Lecture 19. Review

CS 109A/AC 209A/STAT 121A Data Science:

Harvard University

Fall 2016

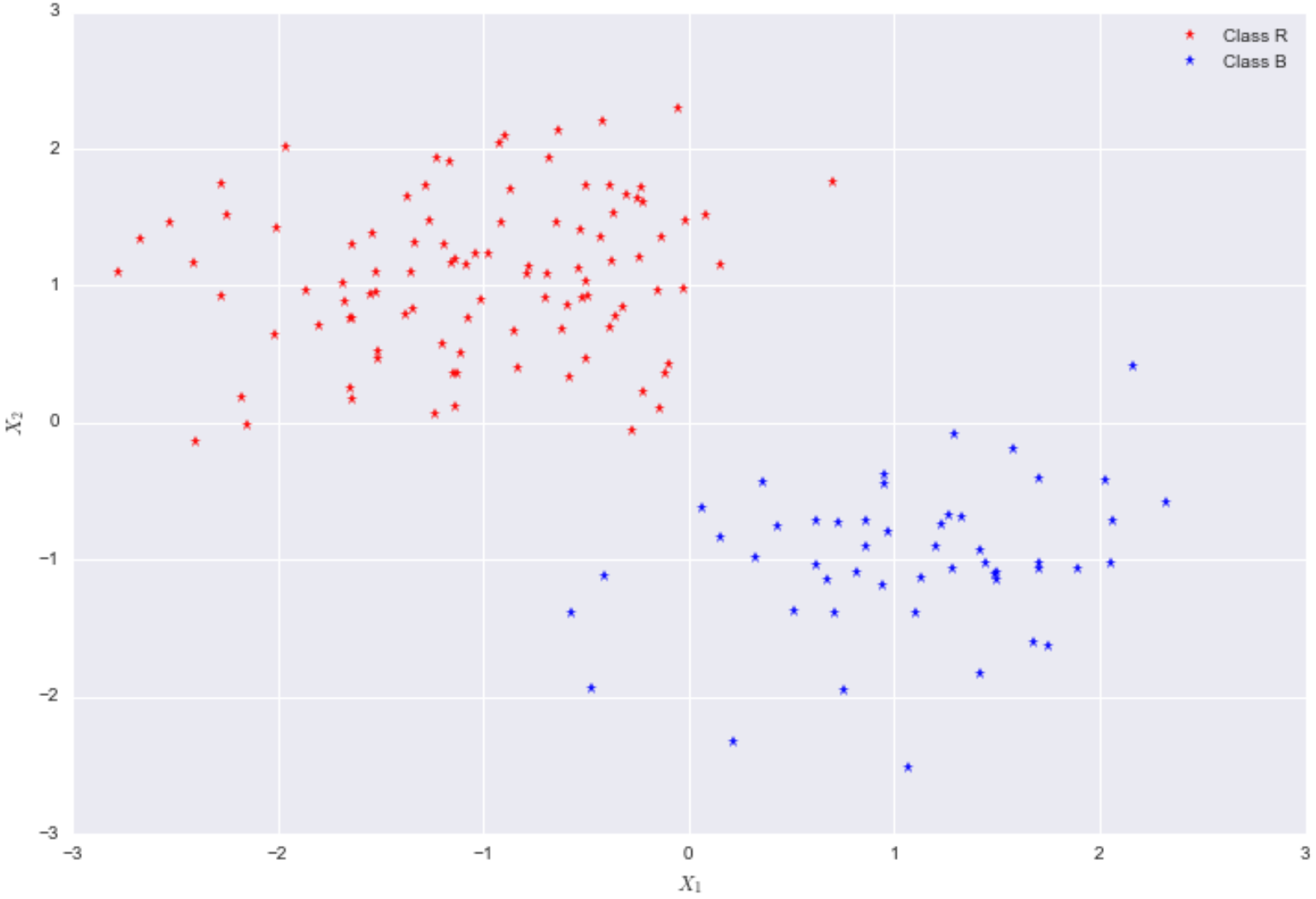
Instructors: P. Protopapas, K. Rader, W. Pan

Announcements

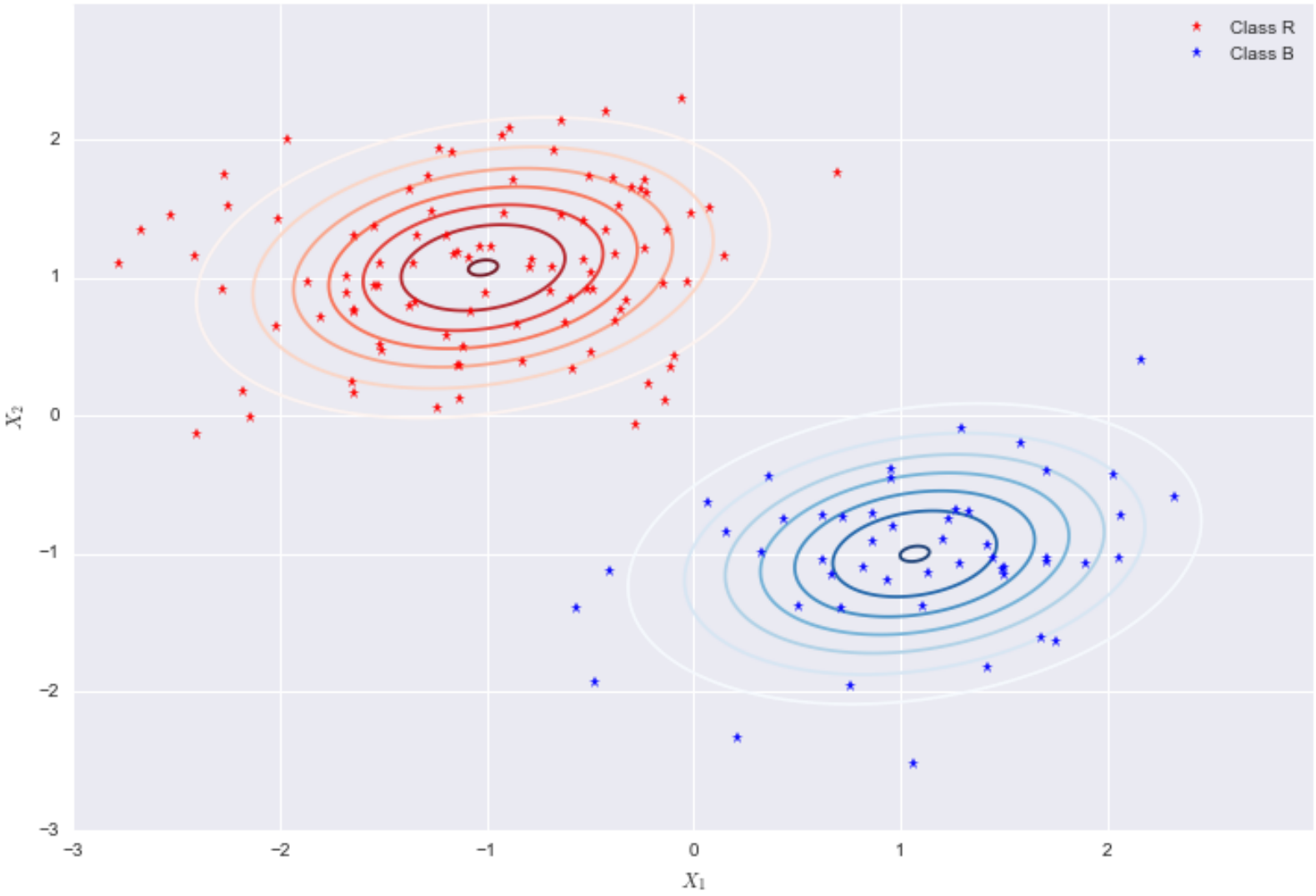
- Lecture on Wednesday is on Experimental Design and more reviews
- Labs on Thursday and Friday will be reviews on material for midterm #2
- No quizzes

Monday Nov. 28: Guest Lecture by Jeff Palmucci, Director, Machine Intelligence at TripAdvisor

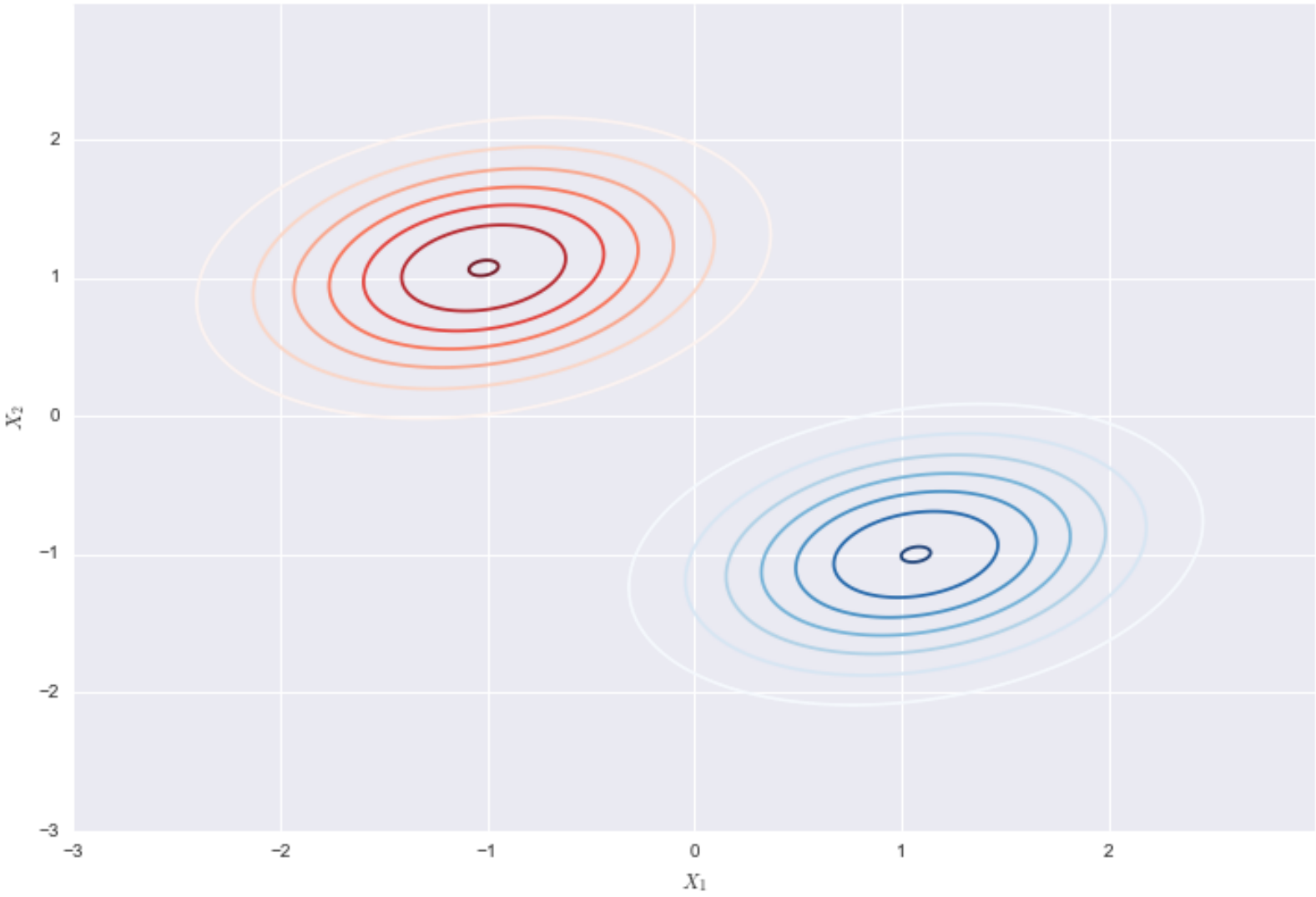
Start with some data



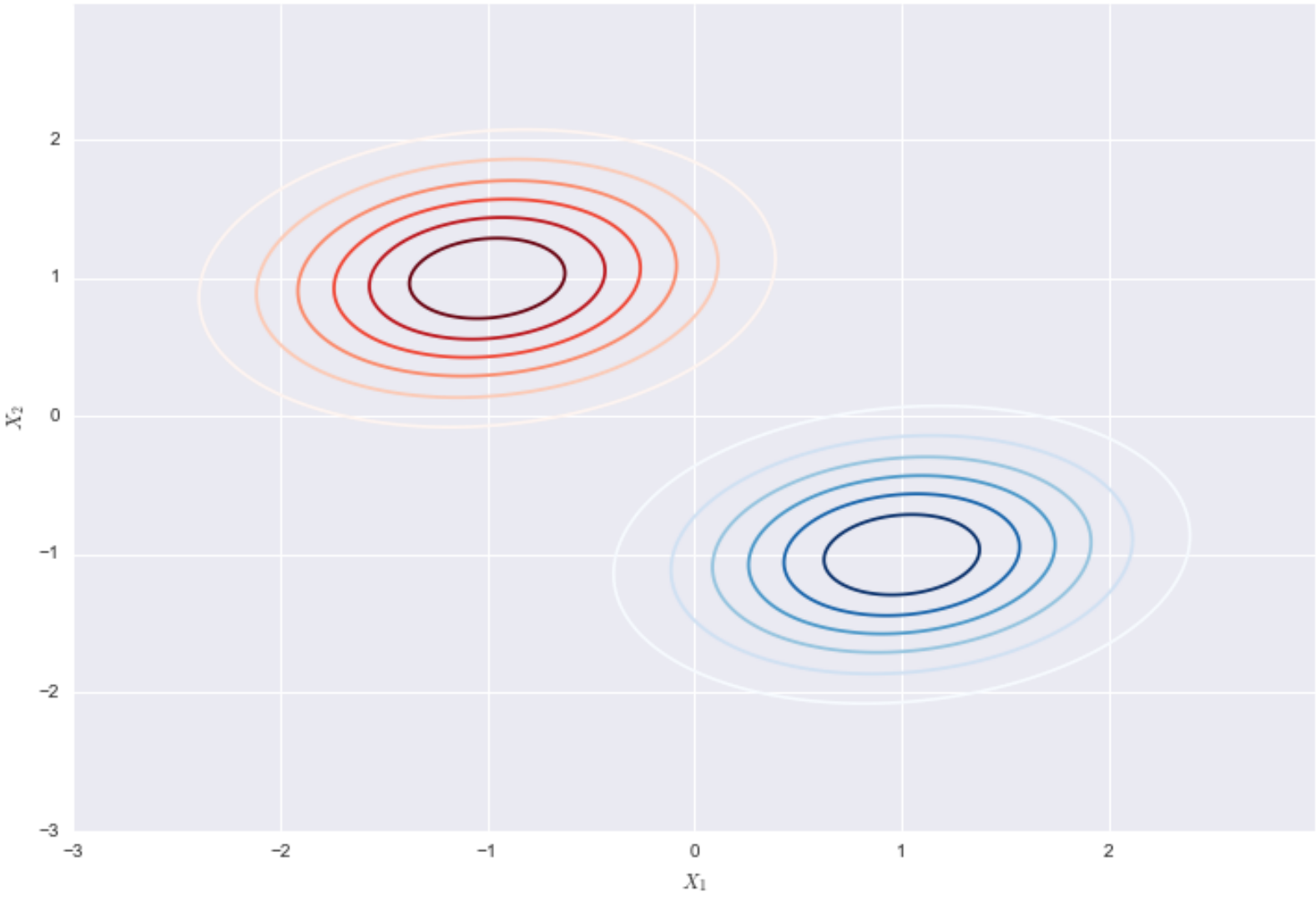
Calculate mean and covariances for each class assuming Normal distributions.



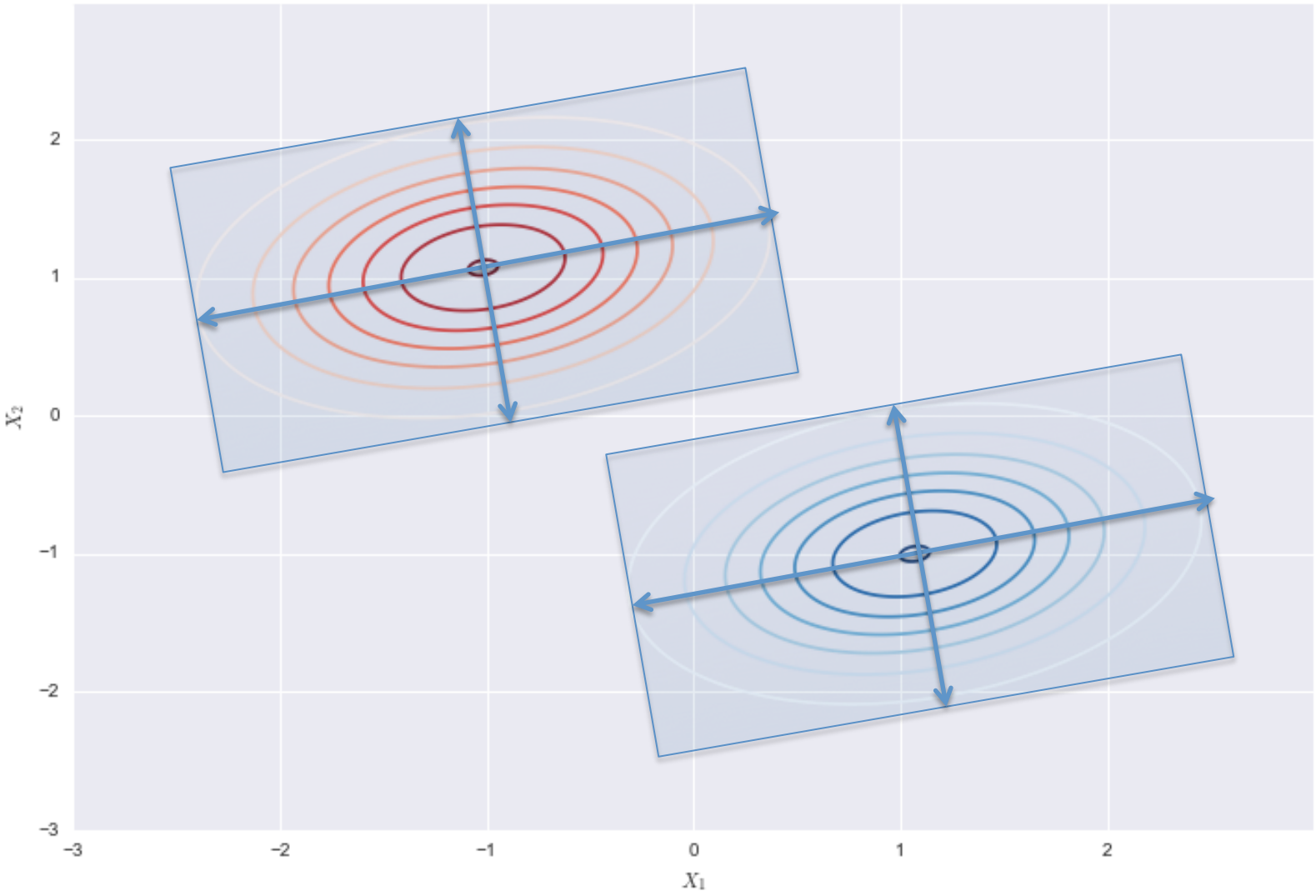
Now forget training data



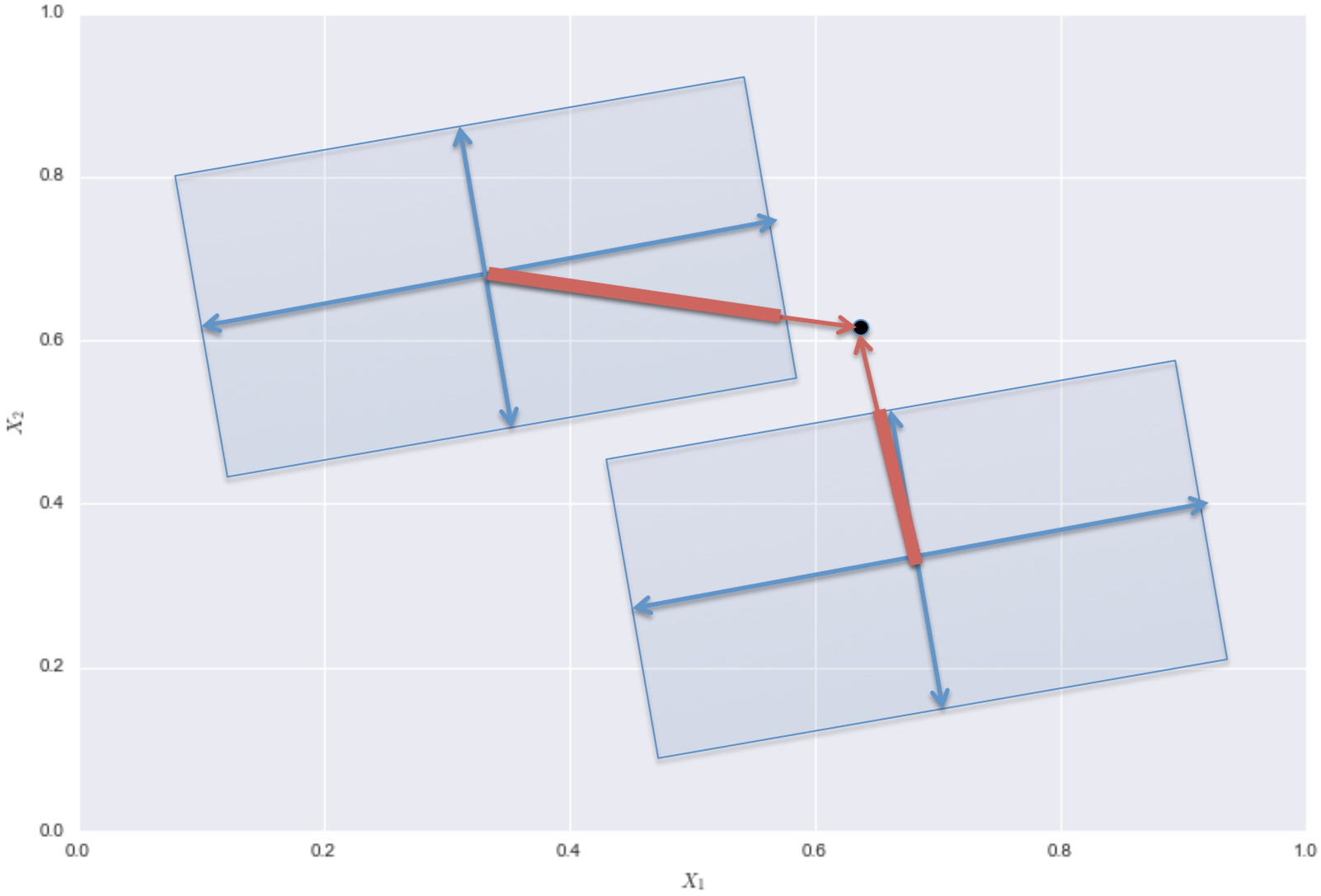
Original distributions' mean and cov for each class.

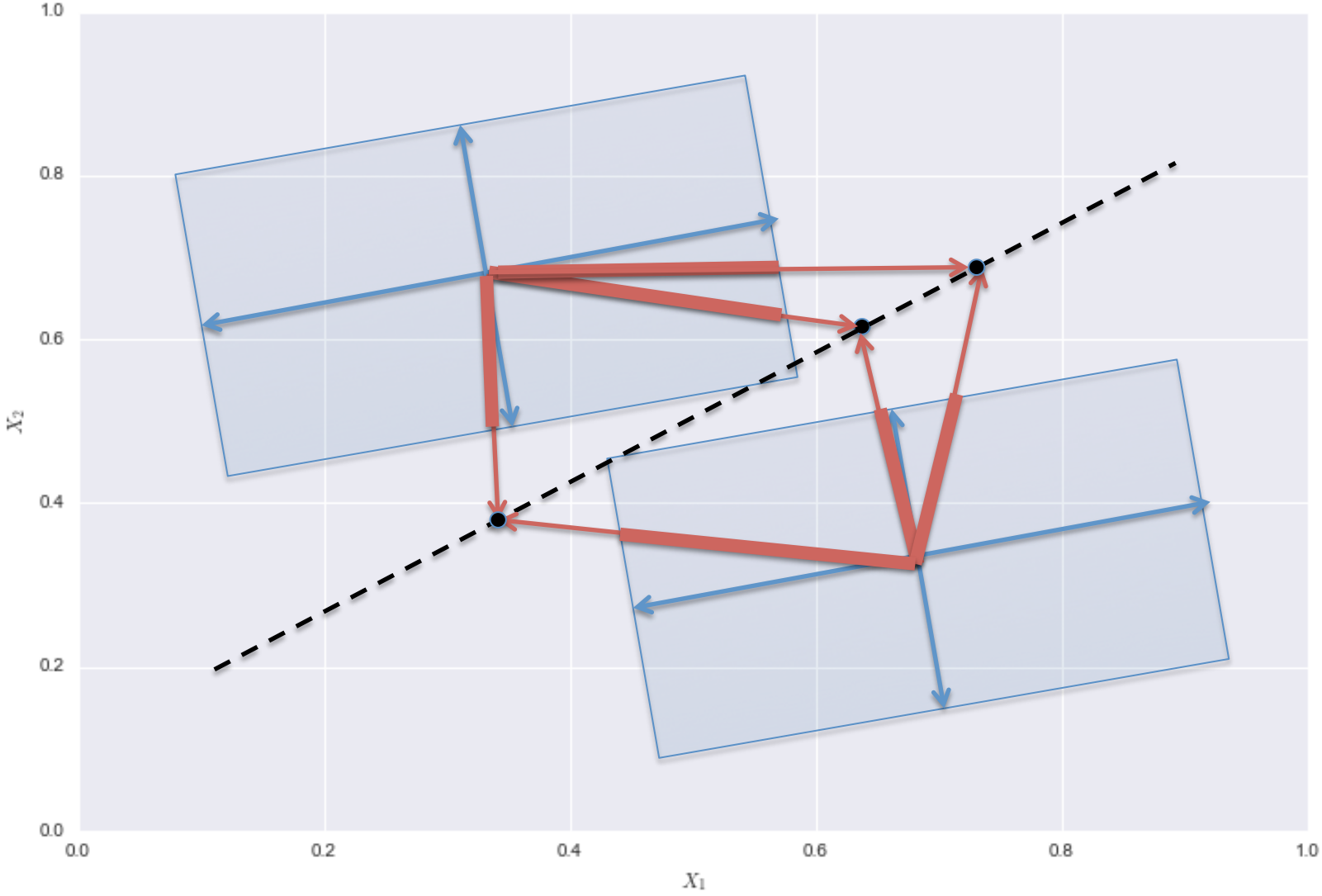


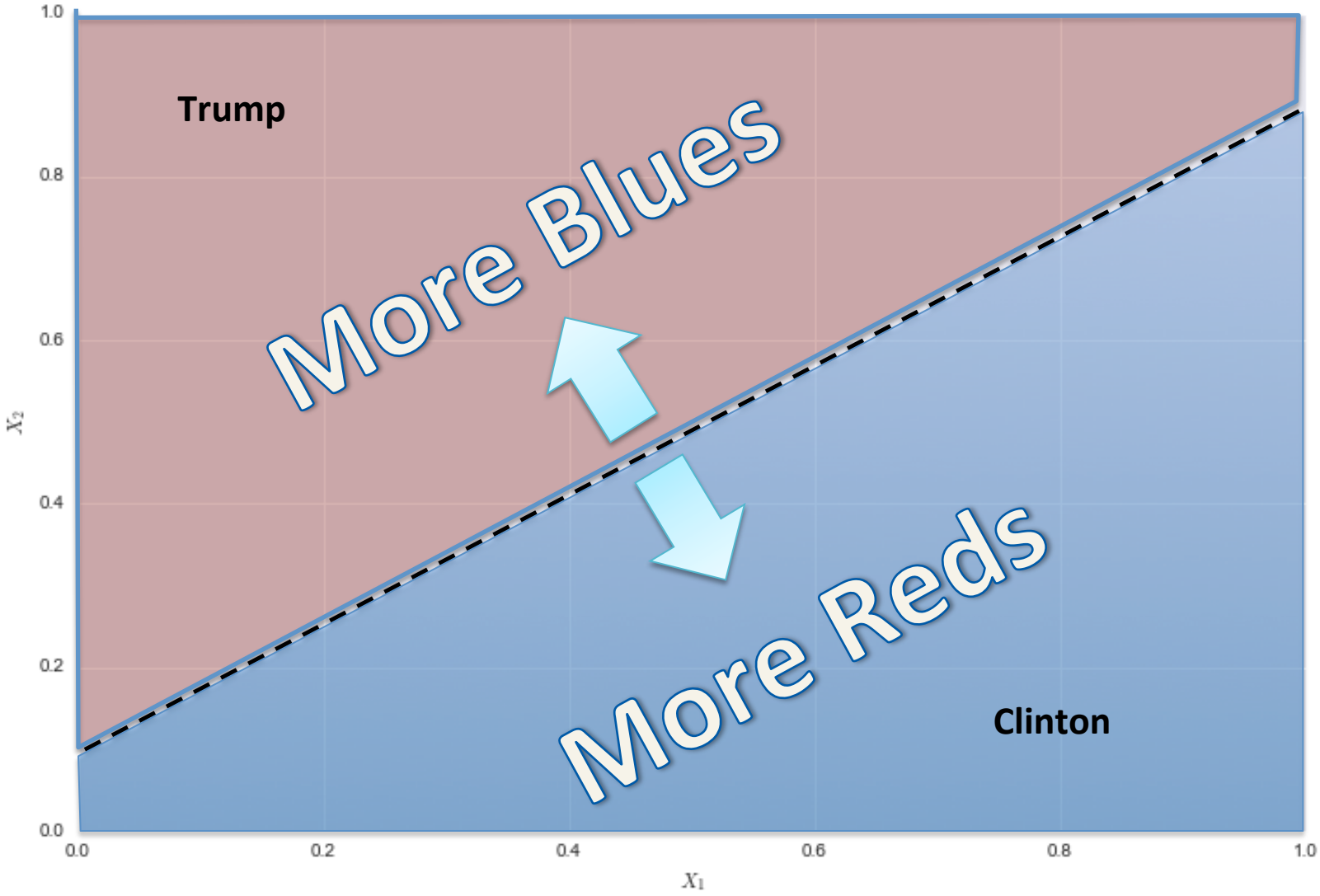
We only care for means and covariances.



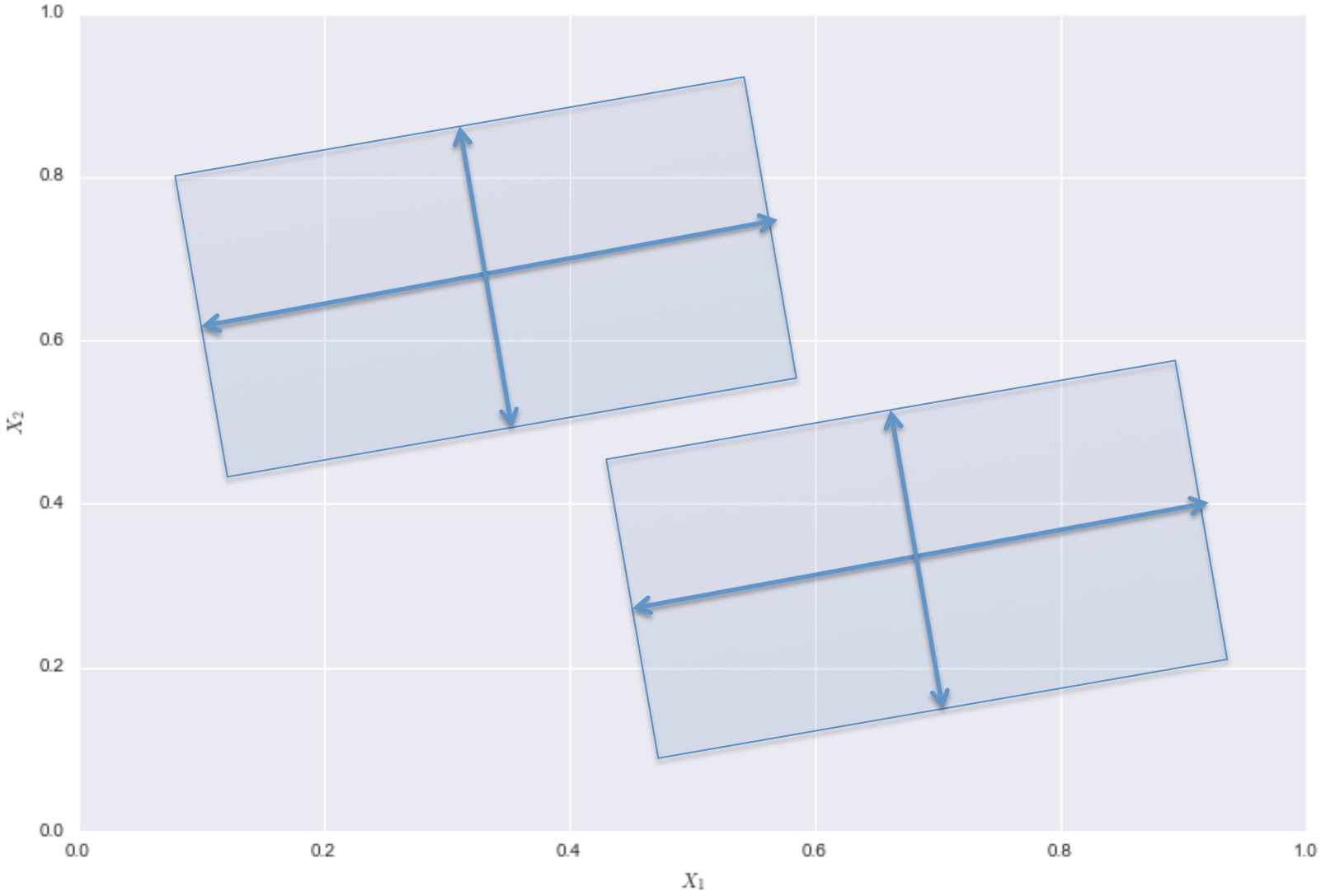
Prediction: measure distance from point in units of cov

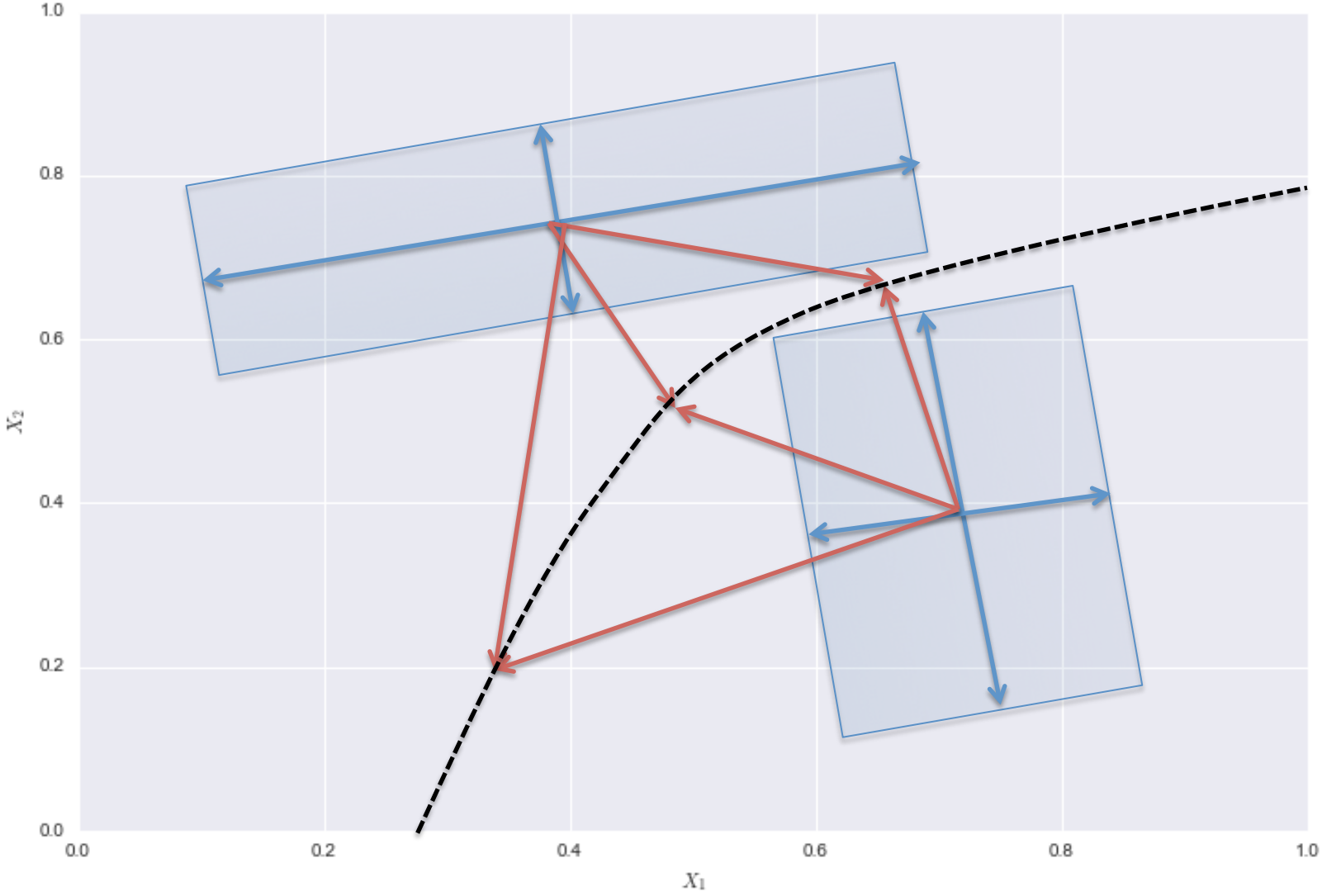



1.0
0.8
0.6
0.4
0.2
0.0
X_2
0.0
0.2
0.4
0.6
0.8
1.0
X_1


Trump
More Blues
More Reds
Clinton
1.0
0.8
0.6
0.4
0.2
0.0
X_2
0.0
0.2
0.4
0.6
0.8
1.0
X_1

Not same Σ: QDA




1.0
0.8
0.6
0.4
0.2
0.0
X_2
0.0
0.2
0.4
0.6
0.8
1.0
X_1

When do we need to regularize?

- Overfitting
- Colinearity
- Lasso/Ridge/Random Sampling/Dropout
- Model Selection

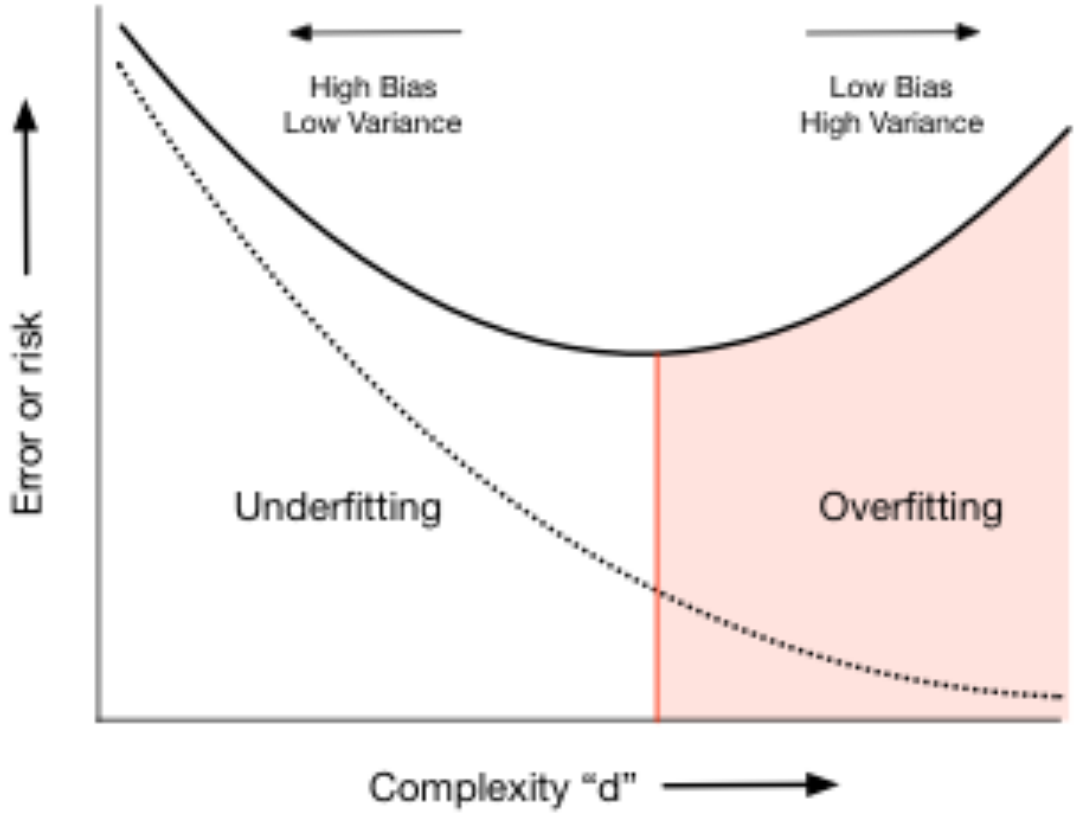
when, what, how?

Overfitting

How do we know we overfit?

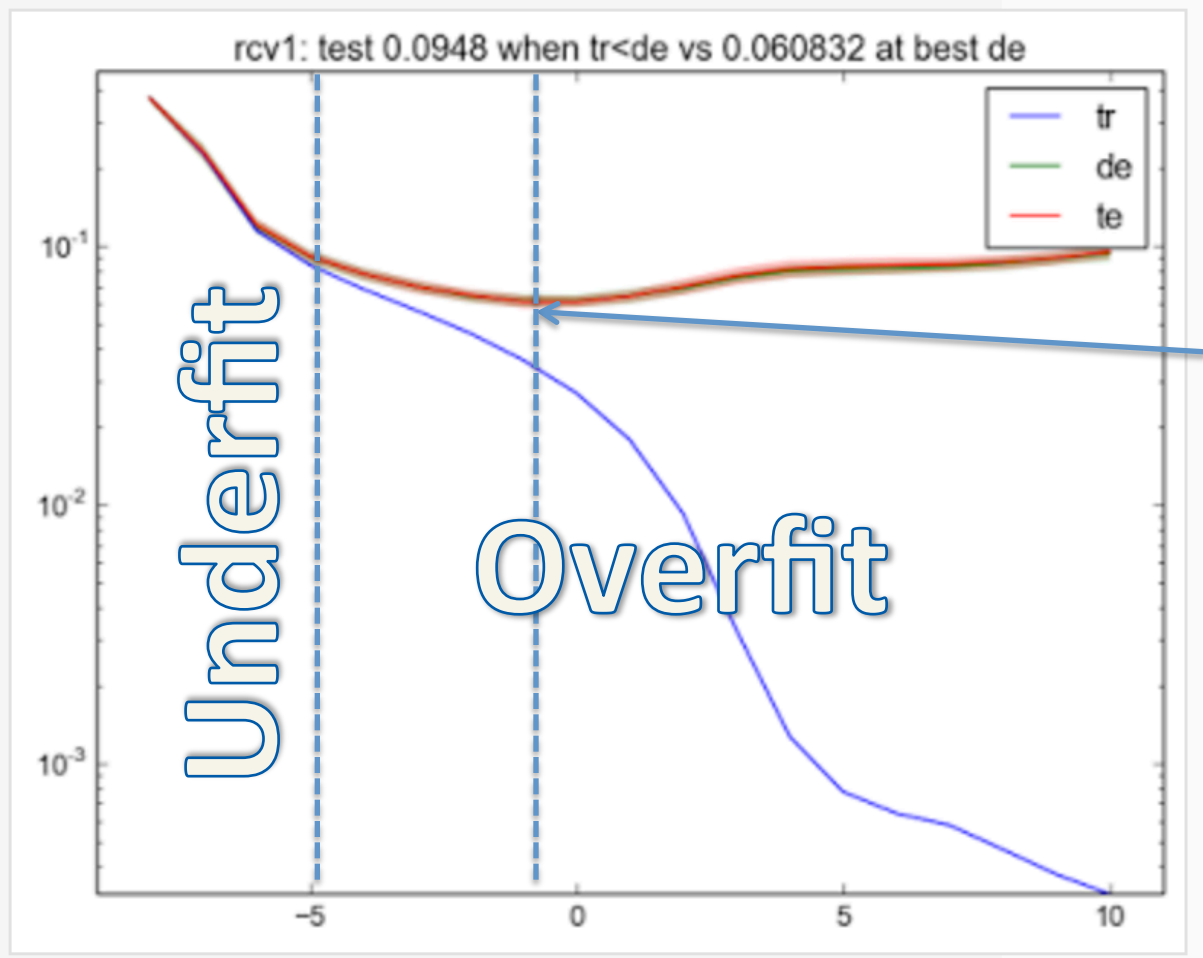
- Test error is rising as a function of some tuning parameter that captures the flexibility of the model
- A model has overfit the training data when its test error is larger than its training error
- Model is too complex
  - How many parameters vs how many training points
- Variance of the training or testing error is too large
- Trade-off bias and variance

Test error is rising as a function of some tuning parameter that captures the flexibility of the model



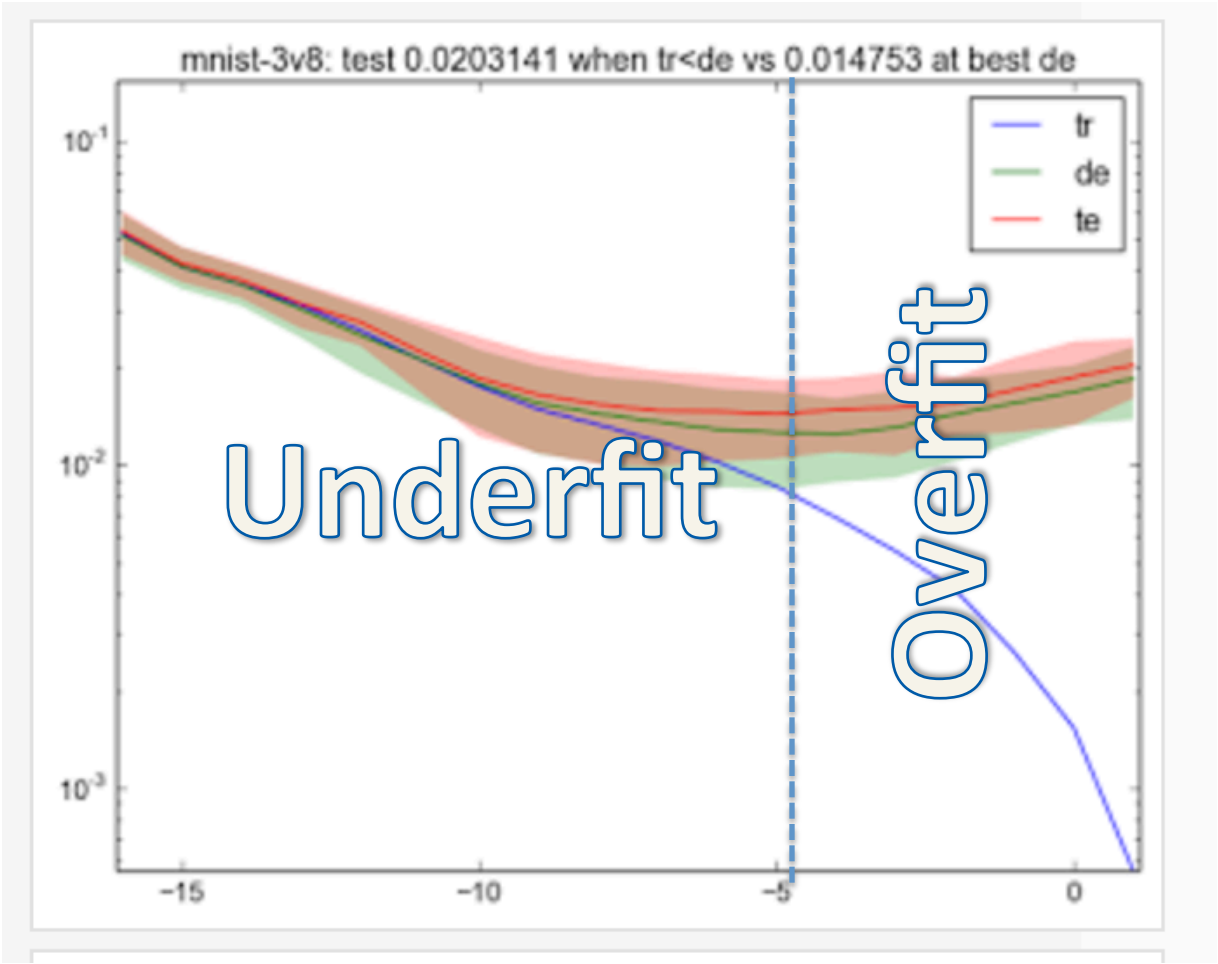
This looks good and it is almost everywhere but the problem with this definition is that it's not testable for a single model. That is, I hand you a trained model f, can you tell me if it's overfit. Do have to sweep hyperparameters (which may or may not be swappable)?

A model has overfit the training data when its test error is larger than its training error

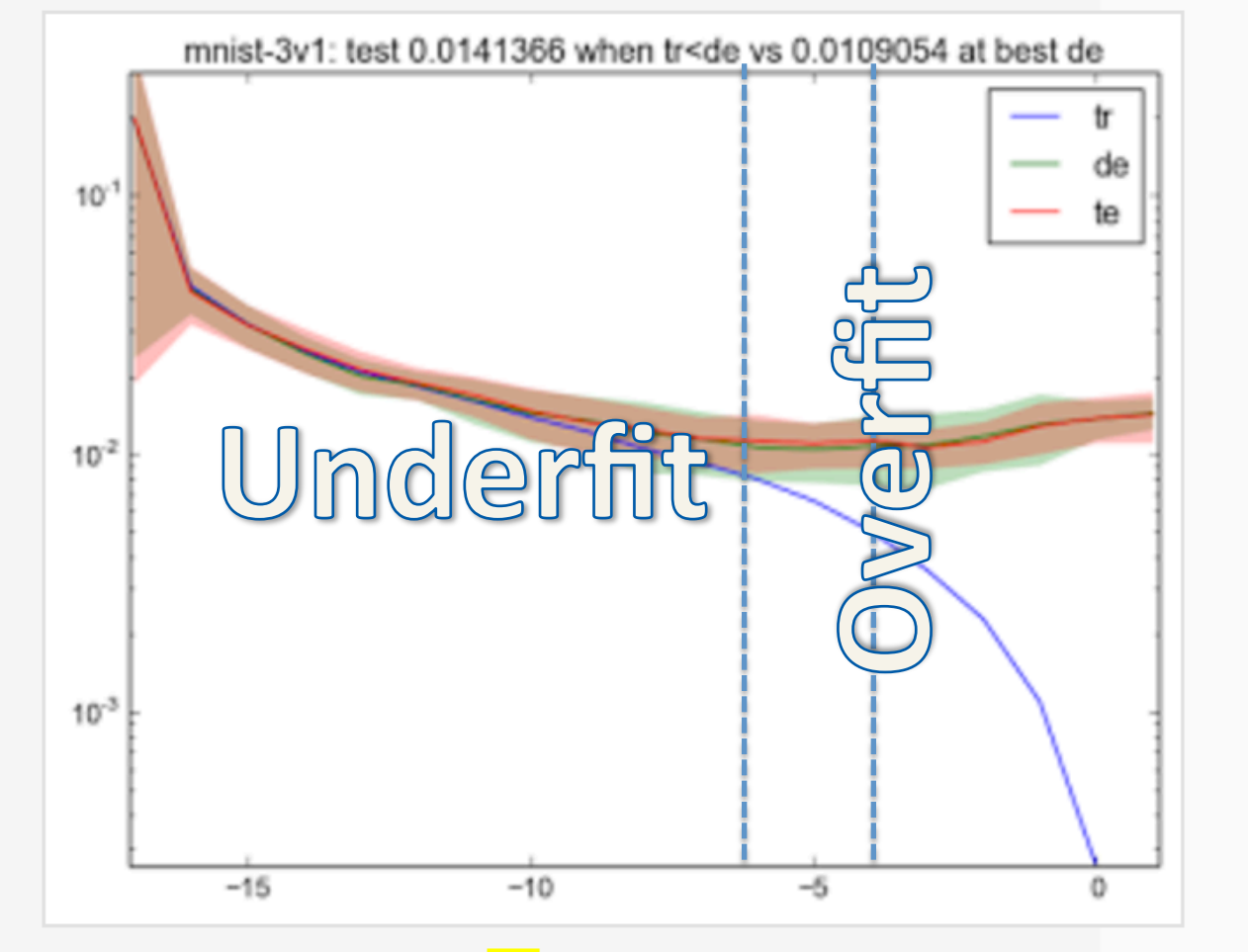


Taken from: http://nlpers.blogspot.com/2015/09/overfitting.html

A model has overfit the training data when its test error is larger than its training error



Taken from: http://nlpers.blogspot.com/2015/09/overfitting.html



Problem:
Distributions of errors are estimates and their variances are overestimated for test set and underestimated for the train set.

How not to overfit?

Method	Why	How
Linear Regression	Too many predictors	Shrinkage or subset selection
Linear Regression	Degree of polynomial	Shrinkage or subset selection
KNN	Small K	Control K
Logistic Regression	Same as linear regression	Same as linear regression just in the odds space
QDA	Different covariance matrices	Control of how much common cov and individual cov
Decision Trees	Depth of the tree	Control max_depth, max_num_cells via pruning
Boosting	# iteration (B)	Control B or learning rate.

Model selection using AIC/BIC VS Regularization

We know that the more predictors, or more terms in the polynomial we have, the more likely is to overfit.

Subset selection using AIC/BIC just removes terms with the hope that overfitting goes away (makes sense). No testing on training, no testing on overfitting (no matter the definition of overfitting)

Shrinkage methods on the other hand hope that by restricting the coefficients we can avoid overfitting.

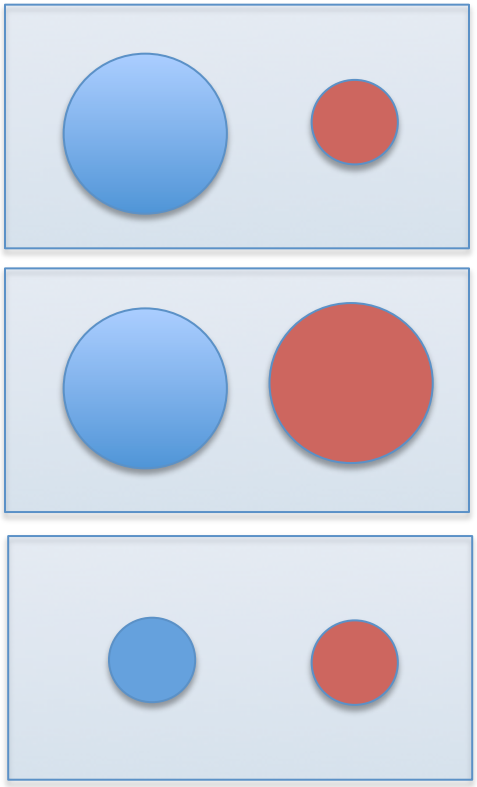
At the end: We ignore overfitting and just find the best model on the test set.

Imbalanced Data

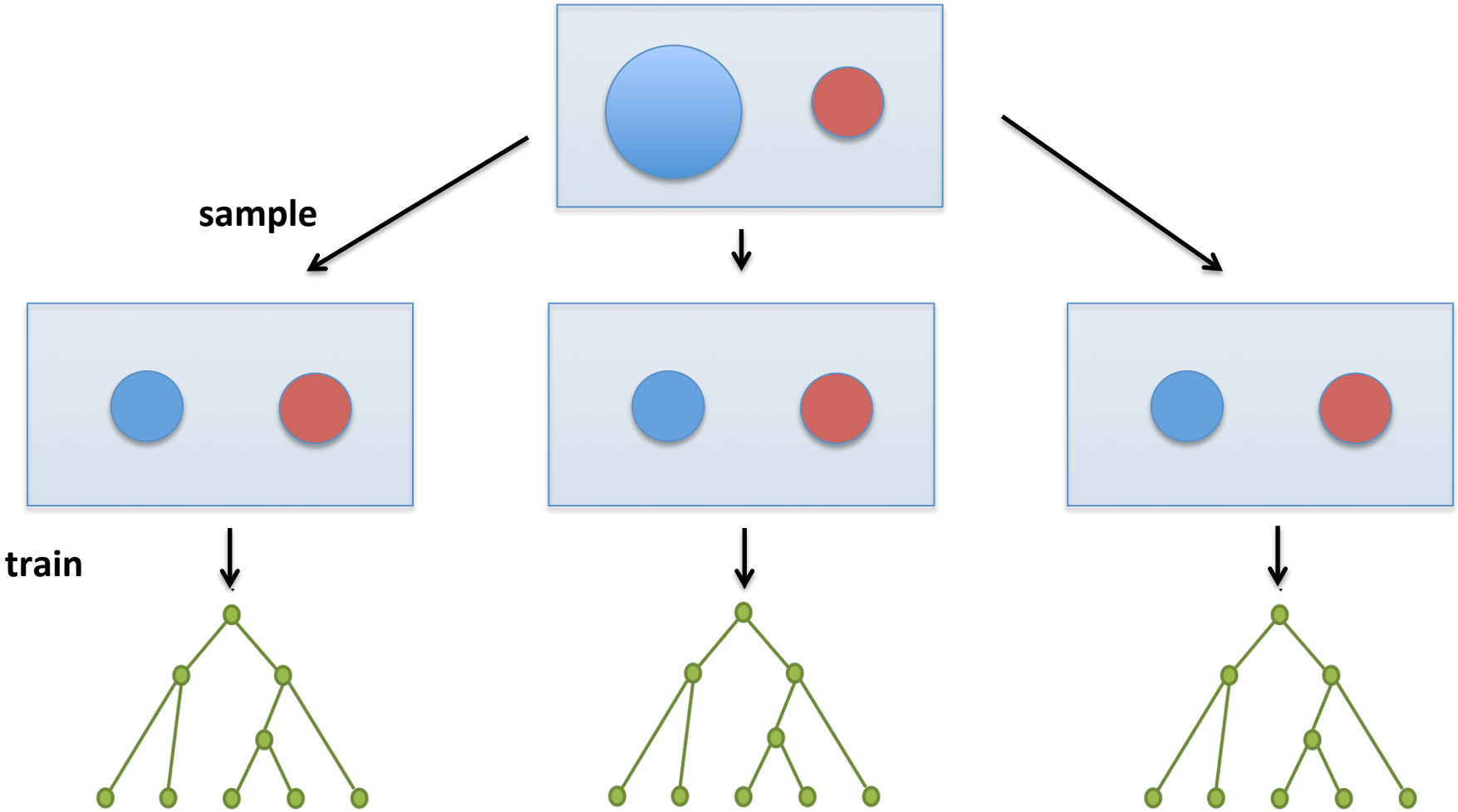
- Subsample
- Oversampe
- Re-weight sample points
- Use clustering to reduce majority class
- Re-calibrate classifier output

Imbalanced Data

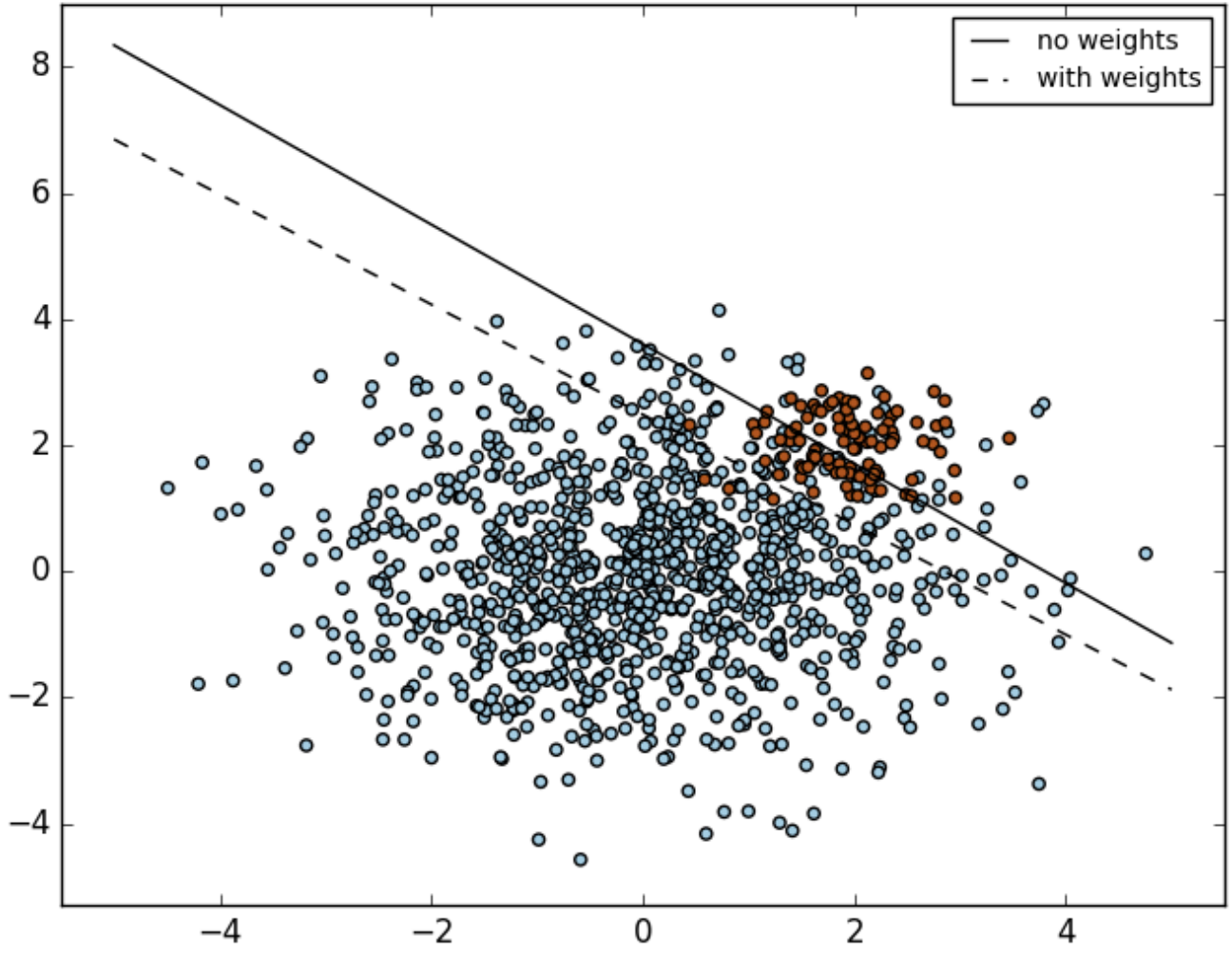
- The problem:
- Oversample:
- Subsample



Example: Random Forest Subsampling



Class weights



http://scikit-learn.org/stable/_images/sphx_glr_plot_separating_hyperplane_unbalanced_001.png