GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

ENSEMBLE METHODS & RANDOM FORESTS

LAST TIME:

Table 4.1. The vertebrate data set.

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

I. DECISION TREES

II. LAB ON DECISION TREES

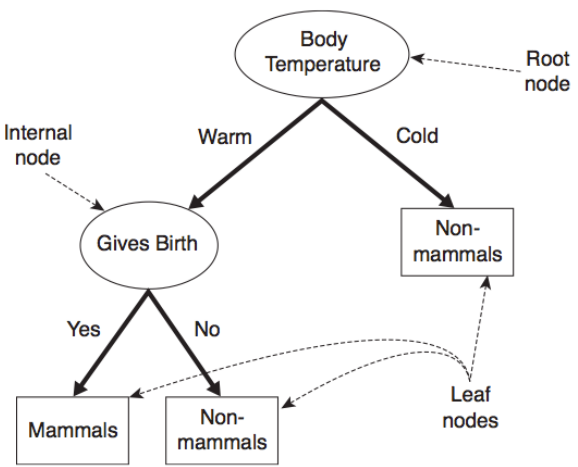


Figure 4.4. A decision tree for the mammal classification problem.

QUESTIONS?

WHAT WAS THE MOST INTERESTING THING YOU LEARNED?

WHAT WAS THE HARDEST TO GRASP?

I. ENSEMBLE TECHNIQUES
II. BOOTSTRAP & BAGGING
III. RANDOM FORESTS

LAB:
IV. RANDOM FOREST IN SCIKIT-LEARN

So far, we have only discussed individual classifiers

What are these?

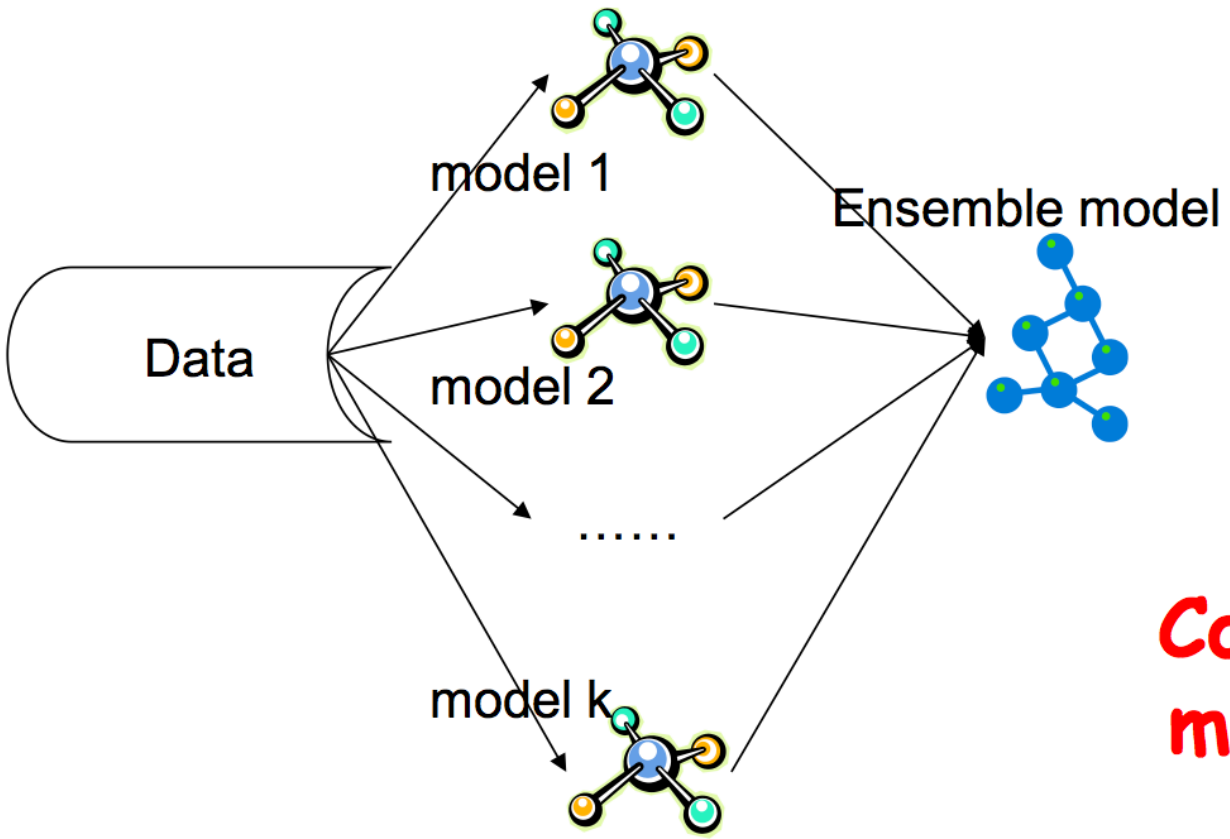
What is your favorite?

-or-

What do you think is Justin's favorite?

How can we come up with a better predictive model?

Can we combine multiple classifiers to produce a better prediction (of class membership)?



Combine multiple models into one!

source: http://www.cse.buffalo.edu/~jing/sdm10ensemble.htm

MOTIVATING EXAMPLE


2009
NETFLIX
DATE 09.21.09
PAY TO THE ORDER OF BellKor's Pragmatic Chaos
$ 1,000,000 00
AMOUNT ONE MILLION
00/100
FOR The Netflix Prize
Reed Hastings

award $1 million to anyone who can improve movie recommendation by 10%

Supervised learning task

– Training data is a set of users and ratings (1,2,3,4,5 stars) those users have given to movies.

source: http://www.cse.buffalo.edu/~jing/sdm10ensemble.htm

Supervised learning task

– Construct a classifier that given a user and an unrated movie, correctly classifies that movie as either 1, 2, 3, 4, or 5 stars

source: http://www.cse.buffalo.edu/~jing/sdm10ensemble.htm

At first, single-model methods are developed, and performances are improved

source: http://www.cse.buffalo.edu/~jing/sdm10ensemble.htm

However, improvements slowed down

source: http://www.cse.buffalo.edu/~jing/sdm10ensemble.htm

Later, individuals and teams merged their results, and significant improvements are observed

source: http://www.cse.buffalo.edu/~jing/sdm10ensemble.htm

Leaderboard

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
12	BellKor	0.8624	9.46	2009-07-26 17:19:11
Progress Prize 2008 - RMSE = 0.8627 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8642	9.27	2009-07-15 14:53:22
14	Gravity	0.8643	9.26	2009-04-22 18:31:32
15	Ces	0.8651	9.18	2009-06-21 19:24:53
Progress Prize 2007 - RMSE = 0.8725 - Winning Team: KorBell				
Cinematch score - RMSE = 0.9525				

"Our final solution (RMSE=0.8712) consists of blending 107 individual results. "

"Predictive accuracy is substantially improved when blending multiple predictors. Our experience is that most efforts should be concentrated in deriving substantially different approaches, rather than refining a single technique. "

source: http://www.cse.buffalo.edu/~jing/sum10ensemble.htm

I. ENSEMBLE METHODS

Q: What are ensemble techniques?

A: Methods of improving classification accuracy by aggregating predictions over several **base classifiers**.

Ensembles are often much more accurate than the base classifiers that compose them.

In order for an ensemble classifier to outperform a single base classifier, the following conditions must be met:

1) *the bc's must be* **accurate***: they must outperform random guessing*

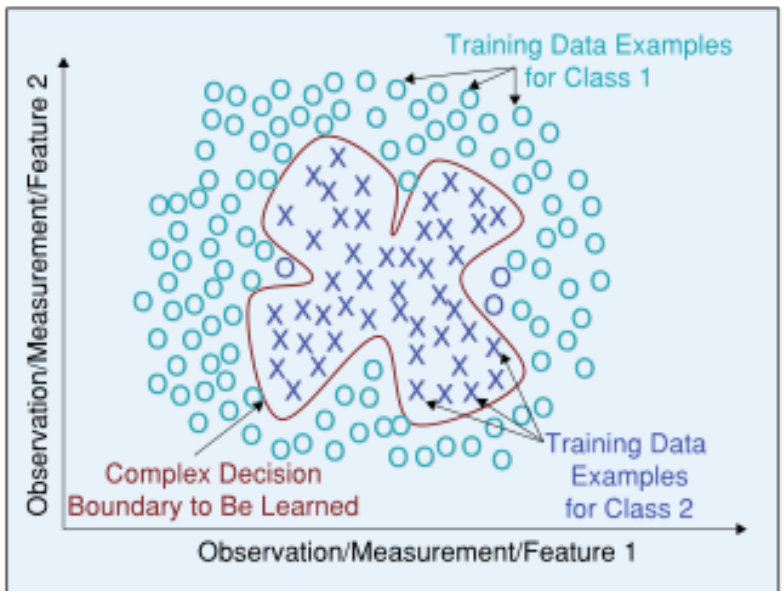
2) *the bc's must be* **diverse***: their misclassifications must occur on different training examples*

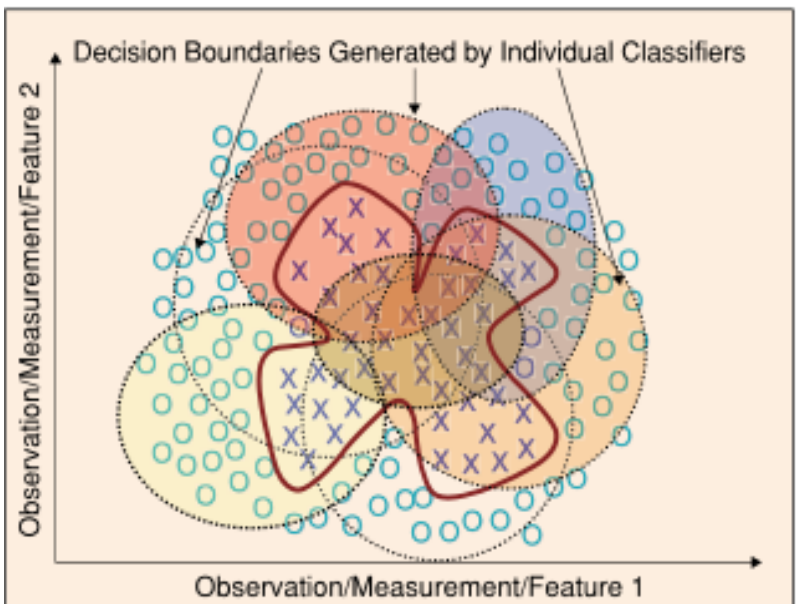
Q: How do you generate several base classifiers?

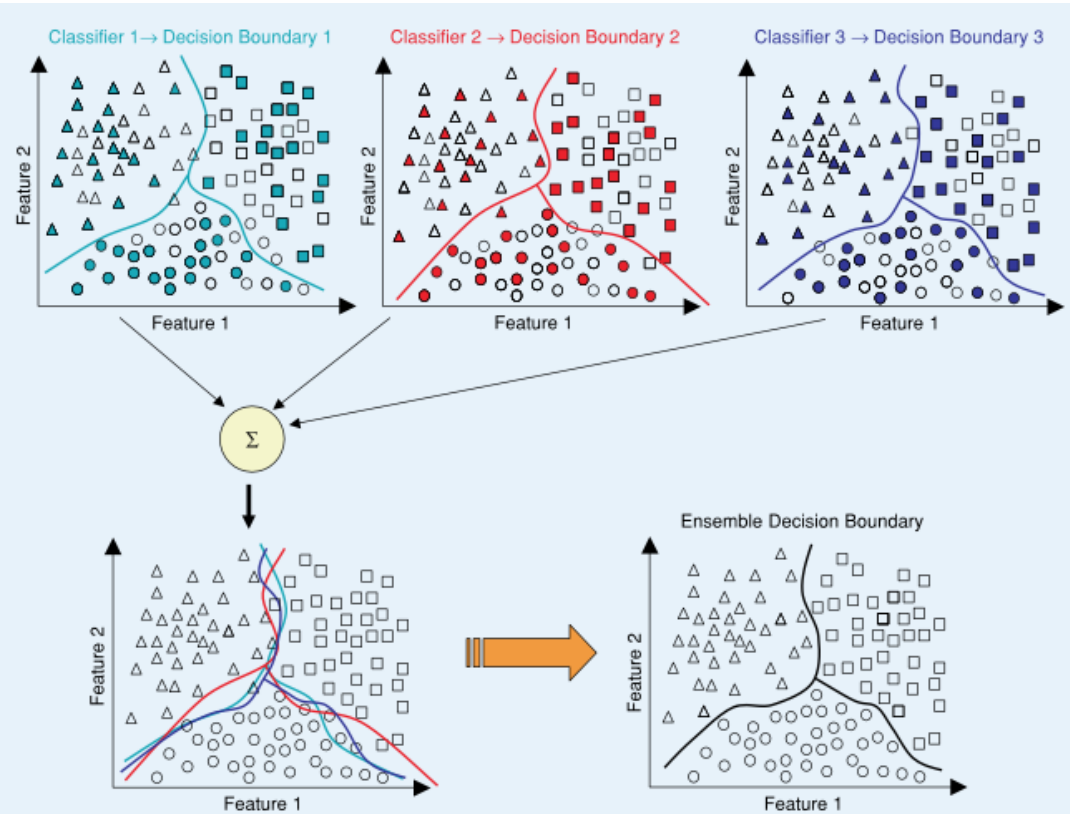
A: There are several ways to do this:

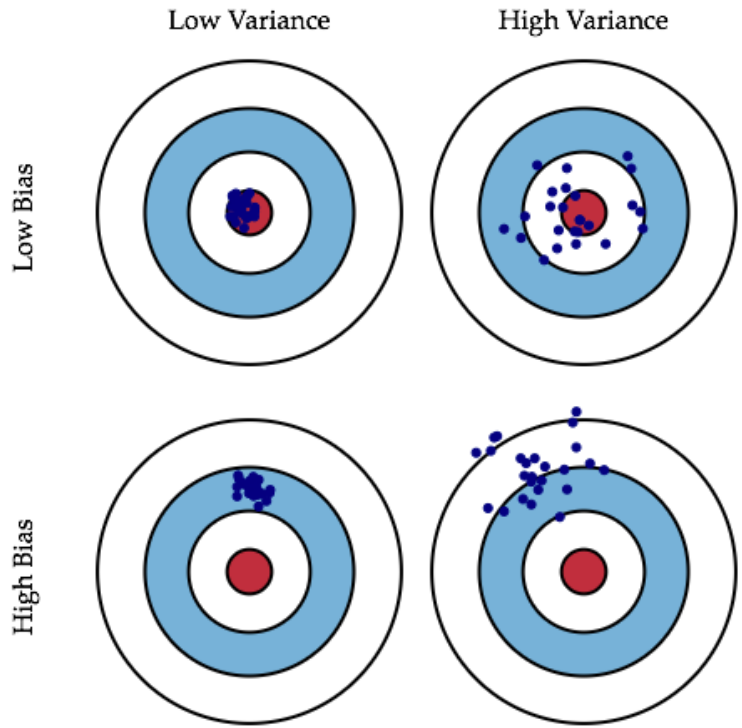
- *manipulating the training set*
- *manipulating the learning algorithm itself*

- *Reasons for using ensemble based systems:*
  - *Improved prediction accuracy, better generalization (committee)*
  - *Too little data (resampling)*
  - *Too much data (downsampling)*
  - *Complex decision boundary*


Training Data Examples
for Class 1
Observation/Measurement/Feature 2
Complex Decision
Boundary to Be Learned
Training Data
Examples
for Class 2
Observation/Measurement/Feature 1


Decision Boundaries Generated by Individual Classifiers
Observation/Measurement/Feature 2
Observation/Measurement/Feature 1


Classifier 1→ Decision Boundary 1
Classifier 2 → Decision Boundary 2
Classifier 3 → Decision Boundary 3
Feature 2
Feature 1
Σ
Ensemble Decision Boundary



- *Reduce bias*
- *Reduce variance*

- *Aggregation methods – averaging of independent estimators*
  - *Bagging (bootstrap aggregation) [Brieman, 1996]*
  - *Random Forest [Brieman, 2001]*

Reduce variance

- *Boosting methods – sequentially combined weak estimators*
  - *AdaBoost [Freud and Schapire, 1995]*
  - *Gradient Tree Boosting (GBDT, GBM) [Friedman, 1999]*

Reduce bias

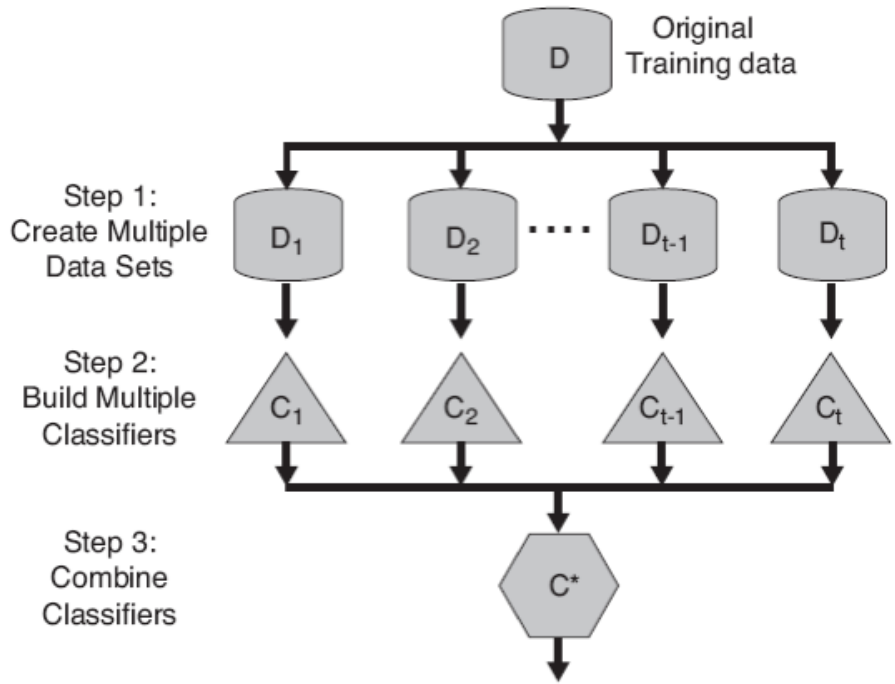
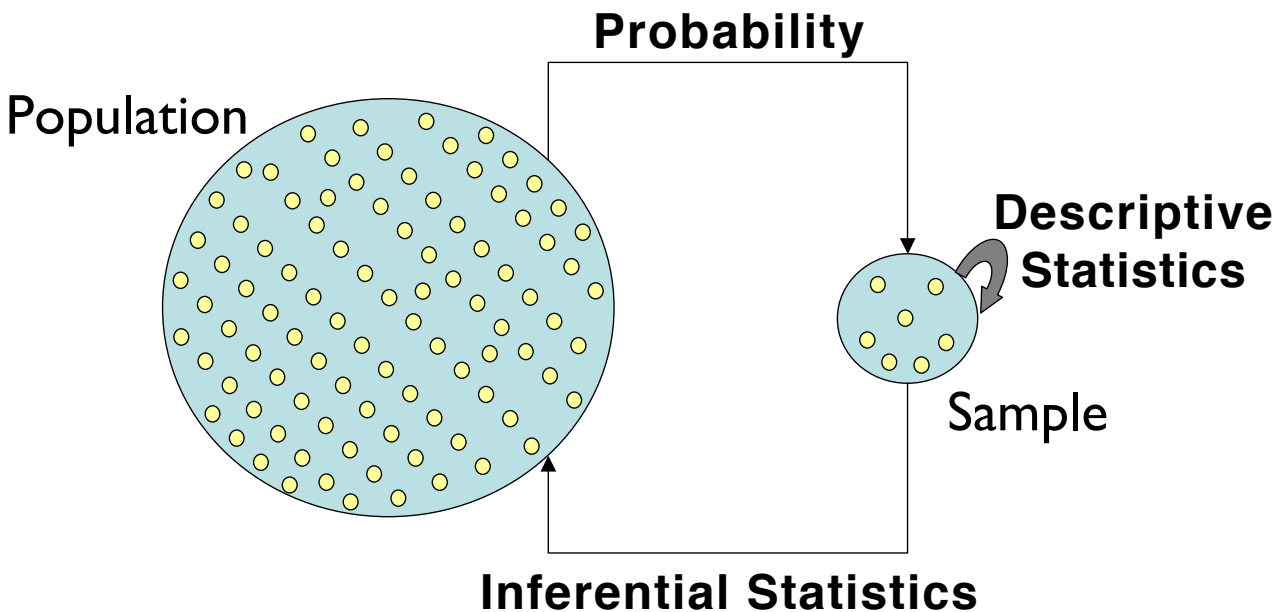


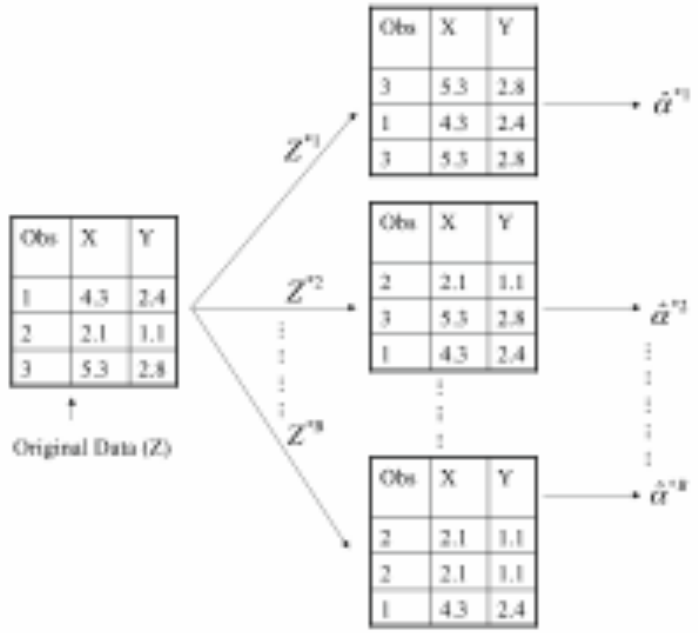
Figure 5.31. A logical view of the ensemble learning method.

II. BOOTSTRAP AND BAGGING



From http://cs109.github.io/2014/

Bootstrapping – is a random sampling with replacement (select N samples from N data points uniformly with replacement)

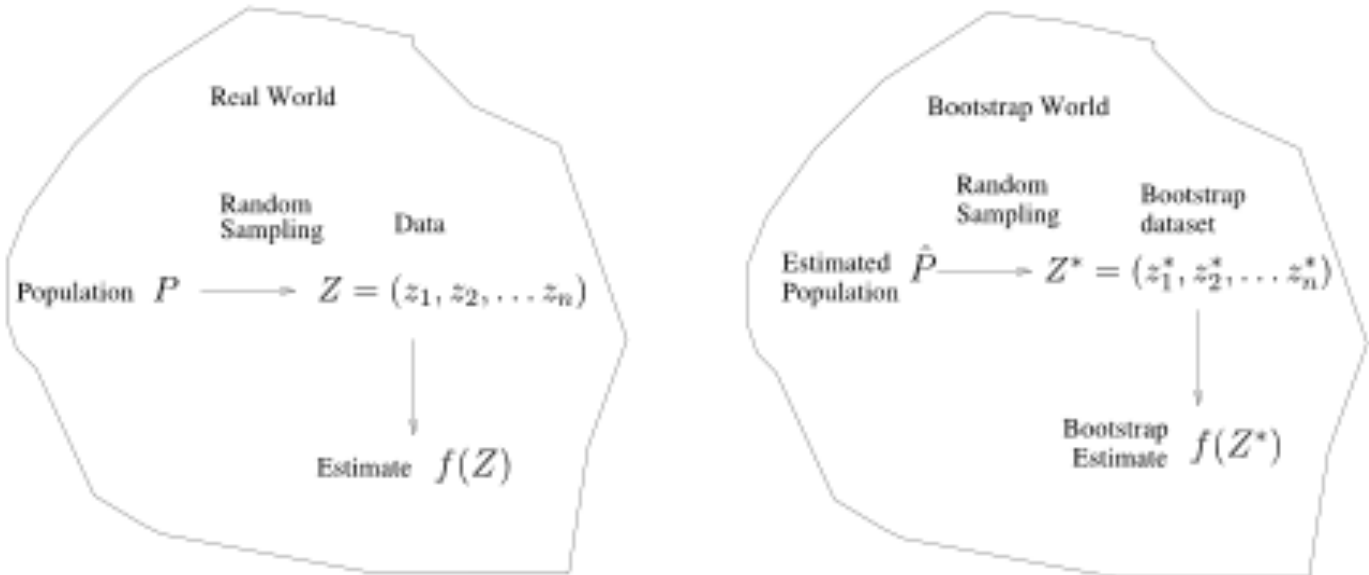


63.2% of unique examples per bootstrapped sample on average

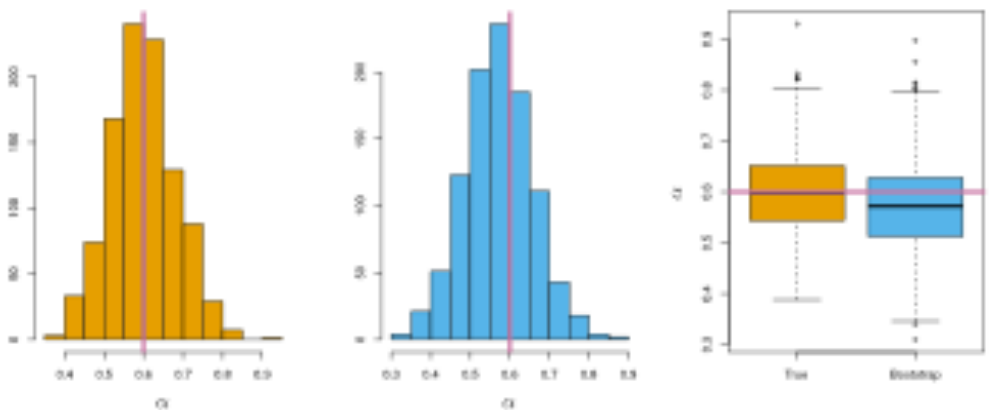
when drawing with replacement n′ values out of a set of n (different and equally likely), the expected number of unique draws is

$$n(1 - e^{-n'/n})$$

Idea: resample the sample data, sample becomes "population"



In statistics it is used to obtain standard errors on an estimate and confidence intervals



Bagging *= bootstrap aggregating*

We learn k base classifiers on k different bootstraped samples of training data.

These samples are independently created by resampling the training data using uniform weights (eg, a uniform **sampling distribution***).*

The final prediction is made by taking a majority vote across bc's.

- *N iid random variables:* $X_1, X_2, \ldots, X_n.$
- *Mean and variance:*

$$\mathrm{E}(X_i) = \mu \quad \mathrm{Var}(X_i) = \sigma^2$$

- *Average of N random variables:*

$$\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n}$$

- *Mean and variance of average:*

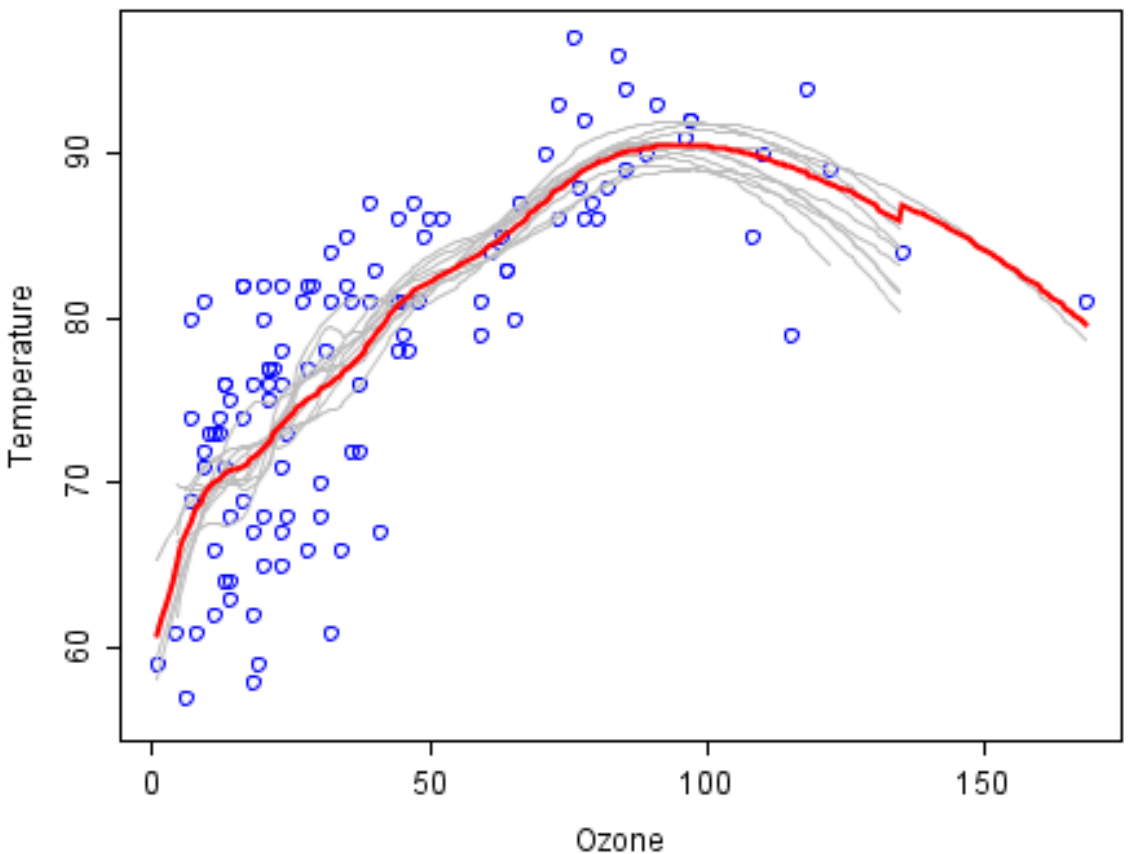
$$\mathrm{E}(\bar{X}) = \frac{1}{n}\mathrm{E}\left(X_1 + X_2 + \ldots + X_n\right) = \left(\frac{1}{n}\right)(n\mu) = \quad \mu$$

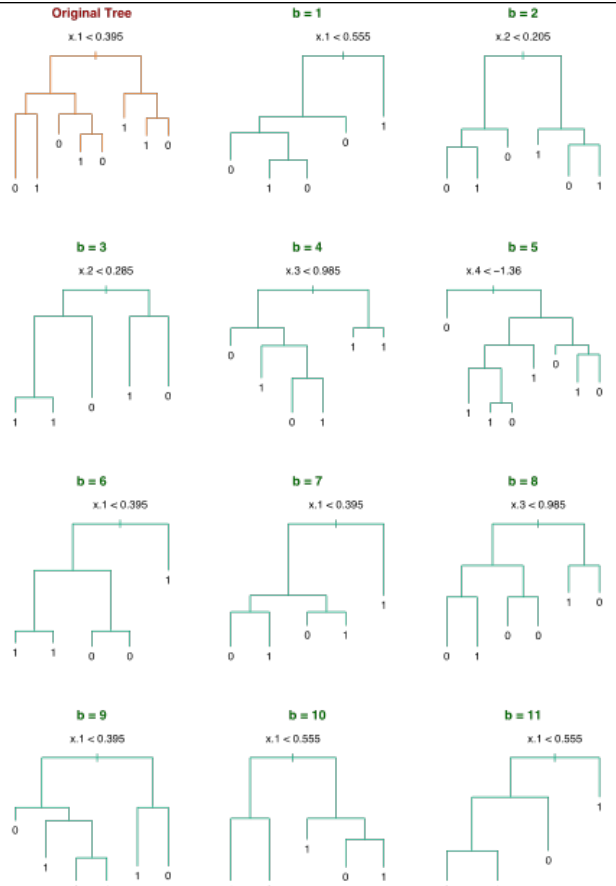
$$\mathrm{Var}(\bar{X}) = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(X_1 + X_2 + \ldots + X_n\right) = \left(\frac{1}{n}\right)^2 (n\sigma^2) = \quad \frac{\sigma^2}{n}$$

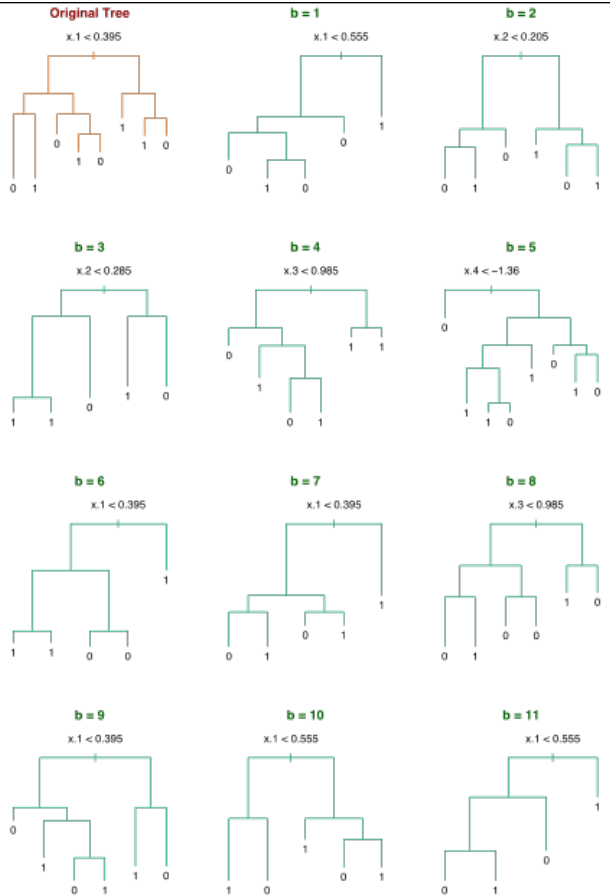
Bagging reduces the variance in our generalization error by aggregating multiple base classifiers together (provided they satisfy our earlier requirements).

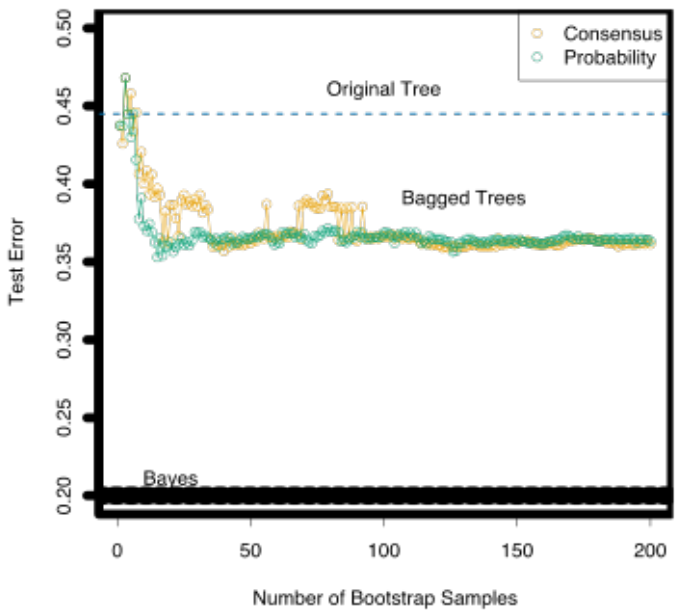
If the base classifier is stable, then the ensemble error is primarily due to bc bias, and bagging may not be effective.

Since each sample of training data is equally likely, bagging is not very susceptible to overfitting with noisy data.


Temperature
60
70
80
90
0
50
100
150
Ozone


Original Tree
x.1 < 0.395
b = 1
x.1 < 0.555
b = 2
x.2 < 0.205
b = 3
x.2 < 0.285
b = 4
x.3 < 0.985
b = 5
x.4 < −1.36
b = 6
x.1 < 0.395
b = 7
x.1 < 0.395
b = 8
x.3 < 0.985
b = 9
x.1 < 0.395
b = 10
x.1 < 0.555
b = 11
x.1 < 0.555





Variance of average of B correlated trees:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

III. RANDOM FORESTS

- *A random forest is an ensemble of uncorrelated decision trees:*
  - *Build a number of decision trees on bootstrapped training samples*
  - *When building decision trees a random selection of m out of p predictors considered for every split in a tree.*

- *Variance of average of B correlated trees :* $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$

A	B	C	D	E	F	G	H	I	J	K	L
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 31	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoorte	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075		S
26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johans	female	38	1	5	347077	31.3875		S
27	0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7.225		C
28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263	C23 C25 C27	S
29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female		0	0	330959	7.8792		Q
30	0	3	Todoroff, Mr. Lalio	male		0	0	349216	7.8958		S
31	0	1	Uruchurtu, Don. Manuel E	male	40	0	0	PC 17601	27.7208		C
32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female		1	0	PC 17569	146.5208	B78	C
33	1	3	Glynn, Miss. Mary Agatha	female		0	0	335677	7.75		Q
34	0	2	Wheadon, Mr. Edward H	male	66	0	0	C.A. 24579	10.5		S
35	0	1	Meyer, Mr. Edgar Joseph	male	28	1	0	PC 17604	82.1708		C
36	0	1	Holverson, Mr. Alexander Oskar	male	42	1	0	113789	52		S
37	1	3	Mamee, Mr. Hanna	male		0	0	2677	7.2292		C
38	0	3	Cann, Mr. Ernest Charles	male	21	0	0	A./5. 2152	8.05		S
39	0	3	Vander Planke, Miss. Augusta Maria	female	18	2	0	345764	18		S
40	1	3	Nicola-Yarred, Miss. Jamila	female	14	1	0	2651	11.2417		C

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

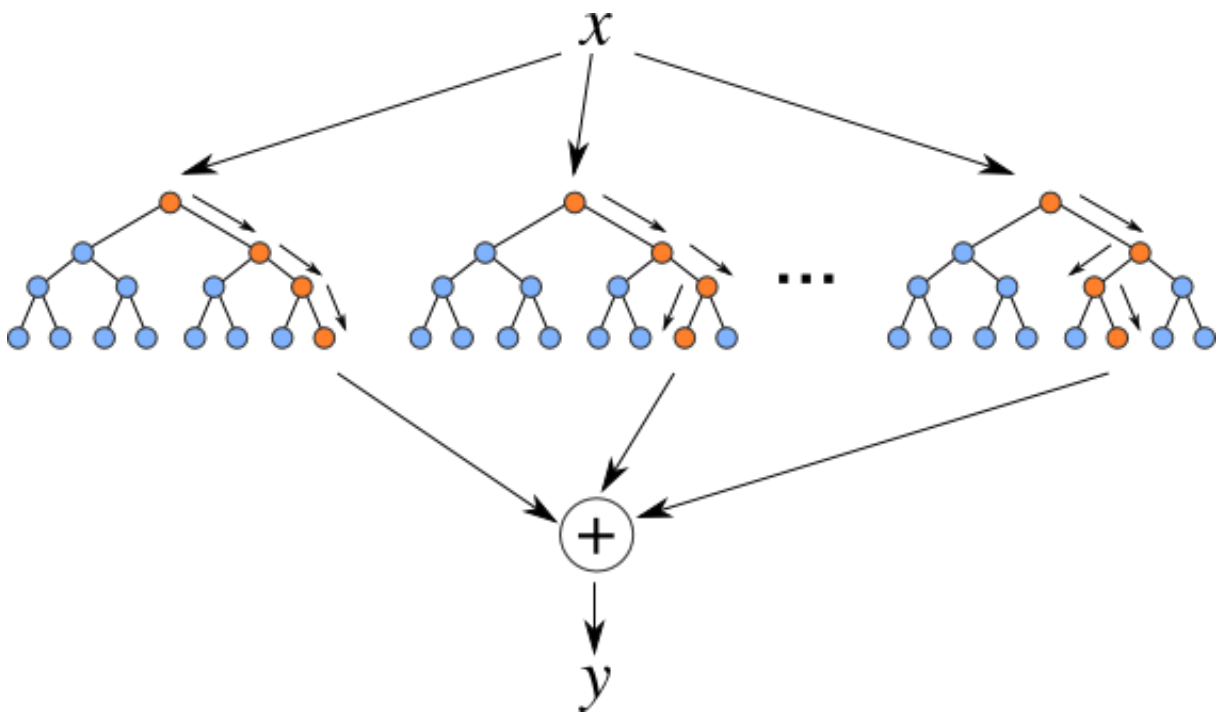
To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

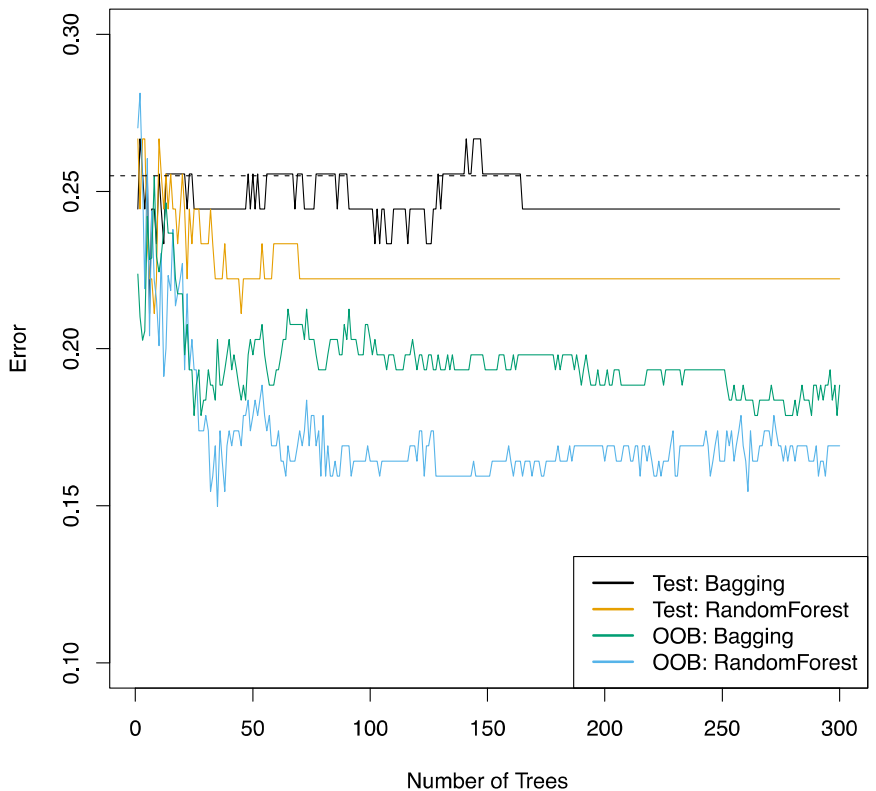
Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

- *Tuning parameters for Random Forest:*
  - *Number of predictors per split m (recommended sqrt(p))*
  - *Number of trees in the forest 100-1000*

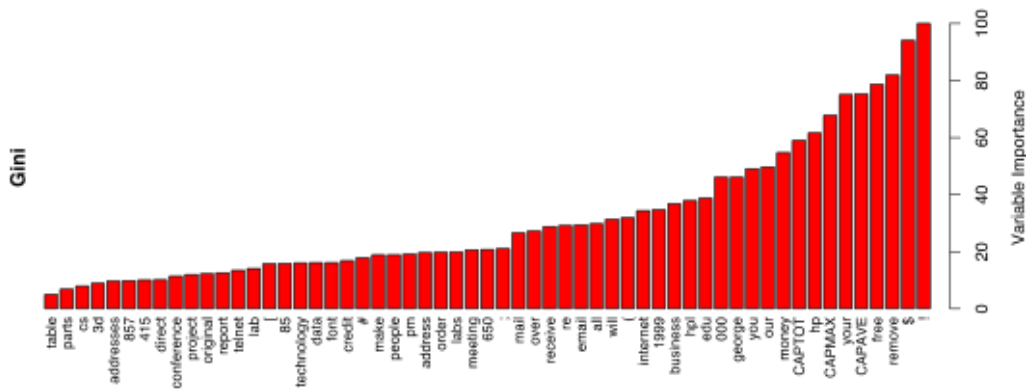
- *Grow complete trees:*
  - *Low bias for every tree (can overfit)*
  - *Low variance for ensemble through averaging*


x
...
+
y

- *Each observation is used in 62.3% trees (bootstrapping)*
- *For each observation (x,y) construct its random forest predictor by averaging only those trees in which (x,y) does not appear*
- *Very close to N-fold cross validation*
- *Random forest can be fit with entire data and perform OOB/CV performed along the way*


Error
0.10
0.15
0.20
0.25
0.30
0
50
100
150
200
250
300
Number of Trees
Test: Bagging
Test: RandomForest
OOB: Bagging
OOB: RandomForest

- *For each tree compute relative improvement of split criteria due to every variable over all splits where it is used*
- *Accumulate over all trees*
- *Sort variables by importance*



LAB: RANDOM FOREST IN SCIKIT-LEARN