Machine Learning for Healthcare

6.7930, HST.956

Lecture 13: Dataset Shift

David Sontag

CSAIL

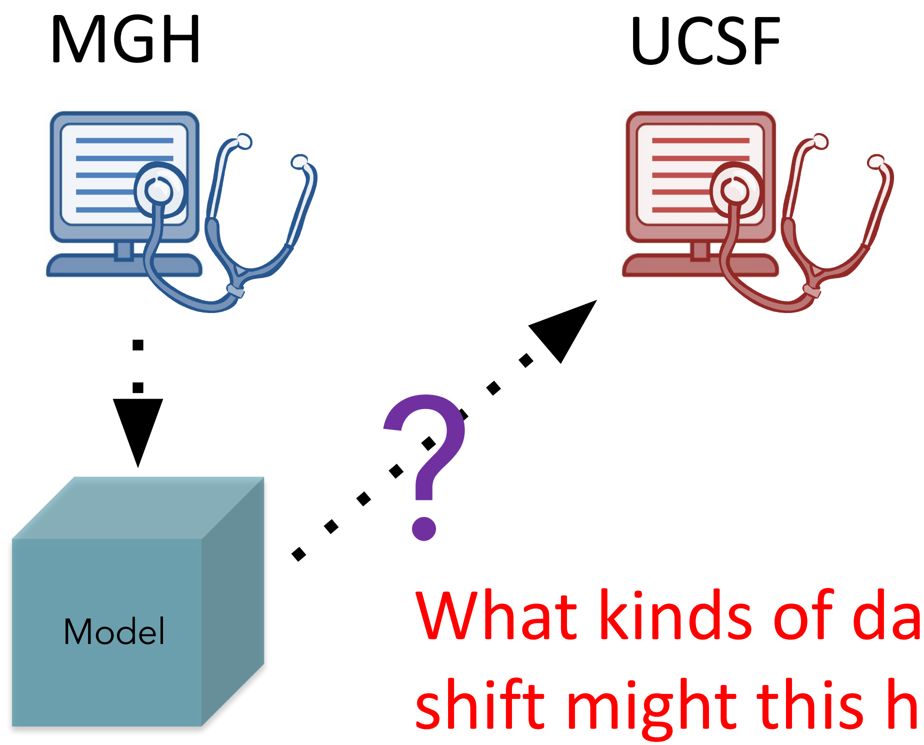
INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

HEALTH SCIENCES & TECHNOLOGY

Acknowledgement: several slides adapted from Monica Agrawal and Michael Oberst

Dataset shift / non-stationarity:

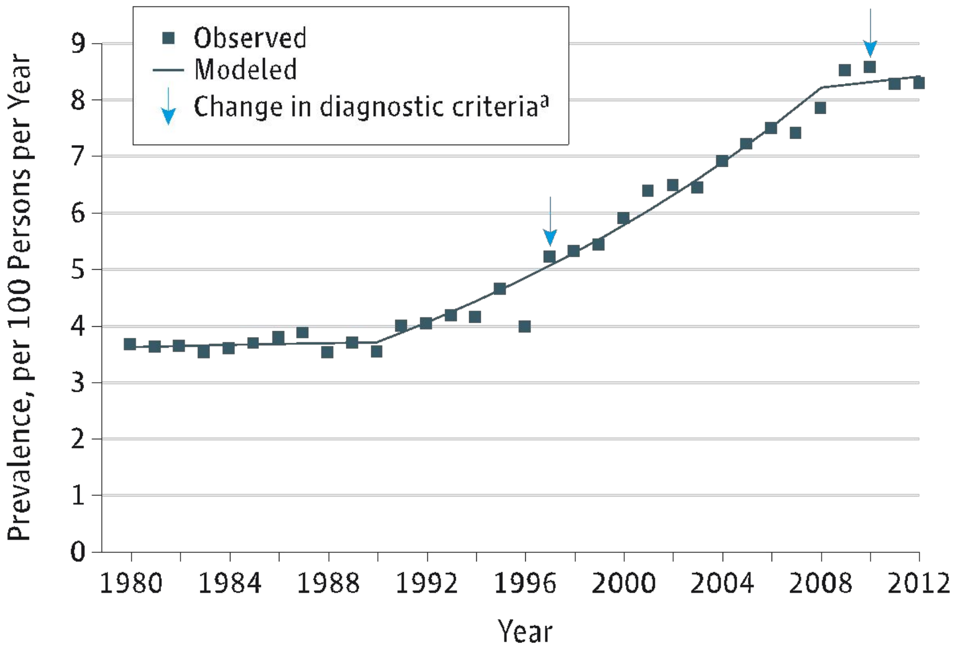
Models often do not generalize



What kinds of dataset shift might this have, and why?

[Figure adopted from Jen Gong and Tristan Naumann]

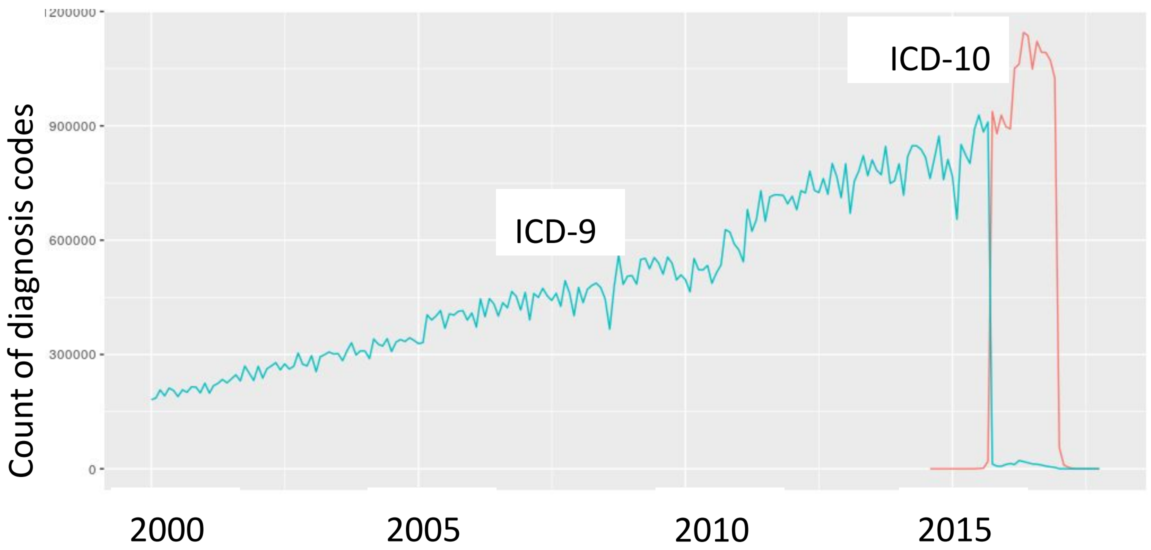
Dataset shift / non-stationarity:
Diabetes Onset After 2009

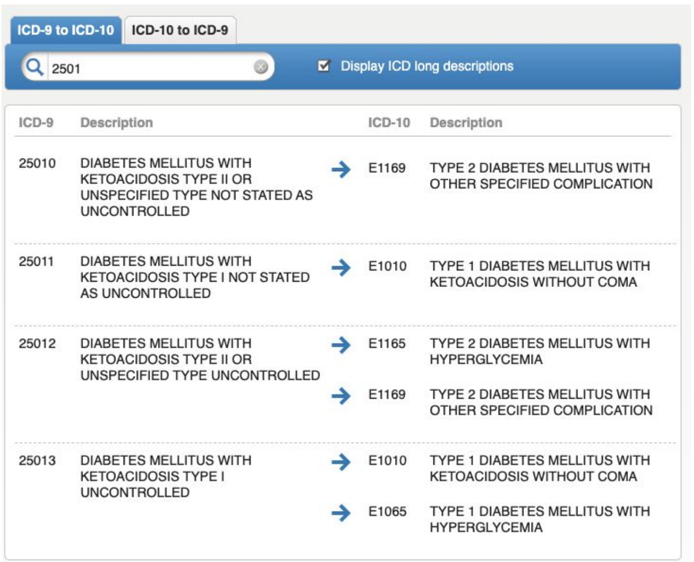


→ Automatically derived labels may change meaning

[Geiss LS, Wang J, Cheng YJ, et al. Prevalence and Incidence Trends for Diagnosed Diabetes Among Adults Aged 20 to 79 Years, United States, 1980-2012. JAMA, 2014.]

Dataset shift / non-stationarity:
ICD-9 to ICD-10 shift





→ Significance of features may change over time

(note, map from ICD10 to ICD9 isn't 1-1)

[Figure credits: (Left) Mike Oberst, (Right) http://www.icd10codesearch.com/]

“Dropsy”

- “Dropsy was a term used to describe generalized swelling and was synonymous with heart failure. Its treatment options were scanty and were aimed to cause ‘emptying of the system’ or to relieve fluid retention. These remedies were rudimentary, erratic in action, and associated with inconvenient side effects.” [J Card Fail]
- “‘Dropsy’ refers to swelling under the skin, and is generally known today as ‘oedema’ or ‘edema’” [U Leeds]
- “Dropsy is the malfunction of the digestive power in the liver” [JAMA]
- Last reported as cause of death in 1949, IIRC.

Outline for today's class

- **Examples & formalization of dataset shift**
- Testing for dataset shift
- Mitigating dataset shift

Formalizing Dataset Shift

- General Task: Perform well on a "target domain" Q
- Assumptions: What is changing vs. what is stable?
  - Covariate Shift / Label Shift / more general shifts

Formalizing Dataset Shift

- General Task: Perform well on a “target domain” Q
- Assumptions: What is changing vs. what is stable?
  - Covariate Shift / Label Shift / more general shifts

An Impossible Problem

- Given $\{X_i, Y_i\}_{i=1}^{n}$ from a *source domain* $P(X, Y)$, find a model that performs well on some *target domain* $Q(X, Y)$

$$\min_{f \in \mathcal{F}} \mathbb{E}_Q[\ell(Y, f(X))]$$

Examples:
- P and Q are two different hospital systems
- P is the past, Q is the future
- …

Not well-posed without further assumptions or information about Q!

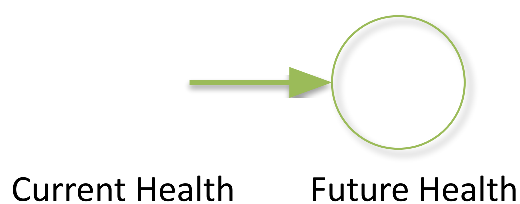
Formalizing Dataset Shift

- General Task: Perform well on a "target domain" Q
- Assumptions: What is changing vs. what is stable?
  - Covariate Shift / Label Shift / more general shifts

Example: Covariate Shift Assumption

- $P(X) \neq Q(X)$
  $P(Y \mid X) = Q(Y \mid X)$

Why might this be true? One rationale: $P(Y \mid X)$ encodes some "causal" mechanism



Example: Risk stratification for different patient populations

Example: Label Shift Assumption

- $P(Y) \neq Q(Y)$

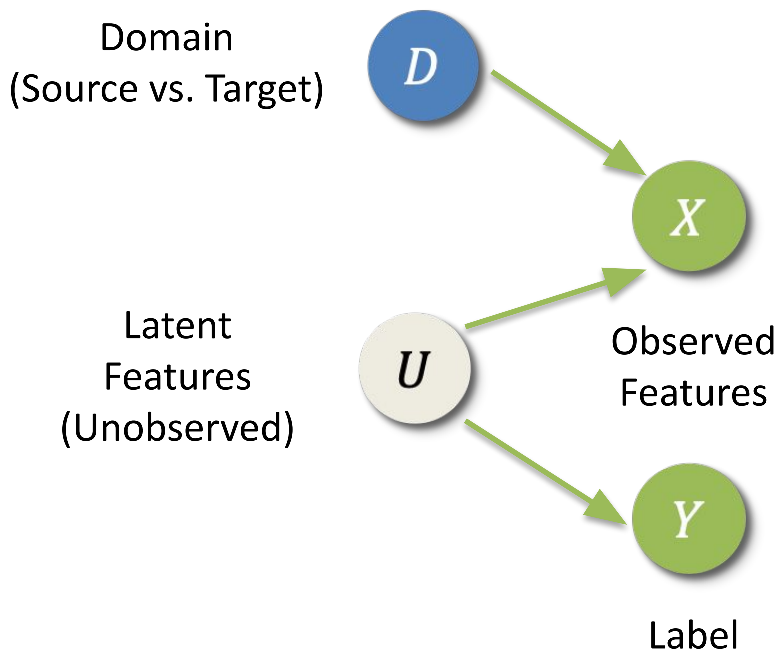
$P(X \mid Y) = Q(X \mid Y)$

Why might this be true? One rationale: $P(X \mid Y)$ encodes some "causal" mechanism

Symptoms $\longleftarrow$ Disease

Example: Diagnostic testing under changes in disease prevalence.

Example: “Domain Shift”



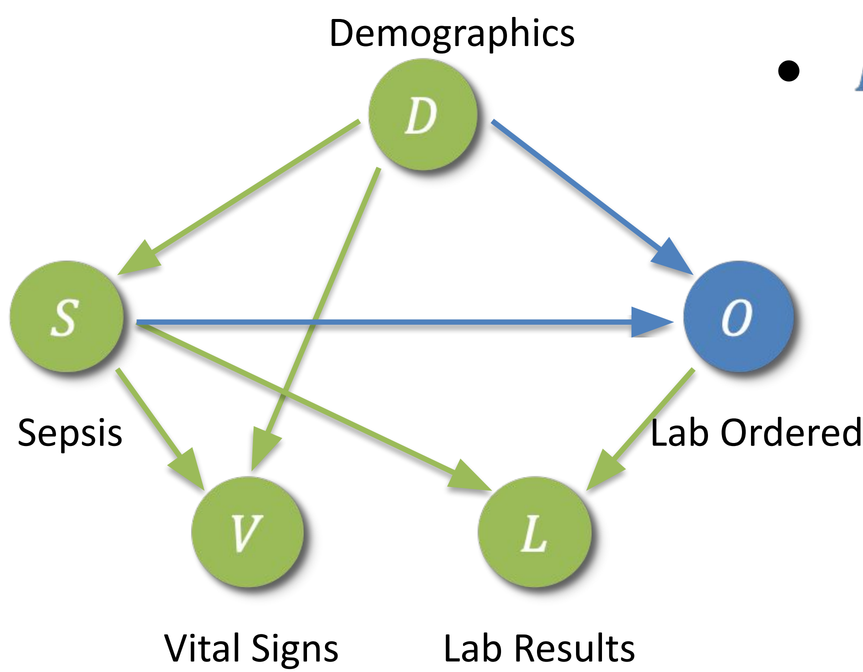
Example: Changes in how features are derived (e.g., ICD-9 versus ICD-10)

We can also view the domain itself as a variable that influences others

Note: So far, we have not discussed how to mitigate these shifts. In this example, more information is required!

Quinonero-Candela et al., (2008). *Dataset Shift in Machine Learning*, MIT Press.

Example: Using causal graphs to reason about shift



- $P(O \mid D, S) \neq Q(\mathrm{O} \mid \mathrm{D}, \mathrm{S})$

More fine-grained shifts can be reasoned about as changes in marginal/conditional distributions

Example: Changes in lab ordering patterns across hospitals

$$P(D,S,O,V,L) = P(D)P(S|D)P(V\ |D,S)P(O|D,S)P(L|O,S)$$

Example from Subbaswany et al. (2021). *Evaluating Model Robustness and Stability to Dataset Shift.* AISTATS

Outline for today's class

- Examples & formalization of dataset shift
- **Testing for dataset shift**
- Mitigating dataset shift

Testing for dataset shift

- Shift in p(y):
  - Plot distributions (across data sets, across time)

Testing for dataset shift

- Shift in p(y):
  - Plot distributions (across data sets, across time)
- Shift in p(x) or p(x|y):
  - Compare feature means

Testing for dataset shift

- Shift in p(y):
  - Plot distributions (across data sets, across time)
- Shift in p(x) or p(x|y):
  - Compare feature means (repeat for each value of Y, assuming discrete)
  - However: means can be identical even if two distributions are different!

Testing for dataset shift

- Shift in p(y):
  - Plot distributions (across data sets, across time)
- Shift in p(x) or p(x|y):
  - Compare feature means
  - Use kernel two-sample test (Gretton et al., JMLR '12)

Integral probability metric:
(Muller, 1997)

$$\mathrm{IPM}_{\mathcal{L}}(p,q) := \sup_{\ell \in \mathcal{L}} |\mathbb{E}_p[\ell(x)] - \mathbb{E}_q[\ell(x)]|$$

Testing for dataset shift

- Shift in p(y):
  - Plot distributions (across data sets, across time)
- Shift in p(x) or p(x|y):
  - Compare feature means
  - Use kernel two-sample test (Gretton et al., JMLR '12)

Integral probability metric:
(Muller, 1997)

$$\mathrm{IPM}_{\mathcal{L}}(p,q) := \sup_{\ell \in \mathcal{L}} |\mathbb{E}_p[\ell(x)] - \mathbb{E}_q[\ell(x)]|$$

Maximum mean discrepancy (MMD): L are functions with norm 1 in a RKHS:
(Gretton et al., 2012)

$$\text{samples } x_1, ..., x_m \sim p,\ x'_1, ..., x'_n \sim q$$

$$\hat{\mathrm{MMD}}_k^2(p,q) := \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, x'_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x'_i, x'_j)$$

Testing for dataset shift

- Shift in p(y):
  - Plot distributions (across data sets, across time)
- Shift in p(x) or p(x|y):
  - Compare feature means
  - Use kernel two-sample test such as maximum mean discrepancy/MMD (Gretton et al., JMLR '12)
  - (Attempt to) learn a classifier to distinguish one dataset from the other

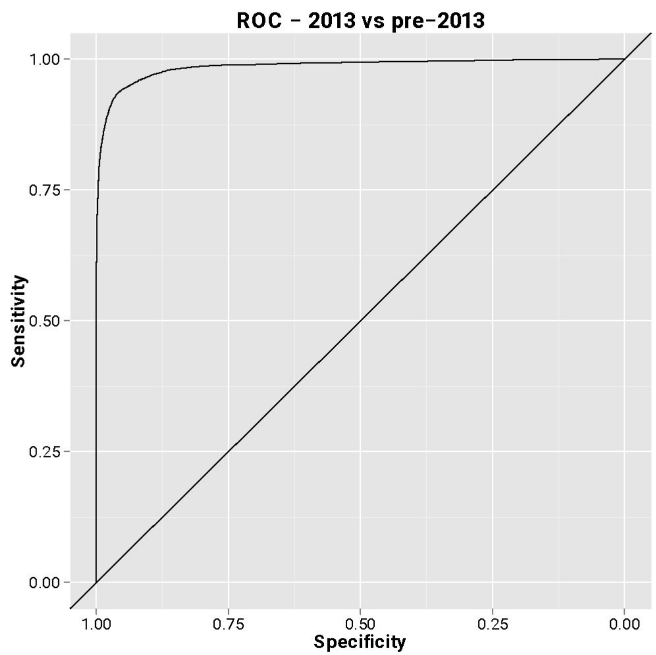
$\text{samples } x_1, ..., x_m \sim p, \; x'_1, ..., x'_n \sim q$

Binary classification (0 vs. 1)

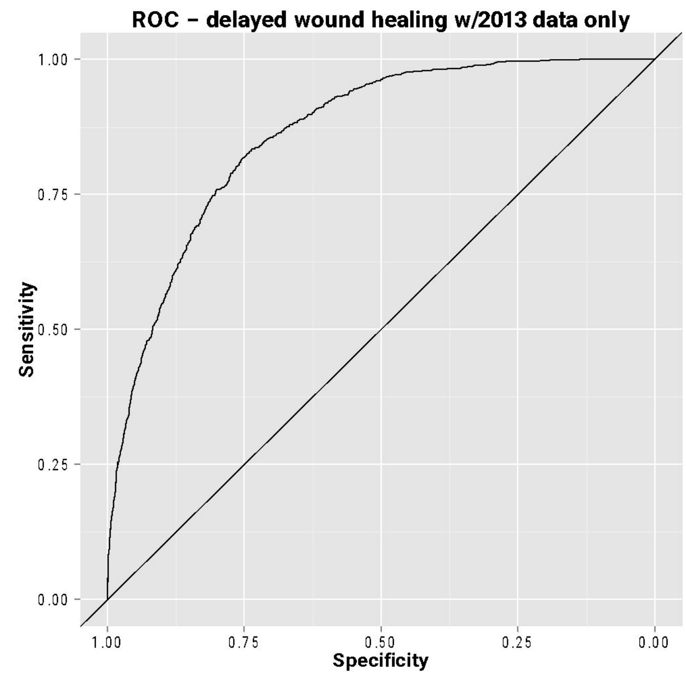
$$\mathcal{D} = \{(x_1, 1), \ldots, (x_m, 1), (x'_1, 0), \ldots, (x'_n, 0)\}$$

Testing for dataset shift

- Testing for covariate shift (wound healing):



Distinguish 2013 from pre-2013



Distinguish first 2/3 of 2013 from last 1/3 of 2013

(Figures from Ken Jung. See also Jung & Shah, Implications of non-stationarity on predictive modeling using EHRs, Journal of Biomedical Informatics, 2015)

Outline for today's class

- Examples & formalization of dataset shift
- Testing for dataset shift
- **Mitigating dataset shift**

Some practical answers

- **Domain shift** – transform features (e.g., imputation of missing values or artificially introduce noise/missingness during training, reprocess images, map to a common space), or drop features that do not transfer
- **Concept drift / non-stationarity** (eg, $p(y|x)$ changes because of new medical treatments) –
  Retrain the model with most recent data

  (Research question: how to automate the above?)

- *Covariate shift?*

Covariate shift: nonparametric regression just "works"

- When can we expect training on p(x,y) and testing on q(x,y) to give good results, for $\mathrm{p} \neq q$?

<u>Theorem:</u> If $\mathrm{p(x)} > 0$ whenever $\mathrm{q(x)} > 0$ and $p(\, y \mid x\,) = q(\, y \mid x\,)$, then in the limit of infinite data from *p*, can achieve Bayes' error on *q*

Covariate shift: nonparametric regression just "works"

- When can we expect training on p(x,y) and testing on q(x,y) to give good results, for $\mathrm{p} \neq q$?

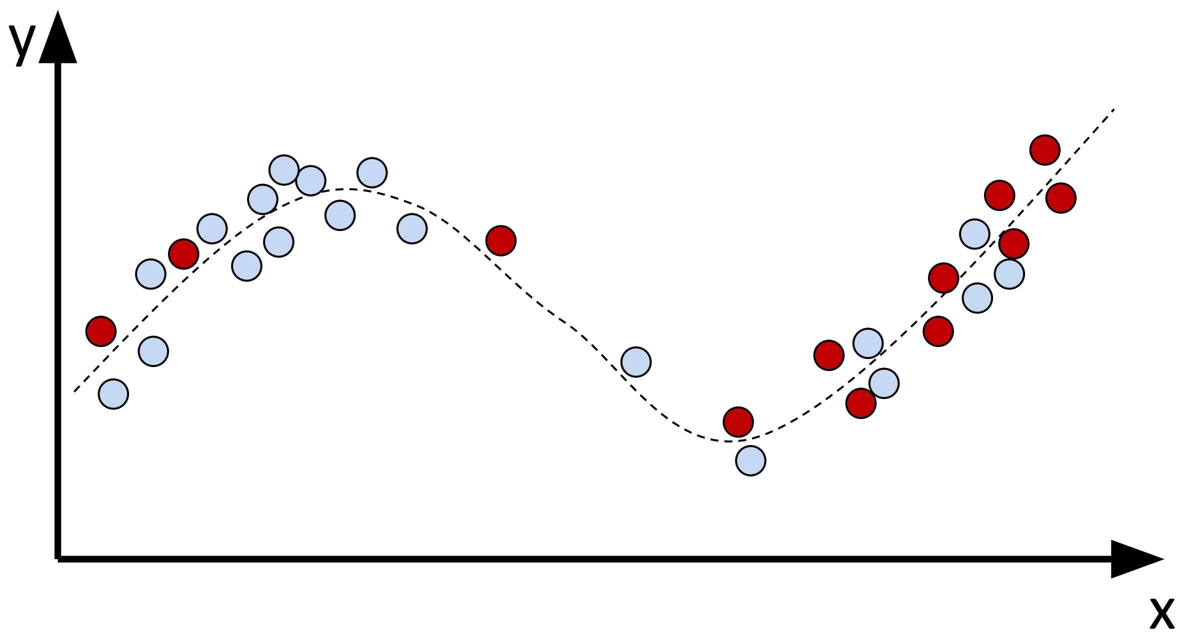
<u>Theorem:</u> If $\mathrm{p}(\mathrm{x}) > 0$ whenever $\mathrm{q}(\mathrm{x}) > 0$ and $p(y \mid x) = q(y \mid x)$, then in the limit of infinite data from *p*, can achieve Bayes' error on *q*

We never have infinite data!

May have to use a more restricted model to prevent overfitting
(e.g. a linear model despite true one being non-linear)

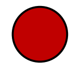
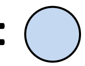
Effect of covariate shift when (naively) learning with misspecified models

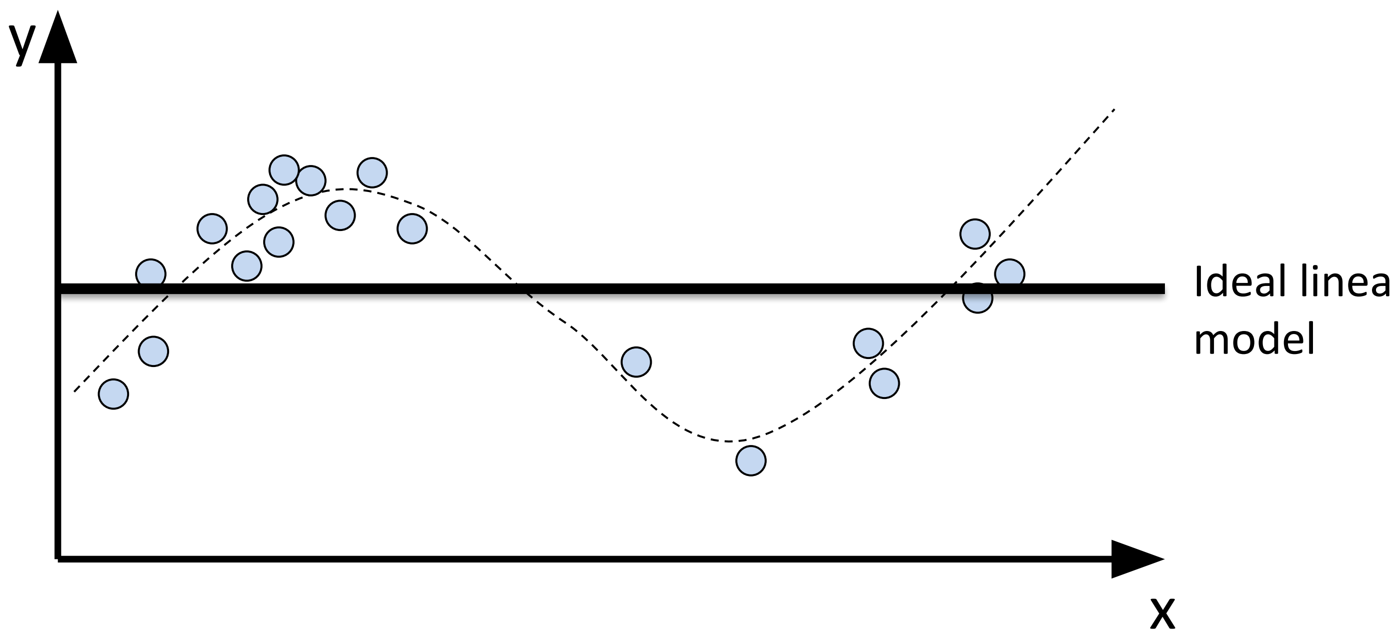
- Training data p(x,y)= and test data q(x,y)=



[Storkey, “When Training and Test Sets are Different”, Dataset in Machine Learning, MIT Press 2009]

Effect of covariate shift when (naively) learning with misspecified models

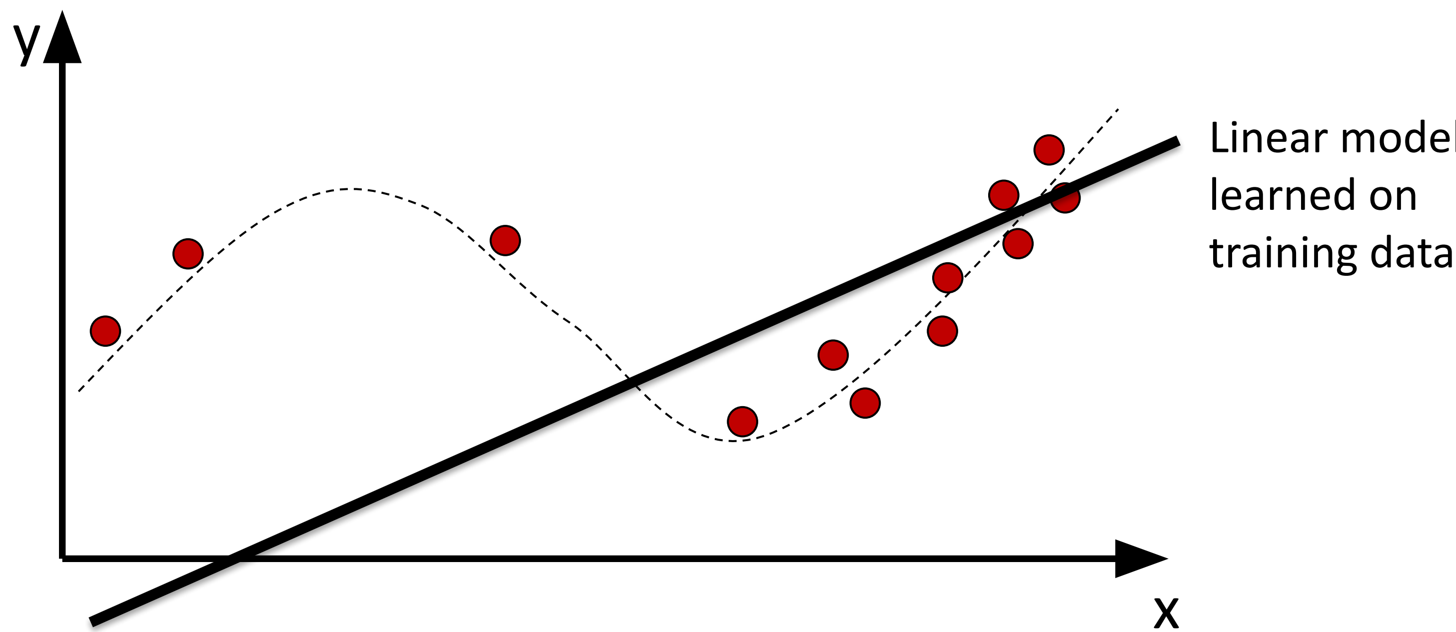
- Training data p(x,y)=  and test data q(x,y)= 



[Storkey, “When Training and Test Sets are Different”, Dataset in Machine Learning, MIT Press 2009]

Effect of covariate shift when (naively) learning with misspecified models

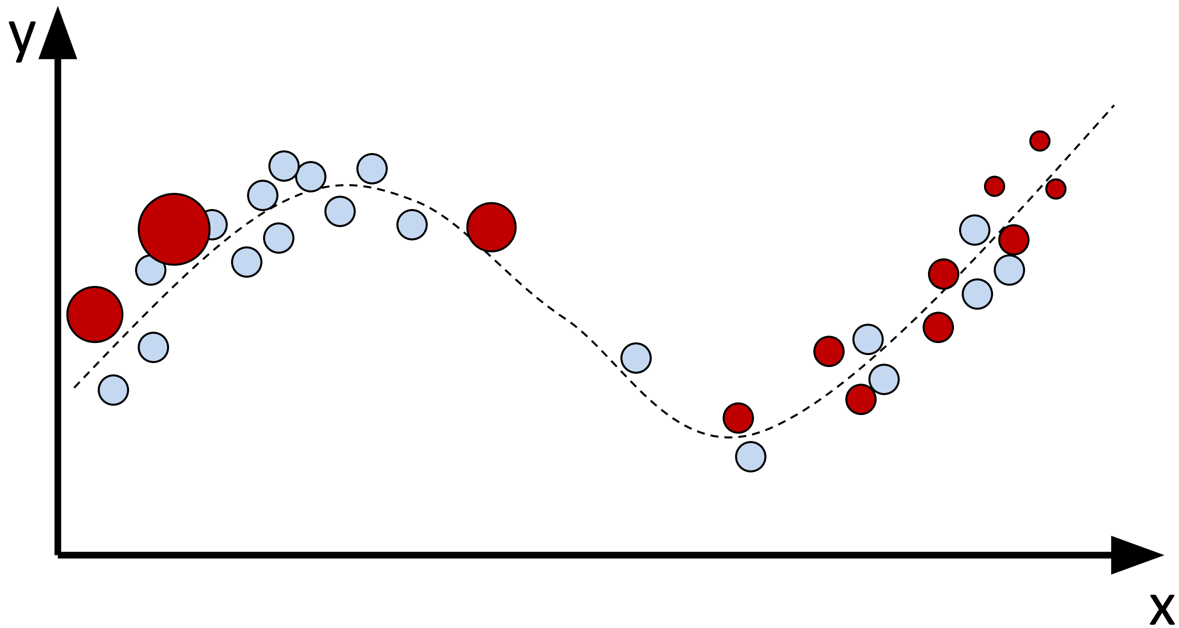
- Training data p(x,y)= and test data q(x,y)=



[Storkey, “When Training and Test Sets are Different”, Dataset in Machine Learning, MIT Press 2009]

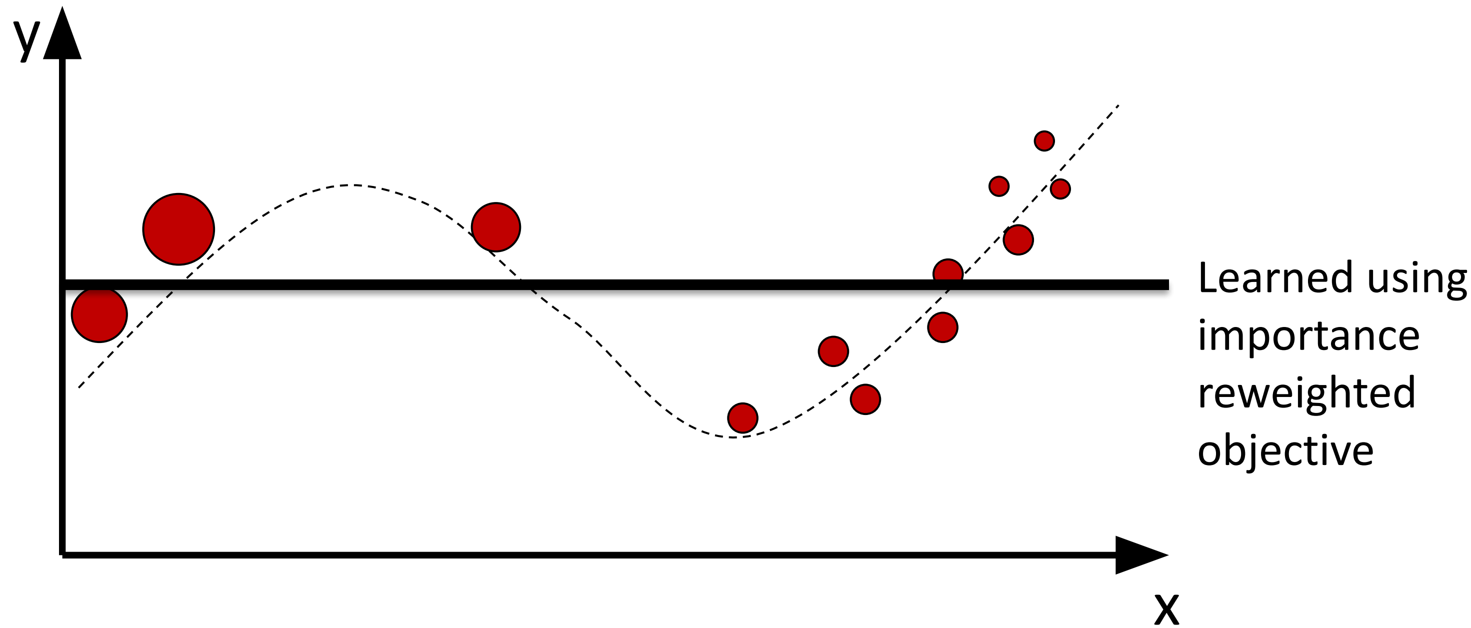
Learning using importance reweighting under covariate shift

- Training data p(x,y)= and test data q(x,y)=



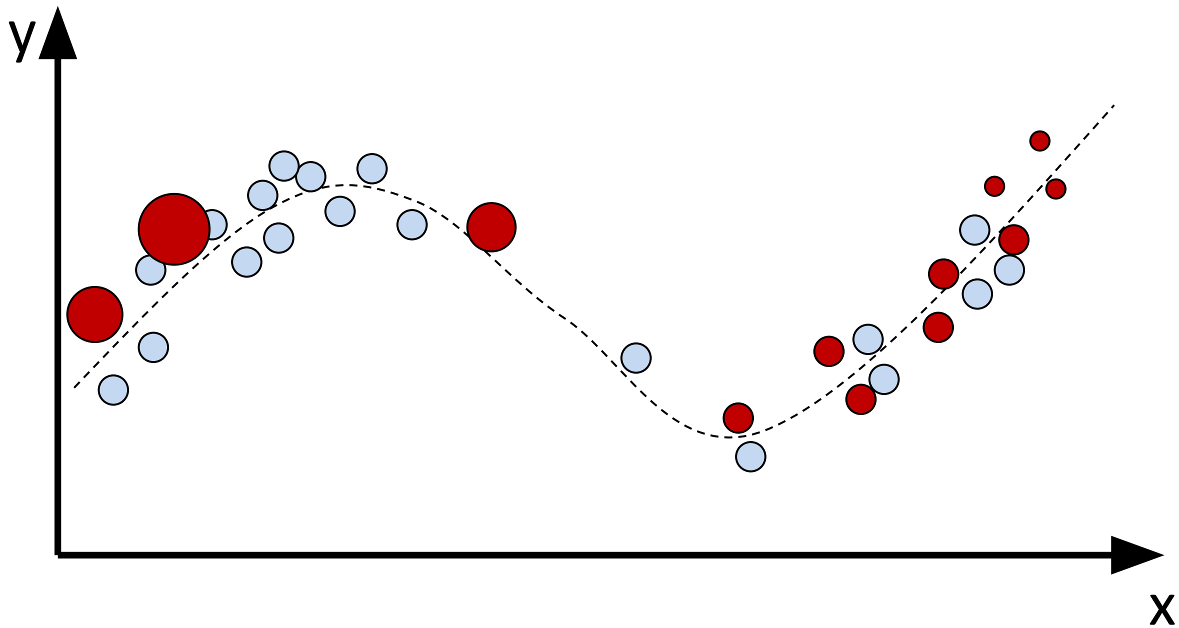
Learning using importance reweighting under covariate shift

- Training data p(x,y)= and test data q(x,y)=



Learning using importance reweighting under covariate shift

- Training data p(x,y)= and test data q(x,y)=



We only needed to know q(x) to figure out how to reweight the training data! Example of *unsupervised* domain adaptation

Learning using importance reweighting under covariate shift

Goal of learning:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim q} L(x, y; \theta)$$

Example – squared loss, linear model

$$L(x, y; \theta) = (y - \theta \cdot x)^2$$

But, suppose all we have are samples $(x_1, y_1), \ldots, (x_m, y_m) \sim p(x, y)$

Learn using: $\displaystyle \frac{1}{m} \sum_{i=1}^{m} \frac{q(x_i)}{p(x_i)} L(x_i, y_i; \theta)$

Learning using importance reweighting under covariate shift

Goal of learning:

$$\min_\theta \mathbb{E}_{(x,y)\sim q} L(x, y; \theta)$$

Example – squared loss, linear model

$$L(x, y; \theta) = (y - \theta \cdot x)^2$$

But, suppose all we have are samples $(x_1, y_1), \ldots, (x_m, y_m) \sim p(x, y)$

Learn using: $\displaystyle \frac{1}{m} \sum_{i=1}^{m} \frac{q(x_i)}{p(x_i)} L(x_i, y_i; \theta)$

How do we obtain q(x)/p(x)?

Approach 1:

Data (*x*, *d*), where *d* denotes the dataset

samples $x_1, \ldots, x_m \sim p$, $x'_1, \ldots, x'_n \sim q$ $\rightarrow$ $\mathcal{D} = \{(x_1, 1), \ldots, (x_m, 1), (x'_1, 0), \ldots, (x'_n, 0)\}$

$$\frac{q(x)}{p(x)} \leftarrow \frac{\Pr(d = 1 \mid x)}{1 - \Pr(d = 1 \mid x)} \frac{n}{m}$$

Learning using importance reweighting under covariate shift

Goal of learning:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim q} L(x,y;\theta)$$

Example – squared loss, linear model

$$L(x,y;\theta) = (y - \theta \cdot x)^2$$

But, suppose all we have are samples $(x_1, y_1), \ldots, (x_m, y_m) \sim p(x,y)$

Learn using: $\frac{1}{m}\sum_{i=1}^{m} \frac{q(x_i)}{p(x_i)} L(x_i, y_i; \theta)$

How do we obtain q(x)/p(x)?

Approach 1:

Data (*x*, *d*), where *d* denotes the dataset

$$\text{samples } x_1, \ldots, x_m \sim p,\ x'_1, \ldots, x'_n \sim q \longrightarrow \mathcal{D} = \{(x_1, 1), \ldots, (x_m, 1), (x'_1, 0), \ldots, (x'_n, 0)\}$$

$$\frac{q(x)}{p(x)} \leftarrow \frac{\Pr(d = 1 \mid x)}{1 - \Pr(d = 1 \mid x)} \frac{n}{m}$$

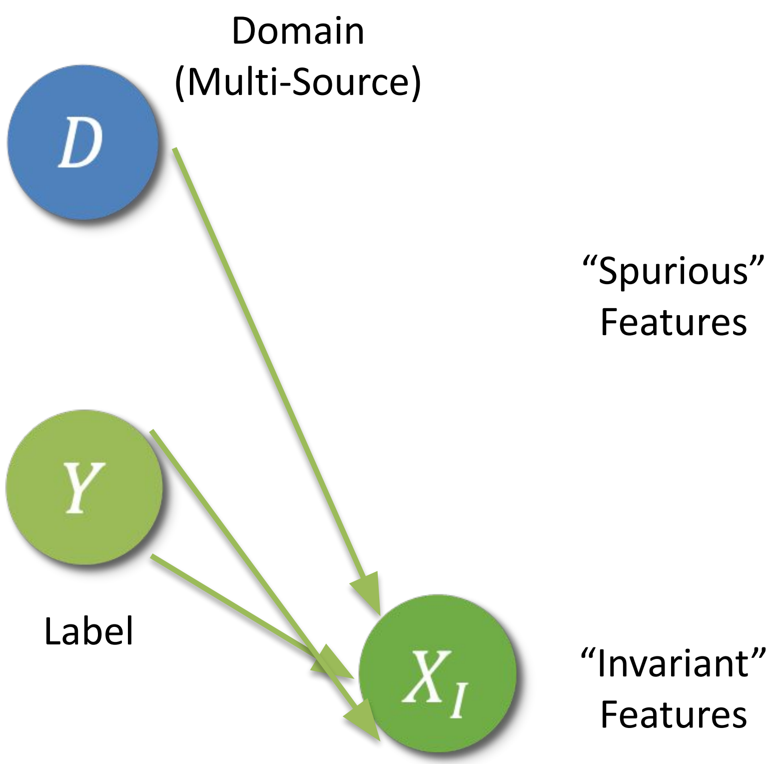
Approach 2: density estimation of *q* and *p*

When importance reweighting is not enough

- Importance reweighted estimator can be high variance
- If there is no *overlap*, then in general impossible – even with infinite data

Current state of research on dataset shift

- Seek "invariant" representations that will work well even after dataset shift



What properties should a representation have?

Here, the domain only influences (some) features. But, how do we know which ones?

Observe: The distribution $P(Y \mid X_I)$ does not depend on D. Can we encourage our representation to recover X_I?

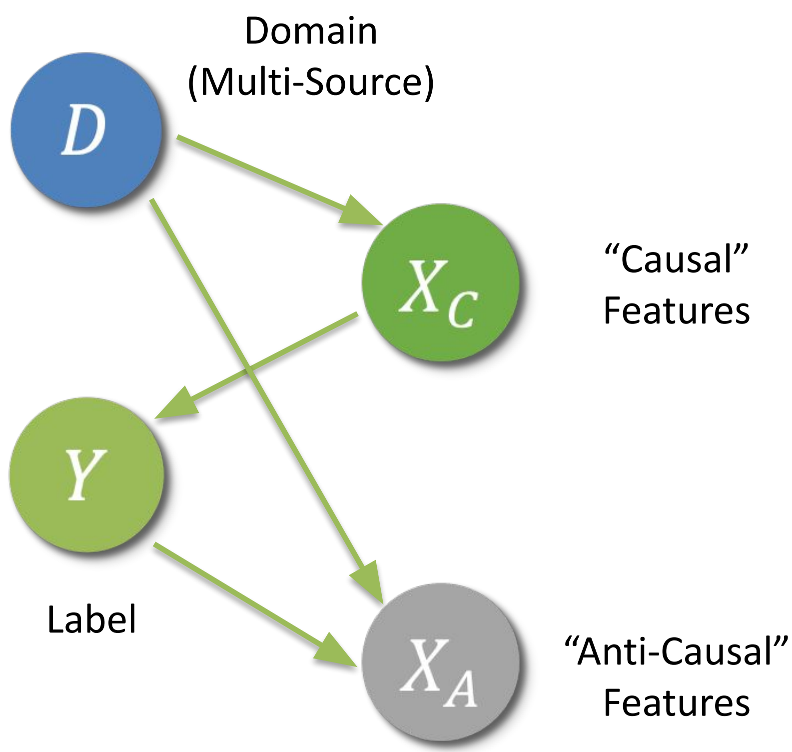
Potential approach: Given multiple source environments, learn a representation such that

$$\boldsymbol{\phi}(\boldsymbol{X}) \perp \boldsymbol{D}$$

Caveat: The right "invariance" depends on the generative structure, and how D impacts X, Y

Current state of research on dataset shift

- Seek "invariant" representations that will work well even after dataset shift



What properties should a representation have?

Here, the domain influences all features.

Observe: The distribution $P(Y \mid X_C)$ does not depend on D. Can we encourage our representation to recover X_C?

Potential approach: Given multiple source environments, learn a representation such that

$$\boldsymbol{Y} \perp \boldsymbol{D} \mid \boldsymbol{\phi}(\boldsymbol{X})$$

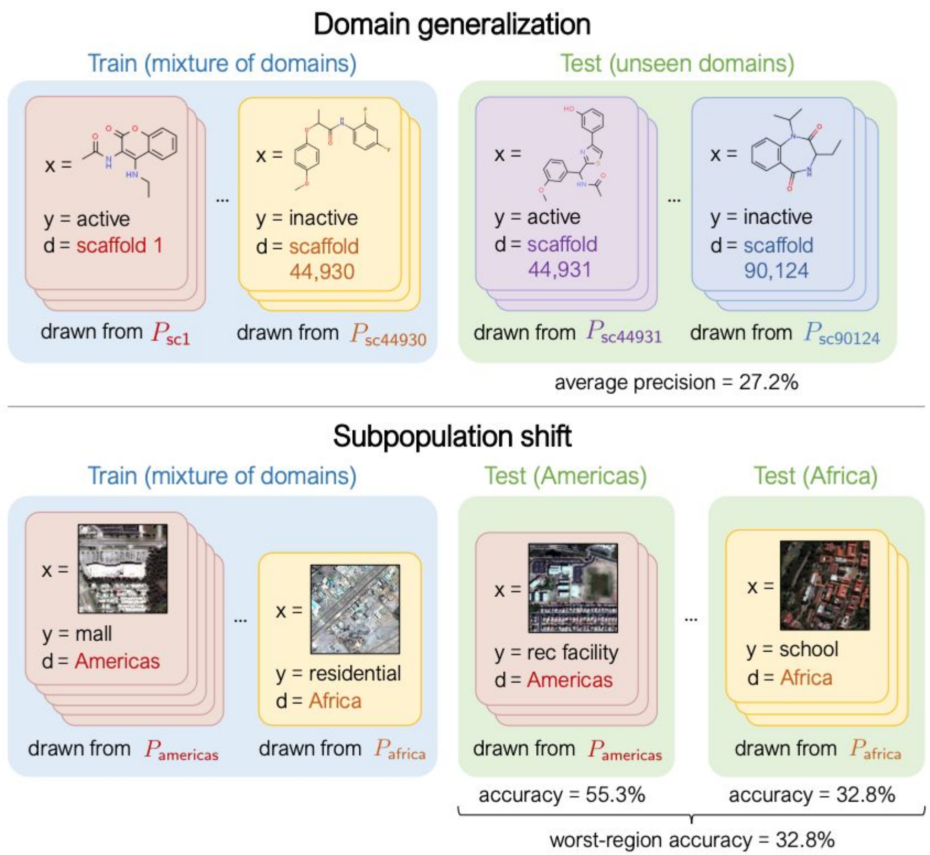
Note: Under this generative structure, it no longer makes sense to seek $\boldsymbol{\phi}(\boldsymbol{X}) \perp \boldsymbol{D}$

Current state of research on dataset shift: benchmarking

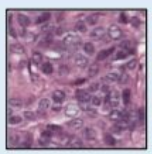
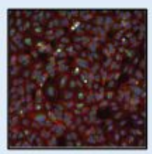
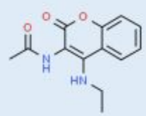
WILDS: A Benchmark of in-the-Wild Distribution Shifts

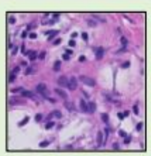
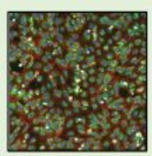
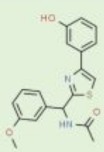
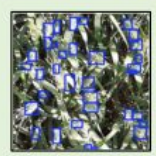
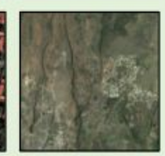
Pang Wei Koh* and Shiori Sagawa* {pangwei, ssagawa}@cs.stanford.edu
Henrik Marklund marklund@stanford.edu
Sang Michael Xie xie@cs.stanford.edu
Marvin Zhang marvin@eecs.berkeley.edu
Akshay Balsubramani abalsubr@stanford.edu
Weihua Hu weihuahu@stanford.edu
Michihiro Yasunaga myasu@stanford.edu
Richard Lanas Phillips richard@cs.cornell.edu
Irena Gao igao@stanford.edu
Tony Lee tonyhlee@stanford.edu
Etienne David etienne.david@inrae.fr
Ian Stavness stavness@usask.ca
Wei Guo guowei@g.ecc.u-tokyo.ac.jp
Berton A. Earnshaw berton.earnshaw@recursionpharma.com
Imran S. Haque imran.haque@recursionpharma.com
Sara Beery sbeery@caltech.edu
Jure Leskovec jure@cs.stanford.edu
Anshul Kundaje akundaje@stanford.edu
Emma Pierson epierson@microsoft.com
Sergey Levine svlevine@eecs.berkeley.edu
Chelsea Finn cbfinn@cs.stanford.edu
Percy Liang pliang@cs.stanford.edu

Correspondence to: wilds@cs.stanford.edu



	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	camera trap photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat head bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	country, rural-urban	user	git repository
# domains	323	5	51	120,084	47	16	16 x 5	23 x 2	2,586	8,421
# examples	203,029	455,954	125,510	437,929	6,515	448,000	523,846	19,669	539,502	150,000

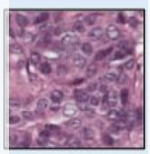
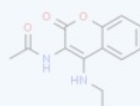
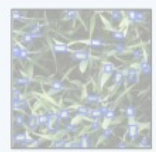
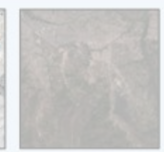
Train example						What do Black and LGBT people have to do with bicycle licensing?			Overall a solid package that has a good quality of construction for the price.	`import numpy as np`
										…
										`norm=np.___`

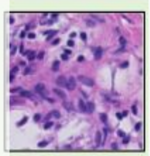
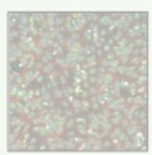
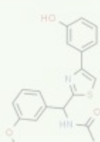
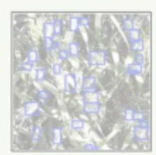
Test example						As a Christian, I will not be patronizing any of those businesses.			I *loved* my French press, it's so perfect and came with all this fun stuff!	`import subprocess as sp`
										`p=sp.Popen()`
										`stdout=p.___`

Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

[Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv:2012.07421, 2021.]

	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	camera trap photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat head bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	country, rural-urban	user	git repository
# domains	323	5	51	120,084	47	16	16 x 5	23 x 2	2,586	8,421
# examples	203,029	455,954	125,510	437,929	6,515	448,000	523,846	19,669	539,502	150,000

Train example						What do Black and LGBT people have to do with bicycle licensing?			Overall a solid package that has a good quality of construction for the price.	`import`
										`numpy as np`
										`…`
										`norm=np.___`

Test example						As a Christian, I will not be patronizing any of those businesses.			I *loved* my French press, it's so perfect and came with all this fun stuff!	`import`
										`subprocess`
										`as sp`
										
										`p=sp.Popen()`
										`stdout=p.___`

Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

[Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv:2012.07421, 2021.]

TL;DR: Existing algorithms don't substantially improve over ERM

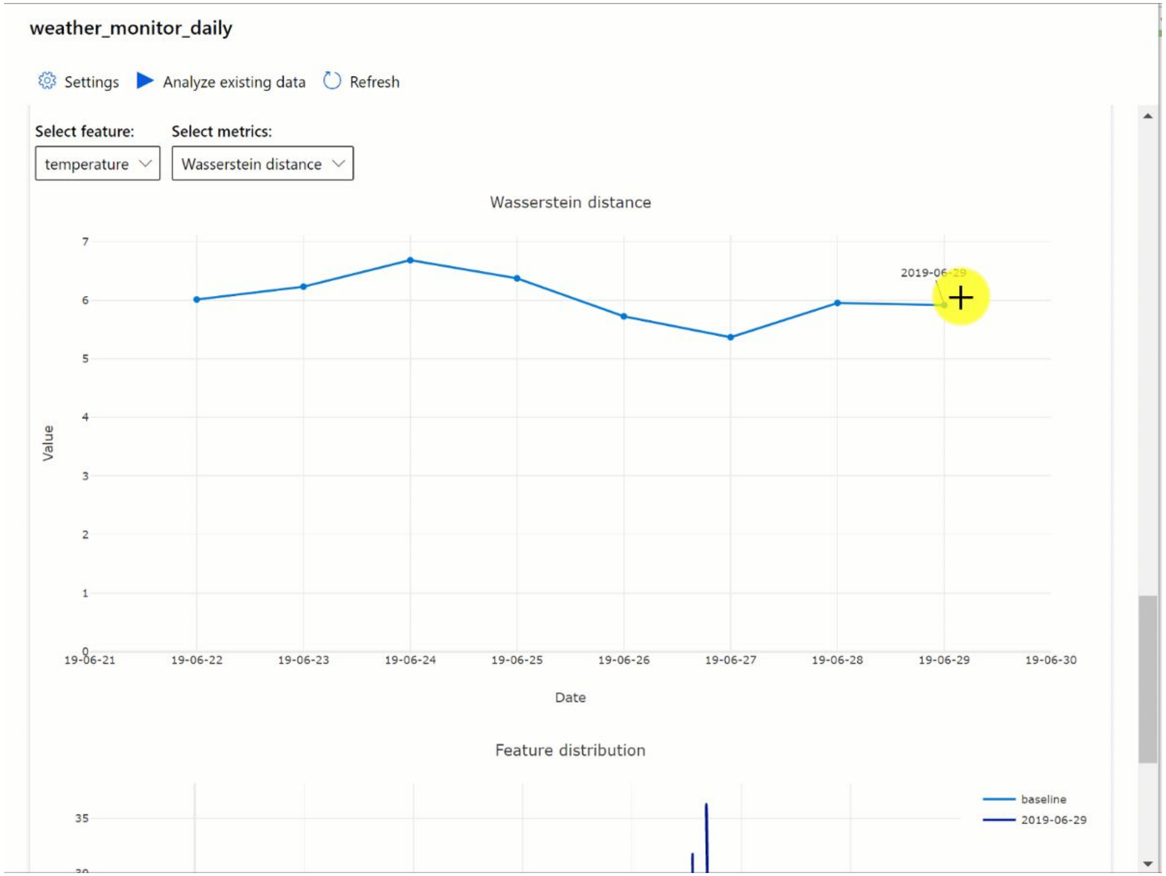
Table 2: The out-of-distribution test performance of models trained with different baseline algorithms: CORAL, originally designed for unsupervised domain adaptation; IRM, for domain generalization; and Group DRO, for subpopulation shifts. Evaluation metrics for each dataset are the same as in Table 1; higher is better. Overall, these algorithms did not improve over empirical risk minimization (ERM), and sometimes made performance significantly worse, except on CivilComments-wilds where they perform better but still do not close the in-distribution gap in Table 1. For GlobalWheat-wilds, we omit CORAL and IRM as those methods do not port straightforwardly to detection settings; its ERM number also differs from Table 1 as its ID comparison required a slight change to the OOD test set. Parentheses show standard deviation across 3+ replicates.

Dataset	Setting	ERM	CORAL	IRM	Group DRO
iWildCam2020-wilds	Domain gen.	31.0 (1.3)	**32.8 (0.1)**	15.1 (4.9)	23.9 (2.1)
Camelyon17-wilds	Domain gen.	**70.3 (6.4)**	59.5 (7.7)	64.2 (8.1)	68.4 (7.3)
RxRx1-wilds	Domain gen.	**29.9 (0.4)**	28.4 (0.3)	8.2 (1.1)	23.0 (0.3)
OGB-MolPCBA	Domain gen.	**27.2 (0.3)**	17.9 (0.5)	15.6 (0.3)	22.4 (0.6)
GlobalWheat-wilds	Domain gen.	**51.2 (1.8)**	—	—	47.9 (2.0)
CivilComments-wilds	Subpop. shift	56.0 (3.6)	65.6 (1.3)	66.3 (2.1)	**70.0 (2.0)**
FMoW-wilds	Hybrid	**32.3 (1.3)**	31.7 (1.2)	30.0 (1.4)	30.8 (0.8)
PovertyMap-wilds	Hybrid	**0.45 (0.06)**	0.44 (0.06)	0.43 (0.07)	0.39 (0.06)
Amazon-wilds	Hybrid	**53.8 (0.8)**	52.9 (0.8)	52.4 (0.8)	53.3 (0.0)
Py150-wilds	Hybrid	**67.9 (0.1)**	65.9 (0.1)	64.3 (0.2)	65.9 (0.1)

Note: These are *blind* implementations (with no domain knowledge injected) that do not attempt to understand the causal nature of the dataset shifts.

[Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv:2012.07421, 2021.]

Current state of industry on dataset shift

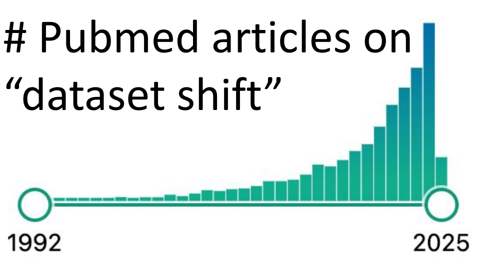


Source: https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets

See also:
https://cloud.google.com/solutions/machine-learning/ml-modeling-monitoring-identifying-training-server
-skew-with-novelty-detection & https://docs.seldon.io/projects/alibi-detect/en/latest/

Conclusion

- Dataset shift happens all the time with healthcare data
- It doesn't always hurt performance
- Interpretability methods can help with detecting and mitigating dataset shift
- Safe deployments should include automated checks for dataset shift
- Active area of research in ML



Additional references

- The Clinician and Dataset Shift in Artificial Intelligence. Finlayson et al., *NEJM* 2021
- Lipton et al. Detecting and Correcting for Label Shift with Black Box Predictors. *ICML*, 2018
- Finlayson et al., Adversarial attacks on medical machine learning, Science, 2019
- Arjovsky et al., Invariant Risk Minimization, arXiv:1907.02893, 2019
- Peters, Bühlmann, Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals, Journal of the Royal Statistical Society 2016
- Veitch et al., Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests, arXiv:2106.00545, 2021