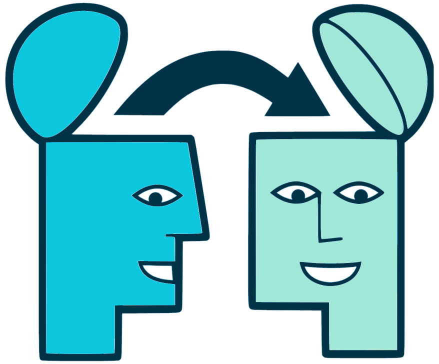
Transfer Learning

All slides adopted from the Internet.

Transfer Learning By Human

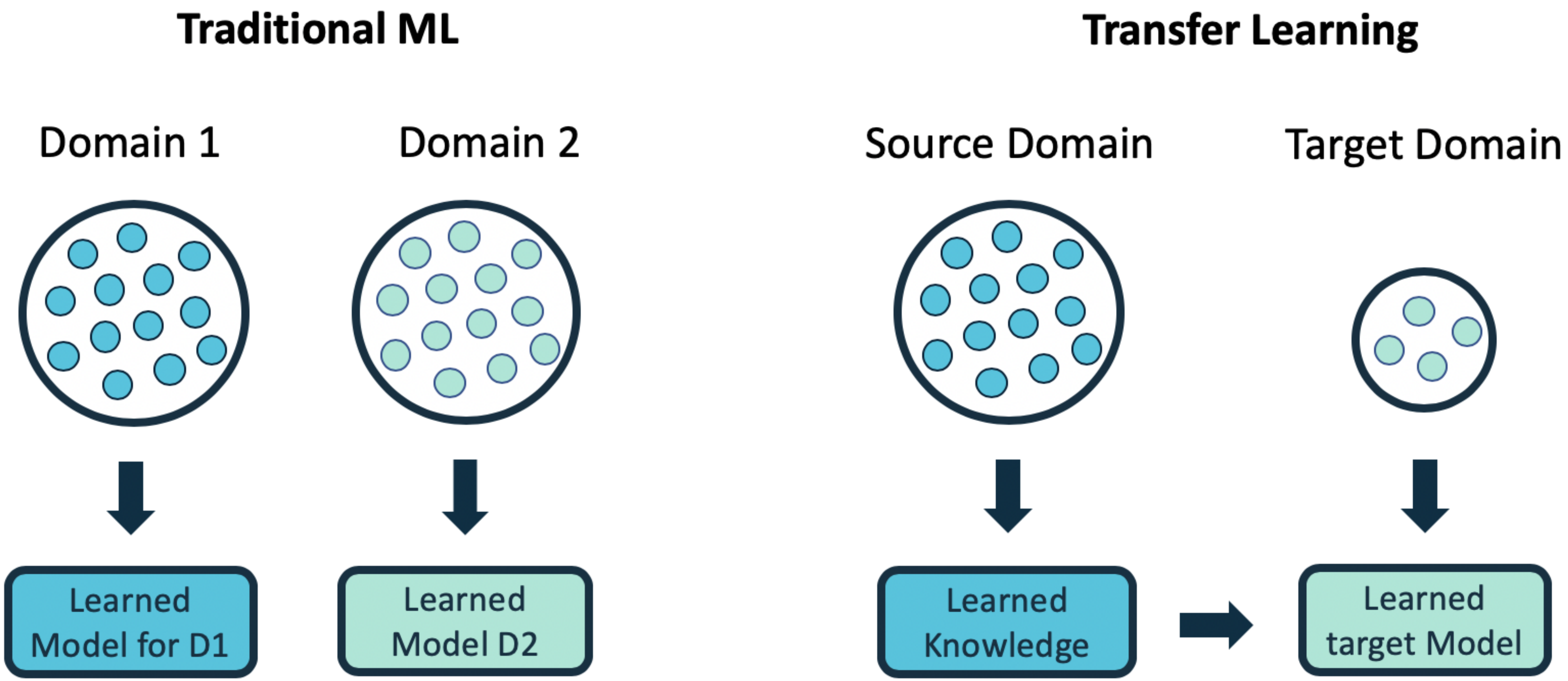


a) Transferring Learned knowledge from Java to Python.

b) Transferring Learned knowledge Table Tennis to Tennis.

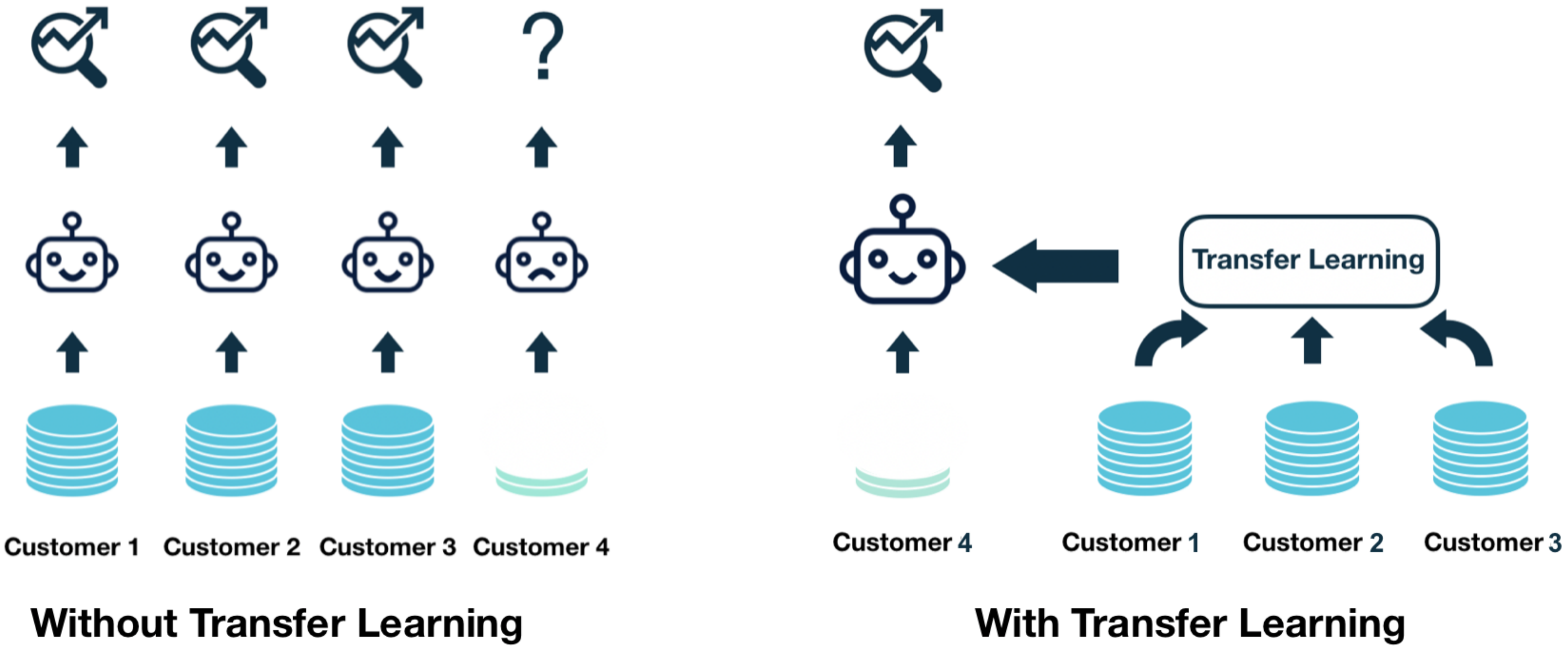
Credit: An Introduction to Transfer Learning, Georgian Impact Blog

Transfer Learning uses knowledge from other existing domains (source) during the learning process for a new domain (target)



Credit: An Introduction to Transfer Learning, Georgian Impact Blog

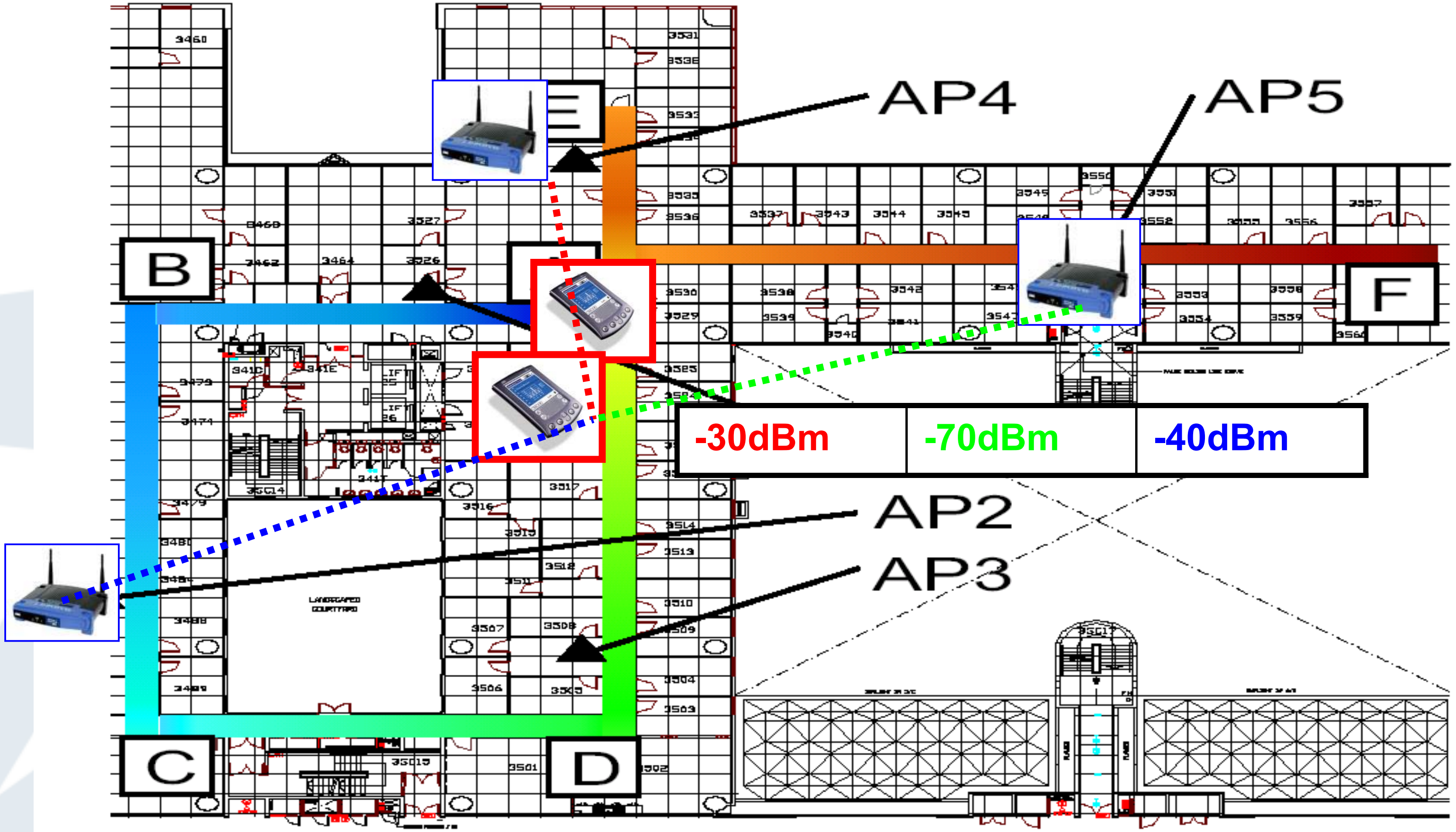
Motivating Example: On-boarding new customer



Credit: An Introduction to Transfer Learning, Georgian Impact Blog

Motivating Example I:

Indoor WiFi localization



Indoor WiFi Localization (cont.)

Average Error Distance

Training

S=(-37dbm, .., -77dbm), L=(1, 3)
S=(-41dbm, .., -83dbm), L=(1, 4)
…
S=(-49dbm, .., -34dbm), L=(9, 10)
S=(-61dbm, .., -28dbm), L=(15,22)

Device A

Localization model

Test

S=(-37dbm, .., -77dbm)
S=(-41dbm, .., -83dbm)
…
S=(-49dbm, .., -34dbm)
S=(-61dbm, .., -28dbm)

Device A

~ 1.5 meters

Drop!

Training

S=(-33dbm, .., -82dbm), L=(1, 3)
…
S=(-57dbm, .., -63dbm), L=(10, 23)

Device B

Localization model

Test

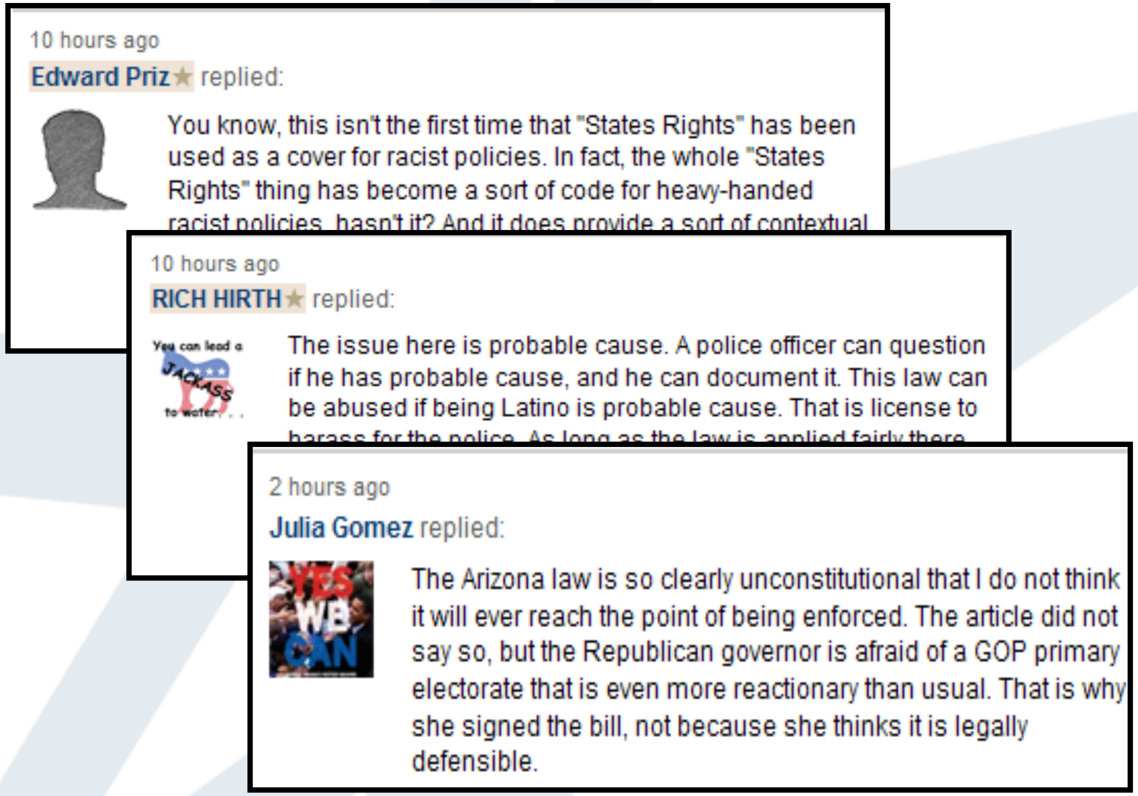
S=(-37dbm, .., -77dbm)
S=(-41dbm, .., -83dbm)
…
S=(-49dbm, .., -34dbm)
S=(-61dbm, .., -28dbm)

Device A

~10 meters

Motivating Example II:

Sentiment classification


10 hours ago

Edward Priz replied:

You know, this isn't the first time that "States Rights" has been used as a cover for racist policies. In fact, the whole "States Rights" thing has become a sort of code for heavy-handed racist policies, hasn't it? And it does provide a sort of contextual

10 hours ago

RICH HIRTH replied:

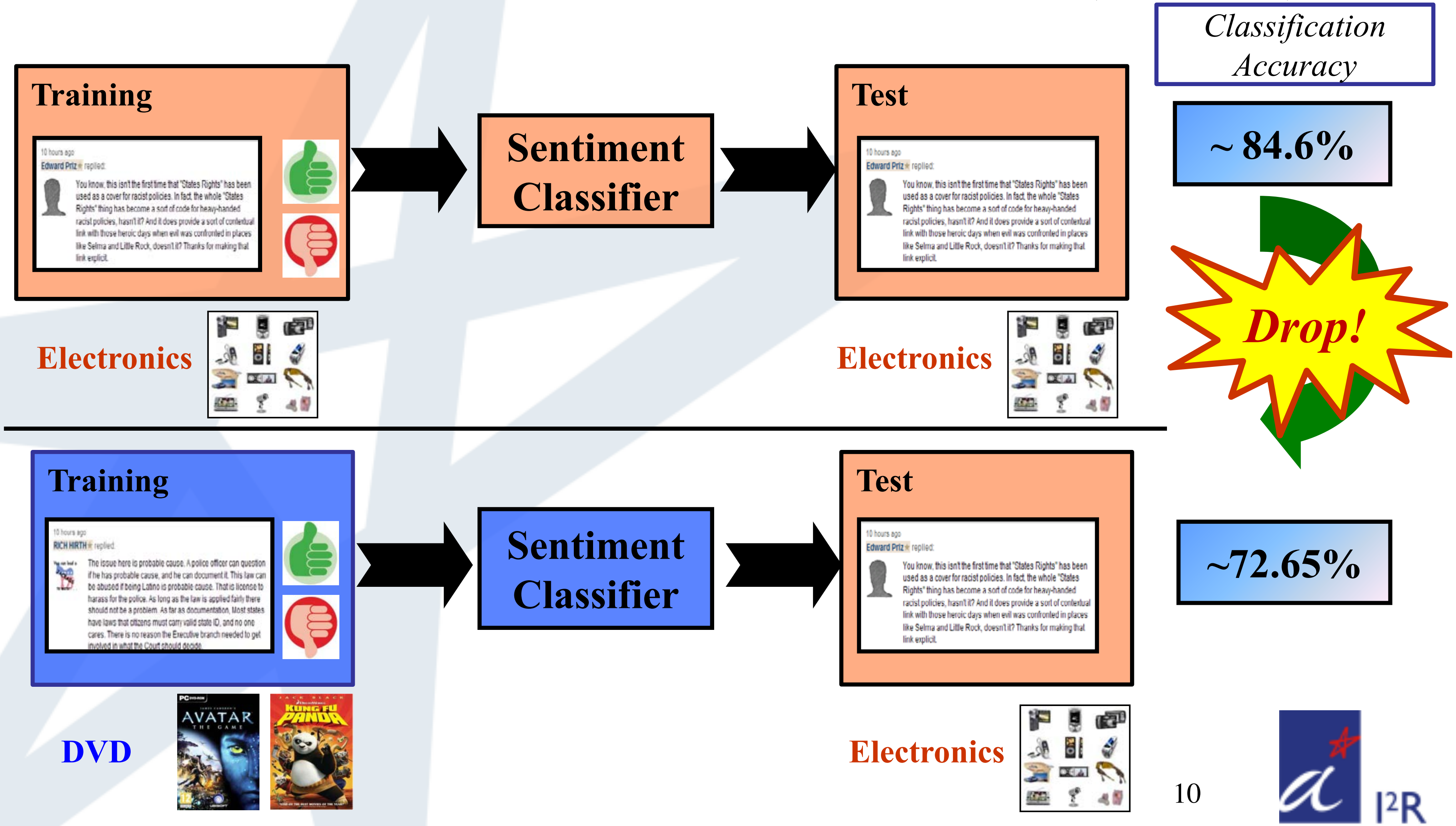
The issue here is probable cause. A police officer can question if he has probable cause, and he can document it. This law can be abused if being Latino is probable cause. That is license to harass for the police. As long as the law is applied fairly there

2 hours ago

Julia Gomez replied:

The Arizona law is so clearly unconstitutional that I do not think it will ever reach the point of being enforced. The article did not say so, but the Republican governor is afraid of a GOP primary electorate that is even more reactionary than usual. That is why she signed the bill, not because she thinks it is legally defensible.

Sentiment Classification (cont.)



Difference between Domains

Electronics	Video Games
(1) **Compact**; easy to operate; very good picture quality; looks **sharp**!	(2) A very good game! It is action packed and full of excitement. I am very much **hooked** on this game.
(3) I purchased this unit from Circuit City and I was very excited about the quality of the picture. It is really nice and **sharp**.	(4) Very **realistic** shooting action and good plots. We played this and were **hooked**.
(5) It is also quite **blurry** in very dark settings. I will never buy HP again.	(6) The game is so **boring**. I am extremely unhappy and will probably never buy UbiSoft again.

Motivation

Why we need Transfer Learning[Tang et al., 2012]?

- **Labeled data are expensive and limited.**
- **Related data are cheap and sufficient.**

Motivation

Why we need Transfer Learning[Tang et al., 2012]?

- **Labeled data are expensive and limited.**
- **Related data are cheap and sufficient.**

Figure : Object detector for static image is easy to obtain. However, the labeled data for video task are limited and expensive.

Transfer learning addresses these three questions

- What information in the source is useful and transferable to target?
- What is the best way of transferring this information?
- How to avoid transferring information that is detrimental to the desired outcome?

Notation: Domain

Domain: A domain $\mathfrak{D} = \{\mathcal{X}, P(X)\}$ is defined by two components:

- A feature space $\mathcal{X}$
- A marginal probability distribution $P(X)$ where $X = \{x_1, x_2, x_3, \ldots, x_n\} \in \mathcal{X}$

If two domains are different, then they either have different feature spaces ($\mathcal{X}_t \neq \mathcal{X}_s$) or different marginal distributions ($P(X_t) \neq P(X_s)$).

Notation: Task

Task: Given a specific domain $\mathfrak{D}$, a task $\mathcal{T}$={Y, $\boldsymbol{f}$(.)} consists of two parts:

- A label space Y
- A predictive function $\boldsymbol{f}$(.), which is not observed but can be learned from training data {(x_i, y_i)| i ∈ {1, 2, 3, …, N}, where $x_i \in X$ and $y_i \in Y$}.
- From a probabilistic viewpoint $\boldsymbol{f}(x_i)$, can also be written as p(y_i|x_i), so we can rewrite task $\mathcal{T}$ as $\mathcal{T}$={Y, P(Y| X)}.

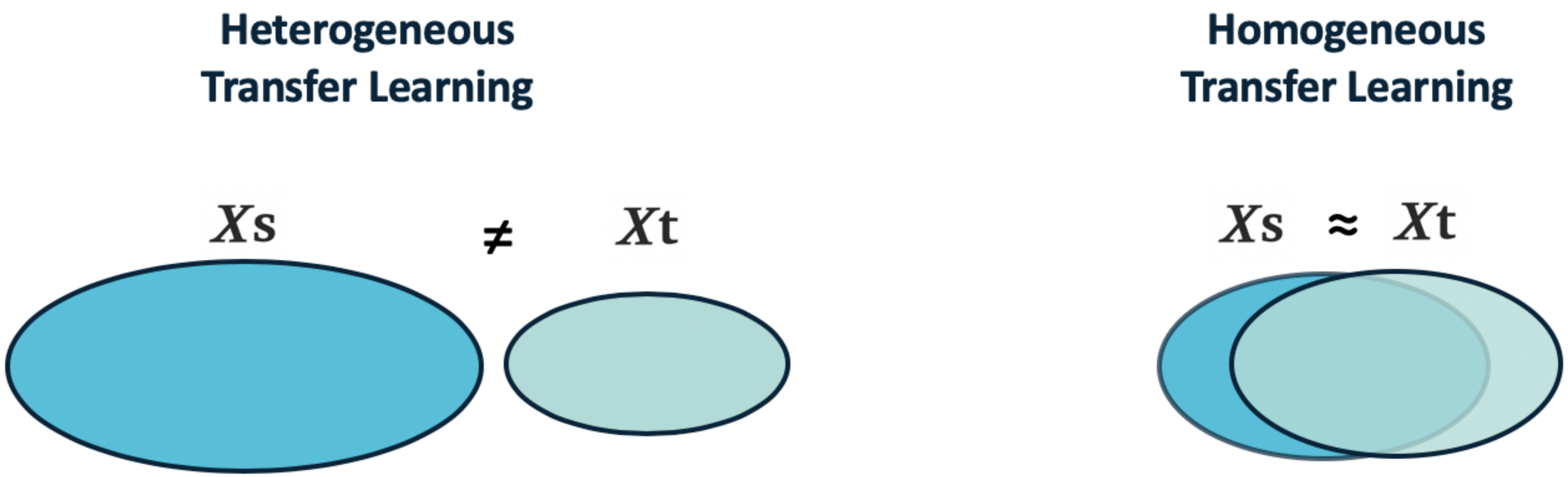
In general, if two tasks are different, then they may have different label spaces($Yt \neq Ys$) or different conditional probability distributions (P(Yt| X t) ≠ P(Ys| X s)).

Definition of Transfer Learning

- Given a source domain $\mathfrak{D}s$ and corresponding learning task $\mathcal{T}s$, a target domain $\mathfrak{D}t$ and learning task $\mathcal{T}t$, **transfer learning** aims to improve the learning of the conditional probability distribution P(Yt| X t) in $\mathfrak{D}t$ with the information gained from $\mathfrak{D}s$ and $\mathcal{T}s$, where $\mathfrak{D}t \neq \mathfrak{D}s$ or $\mathcal{T}t \neq \mathcal{T}s$.

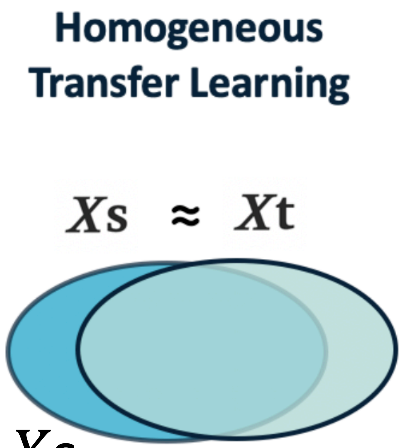
- If we take this definition of domain and task, then we will have either $\mathfrak{D}t \neq \mathfrak{D}s$ or $\mathcal{T}t \neq \mathcal{T}s$
  - $\mathcal{X}$**t ≠** $\mathcal{X}$**s**
  - **P(X t) ≠ P(X s)**
  - $\mathcal{Y}$**t ≠** $\mathcal{Y}$**s**
  - **P(Yt| X t) ≠ P(Ys| X s)**

Homogeneous v.s. Heterogeneous Transfer Learning



Credit: An Introduction to Transfer Learning, Georgian Impact Blog

Homogeneous Transfer Learning

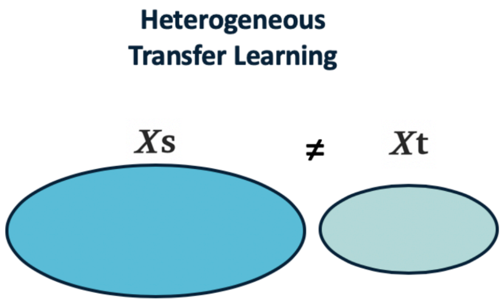


In homogeneous transfer learning， we have the situation where $X_t = X_s$ and $Y_t = Y_s$. Therefore, we want to bridge the gap in the data distributions between the source and target domains, i.e. address $P(X_t) \neq P(X_s)$ and/or $P(Y_t|X_t) \neq P(Y_s|X_s)$. The solutions to homogeneous transfer learning problems use one of the following general strategies:

- Trying to correct for the marginal distribution differences in the source and target ($P(X_t) \neq P(X_s)$).
- Trying to correct for the conditional distribution difference in the source and target ($P(Y_t|X_t) \neq P(Y_s|X_s)$).
- Trying to correct both the marginal and conditional distribution differences in the source and target.

Credit: An Introduction to Transfer Learning, Georgian Impact Blog

Heterogeneous Transfer Learning

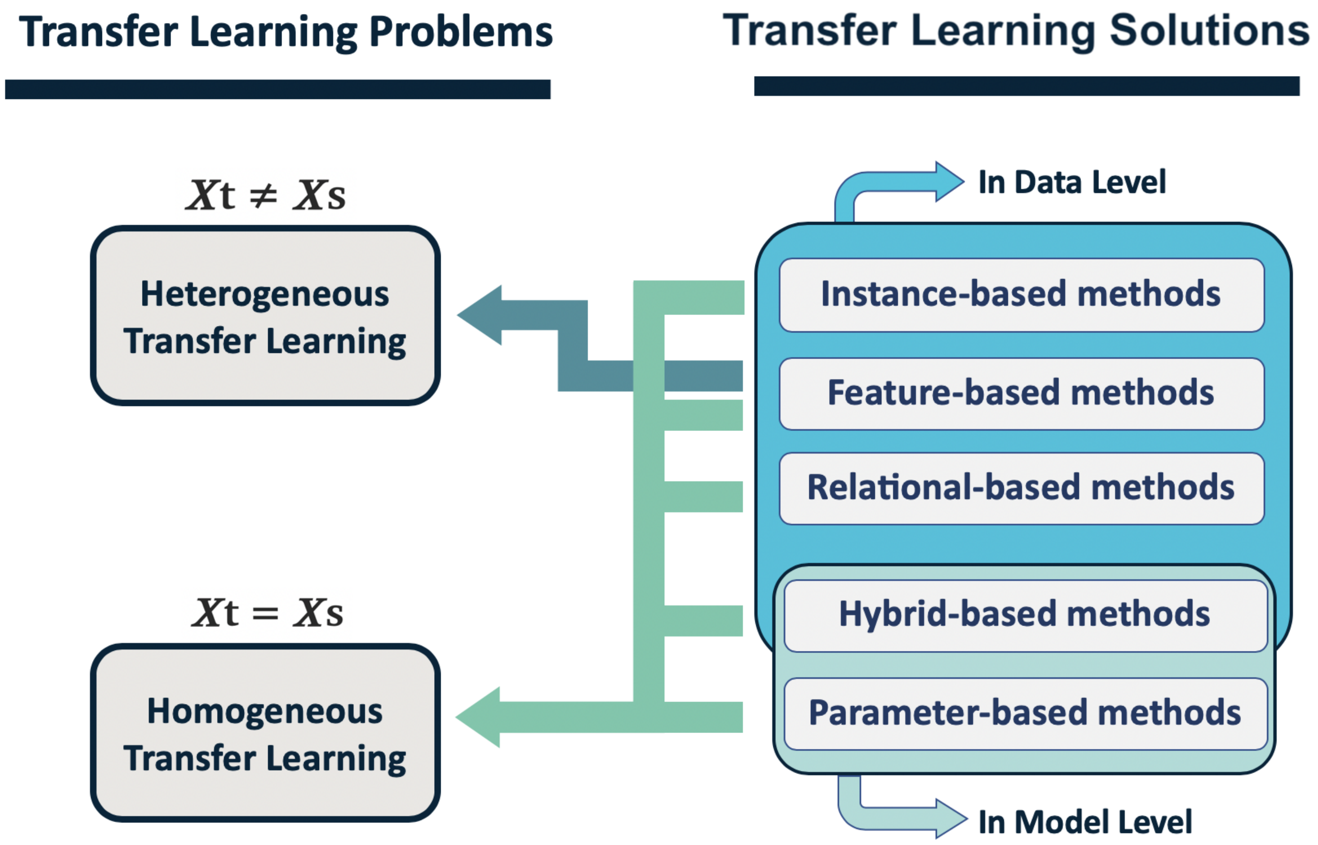


- In heterogeneous transfer learning, the source and target have different feature spaces $X_t \neq X_s$ (generally non-overlapping) and/or $Y_t \neq Y_s$, as the source and target domains may share no features and/or labels.

- Heterogeneous transfer learning solutions bridge the gap between feature spaces and reduce the problem to a homogeneous transfer learning problem where further distribution (marginal or conditional) differences will need to be corrected.

Credit: An Introduction to Transfer Learning, Georgian Impact Blog

Transfer Learning Solutions

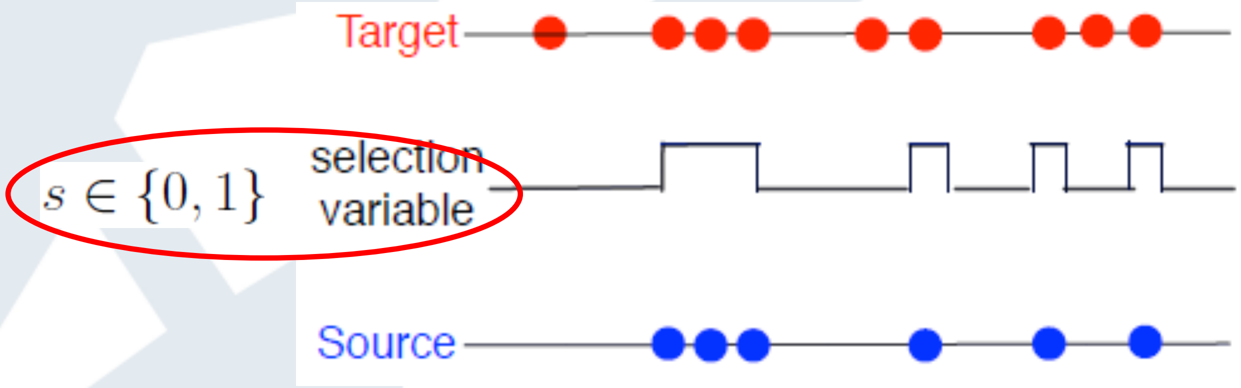


Credit: An Introduction to Transfer Learning, Georgian Impact Blog

Instance-based Approaches

Correcting sample selection bias

- Imagine a ***rejection*** sampling process, and view the source domain as samples from the target domain



Assumption: sample selection bias is caused by the data generation process

Instance-based Approaches

Correcting sample selection bias (cont.)

- The distribution of the selector variable maps the target onto the source distribution

$$P_S(x) \propto P_T(x)P(s=1|x)$$

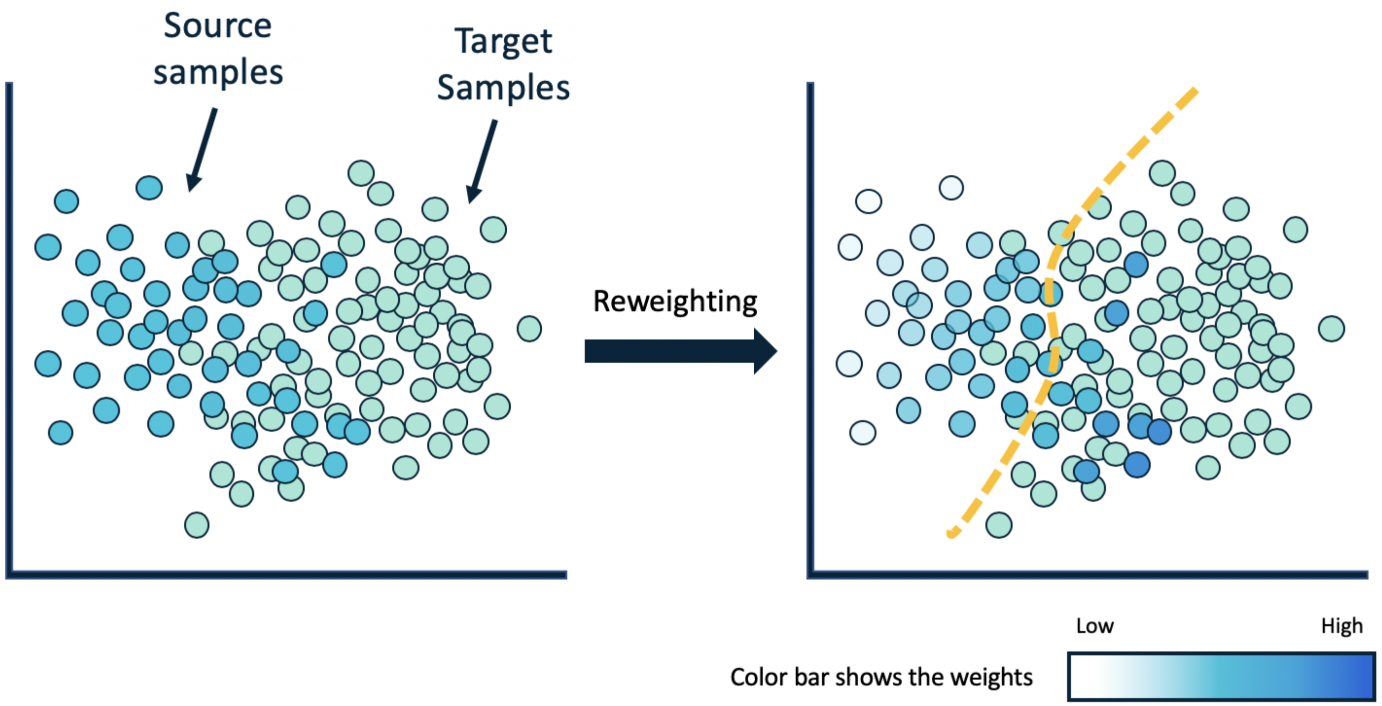
$$\Downarrow$$

$$\beta(x) = \frac{P_T(x)}{P_S(x)} \propto \frac{1}{P(s=1|x)}$$

[Zadrozny, ICML-04]

- ➢ Labeled instances from the source domain with label 1
- ➢ Unlabeled instances from the target domain with label 0
- ➢ Train a binary classifier

1. Instance-Based Transfer Learning



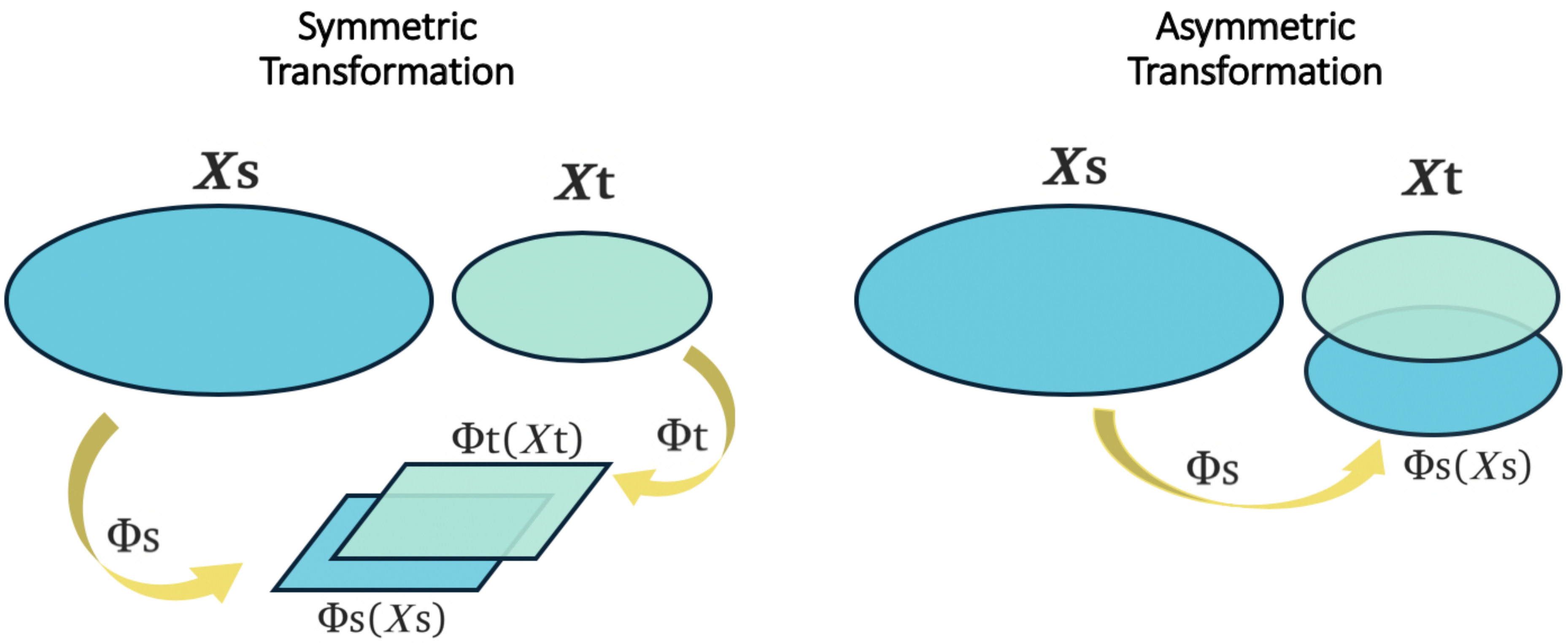
A source sample's probability of being in target domains

Instance-based transfer learning methods try to reweight the samples in the source domain in an attempt to correct for marginal distribution differences. One common solution is to train a binary classifier that separates source samples from target samples and then use this classifier to estimate the source sample weights. This method gives a higher weight to the source samples that are more similar to target samples.

Credit: An Introduction to Transfer Learning, Georgian Impact Blog

2. Feature-based transfer Learning (for both homogeneous and heterogenous transfer learning)

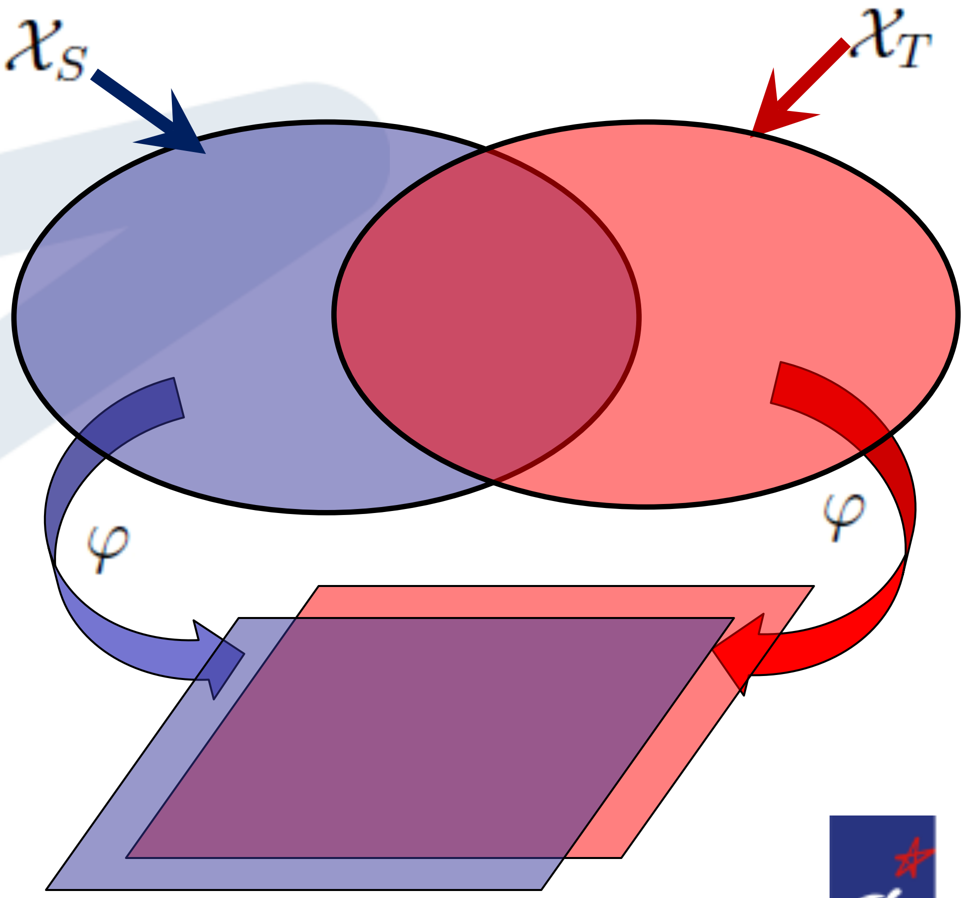
This approach discovers underlying meaningful structures by transforming both of the domains to a common latent feature space — usually of a low dimension — that has predictive qualities while reducing the marginal distribution between the domain.



Credit: An Introduction to Transfer Learning, Georgian Impact Blog

Feature-based Transfer Learning Approaches

When source and target domains only have some overlapping features. (lots of features only have support in either the source or the target domain)



Feature-based Approaches

Encode application-specific knowledge

Electronics	Video Games
(1) **Compact**; easy to operate; very good picture quality; looks **sharp**!	(2) A very good game! It is action packed and full of excitement. I am very much **hooked** on this game.
(3) I purchased this unit from Circuit City and I was very excited about the quality of the picture. It is really nice and **sharp**.	(4) Very **realistic** shooting action and good plots. We played this and were **hooked**.
(5) It is also quite **blurry** in very dark settings. I will never_buy HP again.	(6) The game is so **boring**. I am extremely unhappy and will probably never_buy UbiSoft again.

Feature-based Approaches

Encode application-specific knowledge (cont.)

Electronics

	compact	sharp	blurry	hooked	realistic	boring
👍	1	1	0	0	0	0
👍	0	1	0	0	0	0
👎	0	0	1	0	0	0

Training

$$y = f(x) = \mathrm{sgn}(w \cdot x^T), \qquad w = [1,1,-1,0,0,0]$$

Prediction

Video Game

	compact	sharp	blurry	hooked	realistic	boring
👍	0	0	0	1	0	0
👍	0	0	0	1	1	0
👎	0	0	0	0	0	1

Feature-based Approaches

Encode application-specific knowledge (cont.)

Electronics	Video Games
(1) **Compact**; easy to operate; very ***good*** picture quality; looks **sharp**!	(2) A very ***good*** game! It is action packed and full of ***excitement***. I am very much **hooked** on this game.
(3) I purchased this unit from Circuit City and I was very ***excited*** about the quality of the picture. It is really ***nice*** and **sharp**.	(4) Very **realistic** shooting action and ***good*** plots. We played this and were **hooked**.
(5) It is also quite **blurry** in very dark settings. I will ***never_buy*** HP again.	(6) The game is so **boring**. I am extremely ***unhappy*** and will probably ***never_buy*** UbiSoft again.

Feature-based Approaches

Encode application-specific knowledge (cont.)

- Three different types of features
  - Source domain (***Electronics***) specific features, e.g., ***compact, sharp, blurry***
  - Target domain (***Video Game***) specific features, e.g., ***hooked***, ***realistic***, ***boring***
  - Domain independent features (pivot features), e.g., ***good, excited***, ***nice***, ***never_buy***

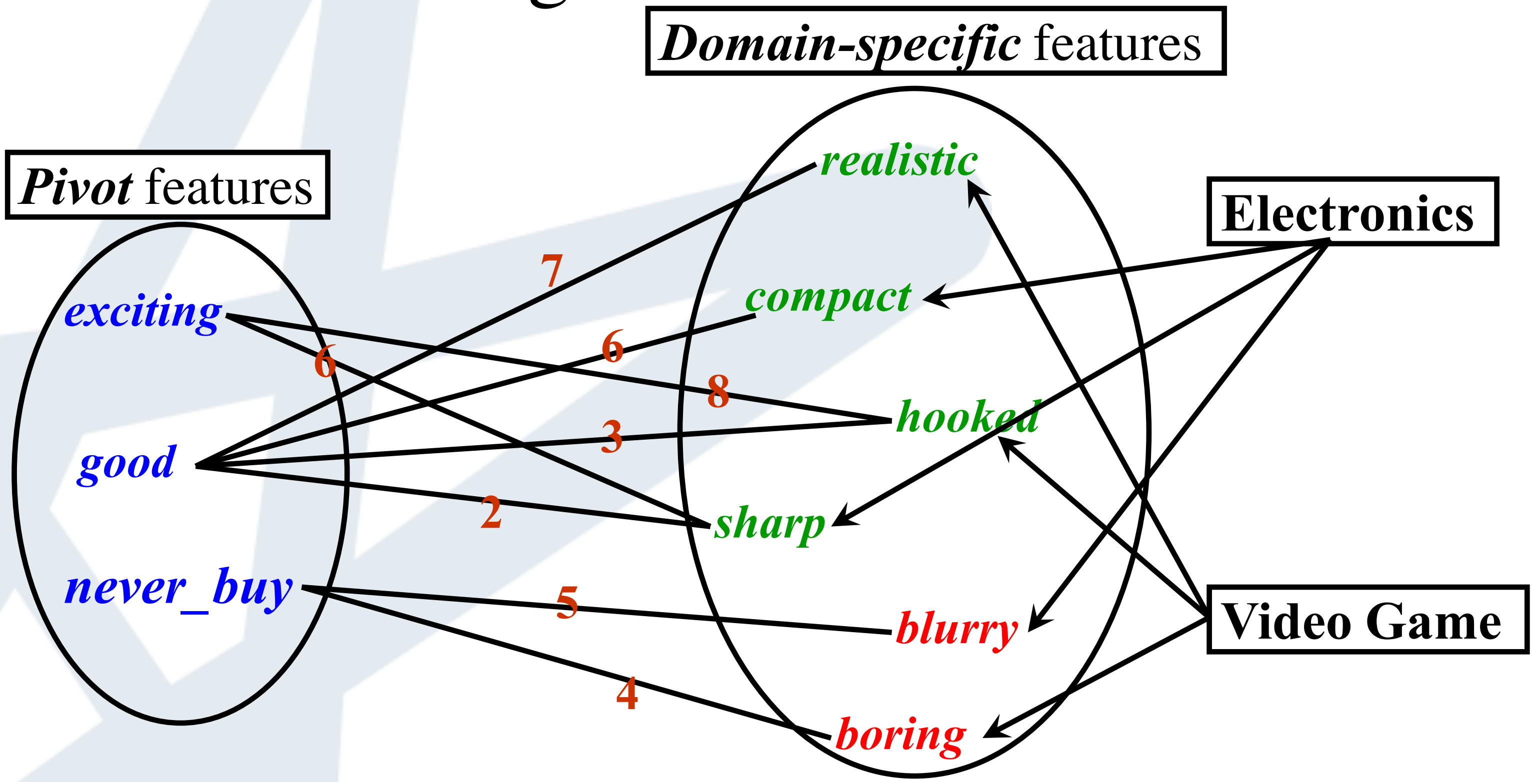
Feature-based Approaches

Spectral Feature Alignment (SFA)

- **Intuition**
  - Use a ***bipartite*** graph to model the correlations between ***pivot*** features and other features
  - Discover new shared features by applying ***spectral clustering*** techniques on the graph

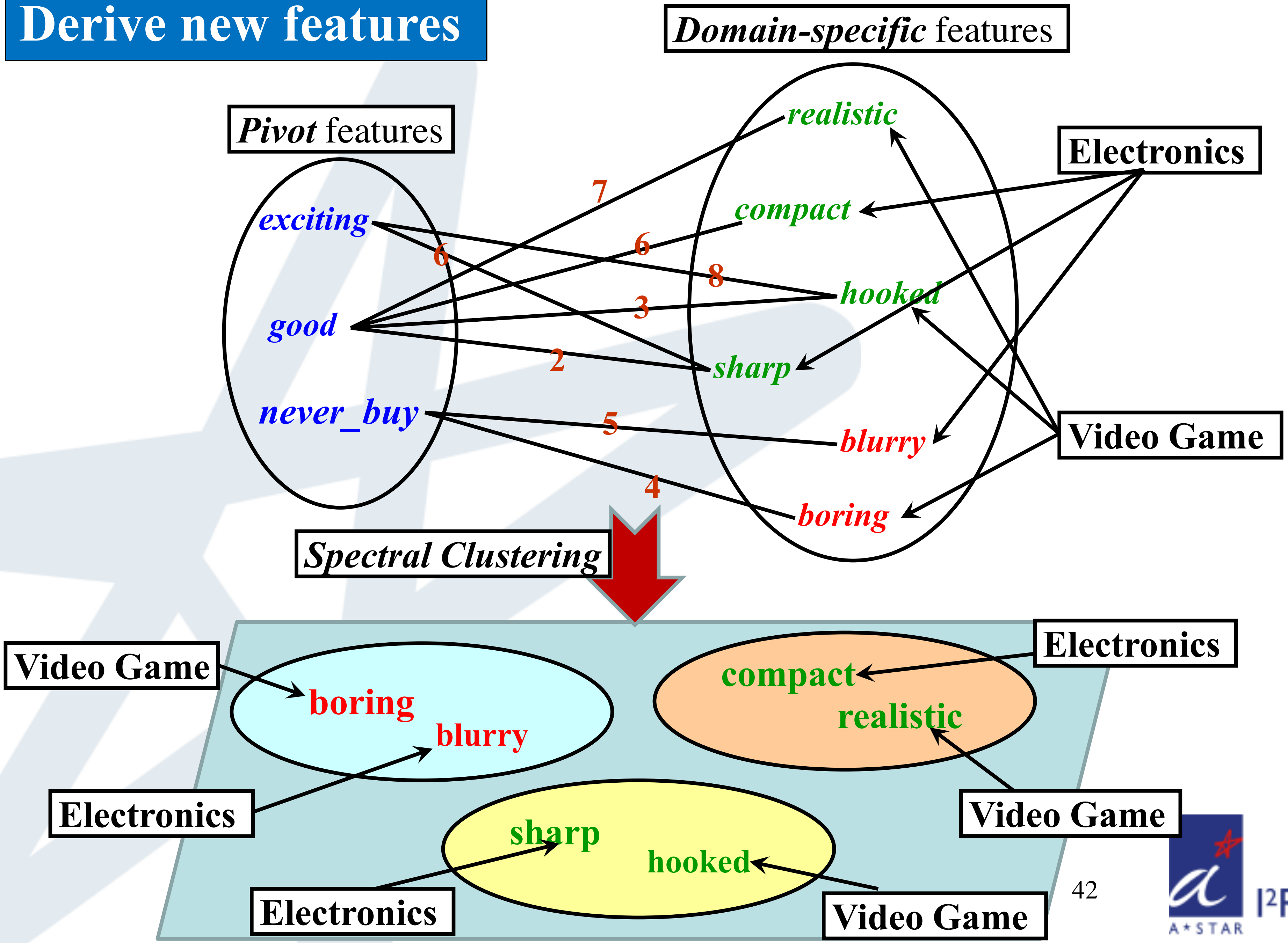
Spectral Feature Alignment (SFA)

High level idea



- ➢ If two ***domain-specific*** words have connections to more common ***pivot*** words in the graph, they tend to be aligned or clustered together with a higher probability.
- ➢ If two ***pivot*** words have connections to more common ***domain-specific*** words in the graph, they tend to be aligned together with a higher probability.

Derive new features



Spectral Feature Alignment (SFA)

Derive new features (cont.)

Electronics

	sharp/hooked	compact/realistic	blurry/boring
	1	1	0
	1	0	0
	0	0	1

Training

$$y = f(x) = \mathrm{sgn}(w \cdot x^T), \qquad w = [1,1,-1]$$

Prediction

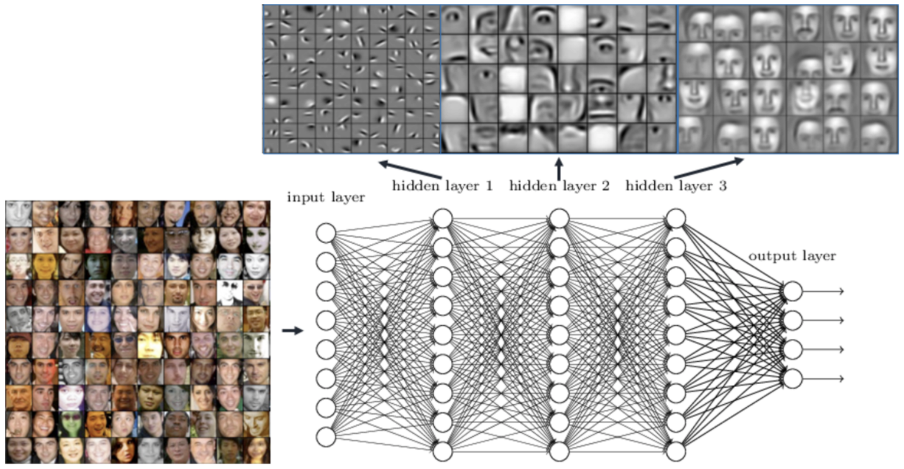
Video Game

	sharp/hooked	compact/realistic	blurry/boring
	1	0	0
	1	1	0
	0	0	1

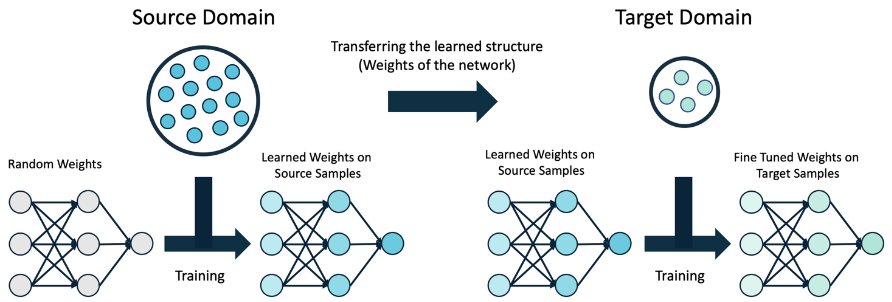
3. Parameter-based Approaches

- Idea: a well-trained model on the source domain has learned a well-defined structure, and if two tasks are related, this structure can be transferred to the target model.

- How: Instead of starting with random weights, start with the previously trained weights from another similar domain (source) and then fine-tune the weights specifically for a new domain (target).
  - Save time
  - Requires much less labeled data.
  - Improve robustness

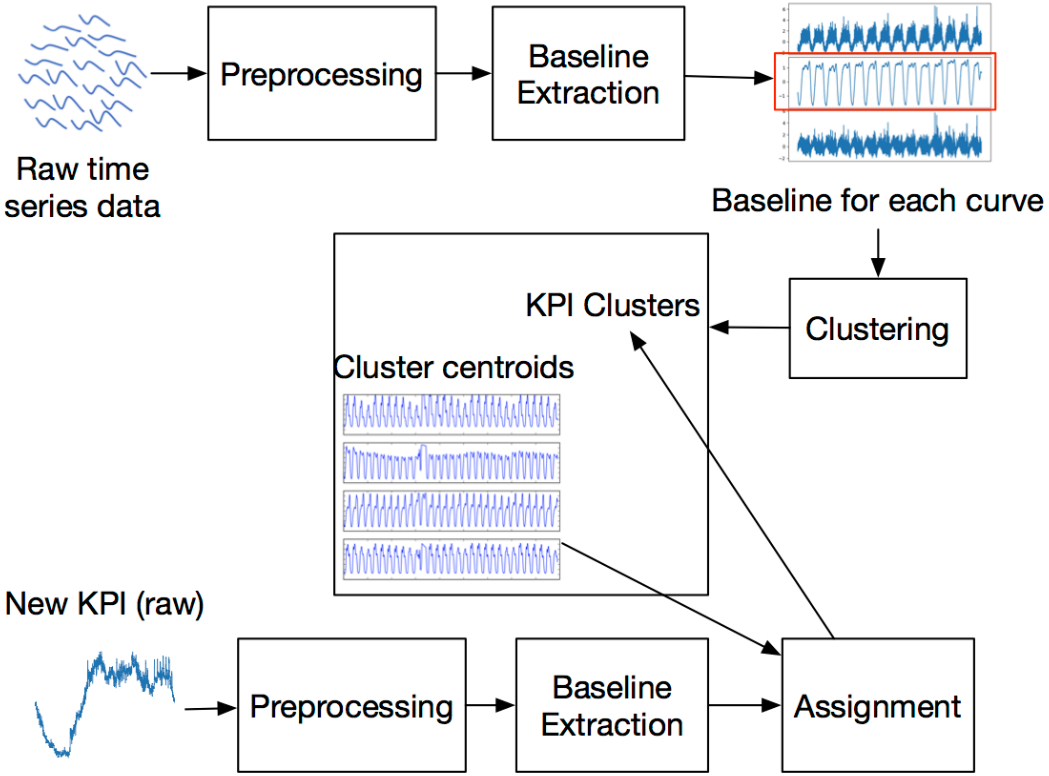


b) Deep neural networks learn hierarchical representations [26]



a) Parameter-based Transfer Learning methods

ROCKA: Clustering + Transfer Learning to reduce training overhead



IWQoS 2018

	Original DONUT [WWW2018]	ROCKA+DONUT+KPI-specific threshold	ROCKA+DONUT
Avg. F-score	0.89	0.88	0.76
Total training time (s)	51621	5145	5145

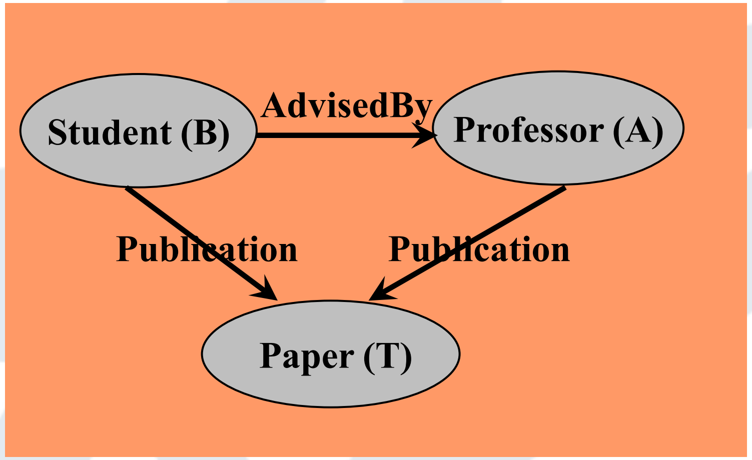
Relational Transfer Learning Approaches

- **Motivation:** If two relational domains (data is non-i.i.d) are related, they may share some similar relations among objects. These relations can be used for knowledge transfer across domains.

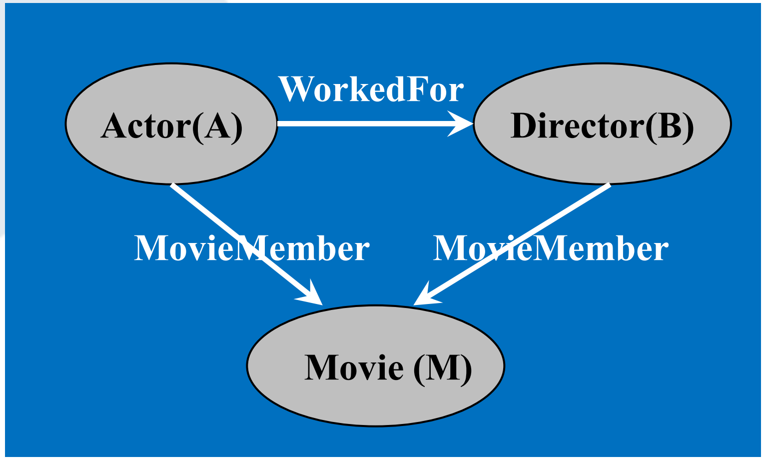
Relational Transfer Learning Approaches (cont.)

[Mihalkova *etal*., AAAI-07, Davis and Domingos, ICML-09]

Academic domain (source)



Movie domain (target)

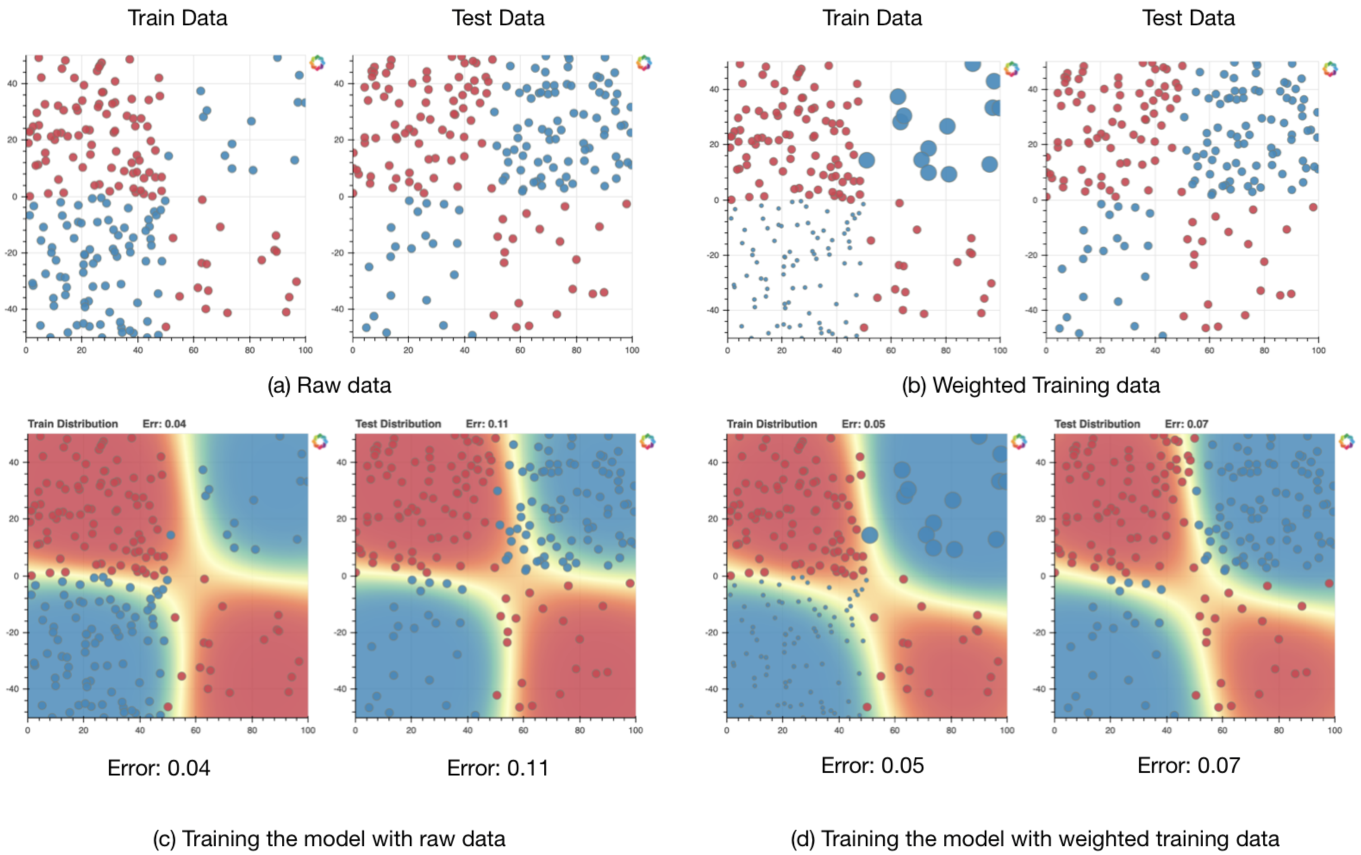


AdvisedBy (B, A) ∧ Publication (B, T) => Publication (A, T)

WorkedFor (A, B) ∧ MovieMember (A, M) => MovieMember (B, M)

P1(x, y) ∧ P2 (x, z) => P2 (y, z)

Real-World Application: Dealing with Domain Shift/Concept Drift



Credit: An Introduction to Transfer Learning, Georgian Impact Blog

StepWise: Robust and Rapid Adaption for Concept Drift in Software System Anomaly Detection