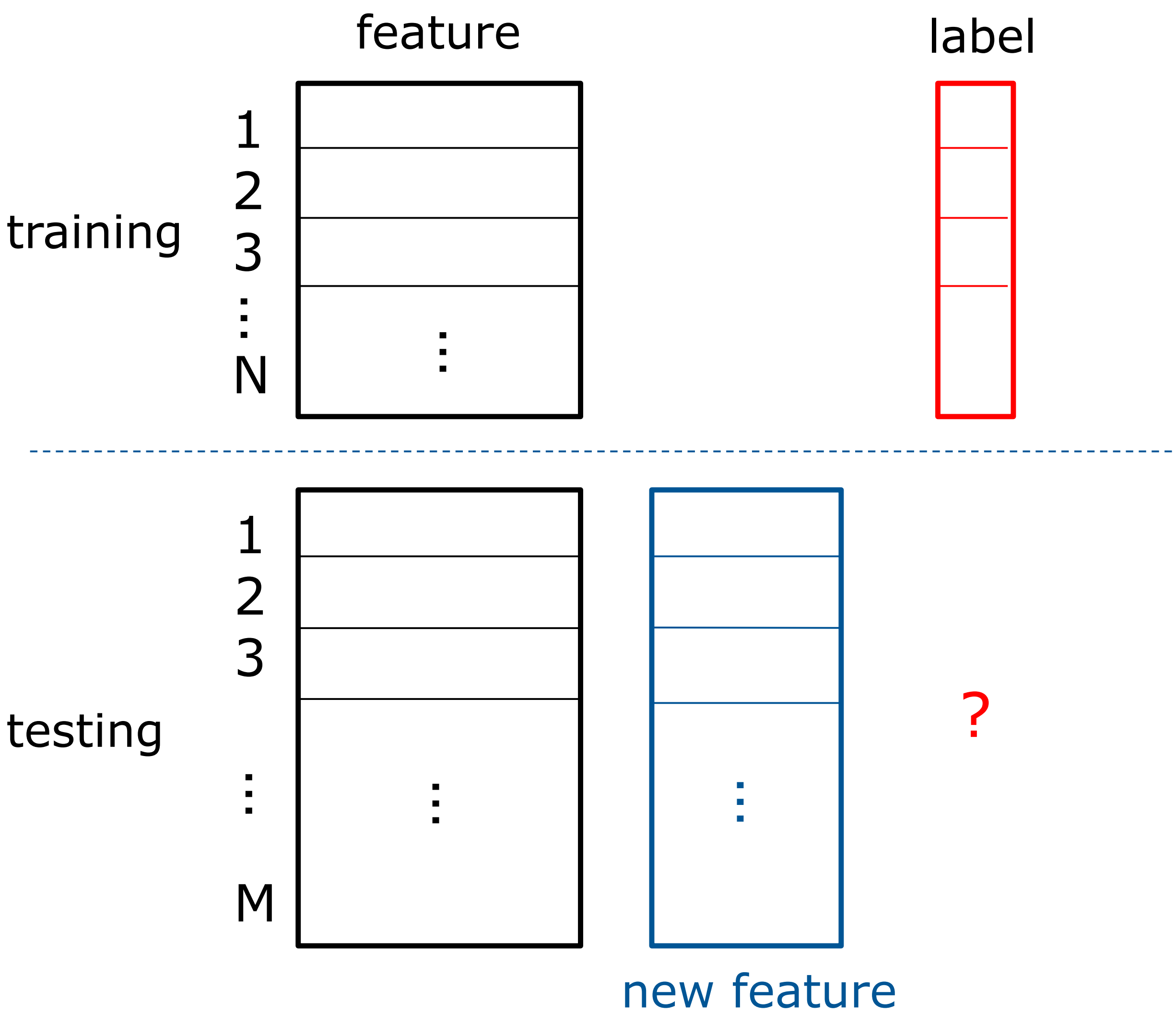
Humidity, dew point and temperature are inter-correlated

Problem

- **Context:** Machine learning systems deployed in an open environment
  - May access features that are previously unseen in the labeled training set

- **Goal:** Improving a machine learning model by automatically identifying and using features that are not in the training set
  - Without additional labels

Problem



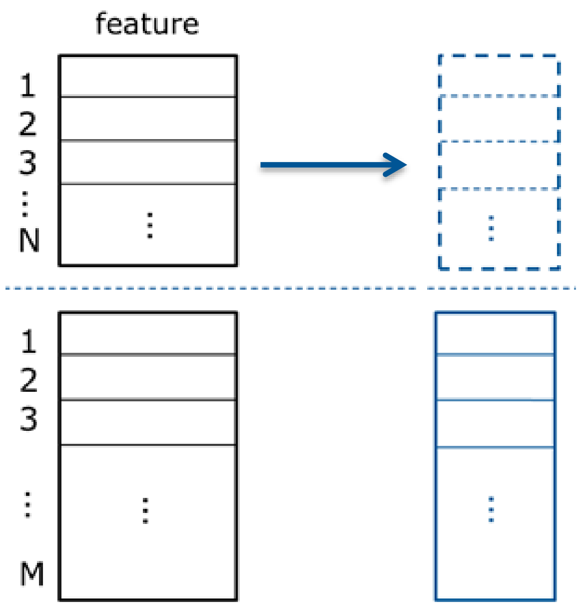
How to Leverage New Features?

- **Approach 1:** Ignore them
- But new features may contain complementary information over original features
    - Extreme case: new feature = label

How to Leverage New Features?

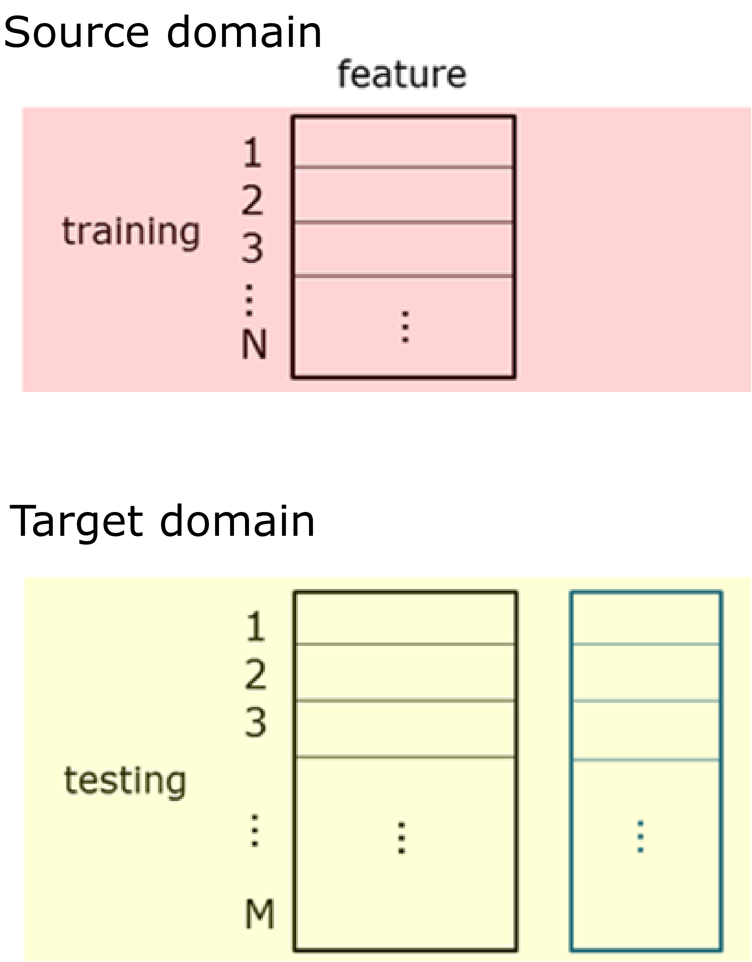
- **Approach 1:** Ignore them
- But new features may contain complementary information over original features
  - Extreme case: new feature = label
- **Approach 2:** Predict new features and combine them with original features

But prediction can be very challenging…



Related Work: Heterogeneous domain adaptation

Heterogeneous domain adaptation [Pan and Yang, 2010]



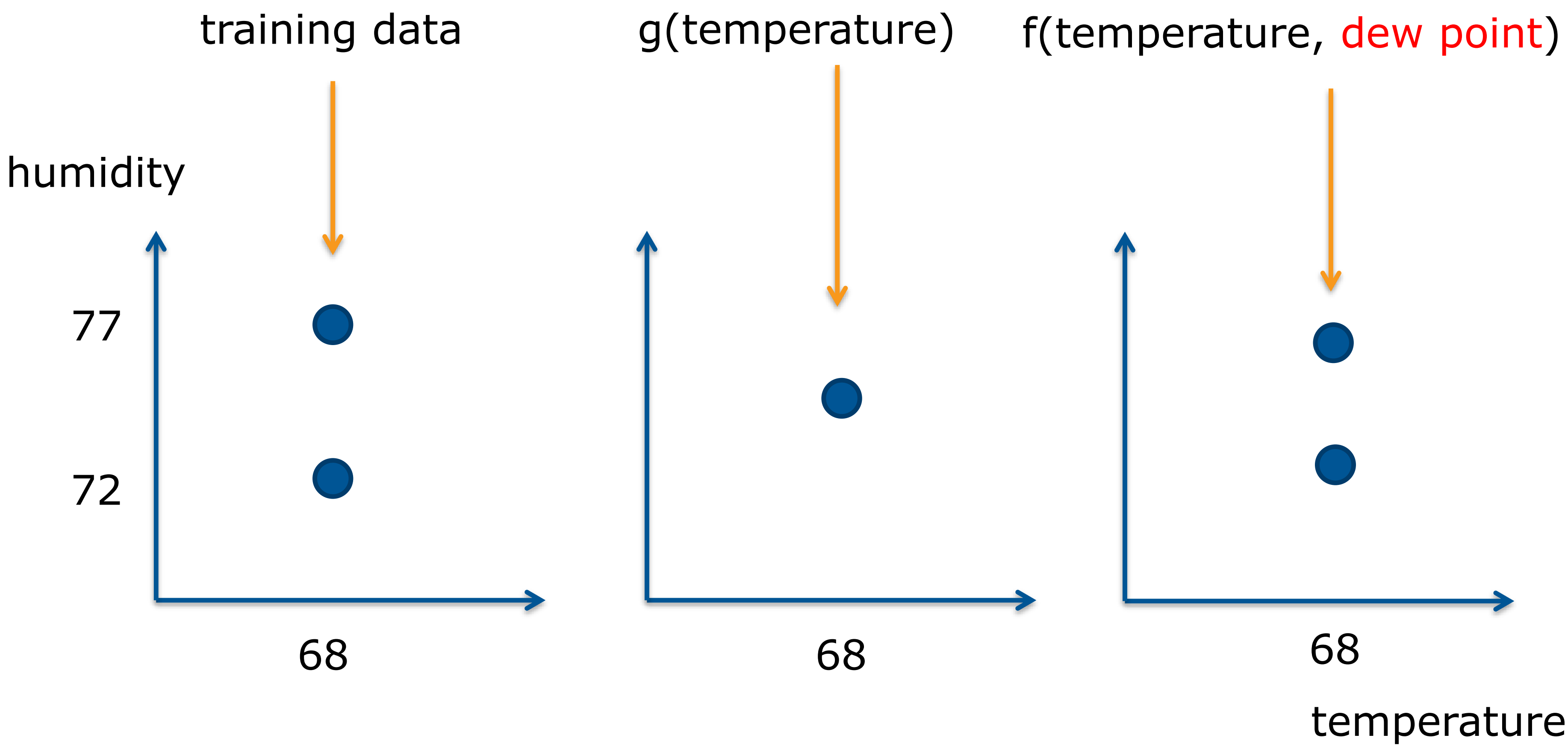
- Most existing approaches attempt to match the feature space of two domains: **ignoring new features** just maximizes the match! [Dai et al., 2008; Socher et al., 2013; Zhou et al., 2014; Kulis et al., 2011;Wang and Mahadevan, 2011; Argyriou et al., 2008; Duan et al., 2012; Shi et al., 2010; Harel and Mannor, 2010; Wei and Pal, 2011; Yeh et al., 2014]

- Some work require additional labels to leverage new features [Zhao and Hoi, 2010; Hou and Zhou, 2016]

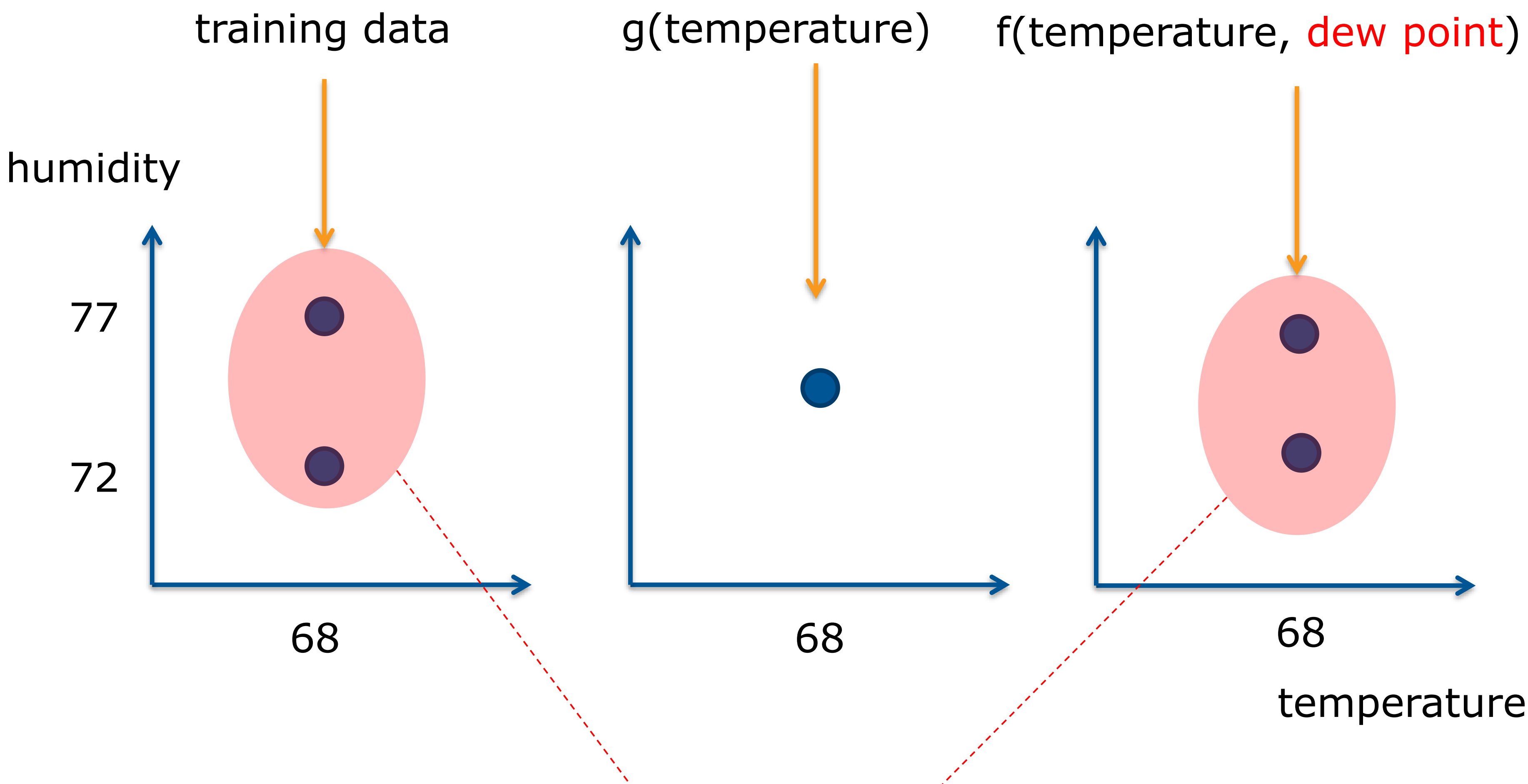
Our Approach:

Learning with previously Unseen Features (LUF)

Our Approach - Intuition

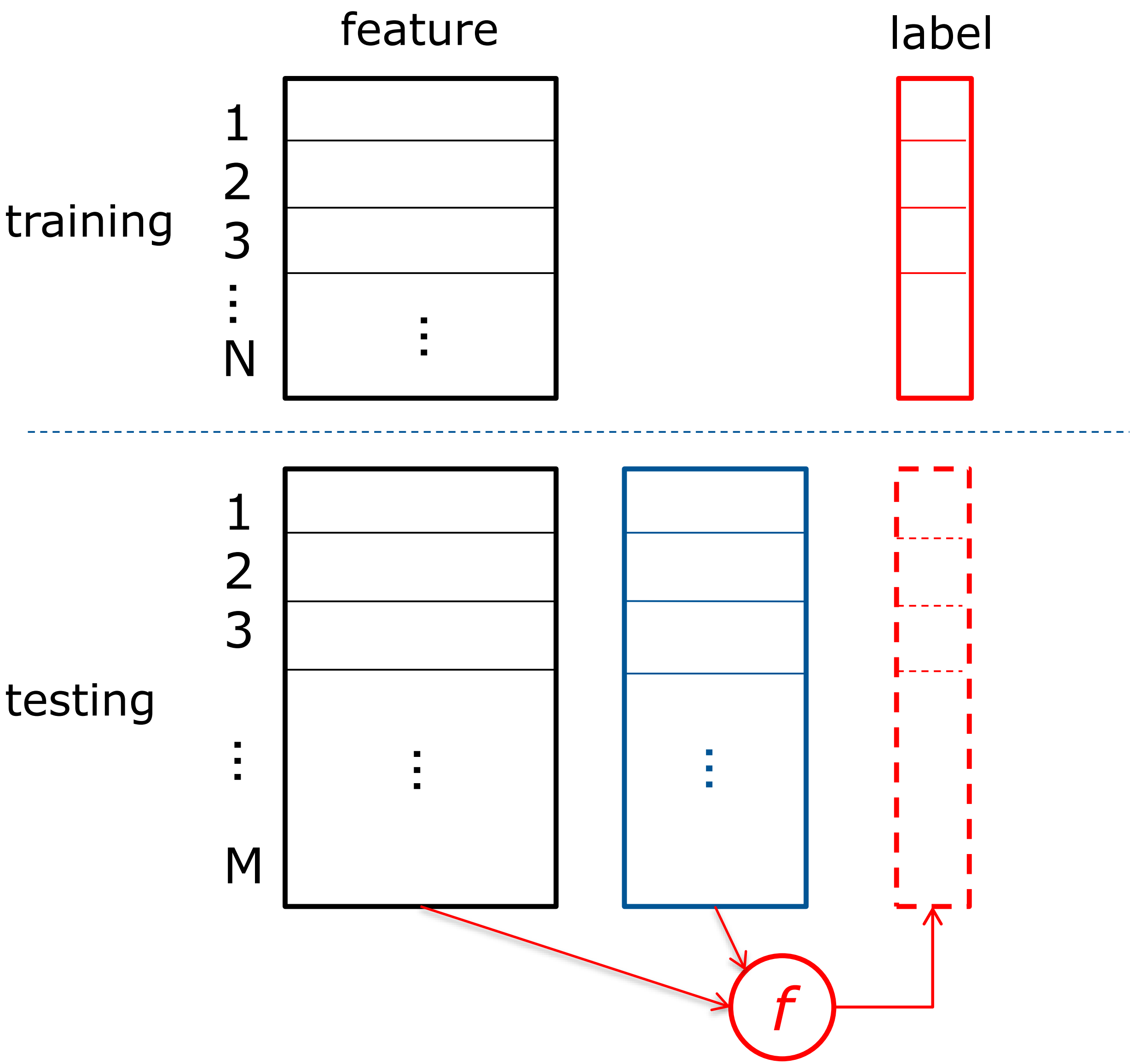


Our Approach - Intuition

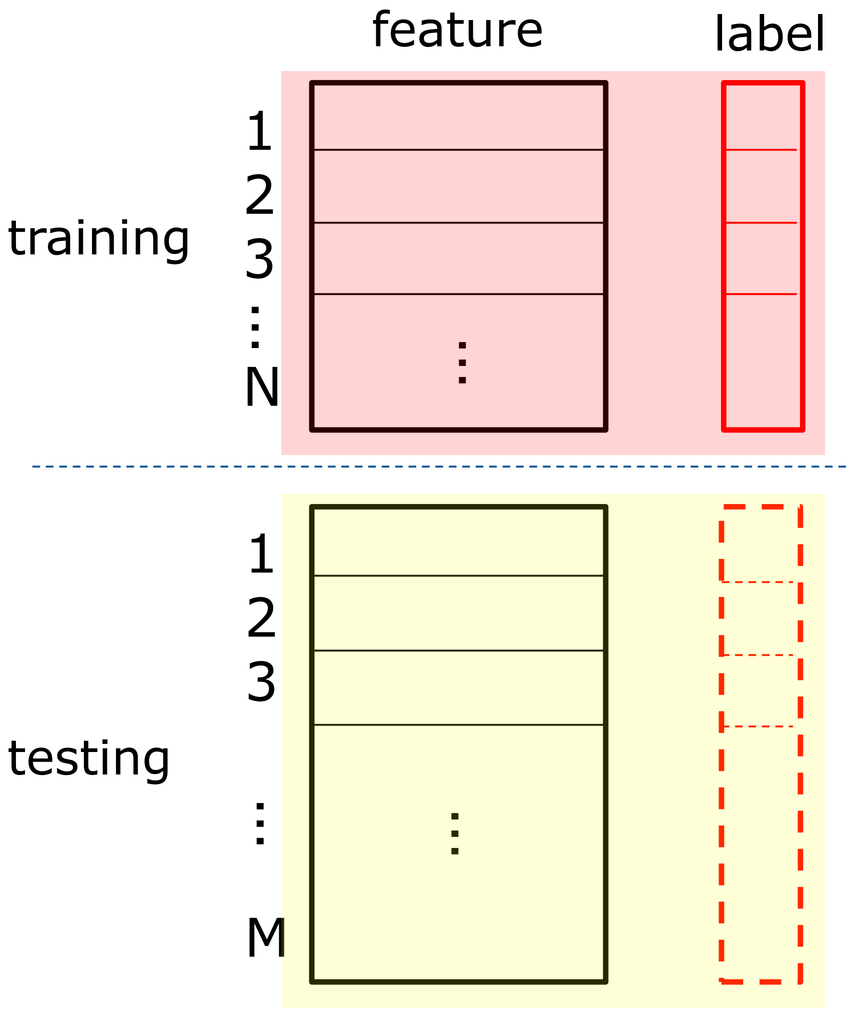


matching the joint distribution p(humidity, temperature)

Our Approach - LUF



Our Approach - LUF



Two sets have the same joint distributions!

Our Approach - LUF

Two sets have the same joint distributions!

- $(\mathbf{x}_s, y_s)$
- $(\mathbf{x}_t, \hat{y}_t)$

$$\hat{y} = f_\theta(\mathbf{x}, \mathbf{z})$$

Our Approach - LUF

Two sets have the same joint distributions!

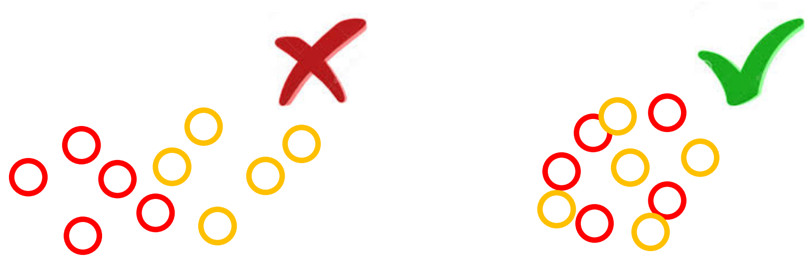
$(\mathbf{x}_s, y_s)$ $(\mathbf{x}_t, \hat{y}_t)$ $\hat{y} = f_\theta(\mathbf{x}, \mathbf{z})$

Two sets of samples mixed as much as possible

Our Approach - LUF

Two sets of samples mixed as much as possible

$(\mathbf{x}_s, y_s)$ $(\mathbf{x}_t, \hat{y}_t)$ $\hat{y} = f_\theta(\mathbf{x}, \mathbf{z})$



Two sets of samples mixed as much as possible

Minimize cross-domain *k*-nearest neighbor distances

Our Approach - LUF

Minimize cross-domain *k*-nearest neighbor distances

$$\text{dist}[(\mathbf{x}_s, y_s), (\mathbf{x}_t, \hat{y}_t)] = \|\mathbf{x}_s - \mathbf{x}_t\|_2^2 + \gamma\Delta(y_s, \hat{y}_t)$$

$$\min_\theta \sum_s \sum_{t \in \mathcal{N}_T^k(s)} \text{dist}[(\mathbf{x}_s, y_s), (\mathbf{x}_t, \hat{y}_t)] + \sum_t \sum_{s \in \mathcal{N}_S^k(t)} \text{dist}[(\mathbf{x}_t, \hat{y}_t), (\mathbf{x}_s, y_s)]$$

$(\mathbf{x}_s, y_s)$'s *k* neighbors in the target domain

$(\mathbf{x}_t, \hat{y}_t)$'s *k* neighbors in the target domain

Our Approach - LUF

Minimize cross-domain *k*-nearest neighbor distances

$$\text{dist}[(\mathbf{x}_s, y_s), (\mathbf{x}_t, \hat{y}_t)] = \|\mathbf{x}_s - \mathbf{x}_t\|_2^2 + \gamma\Delta(y_s, \hat{y}_t)$$

$$\min_{\theta} \sum_{s} \sum_{t \in \mathcal{N}_T^k(s)} \text{dist}[(\mathbf{x}_s, y_s), (\mathbf{x}_t, \hat{y}_t)] + \sum_{t} \sum_{s \in \mathcal{N}_S^k(t)} \text{dist}[(\mathbf{x}_t, \hat{y}_t), (\mathbf{x}_s, y_s)]$$

$(\mathbf{x}_s, y_s)$'s *k* neighbors in the target domain

$(\mathbf{x}_t, \hat{y}_t)$'s *k* neighbors in the target domain

non-smooth in θ, because neighbors are dependent on θ : alternating optimization

Empirical Study

Regression Tasks

Errors in regression tasks

New features

Ignore new features

Dataset	Unseen feat. ID	**R**	**R-Z^{KR}**	**R-Z^{NN}**	**LUF**	Improv.(%)
Abalone	1	2.42 ± 0.08	2.33 ± 0.076	2.31 ± 0.064	$\mathbf{2.28 \pm 0.01}$	1.3
Bank	1	0.12 ± 0.00	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.11 \pm 0.00}$	0.12 ± 0.00	-3.7
	1,2	0.15 ± 0.00	0.16 ± 0.01	0.15 ± 0.01	$\mathbf{0.13 \pm 0.00}$	17.8
	1,2,3	0.15 ± 0.00	0.16 ± 0.00	0.16 ± 0.01	$\mathbf{0.14 \pm 0.00}$	4.6
CPU	1	8.34 ± 0.20	9.17 ± 0.21	6.81 ± 0.13	$\mathbf{5.35 \pm 0.60}$	21.44
	1,2	8.37 ± 0.19	9.15 ± 0.26	6.23 ± 0.18	$\mathbf{5.79 \pm 0.56}$	7.1
	1,2,3	8.59 ± 0.18	8.74 ± 0.24	5.75 ± 0.44	$\mathbf{5.39 \pm 0.48}$	6.3
House	1	10.17 ± 1.12	10.14 ± 1.11	9.55 ± 1.18	$\mathbf{6.82 \pm 0.055}$	28.6
	1,2	10.77 ± 1.29	10.49 ± 1.03	8.52 ± 0.86	$\mathbf{6.90 \pm 0.054}$	19.1
	1,2,3	12.60 ± 1.48	12.22 ± 1.35	9.60 ± 1.15	$\mathbf{7.83 \pm 0.088}$	18.4

Predict new features

Similar trend on classification tasks

Sensor Adaptation for Weather Station

- A weather station contains several sensors
- Sensor failure happens

Time	Temp.	Humidity	Wind Speed
8:50 AM	24.2	16.7	4.3
8:55 AM	24.3	?	3.0
9:00 AM	24.8	?	3.9
9:05 AM	25.2	?	1.4

Sensor Adaptation for Weather Station

- A weather station contains several sensors
- Sensor failure happens
- When a sensor fails, we allow it to access the same sensor from a nearby station
  - But directly using the new sensor may perform poorly!

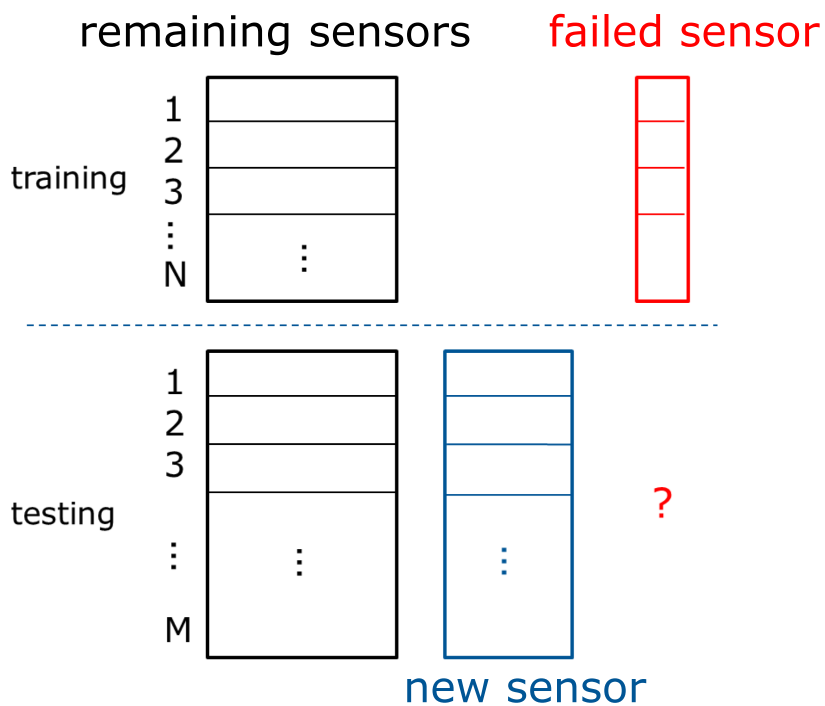
Time	Temp.	Humidity	Wind Speed
8:50 AM	24.2	16.7	4.3
8:55 AM	24.3	?	3.0
9:00 AM	24.8	?	3.9
9:05 AM	25.2	?	1.4

Time	Humidity
8:50 AM	17.6
8:55 AM	16.8
9:00 AM	16.3
9:05 AM	17.9

Nearby station

Sensor Adaptation for Weather Station

Can we reconstruct the failed sensor using the remaining sensors and new sensor?



Sensor Adaptation for Weather Station

Reconstruction errors

Ignore new features

Average improvement: 17.9%

Stations	failed sensor	**R**	**R-Z^{KR}**	**R-Z^{NN}**	**LUF**	Imp.(%)
SF-A : SF-B	wind speed	5.80 ± 0.024	5.94 ± 0.051	**5.76 ± 0.030**	5.93 ± 0.032	-2.9
	wind gust	10.52 ± 0.059	10.76 ± 0.20	10.45 ± 0.18	**9.70 ± 0.068**	7.2
	pressure	4.53 ± 0.23	4.93 ± 0.25	4.60 ± 0.35	**1.52 ± 0.25**	66.4
SJ-A : SJ-B	wind speed	3.76 ± 0.082	3.94 ± 0.15	3.78 ± 0.066	**3.74 ± 0.053**	0.51
	wind gust	4.41 ± 0.083	4.38 ± 0.092	4.37 ± 0.10	**4.16 ± 0.045**	4.8
	pressure	4.01 ± 0.021	4.03 ± 0.10	3.86 ± 0.079	**1.95 ± 0.068**	49.5
	precipitation	0.57 ± 0.034	0.56 ± 0.082	0.61 ± 0.12	**0.46 ± 0.062**	17.9
NY-A : NY-B	pressure	11.40 ± 0.14	11.43 ± 1.17	10.34 ± 0.019	**9.74 ± 0.21**	5.8
	precipitation	3.17 ± 0.092	3.96 ± 0.14	4.19 ± 0.26	**2.82 ± 0.15**	11.5

Predict new features

Thank You!

Question?