Introduction to Transfer Learning

Magda Friedjungová
magda.friedjungova@fit.cvut.cz
1. 11. 2017

Faculty of Information Technology
Czech Technical University in Prague, Czech Republic

Outline

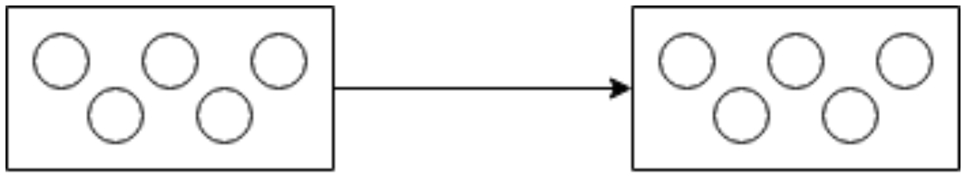
- Traditional Machine Learning
- Motivation
- Definitions
- Research Issues
    - Transfer Learning Hierarchy
- Challenges
- Heterogeneous Transfer Learning
- Actual Work

A Major Assumption in Traditional Machine Learning

Training and test (future) data come from the same task and the same domain:

- **Represented in the same feature and label spaces.**
- **Follow the same distribution.**

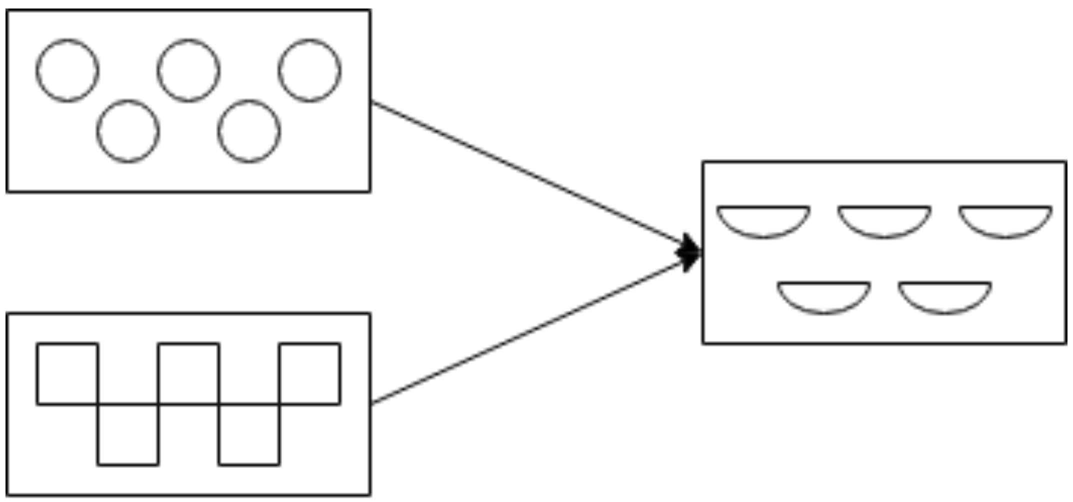
Class	Age	Sex	Survived
3rd	adult	male	0
crew	adult	male	1
3rd	child	male	0
1st	child	female	1
...	...	...	...



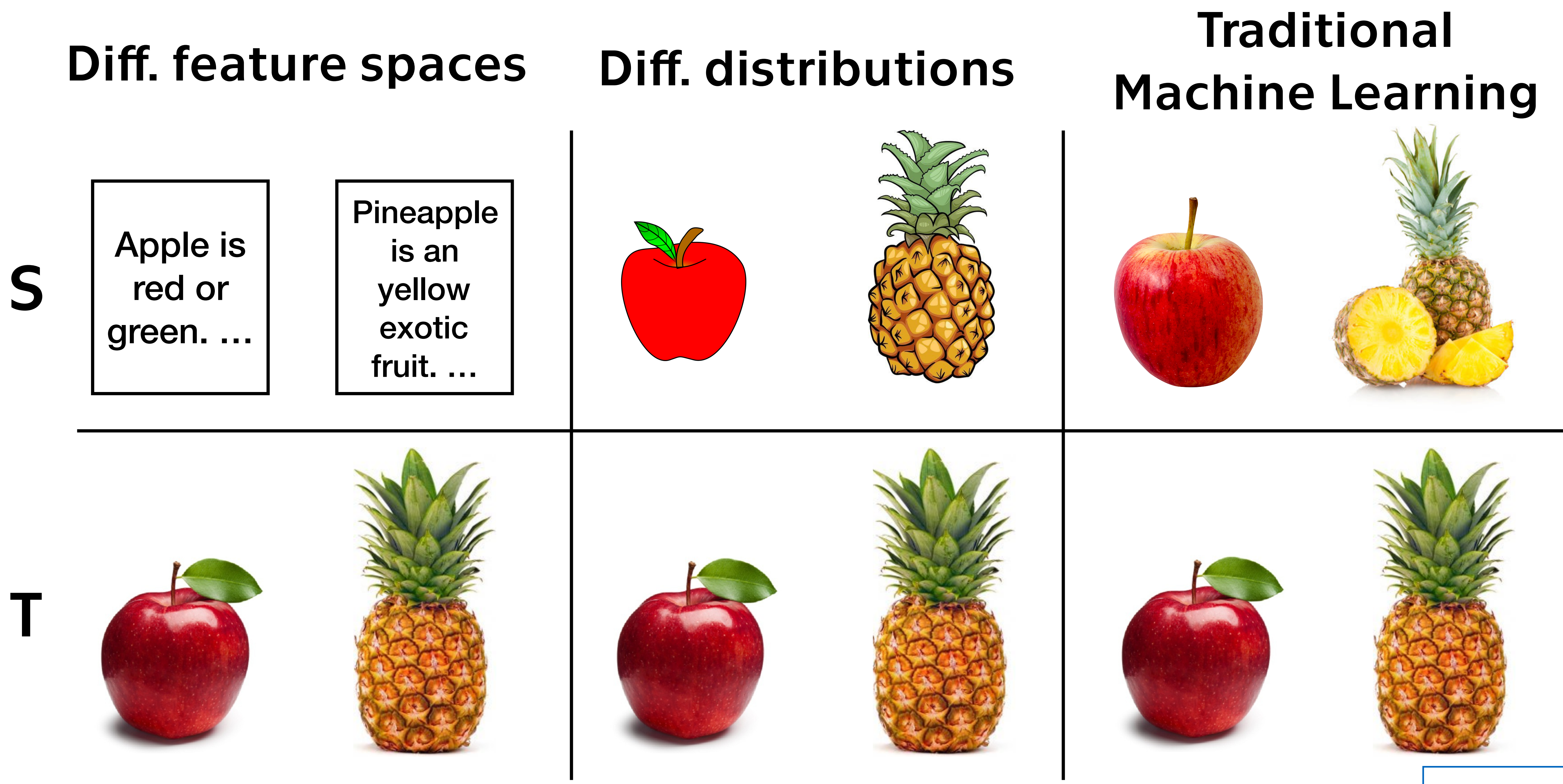
In Real-world Application

Training and test (future) data may come from different domains and different tasks, which have:

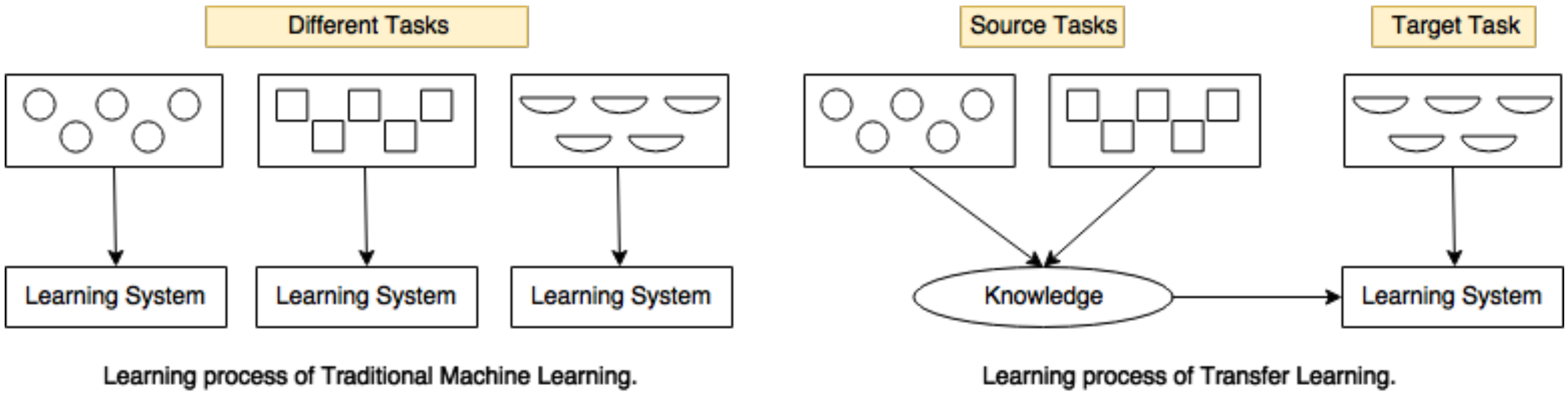
- different marginal distributions or different feature spaces,
- different predictive distributions or different label spaces.



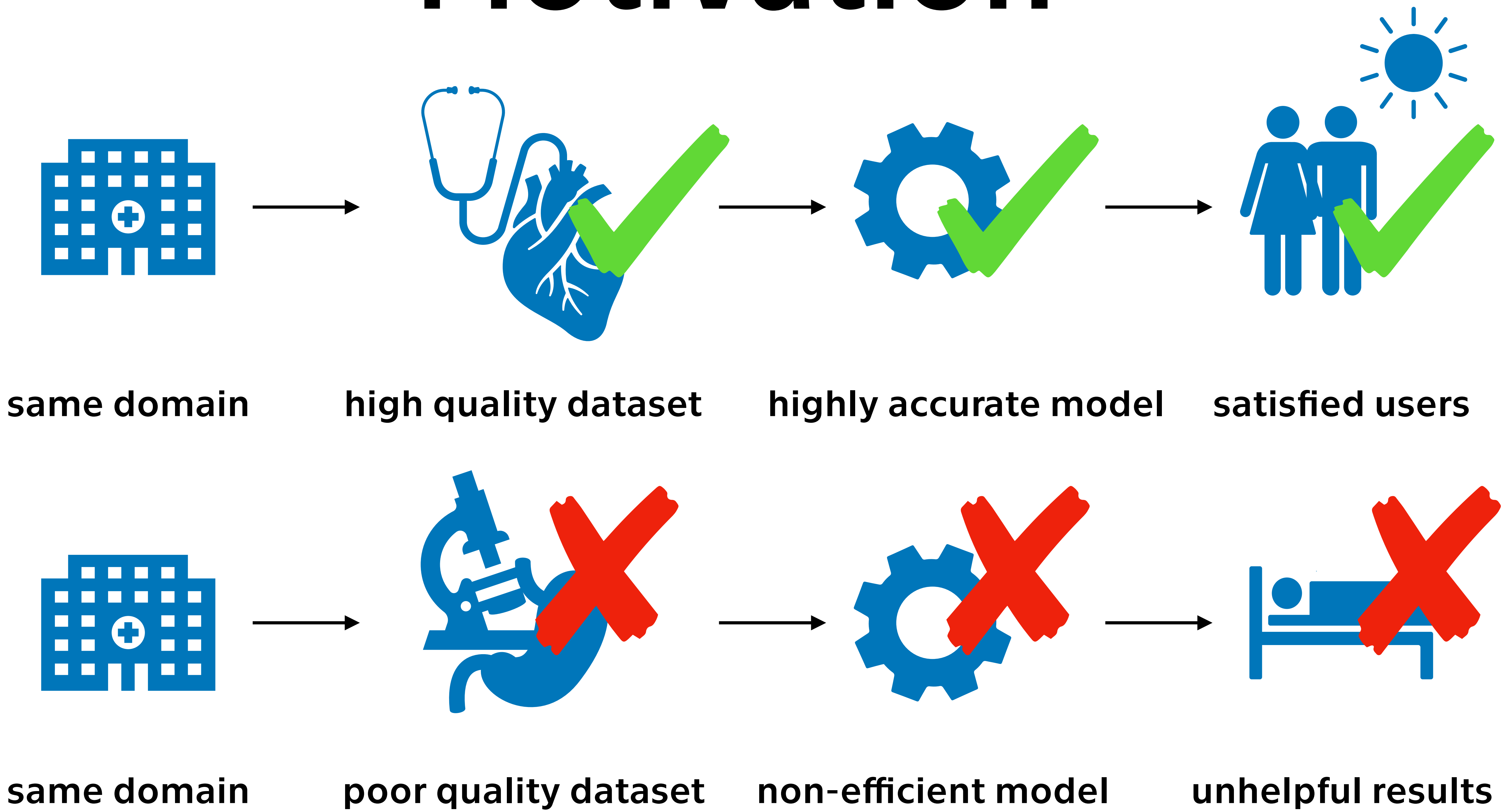
Data Representation



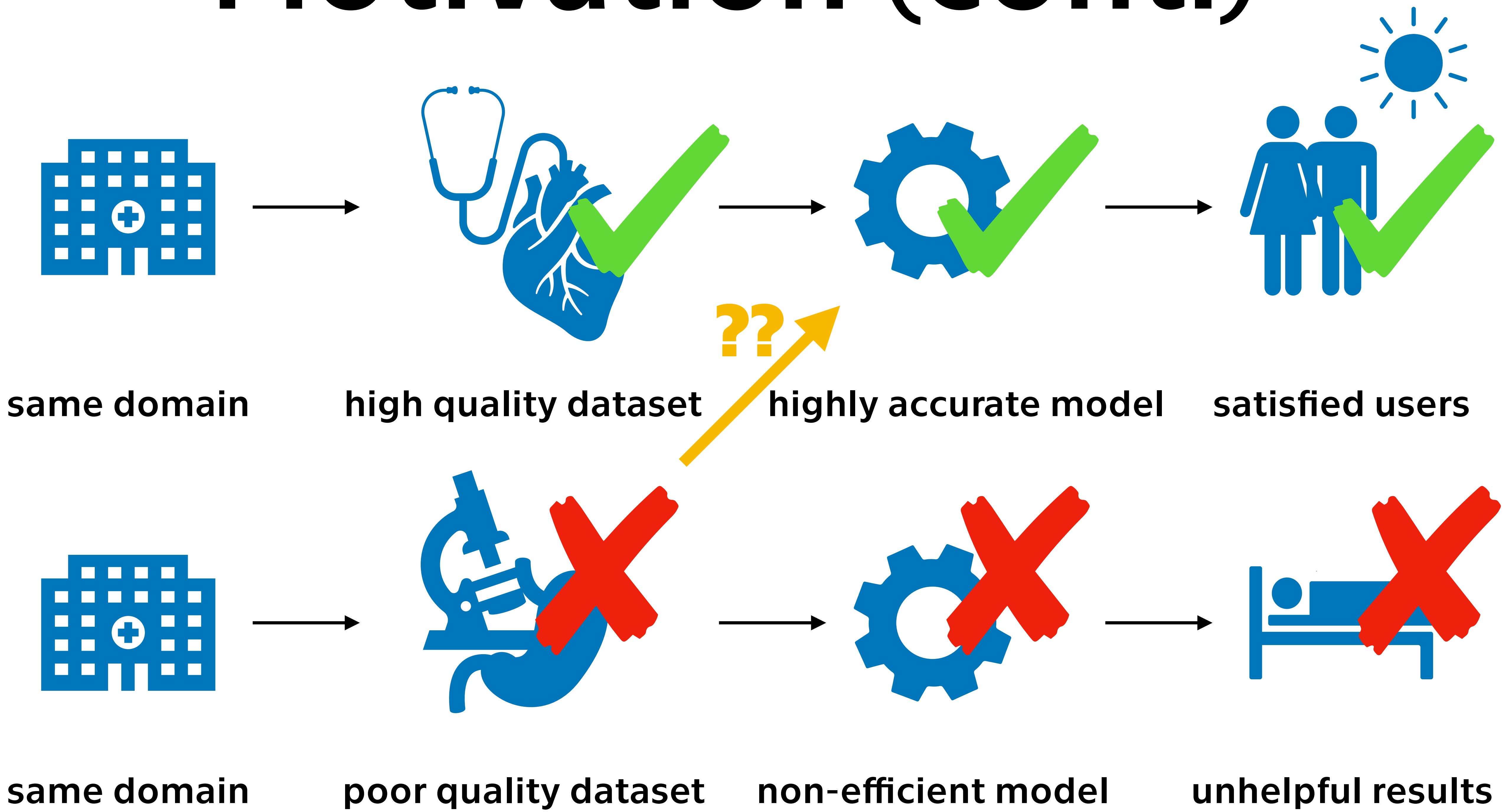
Traditional Machine Learning versus Transfer Learning



Motivation

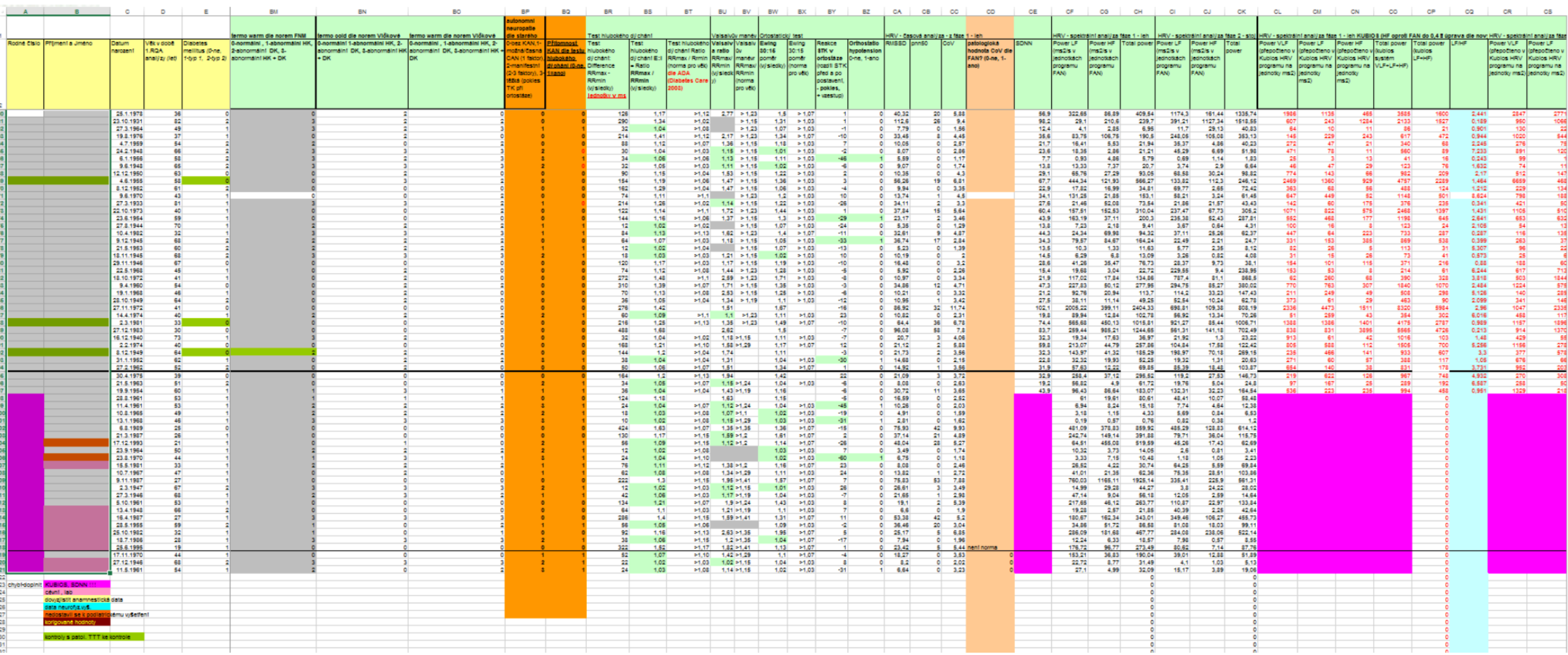

same domain
high quality dataset
highly accurate model
satisfied users
same domain
poor quality dataset
non-efficient model
unhelpful results

Motivation (cont.)



Motivation (cont.)

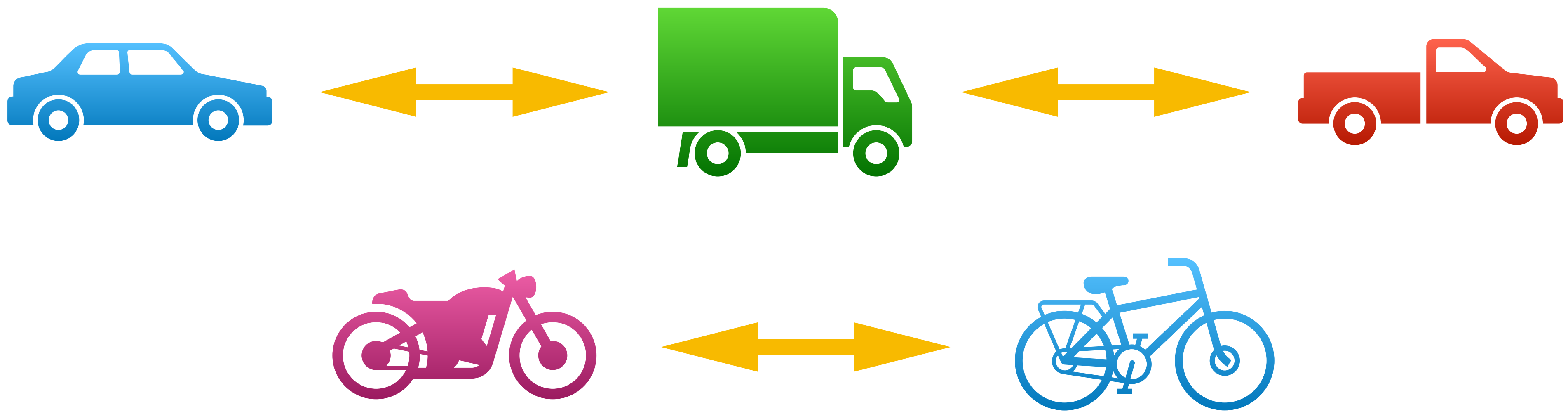
- **real dataset - cancer detection**



Transfer Learning

Transfer learning definition:

The ability of a system to recognize and apply knowledge and skills learned in previous tasks/domains to novel tasks or new domains, which share some commonalities.



Definitions

Definition 1. *A domain D is a two-tuple $(\mathcal{X}, P(X))$, where $\mathcal{X}$ is the feature space of D and $P(X)$ is the marginal probability distribution where $X = x_1, ..., x_n \in \mathcal{X}$.*

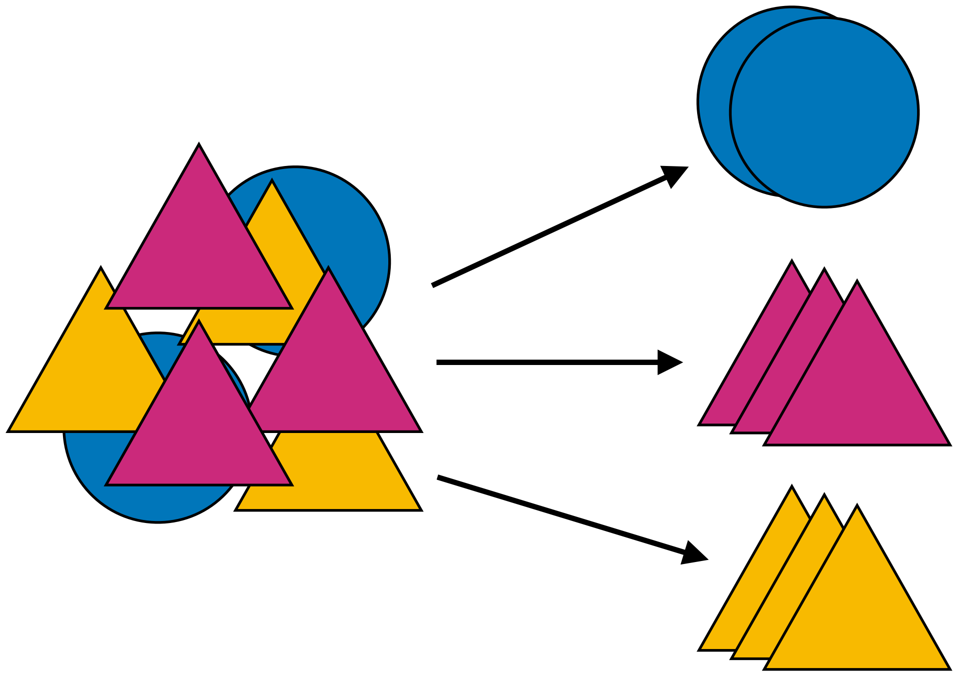
Definition 2. *A task T is a two-tuple $(Y, f())$ for some given domain D. Y is the label space of D and $f()$ is an objective predictive function for D. $f()$ is sometimes written as a conditional probability distribution $P(y|x)$.*

Definition 3. *Given a set of source domains $D_S = D_{s_1}$, ..., D_{s_n} where $n > 0$, a target domain, D_T, a set of source tasks $T_S = T_{s_1}$, ... T_{s_n} where $T_{s_i} \in T_S$ corresponds with $D_{s_i} \in D_S$, and a target task T_T which corresponds to D_T, transfer learning helps improve the learning of the target predictive function $f_T()$ in D_T where $D_T \neq D_S$ and $T_T \neq T_S$.*

Transfer Learning (cont.)

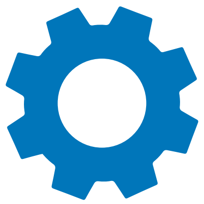
Transfer learning can be very useful in these areas:

- text classification (text clustering),
- reinforcement learning,
- sentiment classification,
- collaborative filtering,
- sensor-based location estimation,
- AI planning,
- metric learning.



Transfer Learning (cont.)

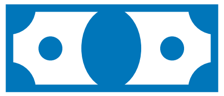
Transfer learning offers benefits:



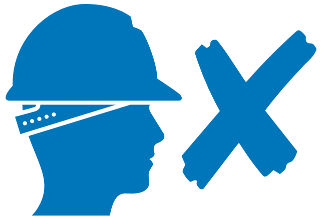
- automatization of mapping process,

- increasing usability of poor data,

- saving time,



- saving human resources,



- no presence of domain expert necessary.

Fields of Transfer Learning

1) Transfer Learning for reinforcement learning:

- Taylor, M. E. and Stone, P. (2009) - Transfer Learning for Reinforcement Learning Domains: A Survey.

2) Transfer learning for **classification and regression problems**:

- Pan, S. J. and Yang, Q. (2009) - A survey on transfer learning.

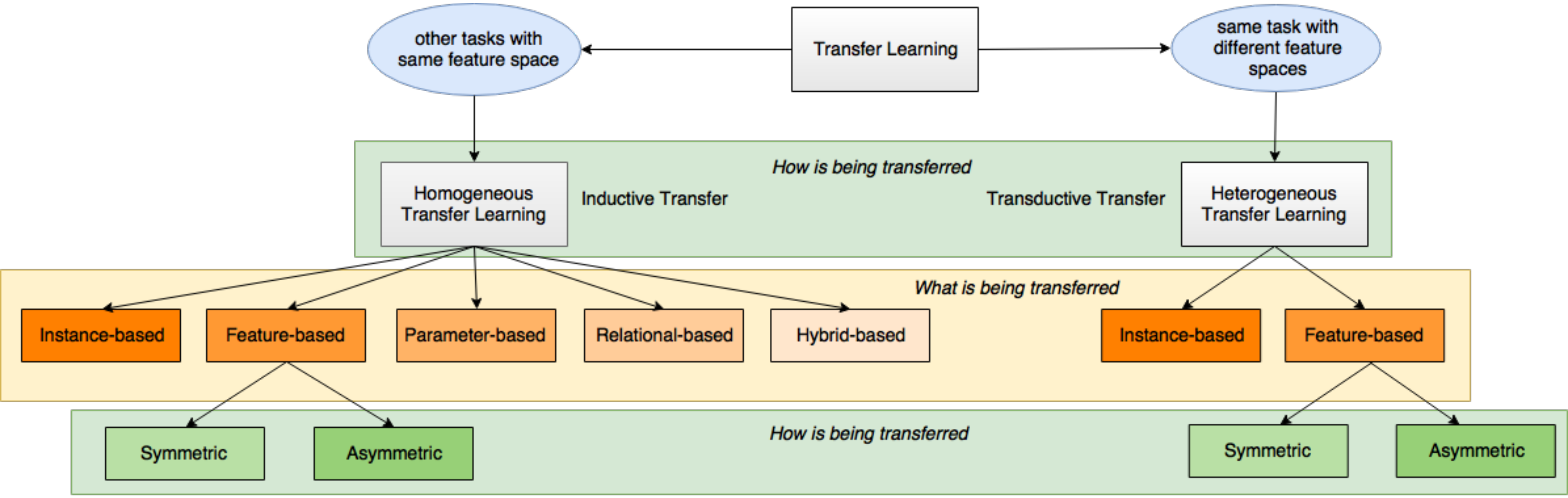
- Weiss, K., Khoshgoftaar, T. M., and Wang, D. D. (2016) - A survey of transfer learning.

Research Issues

Three main research issues:

- **What** to transfer?
  - Asks which part of knowledge can be transferred.
- **How** to transfer?
  - Defines the algorithm that extracts the knowledge.
- **When** to transfer?
  - Defines in which situations knowledge should **not** be transferred (negative transfer learning).

TL Hierarchy



What to Transfer

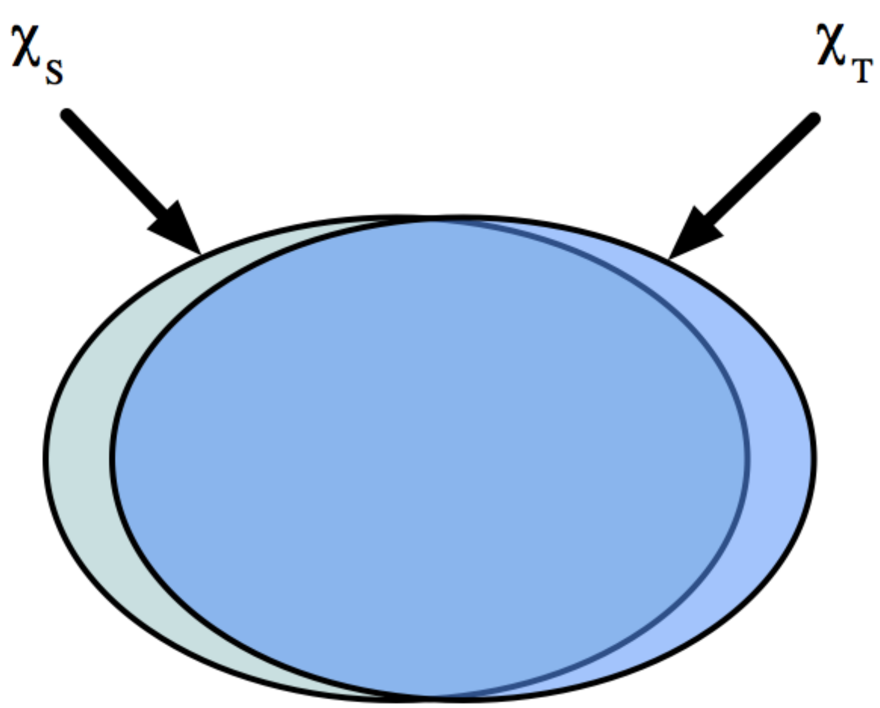
There exists four standard approaches:

- instance-based transfer learning,
- feature-based transfer learning,
- parameter-based transfer learning,
- relational-based transfer learning.

Instance-based TL

The main assumption:

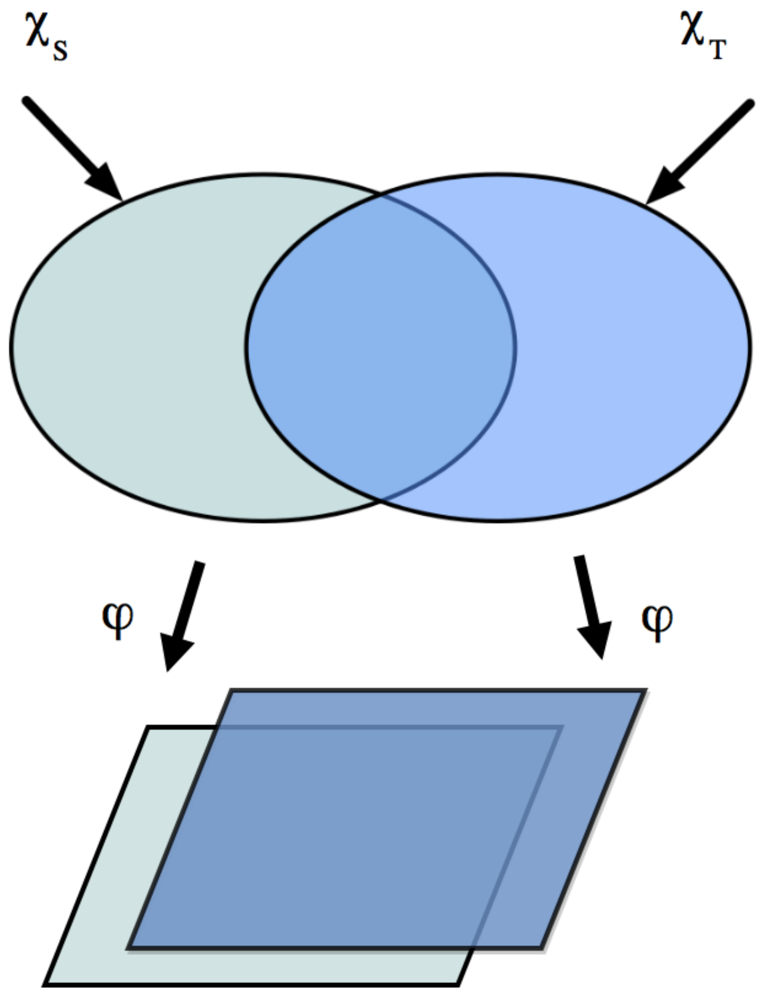
- **Source and target domains have a lot of overlapping features.**



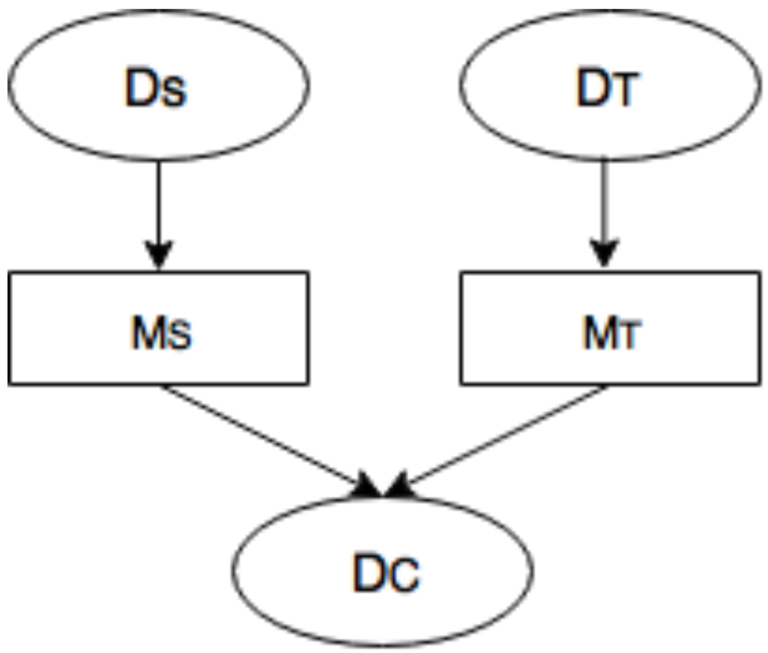
Feature-based TL

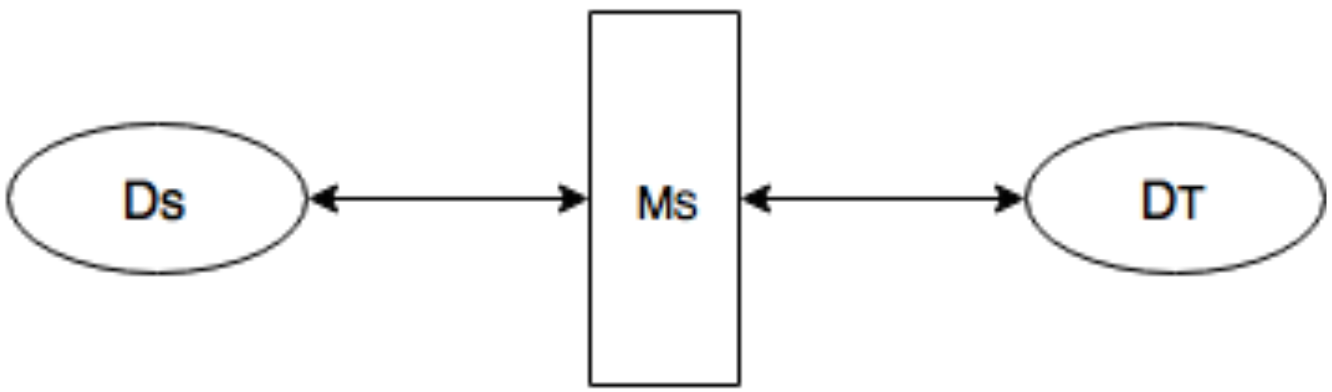
The main assumption:

- **Source and target domains only have some overlapping features.**



How to Transfer





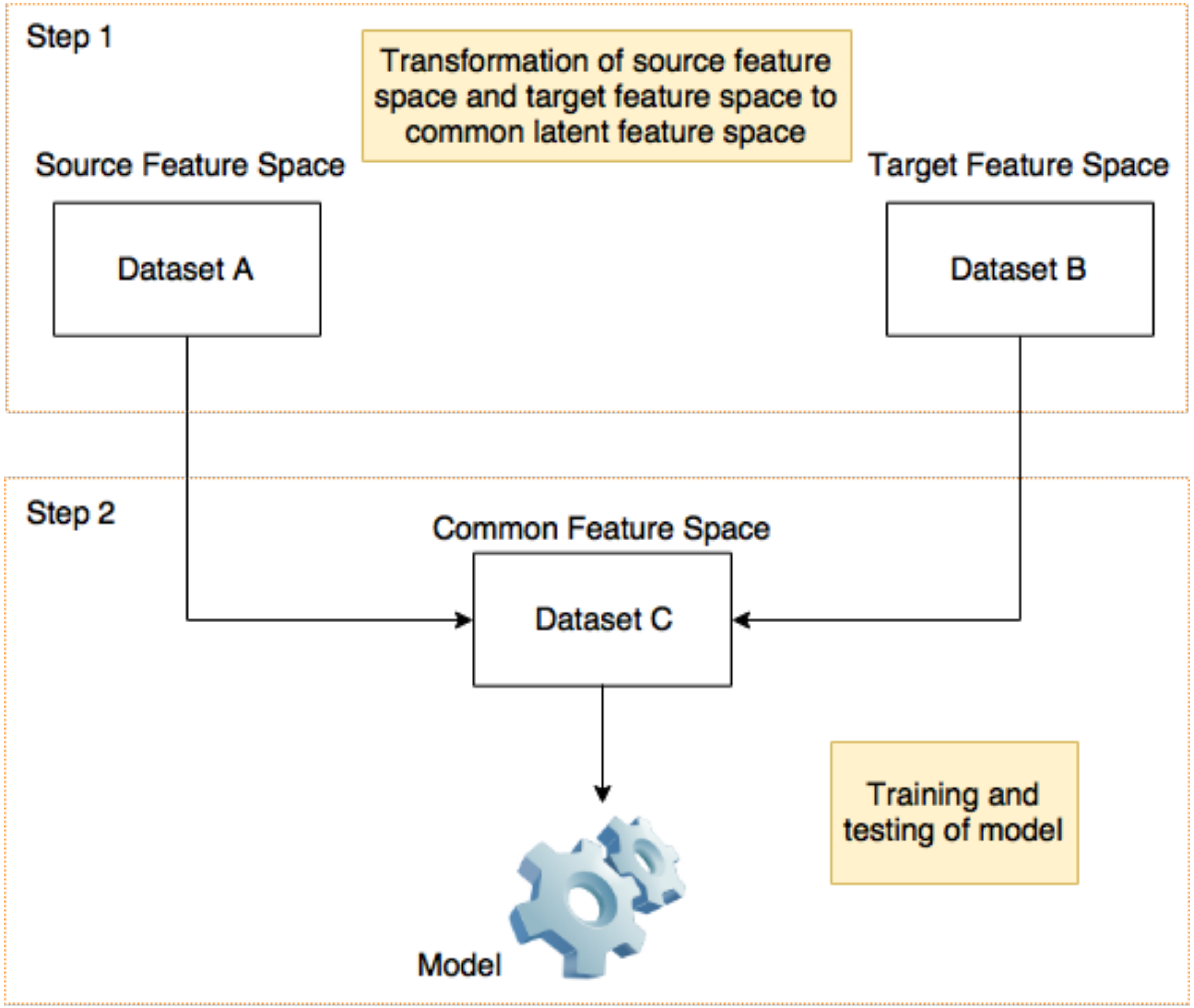
Symmetric Approach

- Source domain D_S
- Target domain D_T
- Specific transformation M for each domain (M_S and M_T)
- Latent space D_C

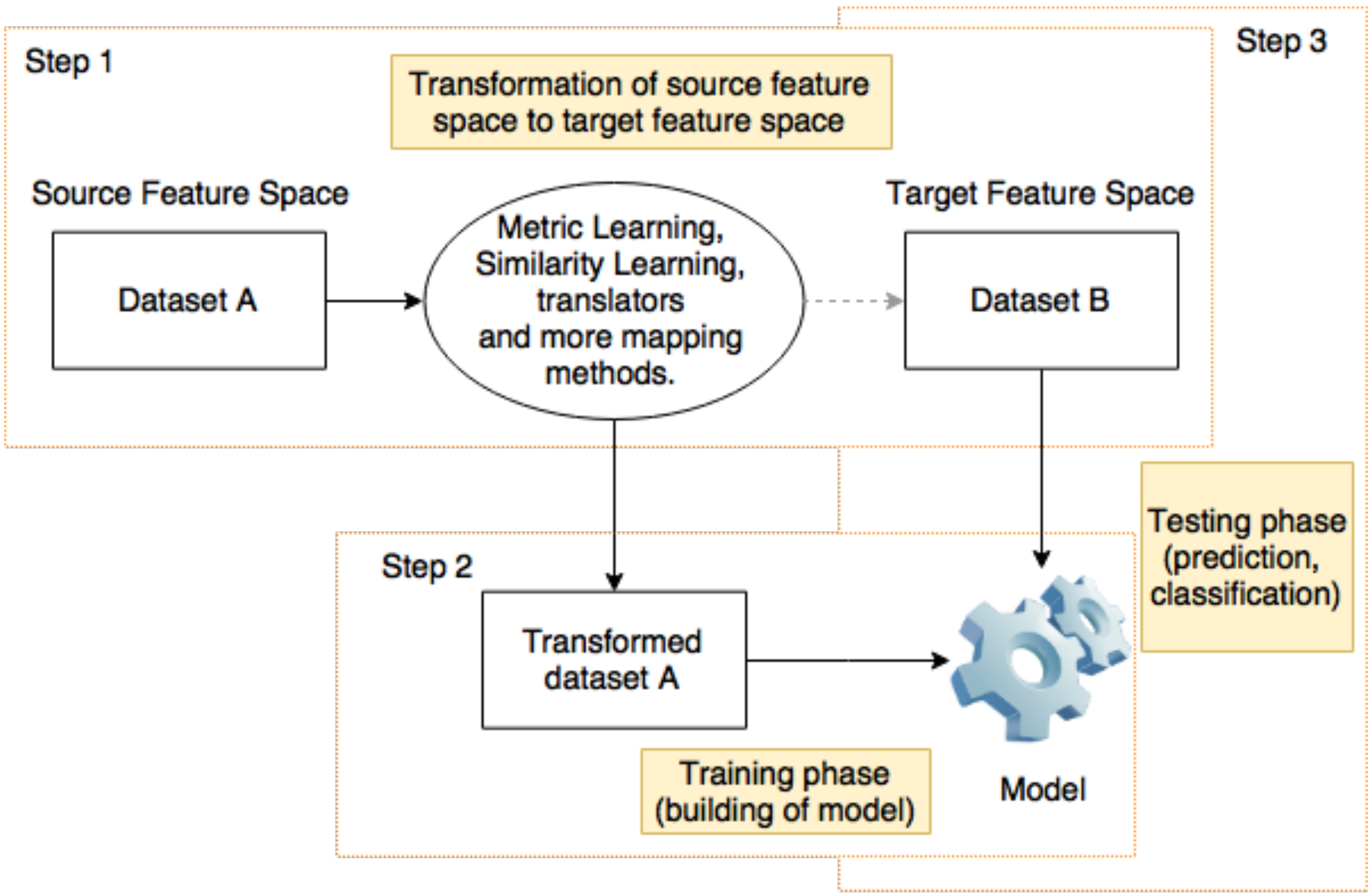
Asymmetric Approach

- Source domain D_S
- Target domain D_T
- Common transformation M

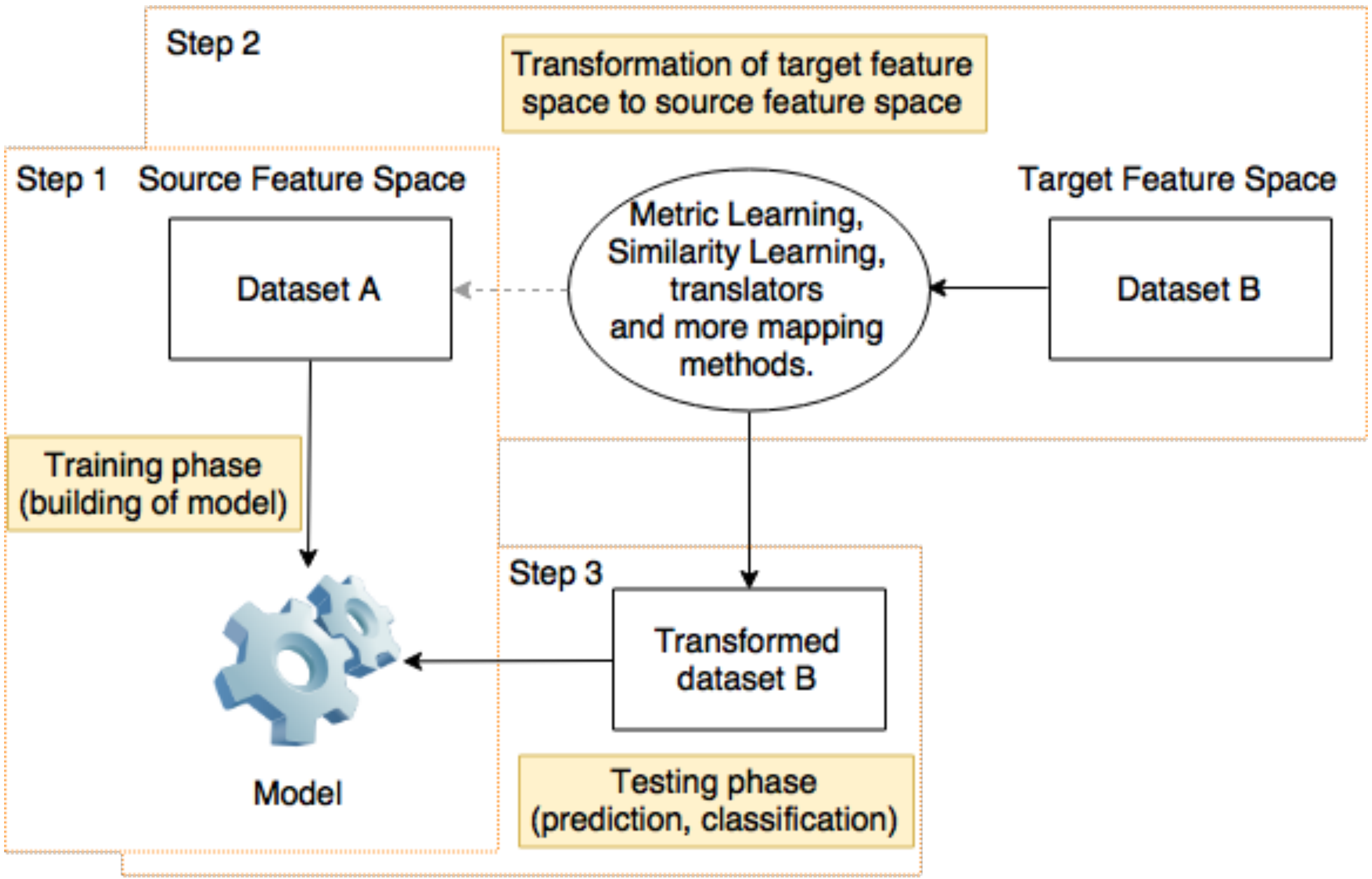
Symmetric Approach



Asymmetric Approach no. 1



Asymmetric Approach no. 2



When to Transfer

Negative Transfer Learning:

- Transfer learning methods assume that the source and target domains are related to each other in some way. If this assumption does not hold, negative transfer may happen.
- We need to first study transferability between source domains or tasks and target domains or tasks and then select relevant source domains or tasks to extract knowledge.
- Fertile area for further research.

Another Settings

1. Inductive TL
    - the target task is different from the source task
    - labeled data in the target domain are required
2. Transductive TL
    - the target domain is different from the source domains
    - no labeled data in the target domain
3. Unsupervised TL
    - focused on unsupervised learning tasks in the target domain
    - no labeled data in both source and target domains

TL Challenges

- Negative transfer learning.
  - How to avoid negative transfer?
  - Robust negative transfer measurements.
- Another area of future work pertains to the scenario where the output label space is different between domains.
- Very few transfer learning solutions addressing the scenario of unlabeled source and unlabeled target data.
- Improved methods for correcting the conditional distribution differences.
- A lack of general solutions (mainly domain specific solutions) (mainly in heterogeneous field).
- How the diversity and large size of sensor data integrates into transfer learning solutions.
- Active learning.
- Cold-start problem.

My Research Area

Asymmetric Heterogeneous Transfer Learning

(e.g. same task, different domains, asymmetric approach)

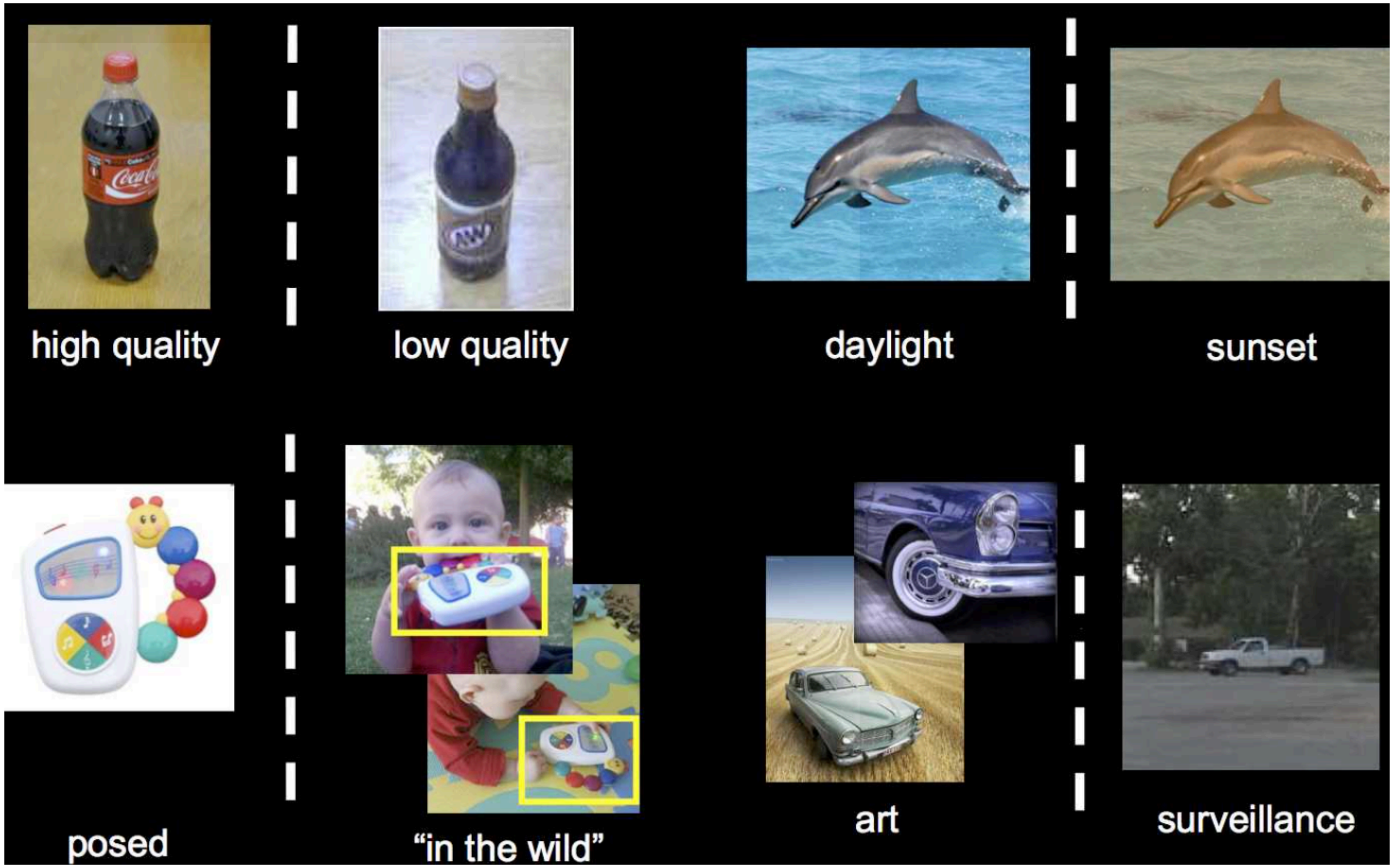
We face to these problems:

- different feature representations,
- different numbers of features,
- different meanings of features,
- a few or no overlapping features.

Definition of HetTL

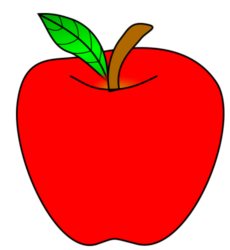
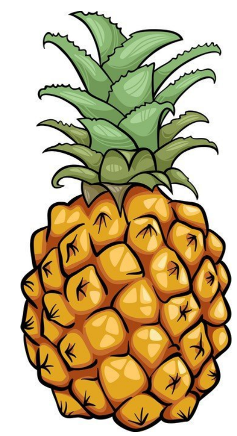
Definition 4. *Given a set of source domains* $D_S = D_{s_1}$ *, ...,* D_{s_n} *where* $n >$ *0, a target domain,* D_T*, a set of source tasks* $T_S = T_{s_1}$*, ...* T_{s_n} *where* $T_{s_i} \in T_S$ *corresponds with* $D_{s_i} \in D_S$*, and a target task* T_T *which corresponds to* D_T*, transfer learning helps improve the learning of the target predictive function* $f_T()$ *in* D_T *where* $\chi_T \bigcap (\chi_{s_1} \bigcup ... \chi_{s_n}) = \emptyset$.

Problems with Data Representation



[Xu, Saenko and Tsang, Domain Transfer Tutorial (2012)]

Problems with Feature Representation

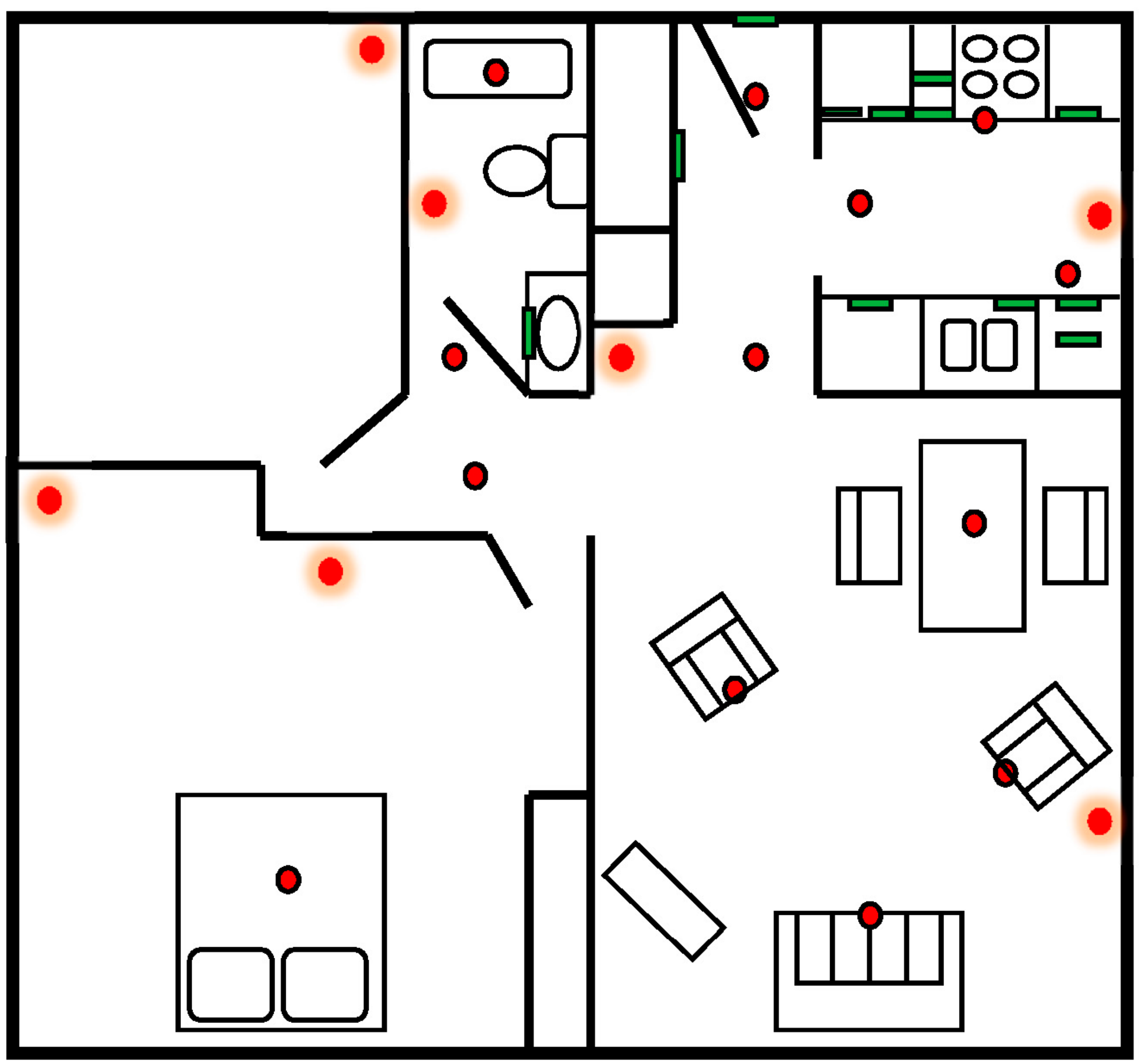
	Heterogeneous TL	TL across diff. distributions	Traditional Machine Learning
S	Apple is red or green. … Pineapple is an yellow exotic fruit. …	 	 
T	 	 	 

Fields of Application

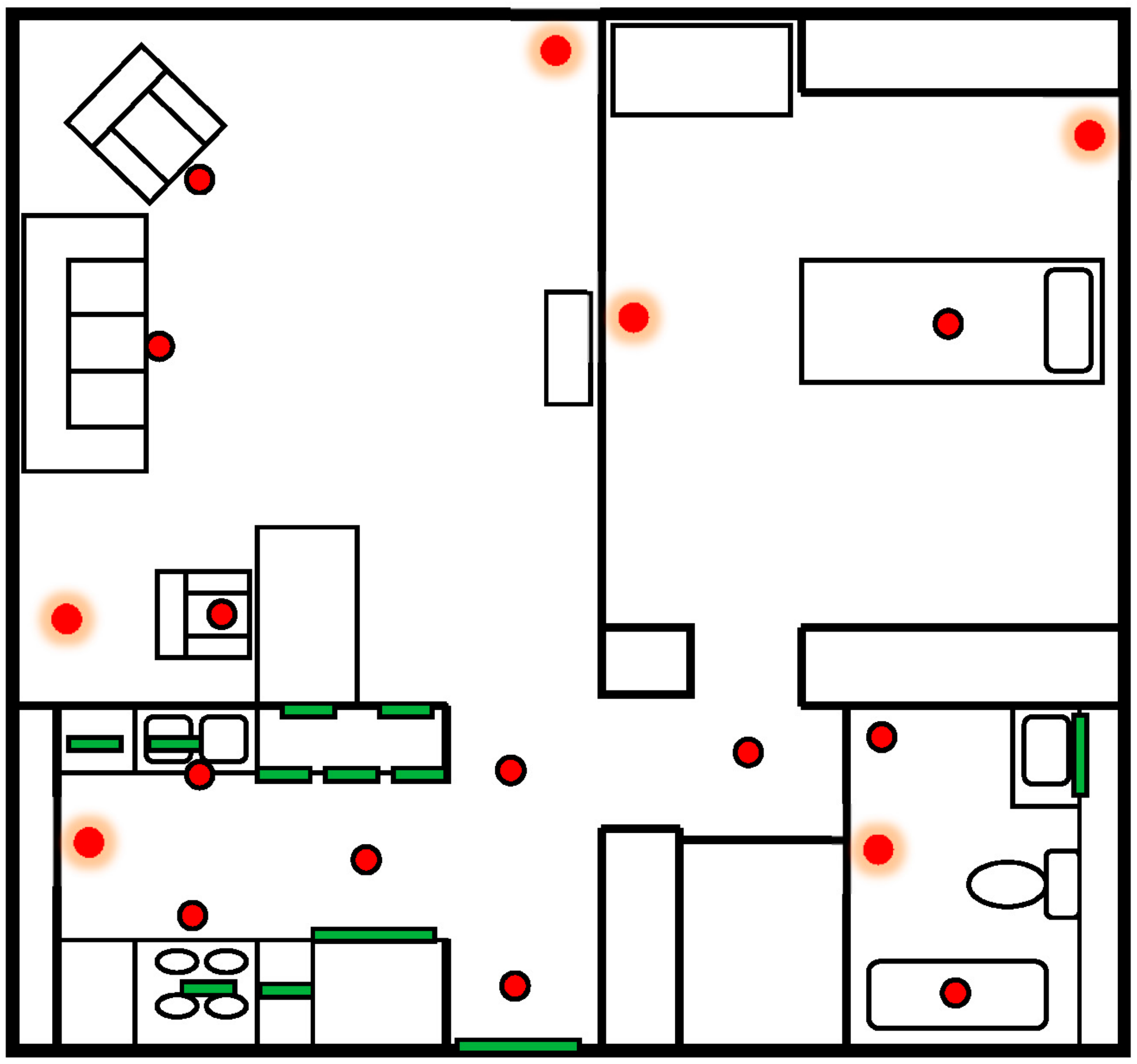
- Computer vision - image classification,
- cross-language classification,
- cross-project defect prediction,
- activity recognition.

[Kulis, B., Saenko, K., and Darrell, T. (2011)]

Activity Recognition Task



(a)



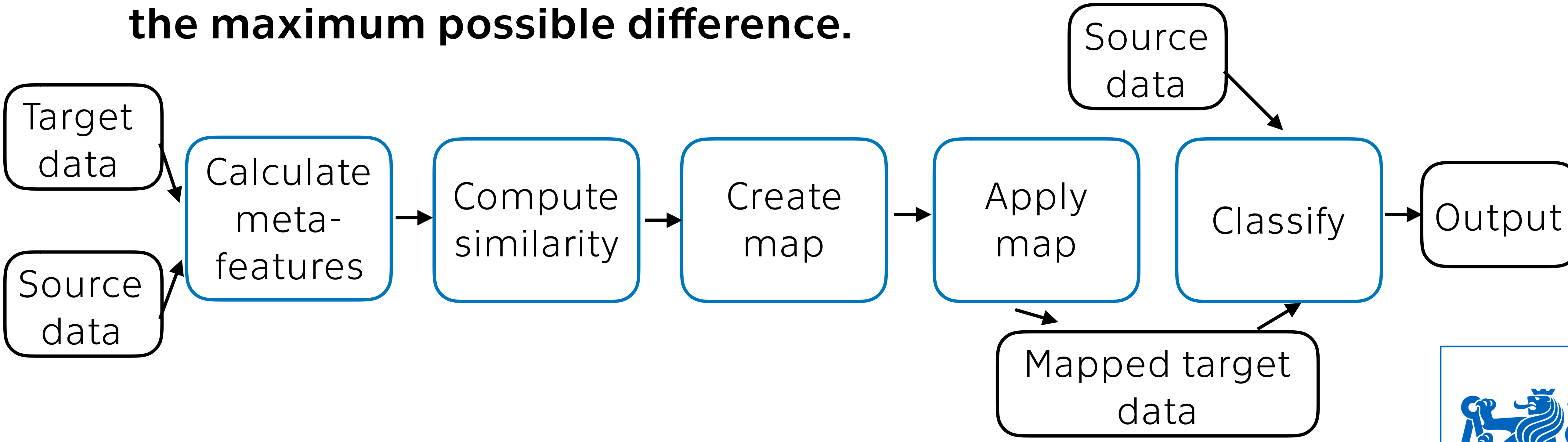
(b)

[Dahmen et al. (2017)]

Activity Recognition Task (cont.)

Feature-Space Remapping method (Feuz and Cook, 2015):

- **Requires a one time manual specification of meta-features.**
- **Computes the average similarity between the source and target meta-features.**
- **The similarity is the absolute value of the difference divided by the maximum possible difference.**

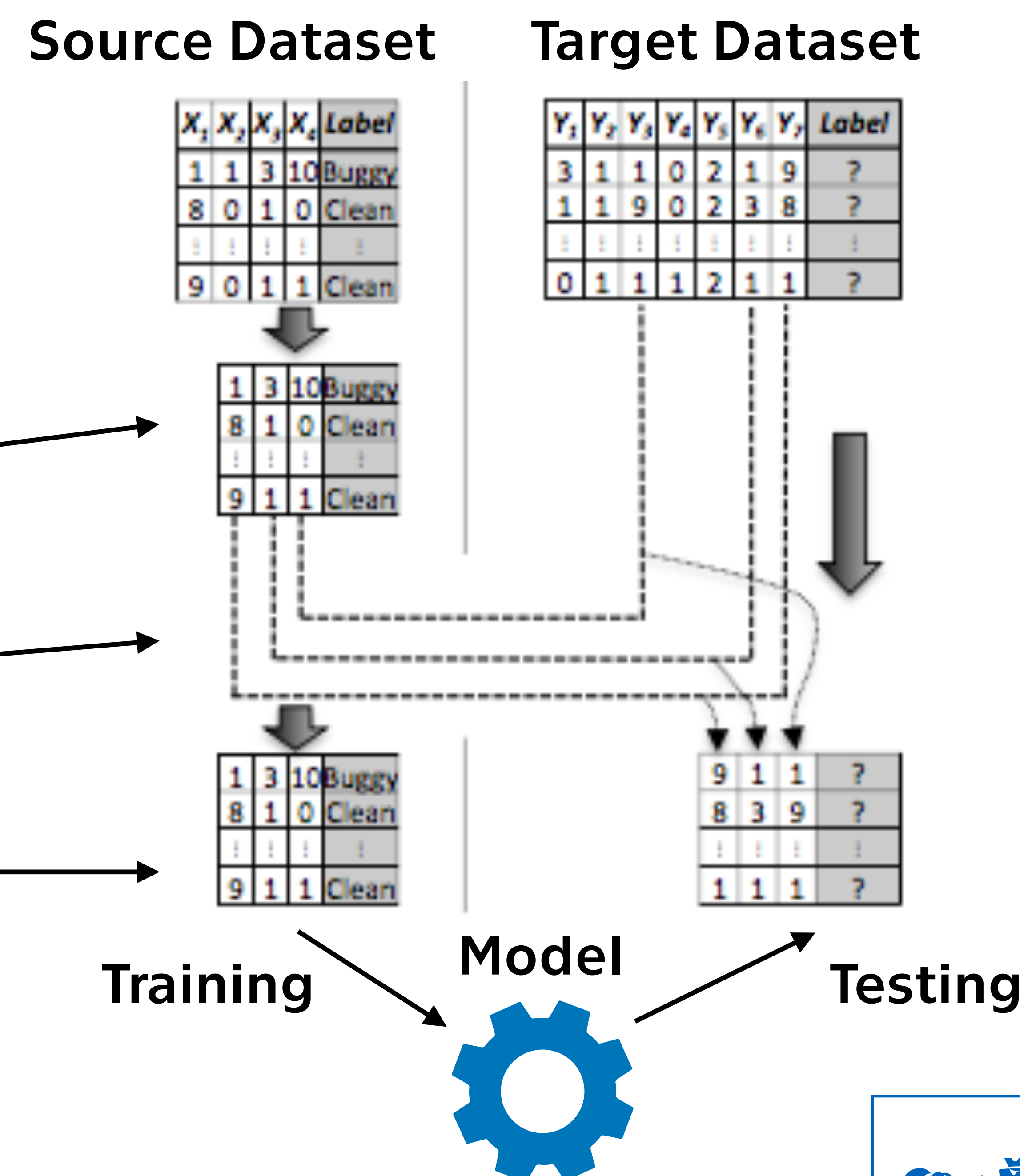


Cross-Project Defect Prediction

Heterogeneous Defect Prediction by Nam and Kim (2015):

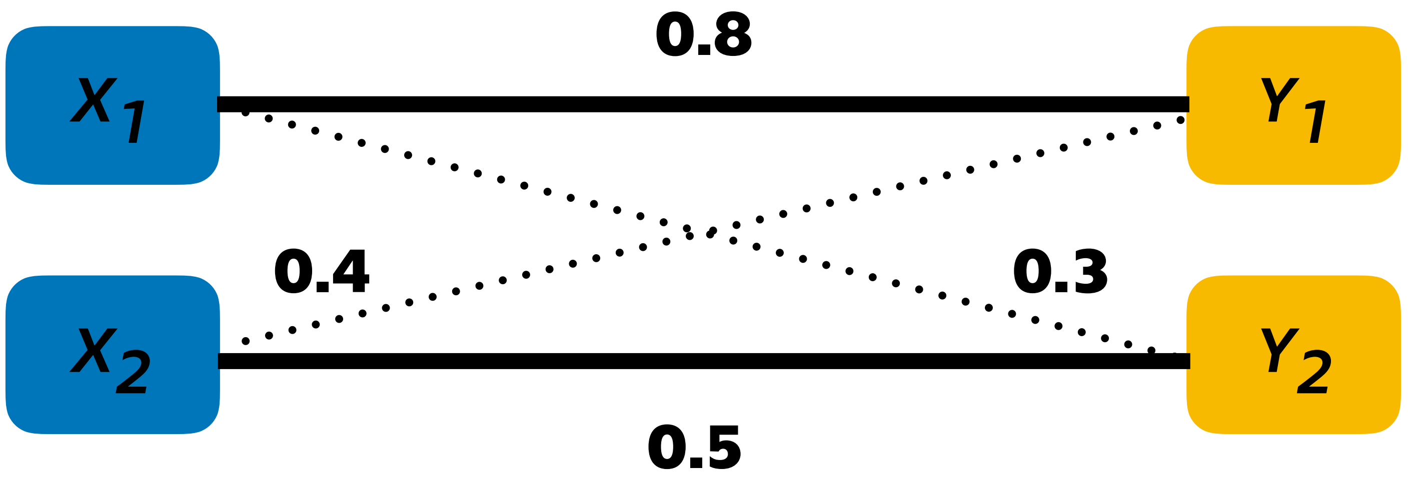
HDP phases:

1. **Metric Selection**
2. **Metric Matching**
3. **Prediction**



Cross-Project Defect Prediction- HDP (cont.)

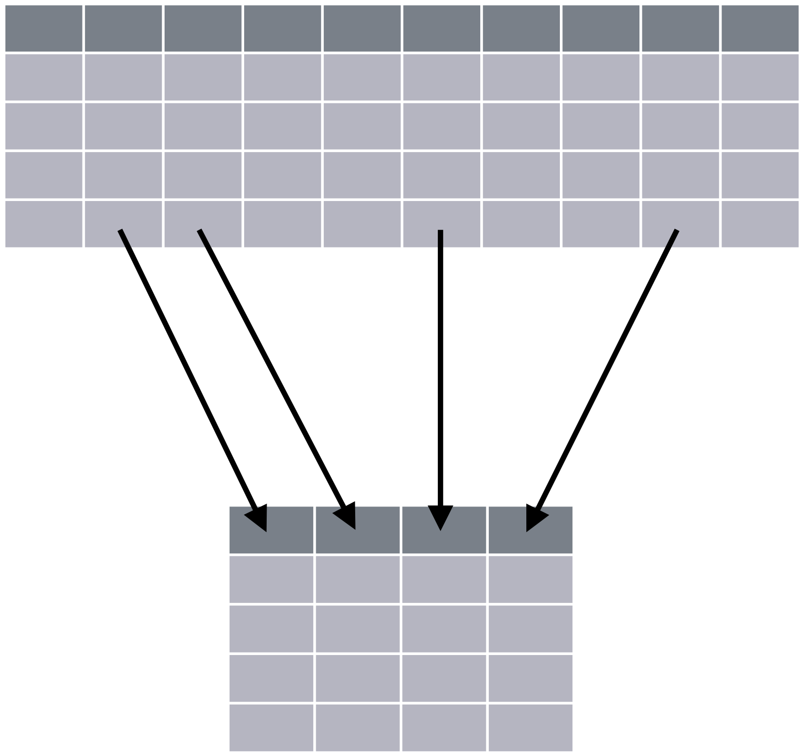
- Used methods for measuring the source and target data similarity:
  - percentiles,
  - **Kolmogorov-Smirnov Test,**
  - Spearman's correlation.



Feature Mapping

State-of-art of feature mapping:

- **preprocessing,**
- **dimensionality reduction,**
- **feature selection.**



Mappings used within transfer learning:

- **statistic methods,**
- **metric methods.**

Actual Work

- Can we use a dataset with a few missing features for classification while preserving model accuracy? If we use such a dataset, how will the model perform?

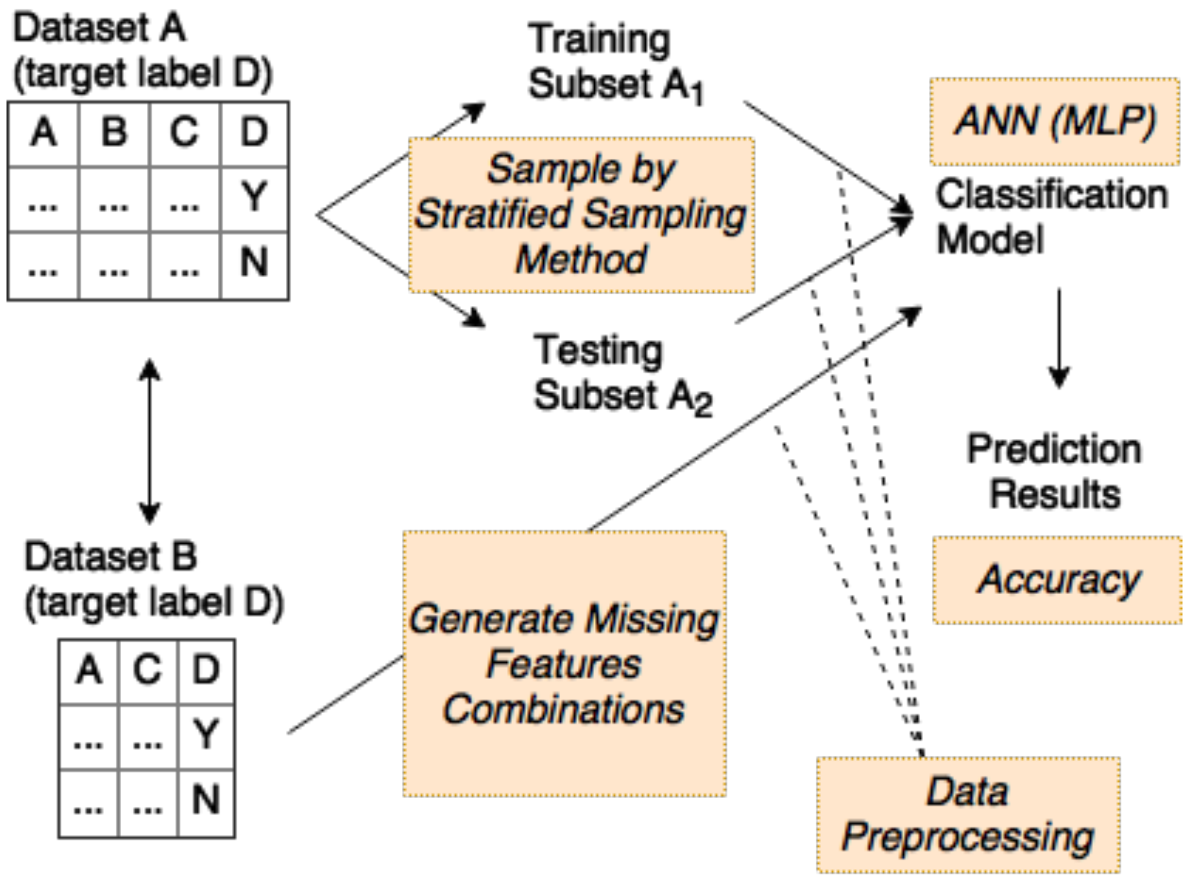
- Is it possible to define a subset of features which are crucial for model accuracy? Can this subset be replaced by a subset of other features?

- Does a method which would determine whether a damaged dataset is usable for model classification without loss of model accuracy exist?

Actual Work (cont.)

- Definition of feature importance and its influence on model performance →Missing Feature Combination method (MFC).

- Count of feature combinations $$C(A_s) = \sum_{r=1}^{q-1} \frac{q!}{(r!(q-r)!)}$$



Actual Work (cont.)

- **MFC method output:**

	0	1	2	3	4	5	6	7	8	9	10	11	12
10	0,71	1,00	0,48	1,00	1,00	1,00	0,71	0,71	0,71	0,81	1,00	1,00	0,90
15	0,50	1,00	0,25	1,00	1,00	1,00	0,00	0,75	0,75	0,00	1,00	1,00	1,00
20	0,88	0,50	0,56	1,00	1,00	1,00	0,38	0,69	0,31	0,50	1,00	1,00	1,00
25	0,68	0,50	0,45	1,00	1,00	1,00	0,36	0,32	0,55	0,27	1,00	1,00	0,82
30	0,41	0,38	0,48	1,00	0,86	1,00	0,17	0,21	0,66	0,45	1,00	1,00	1,00
35	0,89	0,28	0,56	1,00	0,83	1,00	0,39	0,39	0,28	0,28	1,00	1,00	0,56
40	0,46	0,76	0,51	1,00	0,68	1,00	0,49	0,44	0,44	0,15	1,00	0,95	0,51
45	0,63	0,68	0,44	1,00	0,80	1,00	0,29	0,59	0,49	0,17	0,95	0,66	0,61
50	0,52	0,63	0,46	1,00	0,62	0,89	0,23	0,34	0,40	0,23	0,97	0,82	0,57
55	0,58	0,55	0,53	1,00	0,71	0,71	0,19	0,57	0,44	0,13	0,96	0,66	0,58
60	0,53	0,57	0,43	0,97	0,58	0,66	0,22	0,56	0,40	0,30	0,93	0,75	0,54
65	0,49	0,55	0,50	0,90	0,53	0,67	0,21	0,50	0,45	0,25	0,89	0,70	0,57
70	0,53	0,54	0,48	0,86	0,55	0,56	0,25	0,56	0,48	0,29	0,86	0,61	0,60
75	0,54	0,54	0,49	0,72	0,57	0,60	0,35	0,55	0,46	0,34	0,79	0,60	0,62
80	0,50	0,51	0,49	0,61	0,48	0,50	0,33	0,52	0,42	0,38	0,65	0,50	0,57
85	0,48	0,48	0,51	0,35	0,47	0,44	0,62	0,48	0,53	0,60	0,34	0,44	0,44

Thank you!

References

- Bellet, A., Habrard, A., and Sebban, M. (2013). A survey on metric learning for feature vectors and structured data.
- Blitzer, S. J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics.
- Dahmen, J., Thomas, B. L., Cook, D. J., and Wang, X. (2017). Activity Learning as a Foundation for Security Monitoring in Smart Homes. Sensors 2017.
- Feuz, K. D. and Cook, D. J. (2015). Transfer learning across feature-rich heterogeneous feature spaces via feature-space remapping (FSR). In ACM Transactions on Intelligent Systems and Technology. ACM.
- Kulis, B., Saenko, K., and Darrell, T. (2011). What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE.

References (cont.)

- Nam, J. and Kim, S. (2015). Heterogeneous defect prediction. In Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering. ACM.
- Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. In IEEE Transactions on knowledge and data engineering. IEEE.
- Taylor, M. E. and Stone, P. (2009). Transfer Learning for Reinforcement Learning Domains: A Survey. JMLR.
- Torrey L. and Shavlik J. (2009) Transfer learning. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques.
- Weiss, K., Khoshgoftaar, T. M., and Wang, D. D. (2016). A survey of transfer learning. In Journal of Big Data. Springer Berlin Heidelberg.