# Bagging in R

## A Brief Introduction and Practical Exercise for the Data Science Philippines Meetup Group

DataSeer

www.dataseer.com

Mr I Reyes (Isaac.reyes@dataseer.com)

2015 | Dec 11

# About Us

- Manila and New York based
- Analytics and data science training and consulting
- Successful data and analytics projects for multiple Forbes Global 2000 companies

Data Science Association

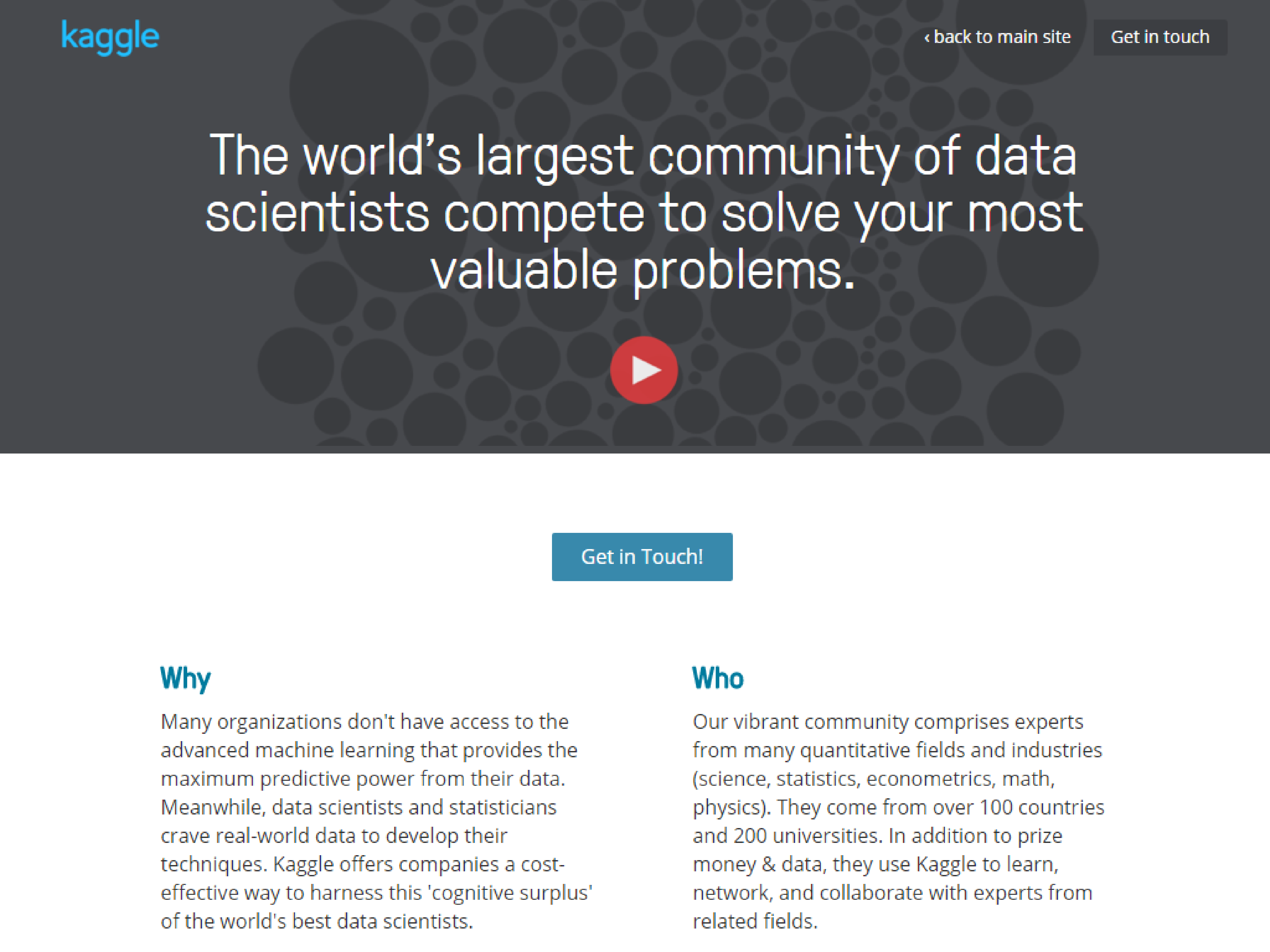
Google Analytics Certified Partner

kaggle

Predictive Analytics World

## Company Training Aim

**Train 1,000 Filipino Data Scientists by 2018**
168 DataSeer graduates in just 4 months
168 down, 832 to go!



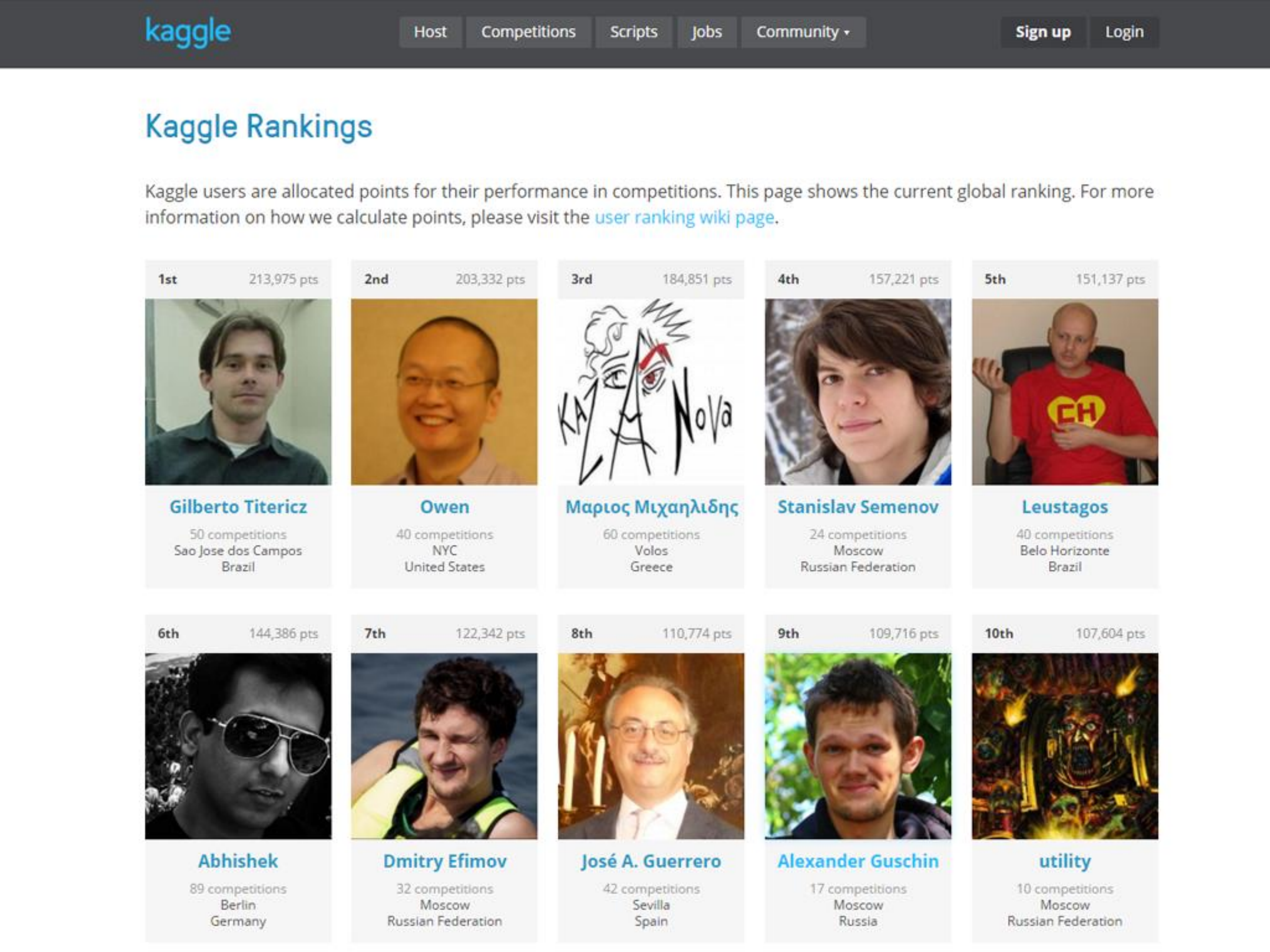
# The world's largest community of data scientists compete to solve your most valuable problems.

## Why

Many organizations don't have access to the advanced machine learning that provides the maximum predictive power from their data. Meanwhile, data scientists and statisticians crave real-world data to develop their techniques. Kaggle offers companies a cost-effective way to harness this 'cognitive surplus' of the world's best data scientists.

## Who

Our vibrant community comprises experts from many quantitative fields and industries (science, statistics, econometrics, math, physics). They come from over 100 countries and 200 universities. In addition to prize money & data, they use Kaggle to learn, network, and collaborate with experts from related fields.



# Kaggle Rankings

Kaggle users are allocated points for their performance in competitions. This page shows the current global ranking. For more information on how we calculate points, please visit the user ranking wiki page.

| | | | | | |
|---|---|---|---|---|---|
| 1st | 213,975 pts | Gilberto Titericz | 50 competitions | Sao Jose dos Campos | Brazil |
| 2nd | 203,332 pts | Owen | 40 competitions | NYC | United States |
| 3rd | 184,851 pts | Μαριος Μιχαηλιδης | 60 competitions | Volos | Greece |
| 4th | 157,221 pts | Stanislav Semenov | 24 competitions | Moscow | Russian Federation |
| 5th | 151,137 pts | Leustagos | 40 competitions | Belo Horizonte | Brazil |
| 6th | 144,386 pts | Abhishek | 89 competitions | Berlin | Germany |
| 7th | 122,342 pts | Dmitry Efimov | 32 competitions | Moscow | Russian Federation |
| 8th | 110,774 pts | José A. Guerrero | 42 competitions | Sevilla | Spain |
| 9th | 109,716 pts | Alexander Guschin | 17 competitions | Moscow | Russia |
| 10th | 107,604 pts | utility | 10 competitions | Moscow | Russian Federation |

kaggle
POINTWEST
DataSeer
DataSeer
Data Science
BOOTCAMP
CLASS OF
2015
INITIAL MODEL 3:30PM
MODEL
training RMSLE
Real Test RMSLE
Modeller
% change
xgboost parameters

## The First Filipino Kaggle Team

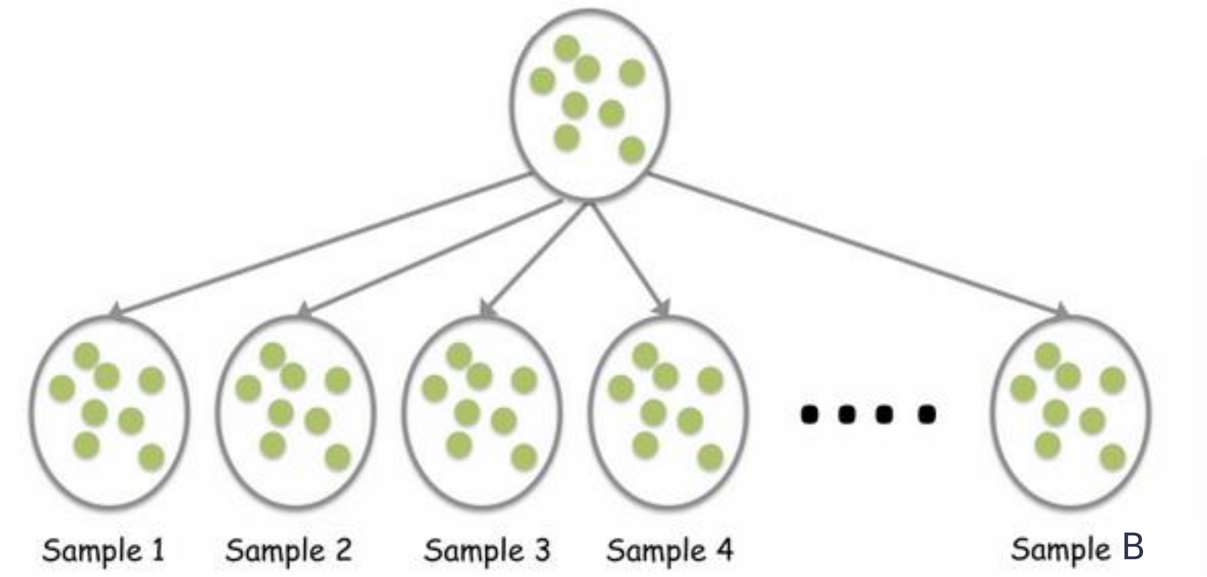
- Team “PointSeerPH”
- Team of trainee Data Scientists from leading companies in the Philippines
- Position finish: 82nd of 1,323 teams in the Caterpillar prize competition
- % Rank: Top 6% of Data Science teams worldwide

# Bagging – Background

# Bagging

- Name comes from a  concatenated contraction of **B**ootstrap **Agg**regat**ing**
- A simple way of increasing the stability and accuracy of a predictive stat/ML model
- L. Breiman, “Bagging predictors,” Machine Learning, 24(2):123-140, 1996.
- Also reduces variance in your predictions and helps to avoid overfitting

# The Bagging Process



- Generate B bootstrap samples of the training data: random sampling with replacement.

- Train a classifier or a regression function using each bootstrap sample.

- For regression: average on the predicted values.

# Let’s Bag in R

## Let’s create some data to be bagged

```
set.seed(10)
y<-c(1:1000)
x1<-c(1:1000)*runif(1000,min=0,max=2)
x2<-c(1:1000)*runif(1000,min=0,max=2)
x3<-c(1:1000)*runif(1000,min=0,max=2)
```

## Let’s view our simulated data

```
plot(x1,y)
plot(x2,y)
plot(x3,y)
```

## Let’s fit a multivariate linear regression to our data

```
lm_fit<-lm(y~x1+x2+x3)
```

## Let’s split our data into a test and training set

```
set.seed(10)
all_data<-data.frame(y,x1,x2,x3)
positions <-
sample(nrow(all_data),size=floor((n
row(all_data)/4)*3))
training<- all_data[positions,]
testing<- all_data[-positions,]
```

## Let’s view our simulated data

```
lm_fit<-
lm(y~x1+x2+x3,data=training)
predictions<-
predict(lm_fit,newdata=testing)
error<-sqrt((sum((testing$y-
predictions)^2))/nrow(testing))
```

The calculated error should be 161.15.

## Let’s implement bagging

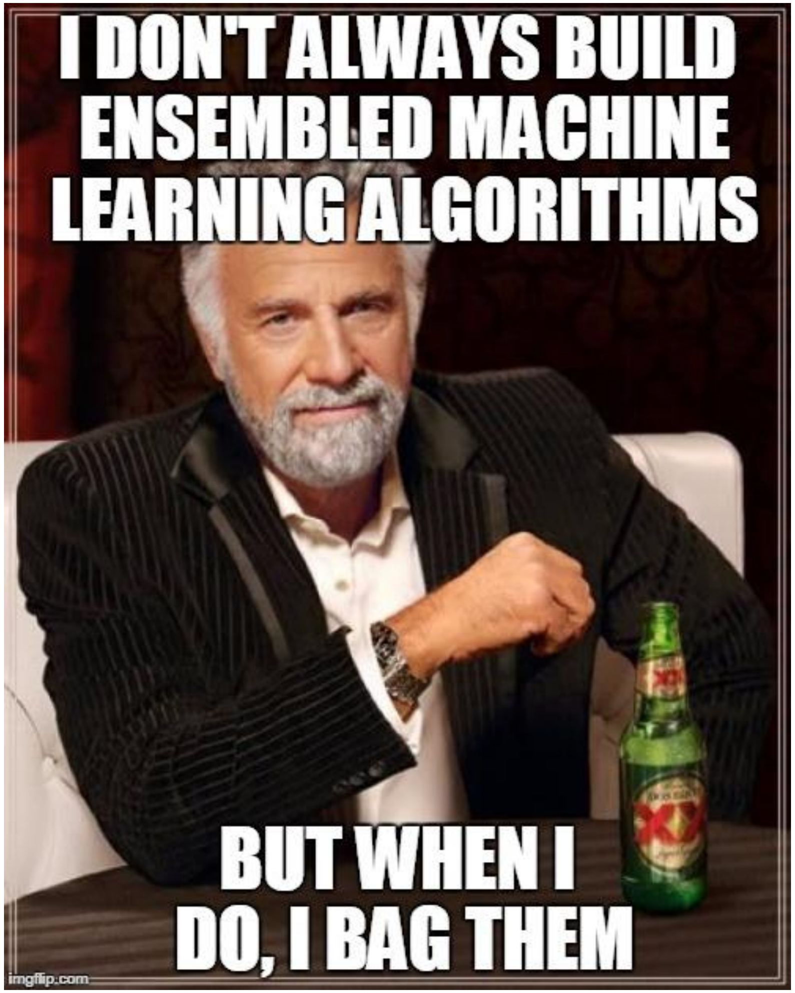
```
library(foreach)
length_divisor<-4
iterations<-1000
```

```
predictions<-foreach(m=1:iterations,.combine=cbind) %do% {
training_positions <- sample(nrow(training),
size=floor((nrow(training)/length_divisor)))
train_pos<-1:nrow(training) %in% training_positions
lm_fit<-lm(y~x1+x2+x3,data=training[train_pos,])
predict(lm_fit,newdata=testing)
}


predictions<-rowMeans(predictions)
error2<-sqrt((sum((testing$y-predictions)^2))/nrow(testing))
```

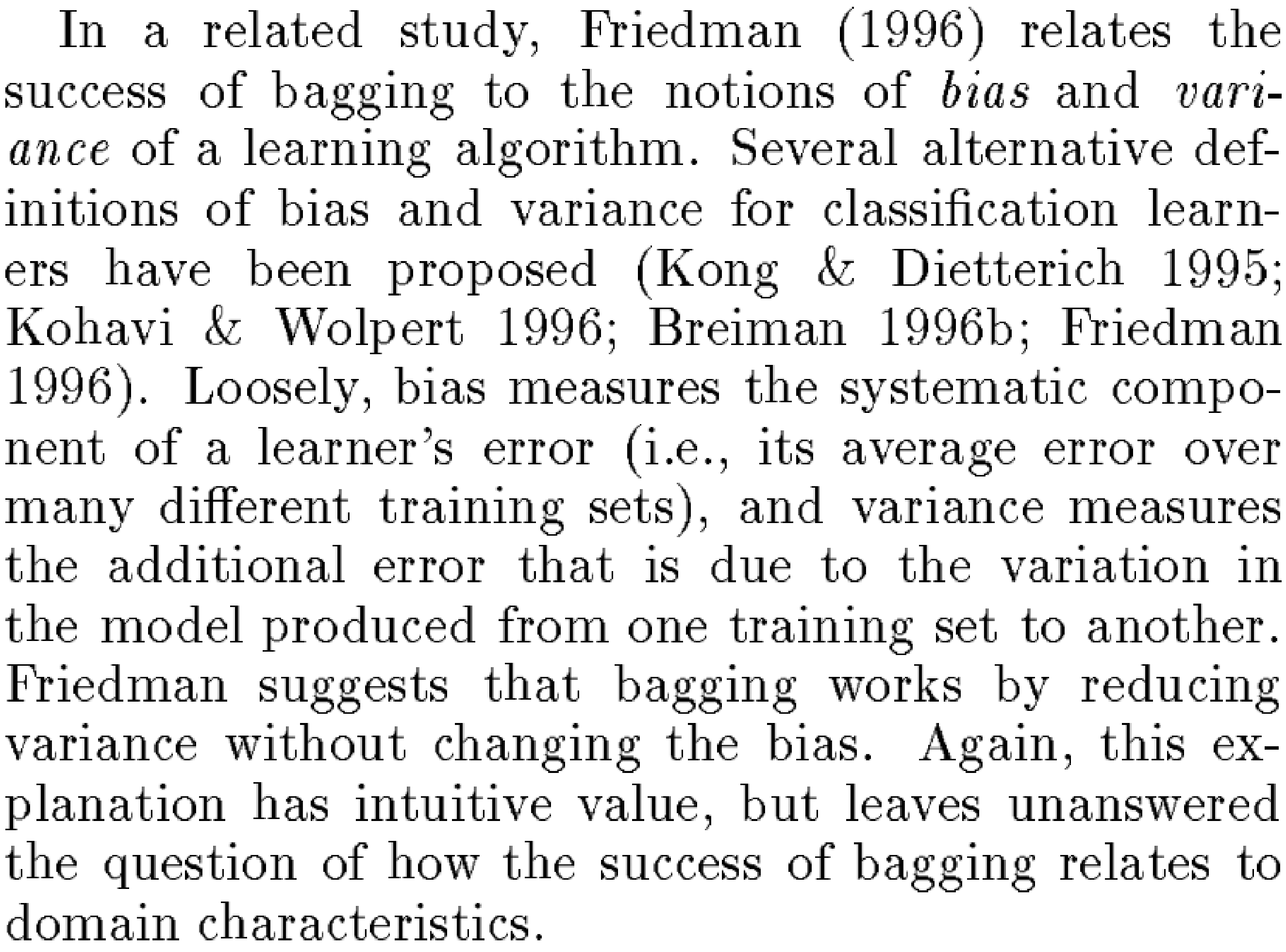
## Bagging Packages

- Bagging in R: *adabag* (fits the Bagging algorithm posed by Breiman in 1996)

- Bagging in Python - sklearn.ensemble.BaggingRegressor (regression problems), sklearn.ensemble.BaggingClassifier


I DON'T ALWAYS BUILD ENSEMBLED MACHINE LEARNING ALGORITHMS
BUT WHEN I DO, I BAG THEM
imgflip.com

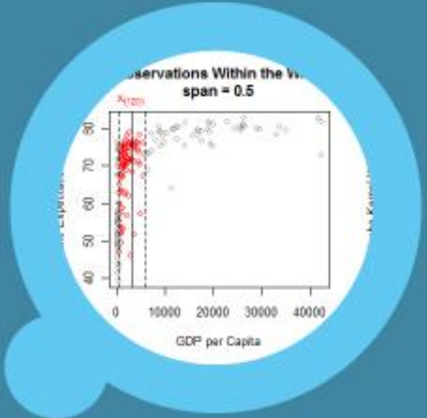
# Why does bagging work?



In a related study, Friedman (1996) relates the success of bagging to the notions of *bias* and *variance* of a learning algorithm. Several alternative definitions of bias and variance for classification learners have been proposed (Kong & Dietterich 1995; Kohavi & Wolpert 1996; Breiman 1996b; Friedman 1996). Loosely, bias measures the systematic component of a learner's error (i.e., its average error over many different training sets), and variance measures the additional error that is due to the variation in the model produced from one training set to another. Friedman suggests that bagging works by reducing variance without changing the bias. Again, this explanation has intuitive value, but leaves unanswered the question of how the success of bagging relates to domain characteristics.

```
nter<-X-mu
-(X-mu)/sig
xs<-seq(-3.5,3.5,.01)
Cs<-dcauchy(xs)
zero<-dnorm(xs,0,1)


hist(x,col='lightgrey',freq=FALSE,las=1,main="Simulated

hist(x,col='lightgrey',freq=FALSE,las=1,main='Hypothetica

hist(Xcenter,col='lightgrey',freq=FALSE,las=1,main='The

hist(Z,col='lightgrey',freq=FALSE,las=1,main='dividing by

hist(Z,col='lightgrey',freq=FALSE,las=1,main='Zooming ba
lines(xs,Zdens,type='l',col='blue',lwd=3)
lines(xs,cs,type='l',col='darkgreen',lwd=1)
abline(v=1,col='red')
```

```
c(0,.485)),sub='Each observation minus
'lightgrey',freq=FALSE,las=1,main=
s SD=1',xlab='Diastolic
```



Thanks!

```
library(faraway)
glmresult <- glm(test ~ pregnant + glucose + diastolic


logregforestplot(glmresult)
logregforestplot(glmresult, logscale=FALSE)
logregforestplot(glmresult, uniform=FALSE)
logregforestplot(glmresult, title="Pima Indians", su
ogregforestplot(glmresult, showrows=c(3,5,7), unif
regforestplot(glmresult, statistics=FALSE, uni
forestplot(glmresult, statistics=FALSE, l
estplot(glmresult, showrows=c(3,5,
plot(glmresult, labels=c("Te
```