Inference vs Prediction	Inference is understanding the relationship between the X_i's and Y to answer underlying questions, prediction is finding Y from the X_i's when it is unknown.
Classification vs Regression	Classification problems (logistic regression) produce a categorical response, regression problems (linear regression) produce a quantitative response.
Bias-Variance Tradeoff	Bias and variance are inversely correlated; as one increases, the other decreases. High bias indicates that the model is too simple for the data (underfitting) and high variance indicates the prediction changes significantly with data different than the data used to create the model (overfitting).
Linear Regression	Intercept, slopes, error term (ε), interaction terms, confidence intervals, hypothesis testing (t-statistic, p-value), residual sums of squares (SSE), total sums of squares (TSS), forward selection, backward selection, Mallows's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted r^2, categorical (dummy) variables
Classification	Logistic regression (log odds, logit), multinomial regression, maximum likelihood estimation, Bayes' Theorem, decision boundary, ROC curve (higher area under the curve corresponds to higher correlation)
- KNN	Binary classification type that chooses a point X_0 and assigns it to the class that a majority of the K nearest points are part of, where K is chosen. This process is repeated for all X_i, and all remaining ties are broken in favor of one class or the other (tied points do not appear in both classes per the Bayes classifier).

- LDA	Linear discriminant analysis uses Bayes' Theorem to assign points to classes, with $f_k(x)$ varying with the number of predictors p. The discriminant score $\delta_k(x) = \log(p_k(x))$. If $p = 1$, $f_k(x)$ is the normal distribution.
- QDA	Quadratic discriminant analysis is linear discriminant analysis with $p = 2$ and differing covariance matrices.
- Naive Bayes	Assumes independent variables are independent in each class
Cross Validation	Training error vs. test error, validation set vs. k-fold cross-validation
- LOOCV	K-fold cross-validation with $k = n$. This can potentially result in high variance (overfitting).
- Validation set approach	K-fold cross-validation with $k = 2$.
- Holdout Test set	Partitioned set of data that is left out of model-fitting process (validation set in validation set approach, part i in k-fold cross-validation).
- K-fold cross validation	Data are divided into k parts, part 1 is left out while the other parts are fit into the model, process is repeated while leaving out parts 2, 3, … , k.
Bootstrapping	Bootstrapping is taking B samples of size n with replacement (each individual observation in the data can be chosen once, more than once, or not at all) for some large value of B. The mean of each of the samples Z^{*1}, Z^{*2}, … , Z^{*B} are considered individual observations and form an approximately normal distribution that can be interpreted with a histogram or boxplot.
- Bootstrap percentile	Confidence interval of the means generated by bootstrapping that can be found with a histogram.
Model selection	Subset selection, tuning parameter, dimension reduction

- Forward selection	Forward selection begins with a model with no predictors and adds the predictor with the greatest additional improvement to the fit to the model. Predictors are added one-by-one until all of the predictors are in the model.
- Backward selection	Backward selection begins with a model with all predictors and removes the least useful predictor. Predictors are removed one-by-one (generally until all of the remaining predictors are statistically significant in the model).
- Best subset selection	Best subset selection fits every single possible model (all p models with one predictor, all $p(p\text{-}1)/2$ models with two predictors, etc.) and chooses the model with the smallest SSE (largest r^2) for each number of predictors, leaving $p+1$ models (p models, one for each number of predictors, plus the null model with no predictors). The single best model is chosen from these $p+1$ models based on lowest cross-validated prediction error, lowest Mallows's C_p, lowest Akaike information criterion (AIC) value, lowest Bayesian information criterion (BIC) value, and highest adjusted r^2 (this choice is subjective and may vary with differing interpretations).
- Ridge regression	SSE + $\lambda(\Sigma_{j=1}^{p} \beta_j^2)$, where λ is a tuning parameter selected by cross-validation and $\lambda(\Sigma_{j=1}^{p} \beta_j^2)$ is a shrinkage penalty.
- LASSO	Least absolute shrinkage and selection operator; SSE + $\lambda(\Sigma_{j=1}^{p} \lvert\beta_j\rvert)$, where λ is a tuning parameter selected by cross-validation and $\lambda(\Sigma_{j=1}^{p} \lvert\beta_j\rvert)$ is a shrinkage penalty.
- Elastic net	Elastic net

- PCR	Principal component regression; first principal component is a straight line along which the observations vary the most.
- PLS	Partial least squares