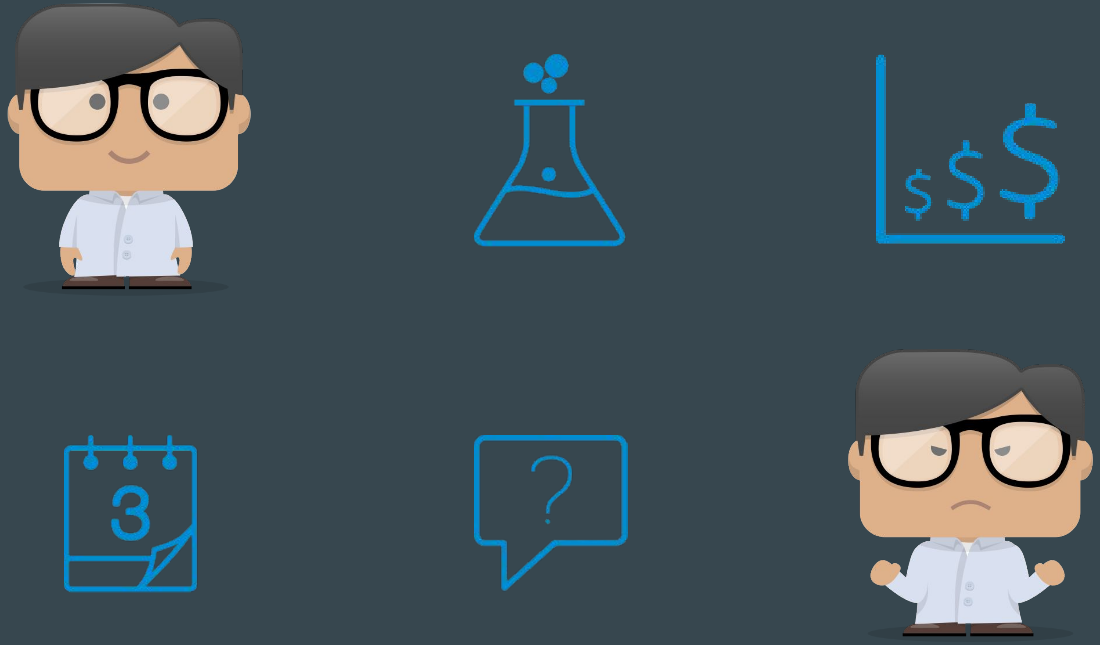
Dataset Shift in Machine Learning

...

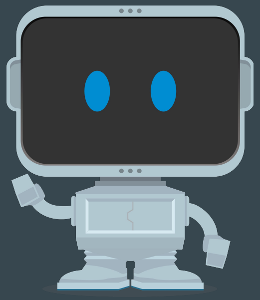
Peter Prettenhofer, OSDC London 2016

Motivation



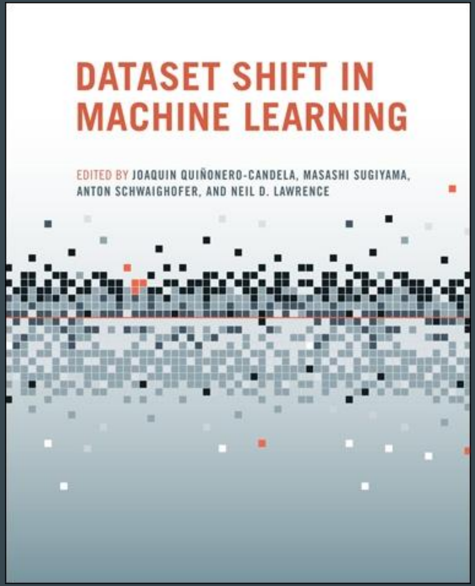
About me

- Data Scientist / Software Engineer @ DataRobot
- Former contributor to scikit-learn



Agenda

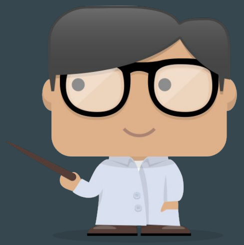
1. Introduction
2. Characterizing Dataset Shift
3. Identifying Dataset Shift
4. Correcting Dataset Shift



https://github.com/pprett/dataset-shift-osdc16

Introduction

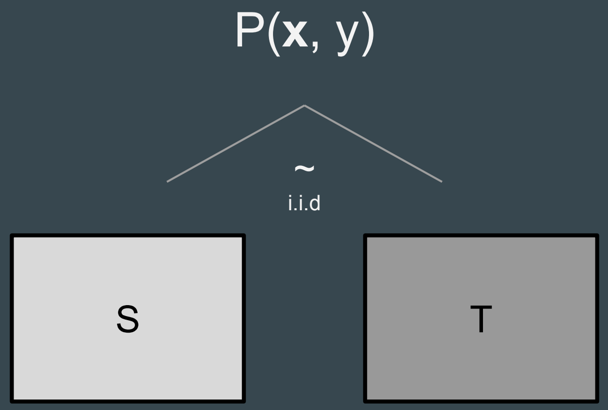
Problem Definition



- Given:
  - Training set $S = \{(\mathbf{x}_i, y_i) \sim P(\mathbf{x}, y)\}$
  - Loss function: $L: (Y, Y) \rightarrow R$
- Goal:
  - Find function $h: X \rightarrow Y$ with minimal error on new set $T = \{(\mathbf{x}_j, y_j) \sim P(\mathbf{x}, y)\}$
- Example:
  - Spam detection
  - Credit risk

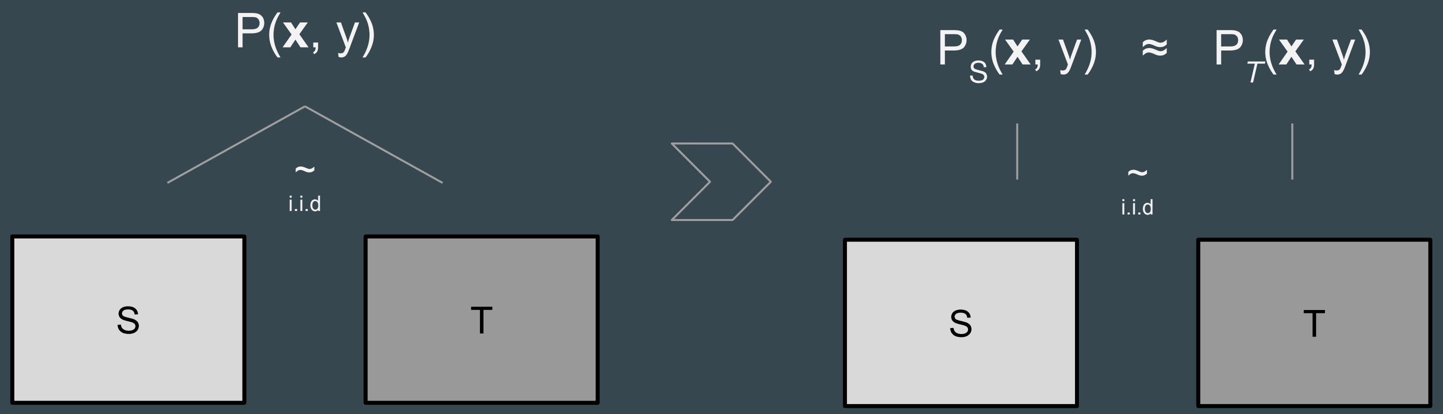
Empirical Risk Minimization

- Question: How do we pick *h* ?
- Answer: Pick the one that minimizes the loss on the training data.
- Assumptions: Train and test set are drawn *i.i.d.* from P(x, y).



Characterizing Dataset Shift

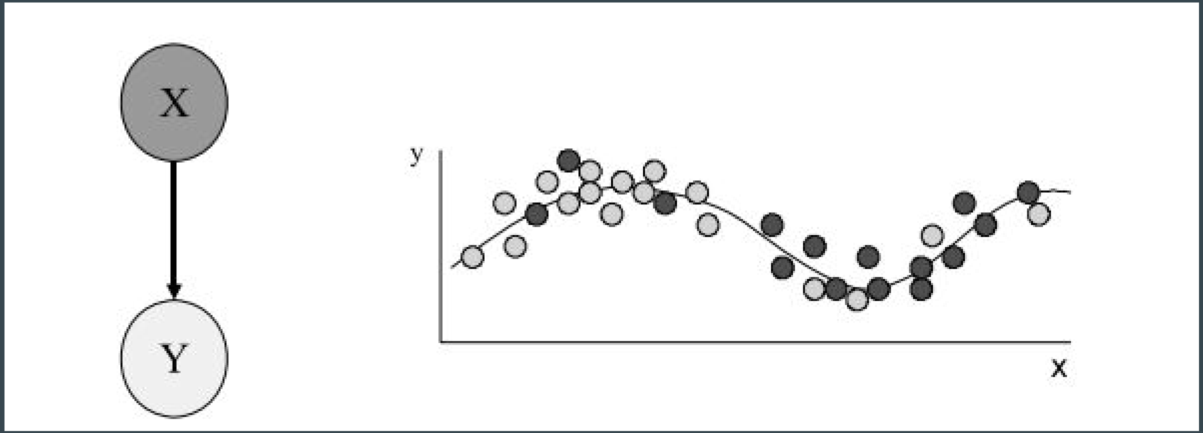
Classical vs. Dataset Shift



- Characterizing distributional change
  - $P(x, y) =$ $P(y|x)\,P(x)$
    $P(x|y)\,P(y)$

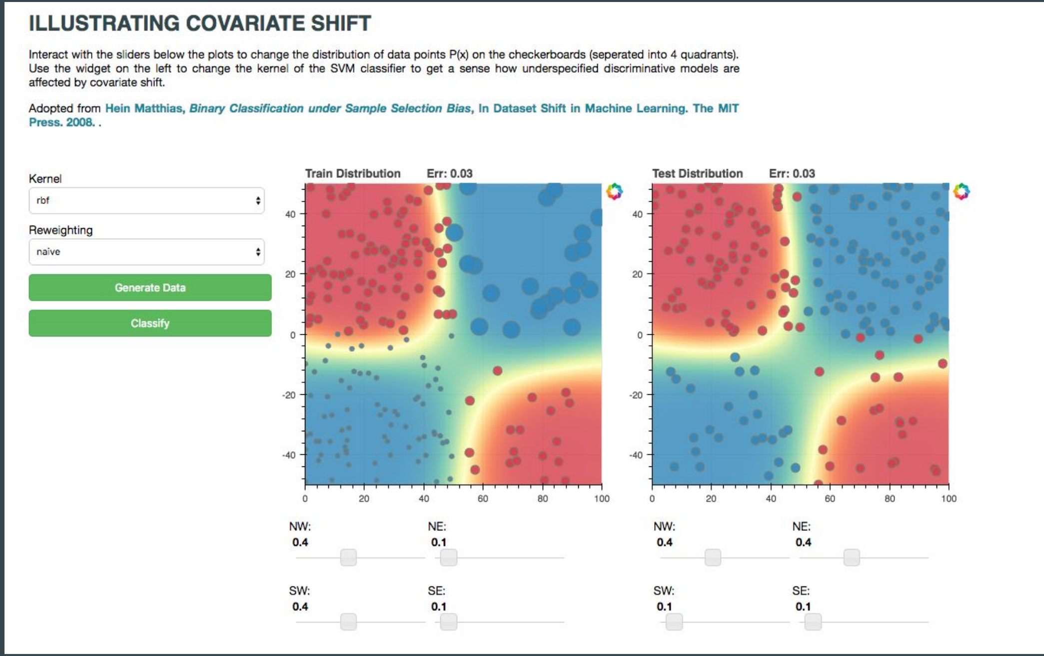
Covariate Shift

- Let: $P_S(y|\mathbf{x}) = P_T(y|\mathbf{x})$
  $P_S(x) \neq P_T(\mathbf{x})$
- Example:
  - Medical testing
- Lots of research
  - J. Heckman in Econometrics



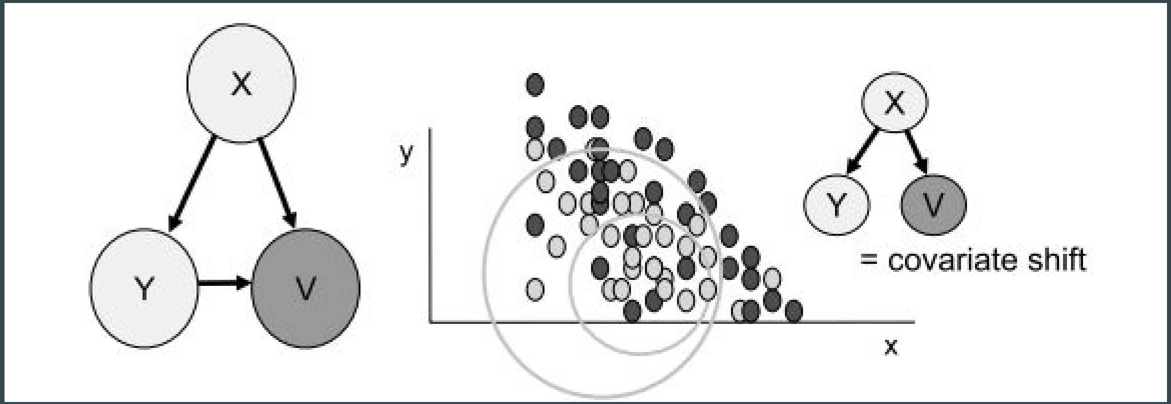
Simple Covariate Shift, From [Storkey, A, 2009].

Illustrating Covariate Shift



Sample selection bias

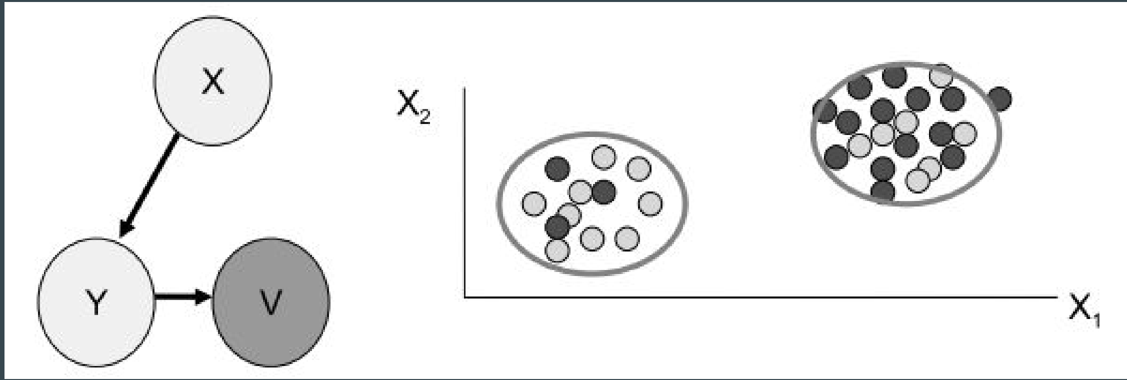
- Distributions differ due to an (unknown) sample rejection process .



Sample selection bias , From [Storkey, A, 2009].

Imbalanced Datasets

- Intentional dataset shift to better deal with class imbalance or large volumes
    - e.g. ad-targeting [He, X. et al, 2014]
- Correction is often trivial
    - Recalibrate probabilities!



Imbalanced data, From [Storkey, A, 2009].

Domain Adaptation

- Intentional Dataset Shift in Natural Language Processing
- Resource rich *source* domain, resource poor *target* domain
  - Sentiment classification, train on product reviews, predict financial news
- Closely linked to *Transfer Learning* .

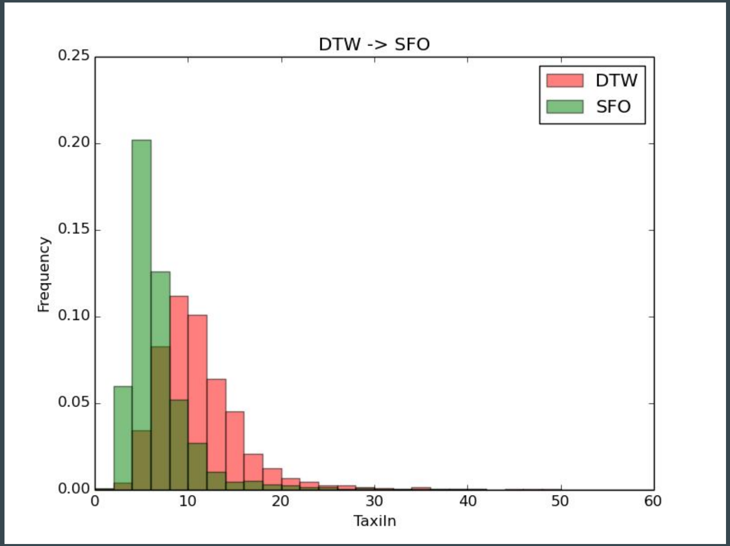
Identifying Dataset Shift

Identifying Dataset Shift

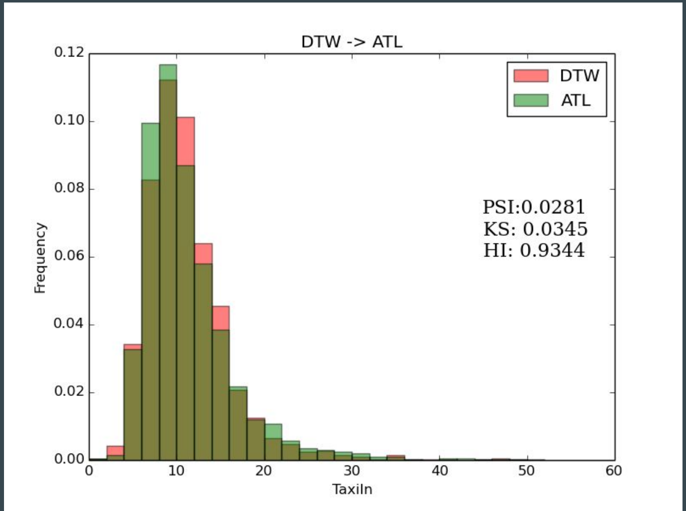
- Identification
  - Supervised
    - Progressive Error
    - External Holdout
  - Unsupervised
    - Statistical Distance
    - Novelty Detection
    - Discriminative Distance

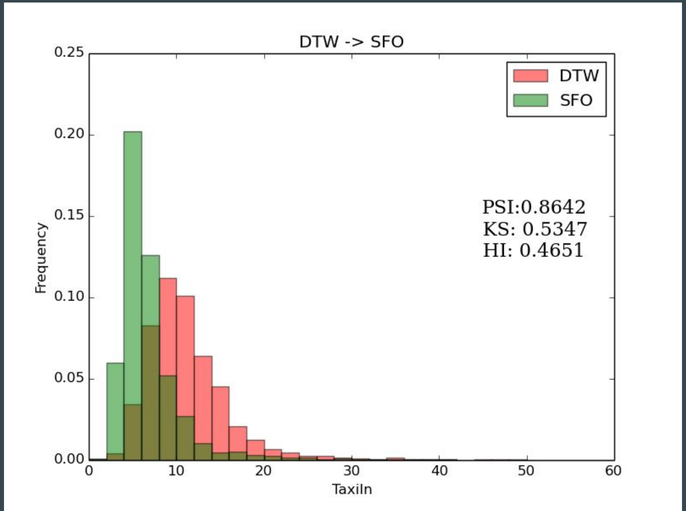
Statistical distance

- Change detection with histograms
  - Mostly uni-variate, sometimes bi-variate
- Advantages:
  - Widely applicable (features and predictions)
  - Simple
  - Easy to spot what changed
- Disadvantages:
  - Not great for high-dimensional or sparse features



Statistical distance cont'



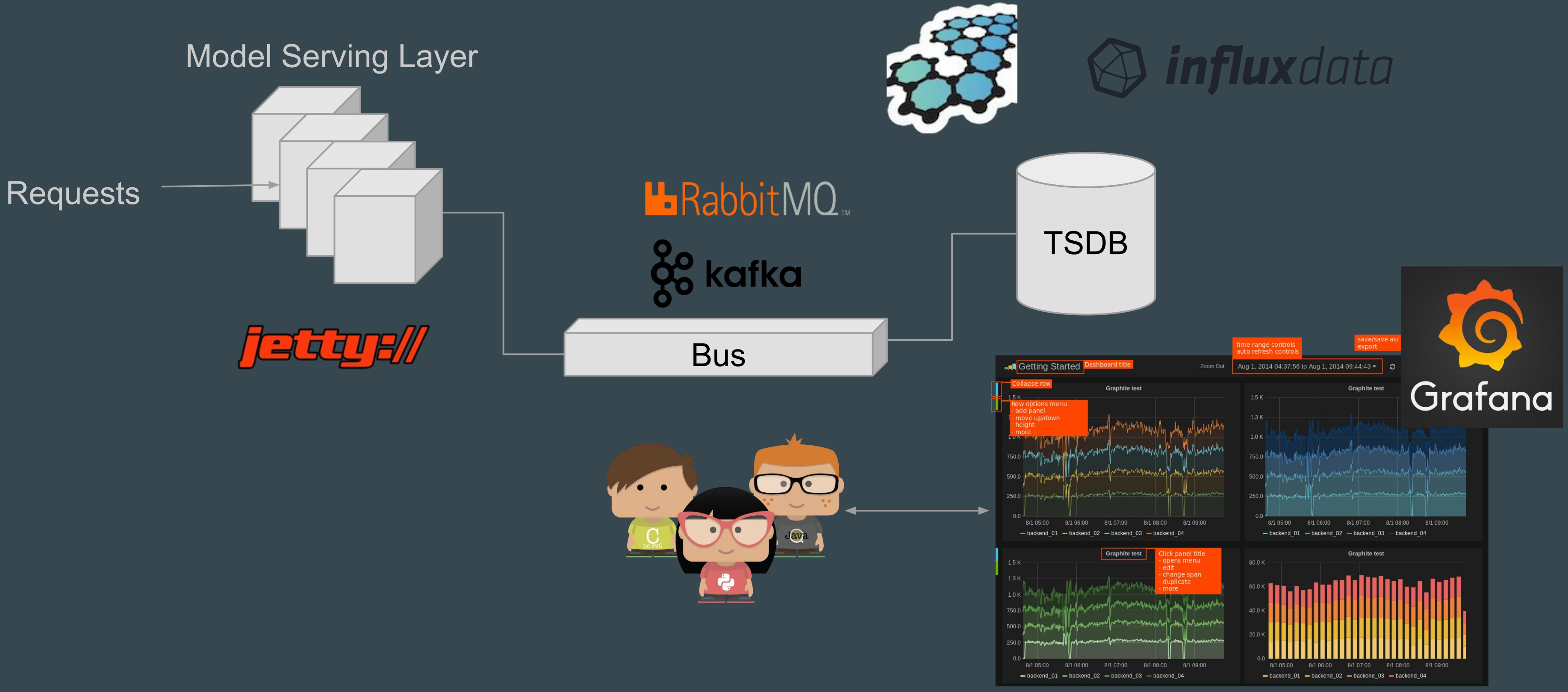


- Population Stability Index (PSI)
- Kolmogorov-Smirnov statistic
- Kullback-Leibler divergence
- Histogram intersection

Where is it used?

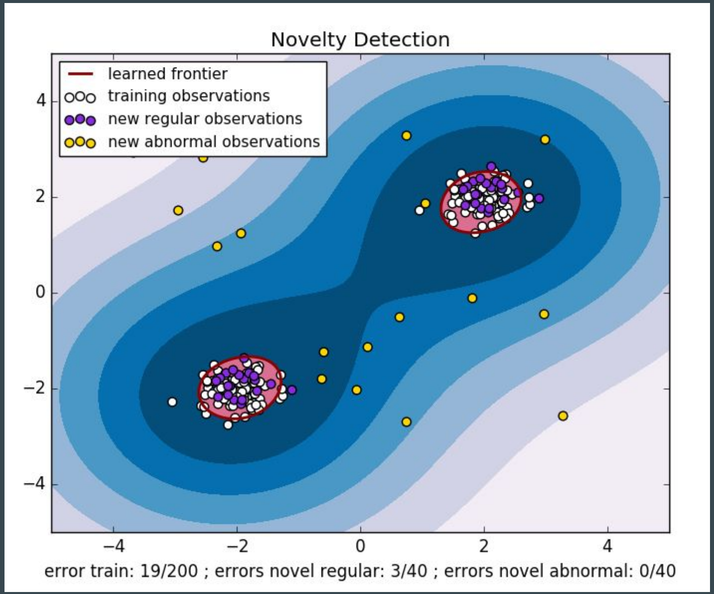
- Risk Management
  - Population Stability Index of score
    - PSI values greater than 0.25 are deemed major shift
- Tech Debt in ML-Systems [Sculley, D et al, 2014]
  - Prediction Bias
    - *"In a system that is working as intended, it should usually be the case that the distribution of predicted labels is equal to the distribution of observed labels."* [Sculley, D et al, 2014]
  - Upstream Changes
    - Tight coupling with upstream components; monitor if invariants hold

Change monitoring architecture



Novelty Detection

- Model $P_S(\mathbf{x})$ & test new $\mathbf{x}$
  - Via Density Estimation techniques
- Example:
  - One-class SVM
- Advantages:
  - Handles many features & complex interactions (eg. vision, audio, remote sensing)
- Disadvantages:
  - Can't tell you what changed



Discriminative Distance

- Intuition: Train a classifier to detect whether an example is from P_S or P_T
  - Use training error as a proxy for distance - the higher the error the closer
- Advantages:
  - Widely applicable (high dimensional, sparse data)
  - Feature importance shows what changed
- Disadvantages:
  - Offline as it requires a batch of data and time to build the model
  - Complicated

Comparison

Method	Data Characteristics			Latency			Insights
	Heterogeneous	Homogenous Dense	Homogenous Sparse	Low	Medium	High	What changed?
Statistical Distance	3	1	1	x			3
Novelty Detection	0	3	2	x			0
Discriminative Distance	3	3	3			x	1

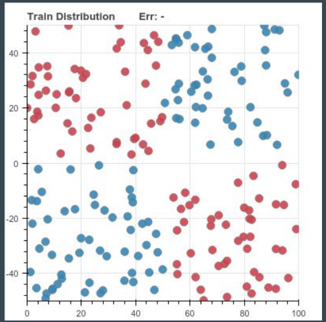
Correcting Dataset Shift

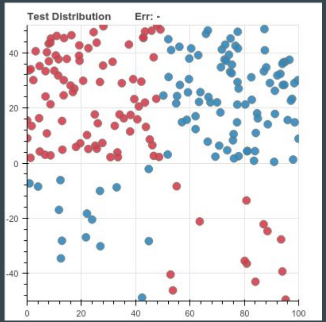
~~Correcting Dataset Shift~~

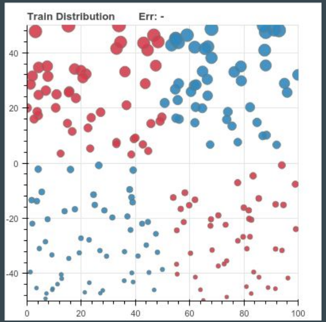
You don’t - you retrain!

Importance Reweighting

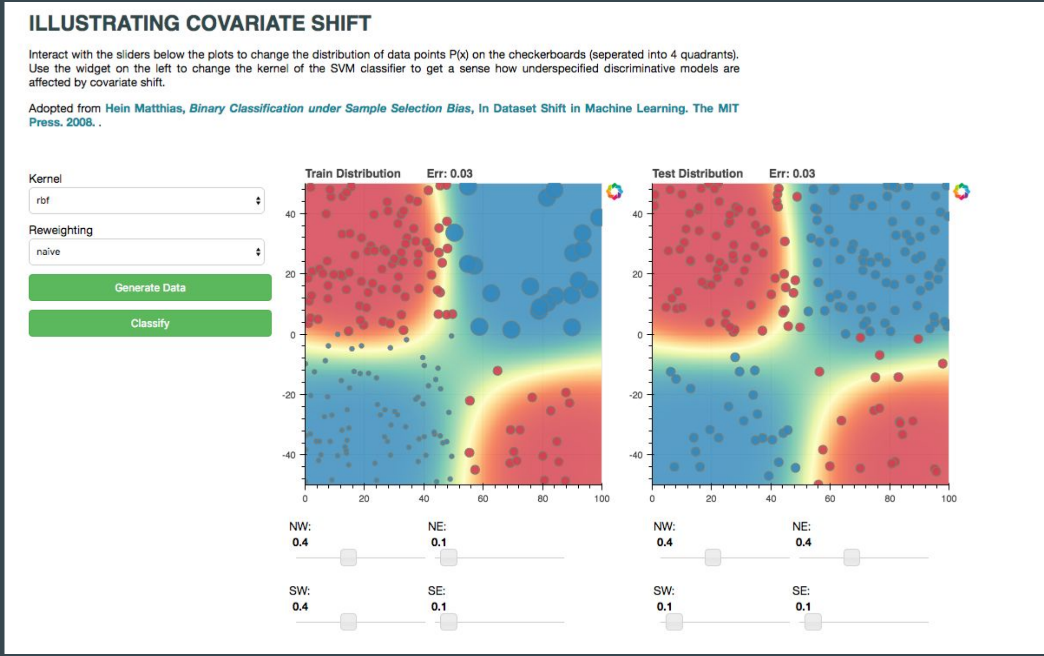
- Upweight training instances that are similar to test instances
  - Weight each training sample by $P_T(\mathbf{x}) / P_S(\mathbf{x})$
- Requires unlabeled data from $P_T(\mathbf{x})$
- How to obtain weights?
  - Density Estimation / Kernel Methods (Kernel Mean Matching)
  - Discriminative Reweighting







Illustrating Importance Reweighting



Discriminative Reweighting

- Estimate $P_T(\mathbf{x}) / P_S(\mathbf{x})$ using a Logistic Regression **

```
X_s, y_s, X_u, X_t = …
X_u = np.r_[X_s, X_u]
y_u = np.r_[-1 * ones_like(y_s), ones(X_u.shape[0])]
st_est = LogisticRegression().fit(X_u, y_u)
weights = np.exp(st_est.decision_function(X_s))
est = LogisticRegression().fit(X_s, y_s, sample_weight=weights)
est.predict(X_t)
```

** http://blog.smola.org/post/4110255196/real-simple-covariate-shift-correction

Changing Representations

- Find a mapping $\mathbf{z} = \phi(\mathbf{x})$ s.t.
  - $P_S(\mathbf{z}, y) = P_T(\mathbf{z}, y)$ distributions are similar
  - Bayes error rate on $P_S(\mathbf{z}, y)$ still acceptable
- How to find the mapping ϕ?
  - Feature selection [Satpal, Sarawagi, 2008]
  - Structural Correspondence Learning [Blitzer et al, 2006]

Summary

- Supervised ML techniques can be negatively affected by Dataset Shift
- Identifying Dataset Shift
  - Simple histogram-based techniques are widely applicable (esp. model scores)
  - Novelty detection & discriminative distance relevant in certain scenarios
- Correcting Dataset Shift
  - Discriminative Reweighting is simple & effective
  - Changing representation necessary in certain situation (eg. no support)

Thanks

References

- Storkey, A., When Training and Test Sets Are Different: Characterizing Learning Transfer, 2009.
- He, X. et al, Practical Lessons from Prediction Clicks on Ads at Facebook, ADKDD'14, 2014.
- Sculley, D et al, Machine Learning: The high-interest card of technical debt, SE4ML'14, 2014.
- Jiang, J., A Literature Survey on Domain Adaptation of Statistical Classifiers, 2008.