Quantification Under Class-Conditional Dataset Shift

David Spence
PyData Edinburgh
6th December 2018

Counting Dogs and Bicycles

....when all the dogs are labradors and all the bikes are mountain bikes

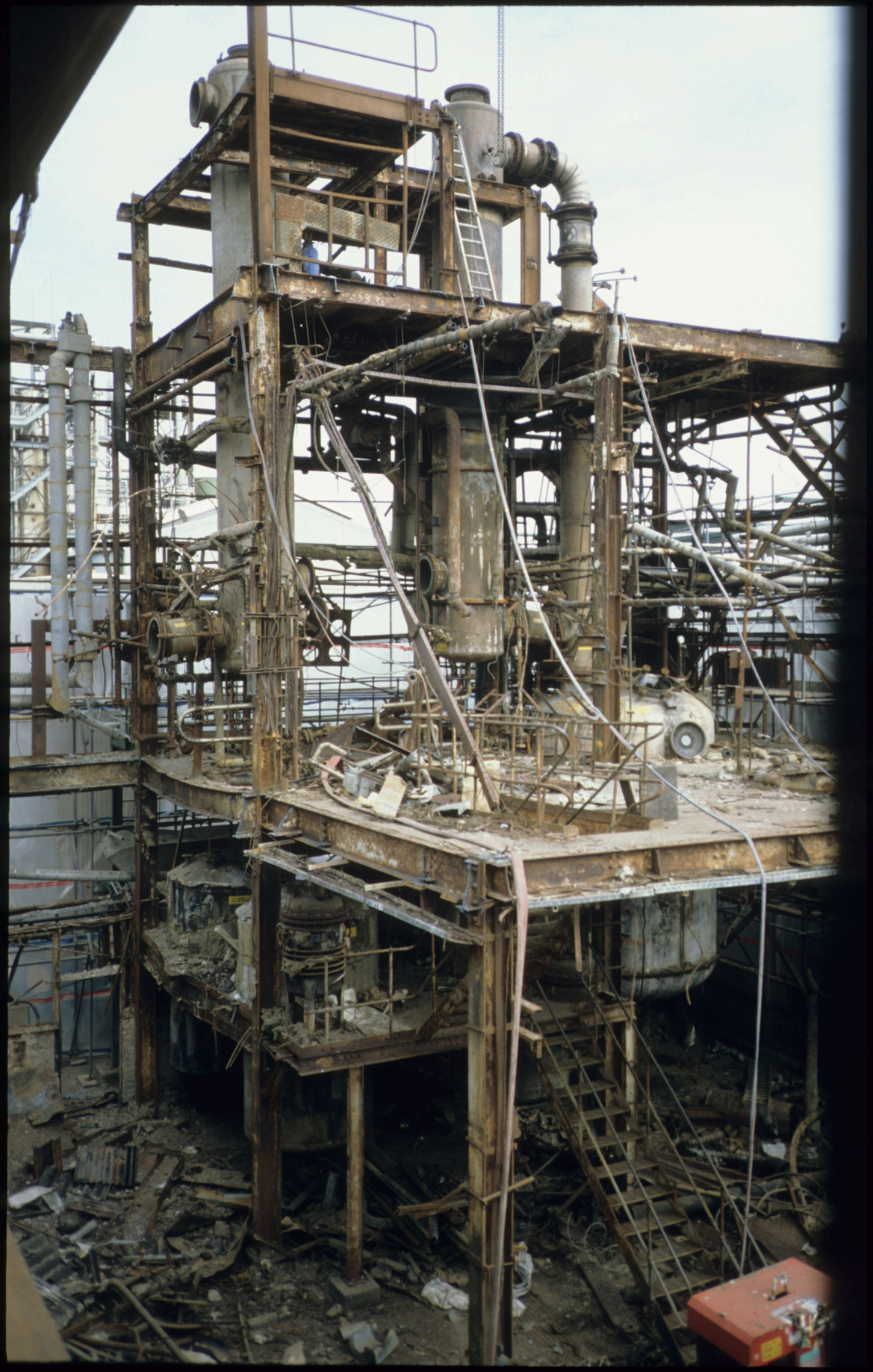
David Spence

PyData Edinburgh

6th December 2018

- NO CODE
- SOME EQUATIONS

Me



LANCASTER
UNIVERSITY

ARTHUR ANDERSEN

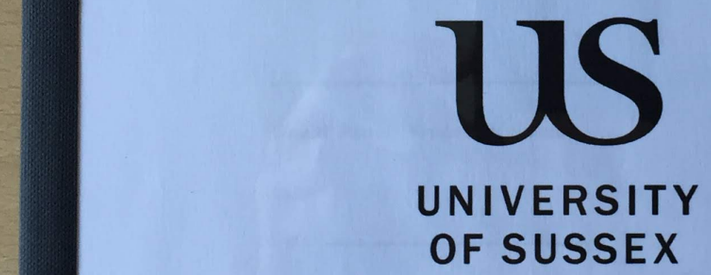

UNIVERSITY
OF SUSSEX

Quantification Under Class-Conditional Dataset Shift

David James Frederick Spence

A thesis presented for the degree of

Doctor of Philosophy

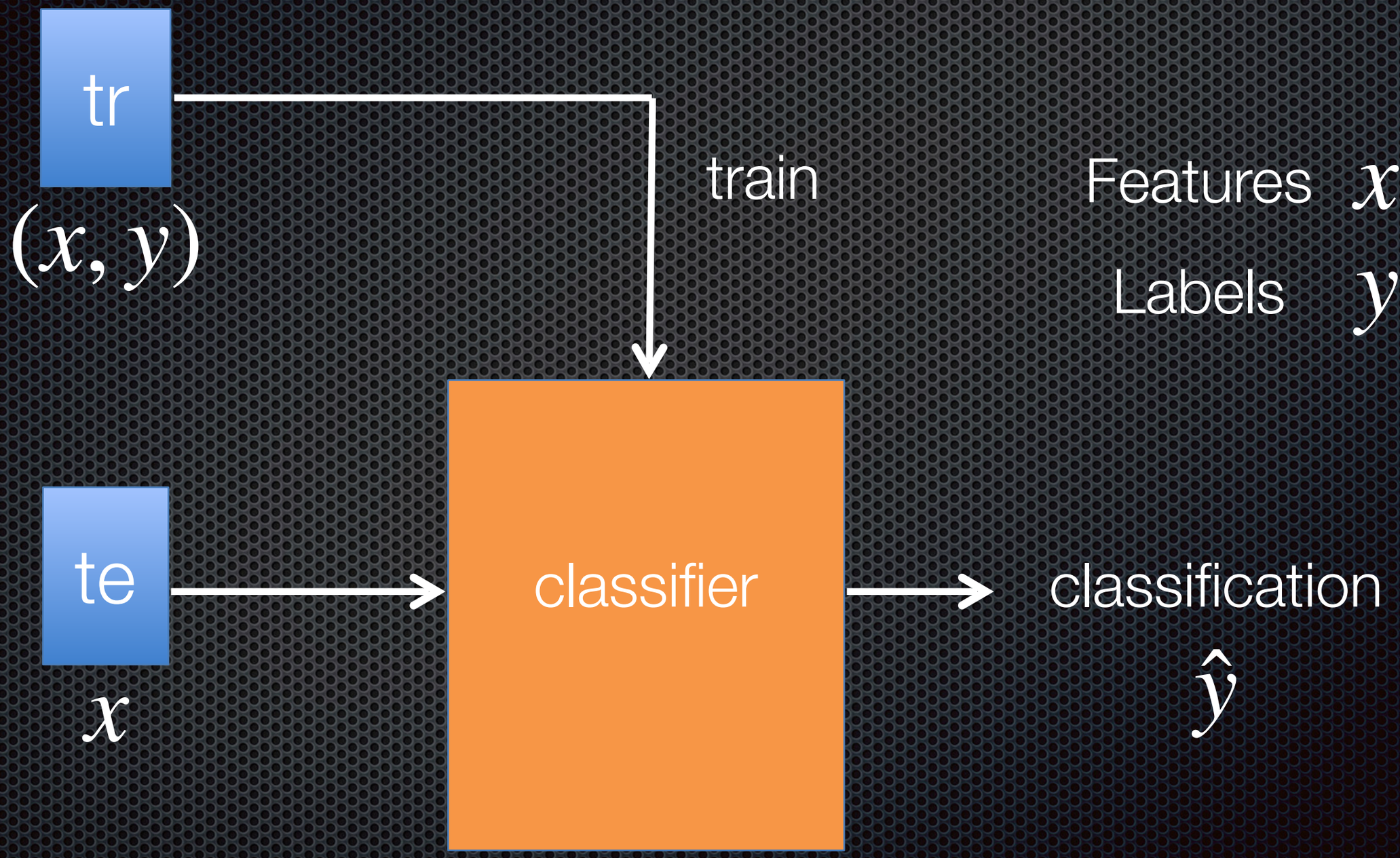


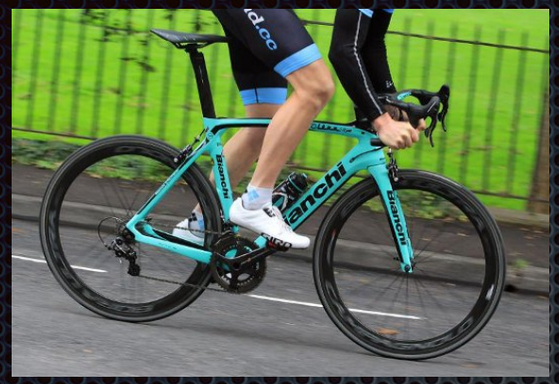
October 2018

Data Scientist?

Supervised Classification

Supervised Classification





$$x \in \mathbb{R}^{(256x256x3)}$$

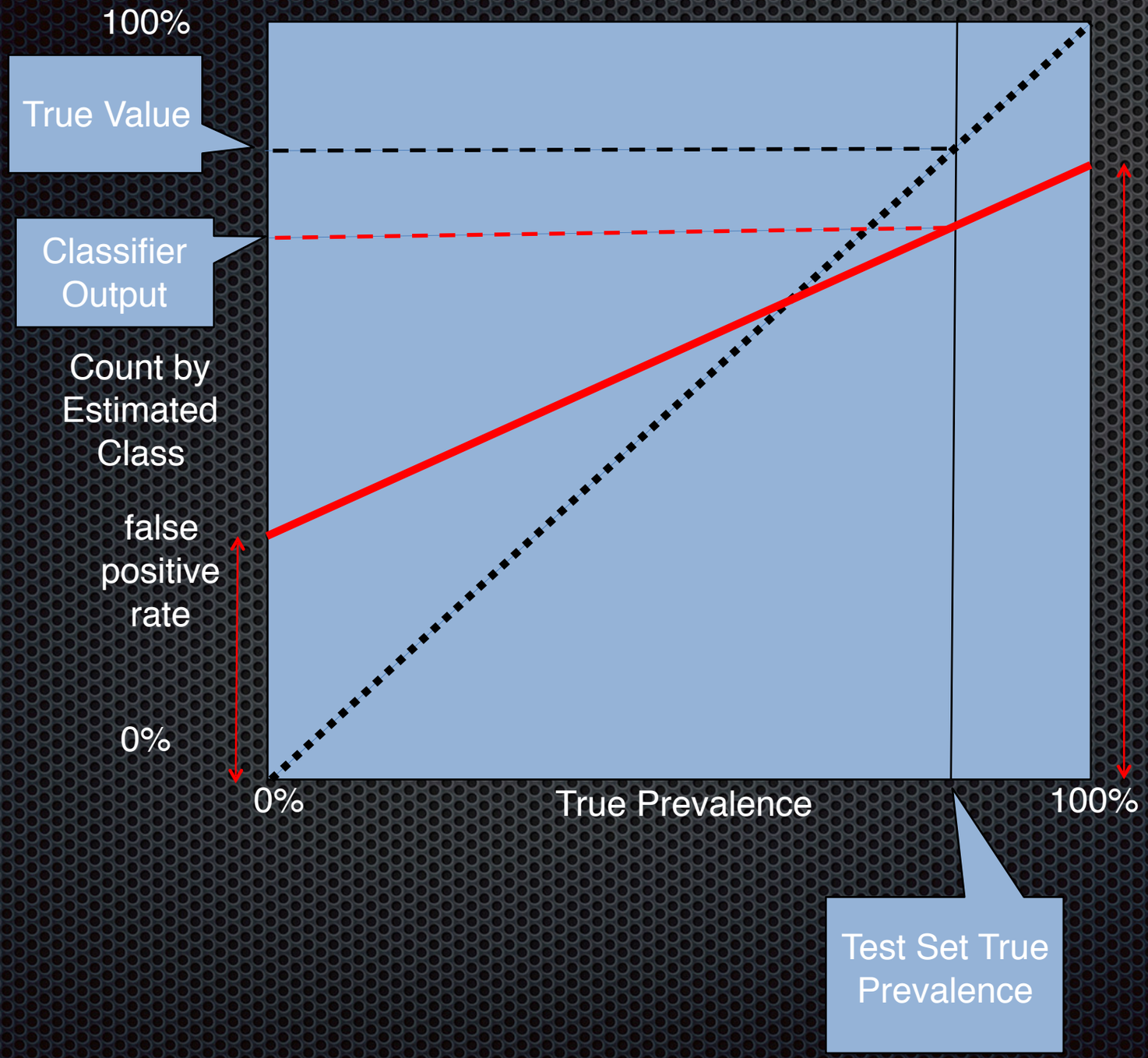
$$y \in \{dog, bicycle\}$$

Quantification

$$P(y = dog) = 0.6$$

$$P(y = bicycle) = 0.4$$

Isn’t quantification trivial?



Still Trivial?

$$p'' \equiv \frac{p' - fpr}{tpr - fpr}$$

- Forman's `Adjusted Count': 2008
- Rogen and Gladen: 1978
- Gart and Buck: 1960s

However....what if....?

Training Data

Test Data

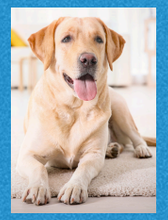
Sample Selection Bias

Training Data









Test Data



However....what if....?

$$tpr(labradors) \neq tpr(dogs_{tr})$$

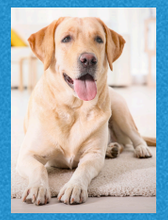
Sample Selection Bias

Training Data









Test Data



Three Possible Approaches

- Quantify the sub-classes separately: divide and conquer
- Weight the training data to match the test data
- Convert everything to a new feature representation

Three Possible Approaches

- Quantify the sub-classes separately: divide and conquer
- Weight the training data to match the test data
- Convert everything to a new feature representation

Training Data

Test Data

Bikes

Dogs

Road Bikes

Town Bikes

Mountain Bikes

Σ

Labradors

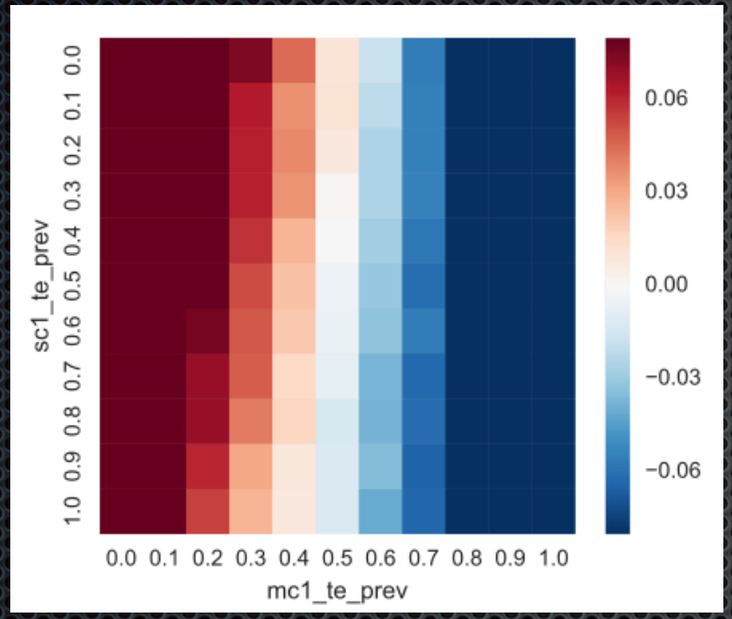
Jack Russels

Beagles

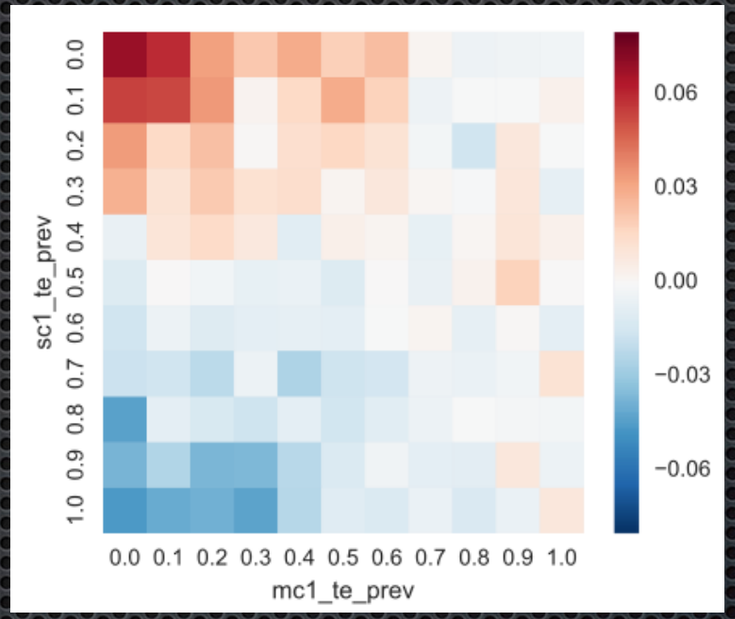
Σ

Explicit Sub-Domains

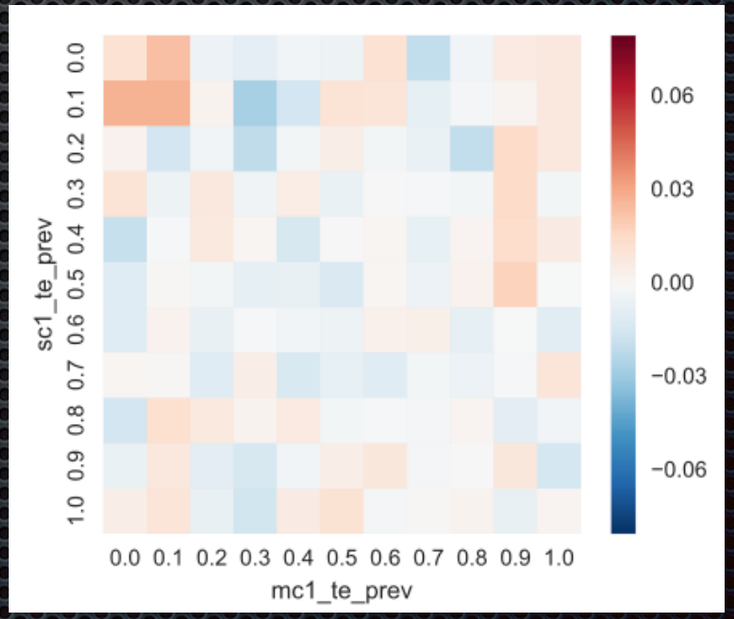
Count by Main-class



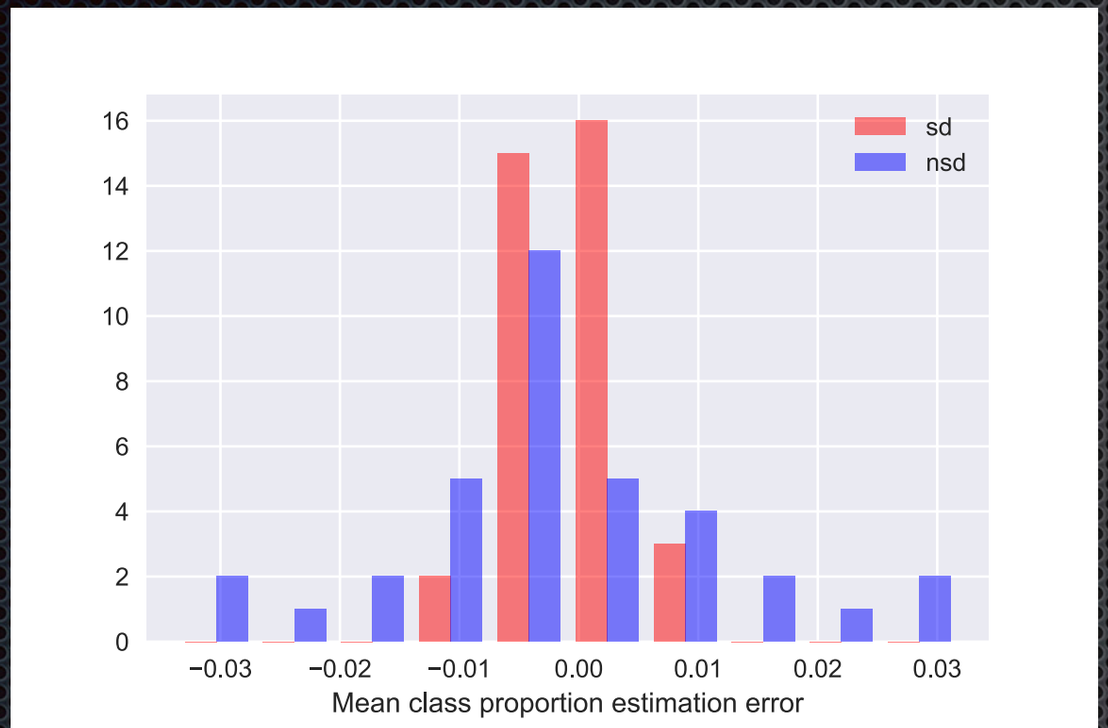
Adjust by Main-class

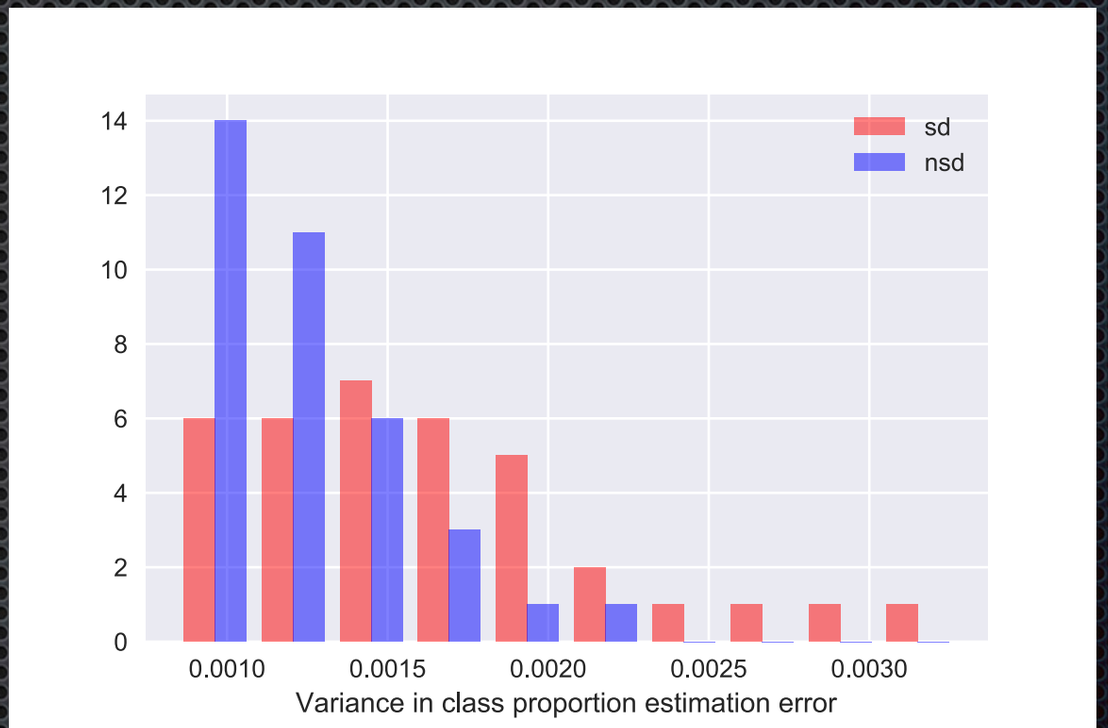


Adjust by Main-class and Sub-domain



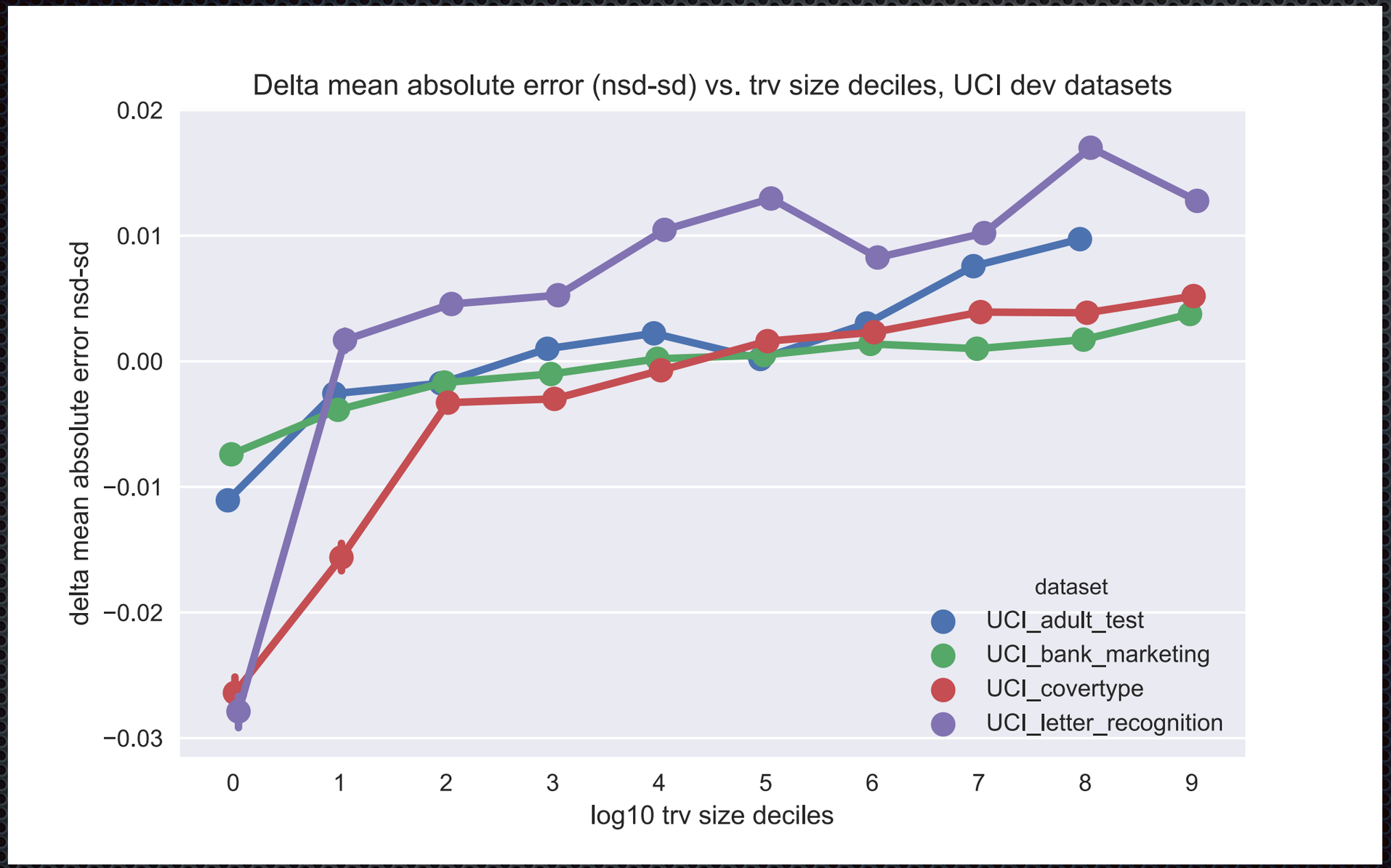
Bias vs. Variance





- Sub-domains reduce bias but increase variance

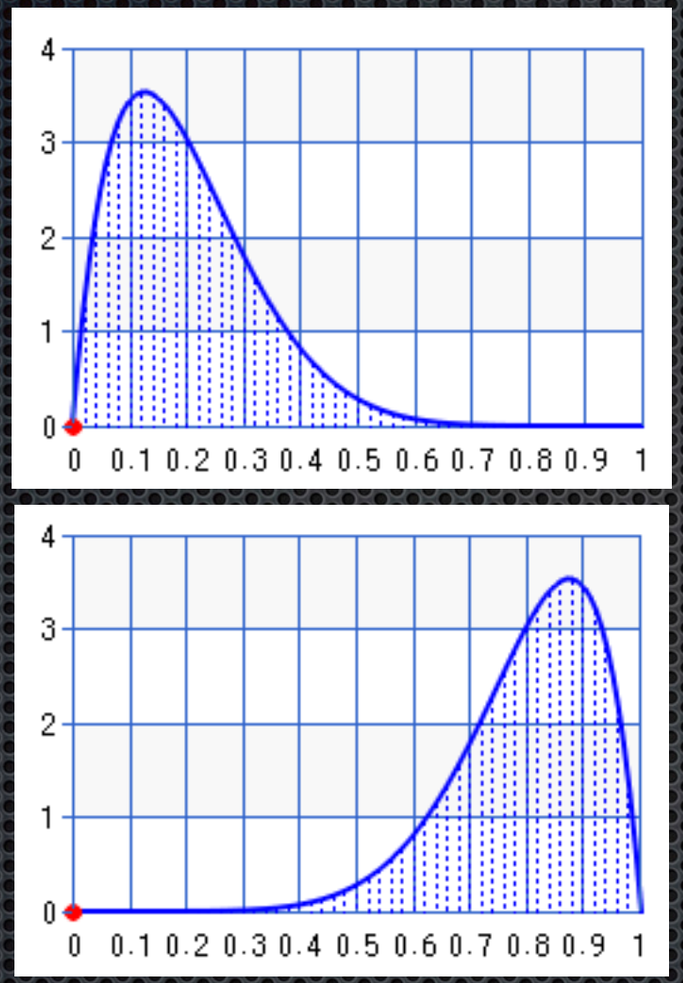
Sub-domains and size of dataset

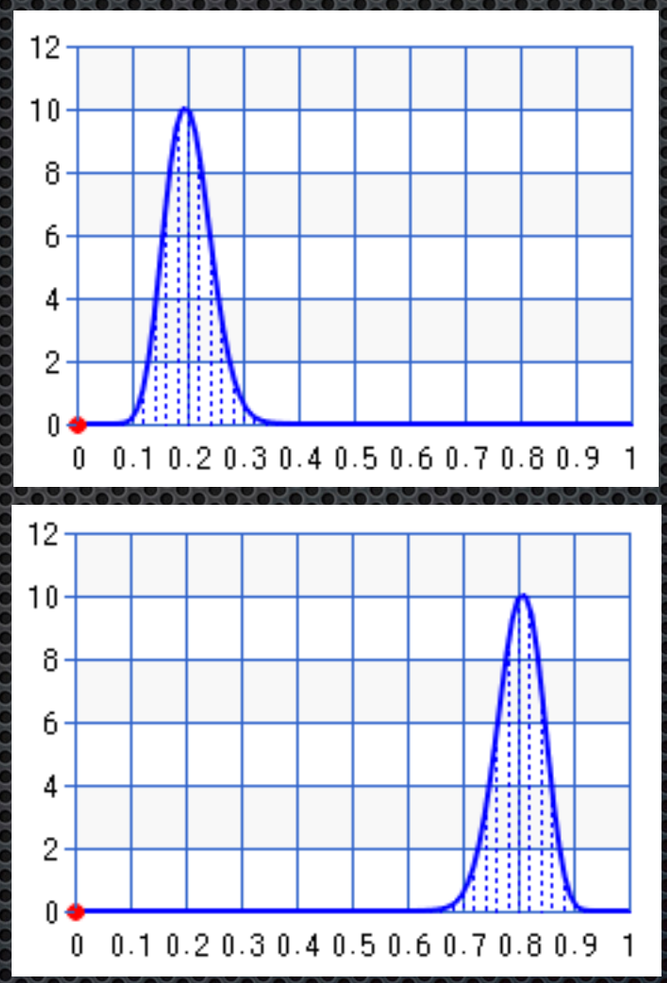


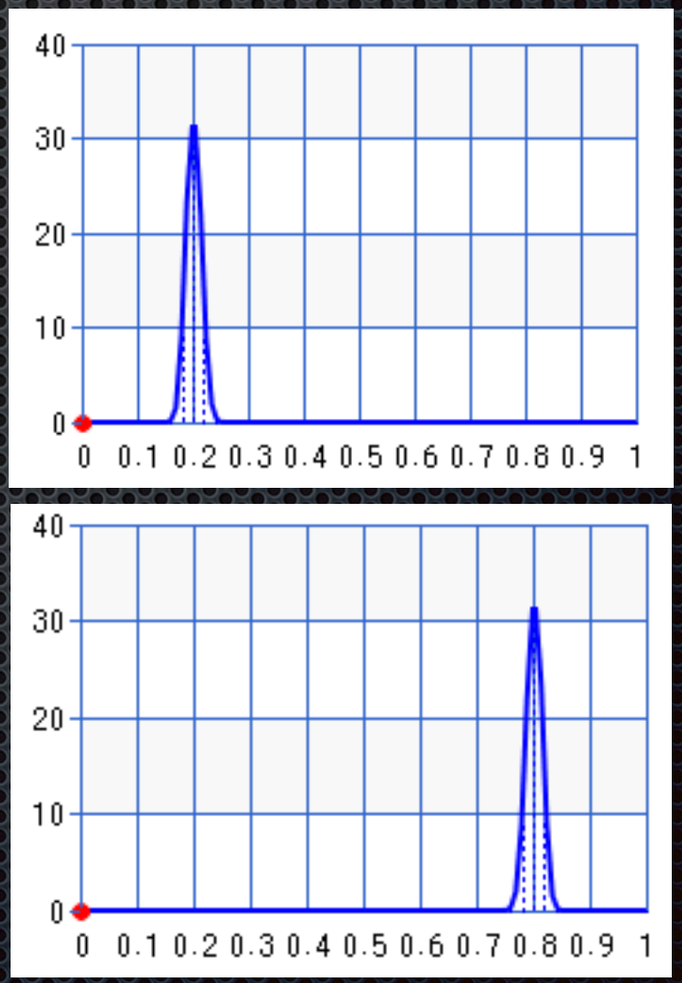
Noise Problem

fpr
0.2

tpr
0.8







n= 10 100 1000

Noise Problem

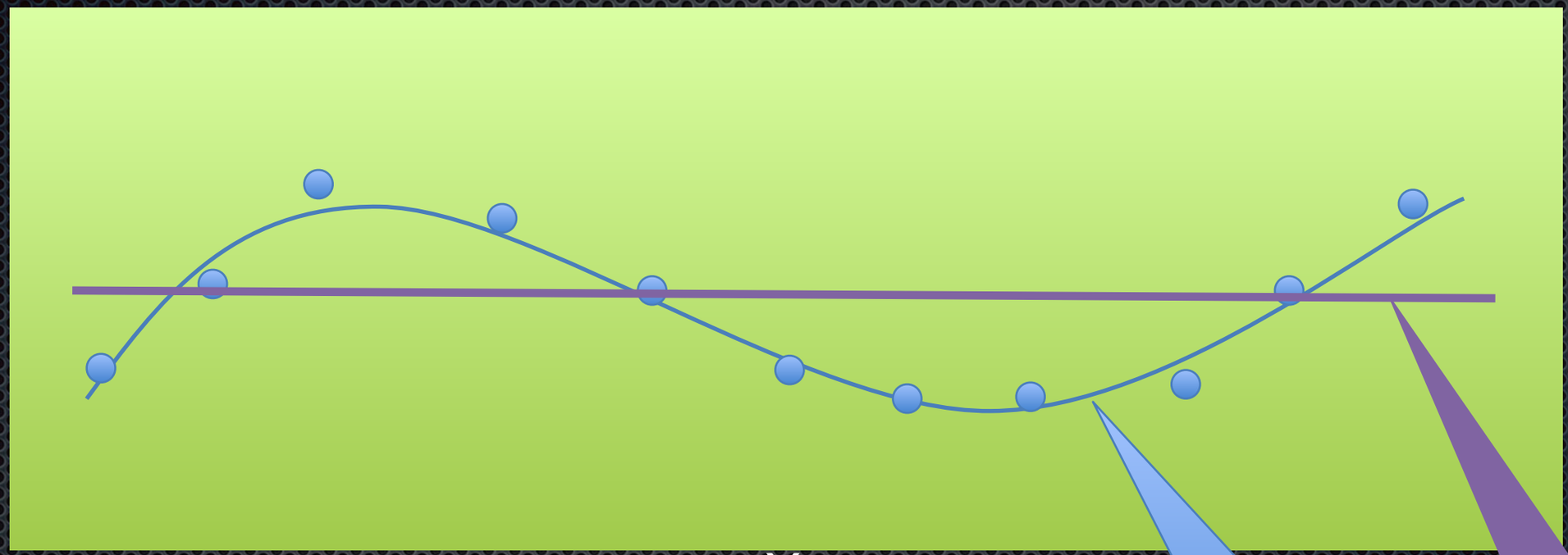
$$p' = (fpr \pm \delta_{fpr})$$

$$p'' = \frac{p'}{(tpr \pm \delta_{tpr}) - (fpr \pm \delta_{fpr})}$$

Three Possible Approaches

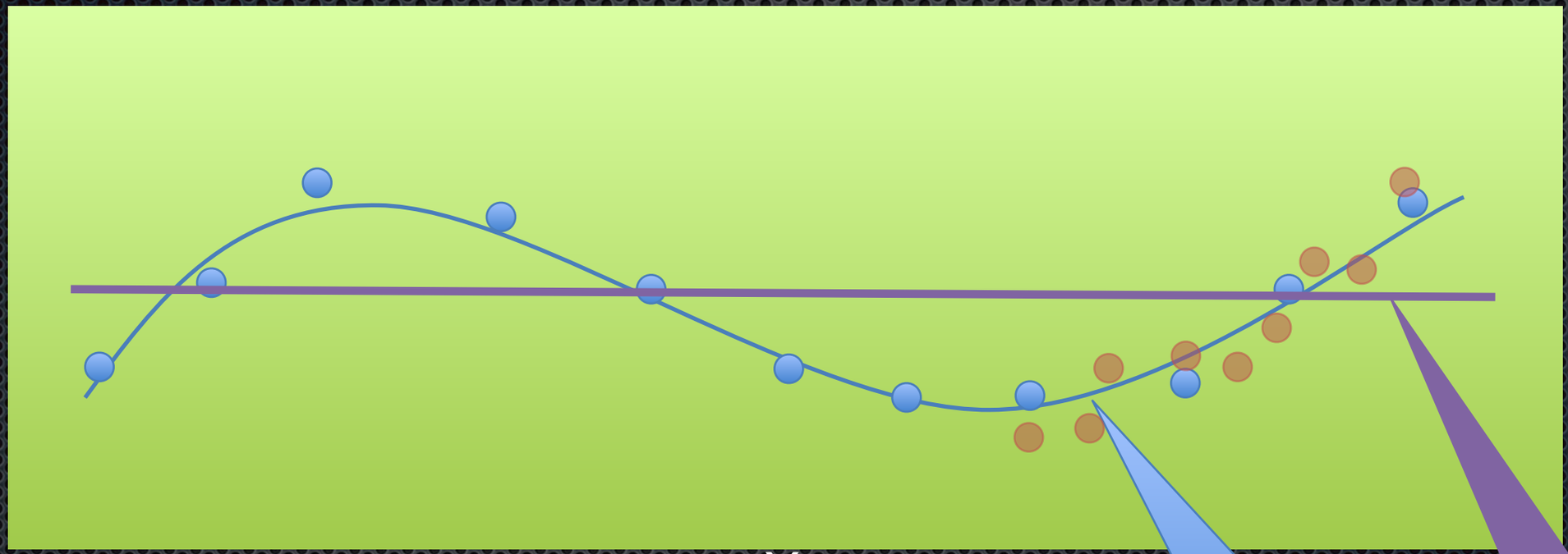
- Quantify the sub-classes separately: divide and conquer
- Weight the training data to match the test data
- Convert everything to a new feature representation

Instance Weighting



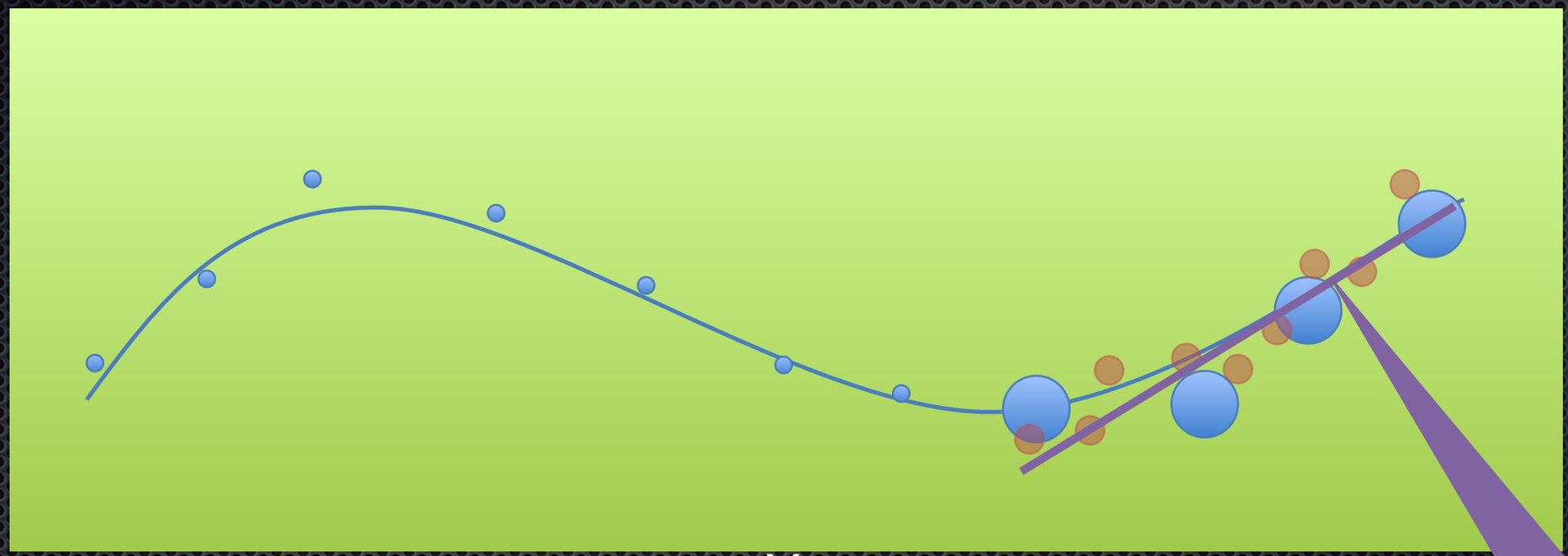
- Training set

Instance Weighting



- Training set
- Test set

Instance Weighting



- Training set
- Test set

Kernel Mean Matching (KMM)

2007 — 2009

"Correcting Sample Selection Bias by Unlabeled data"

Advances in Neural Information Processing Systems (NIPS)

Huang
Gretton
Borgwardt
Scholkopf
Smola

"KMM *always* improves test performance…"

"Covariate Shift by Kernel Mean Matching"

Dataset Shift in Machine Learning

Gretton
Smola
Huang
Schmittfull
Borgwardt
Scholkopf

"KMM…often makes performance slightly worse"

Instance Weighting

- 11% improvement
- All data for training classifier
- ‘Closest’ 50% for *tpr* and *fpr*

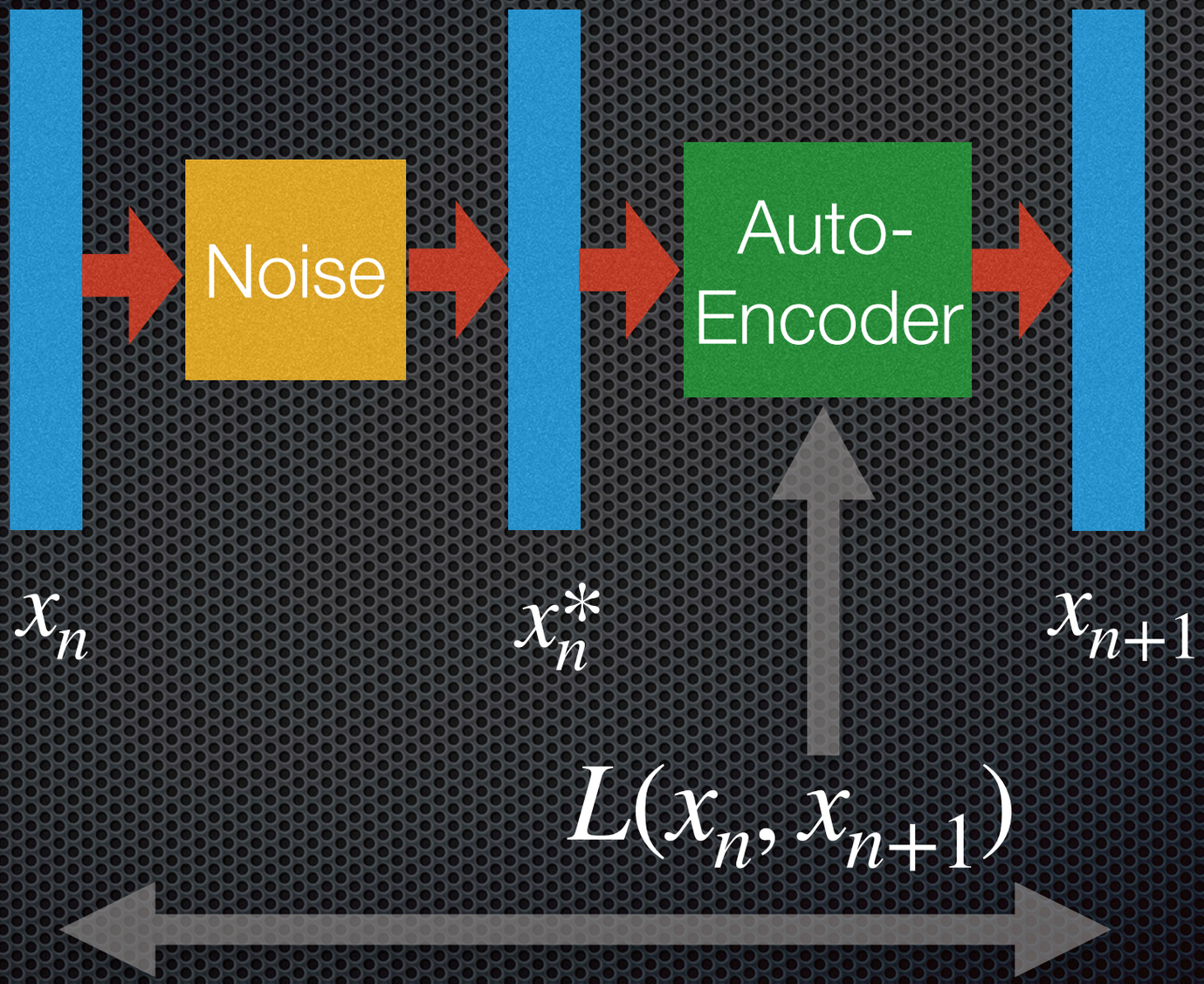
Three Possible Approaches

- Quantify the sub-classes separately: divide and conquer
- Weight the training data to match the test data
- Convert everything to a new feature representation

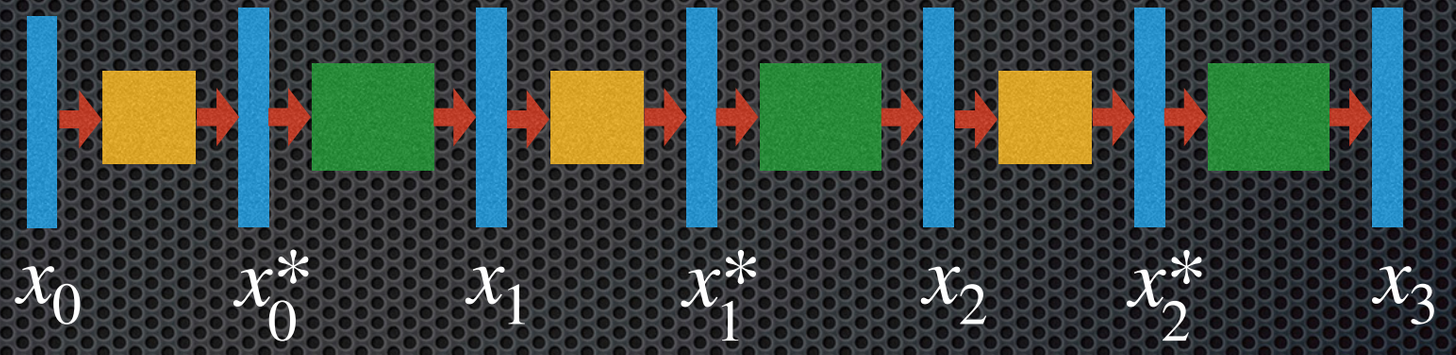
mSDA

- Marginalised
- Stacked
- De-noising
- Autoencoders

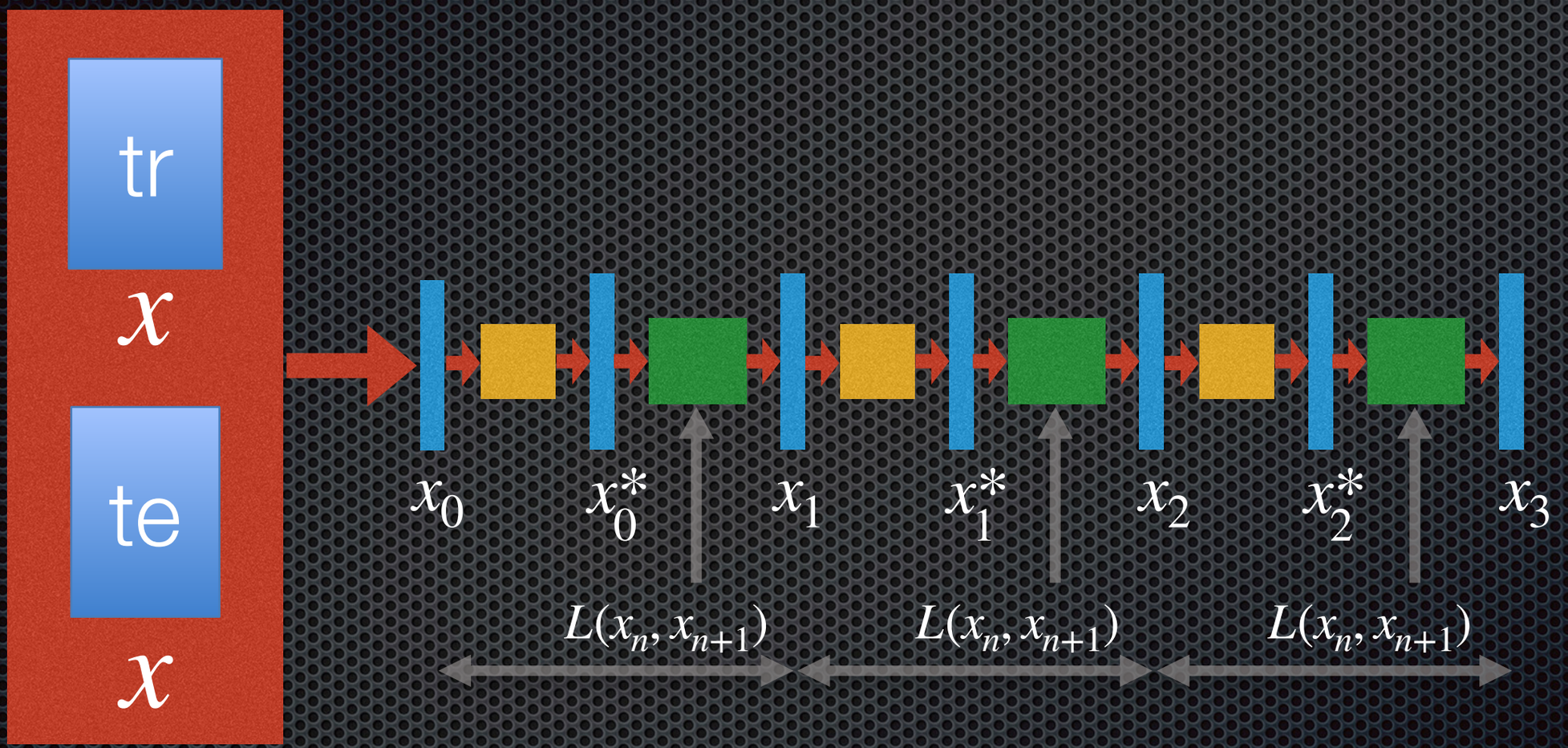
De-Noising Autoencoders



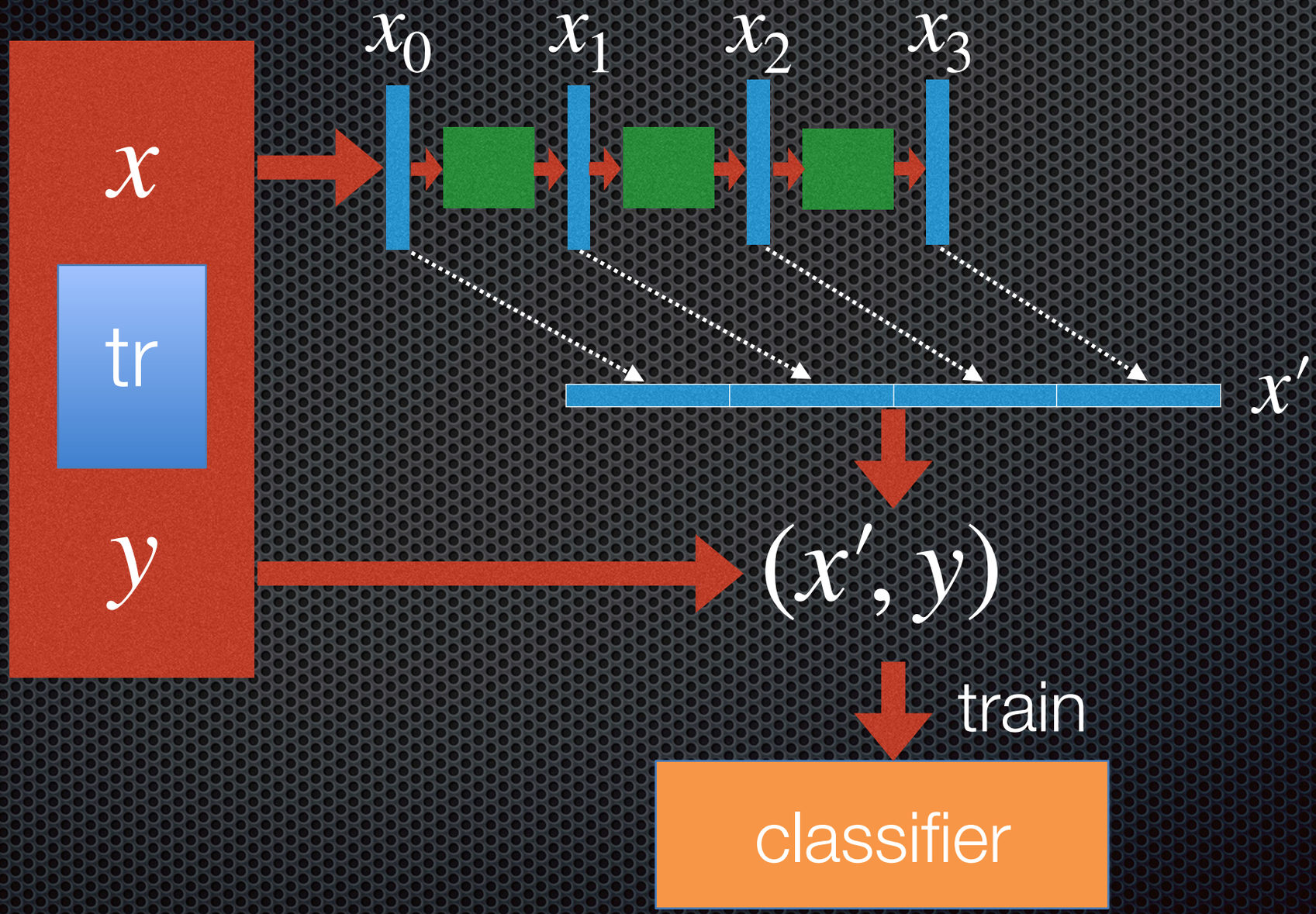
Stacked De-noising Autoencoders (SDA)



Step 1: Train Autoencoders

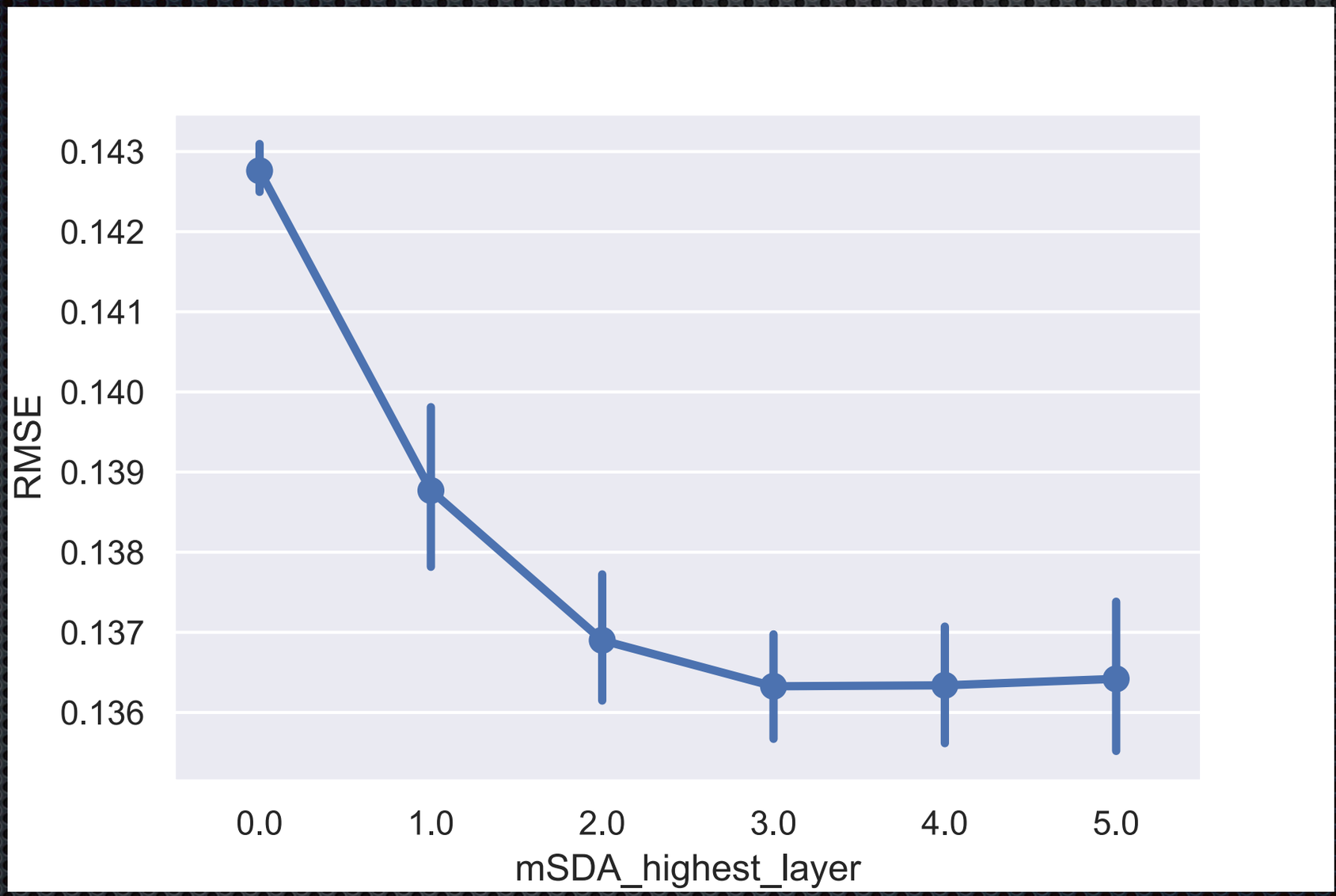


Step 2: Train Classifier



VooDoo

SDA Feature Representation



Further Work?

- Adversarial Neural Networks
- Direct quantification with NNs
- Alternative distance measures

Take Aways

- At least do Foreman's AC for quantification
- Try mSDA for domain adaptation
- Consider biassing your test sets

Thank You!

- david.spence@me.com