



The Ringer/Getty Images

PREDICTING  
THE NCAA  
BASKETBALL  
TOURNAMENT  
WITH  
MACHINE  
LEARNING

ANDREW  
LEVANDOSKI  
AND  
JONATHAN  
LOBO



# THE TOURNAMENT

# MARCH MADNESS

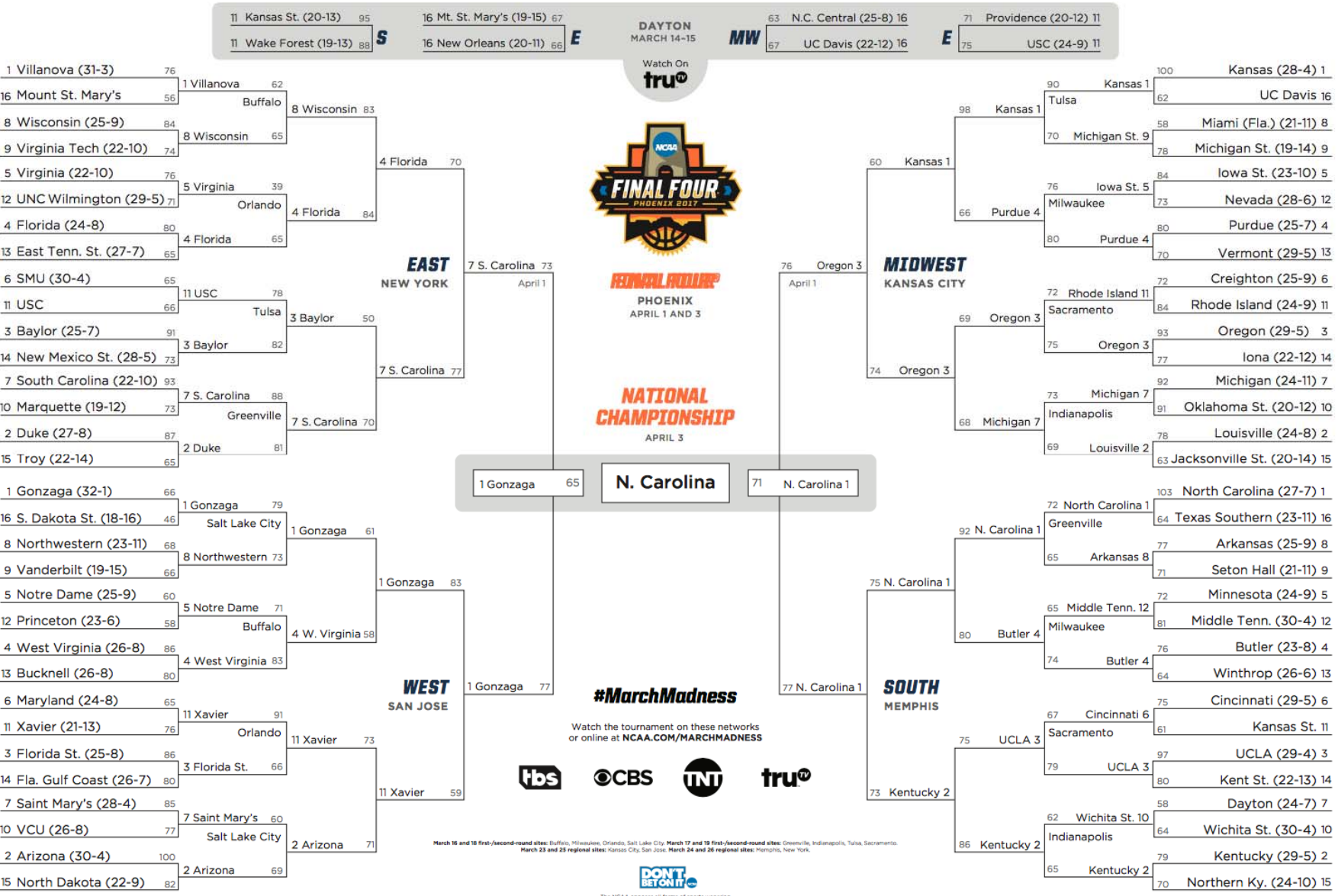
- 68 teams (4 play-in games)
- 63 single-elimination games
- 4 regions, 16 teams each
- Each team seeded 1-16

2017 Champion: **North Carolina**  
(#1 seed in South region)



The Ringer/Getty Images





# THE BRACKET

- 19 million brackets submitted to ESPN
  - $2^{63}$ , or 9.2 quintillion, possible brackets
  - Nobody has ever picked a perfect bracket
- 
- \$3 billion in bracket pools annually
  - Warren Buffet offers \$1 million for a perfect bracket

## BRACKETOLOGY

/ˌbrækəˈtɒləjə/

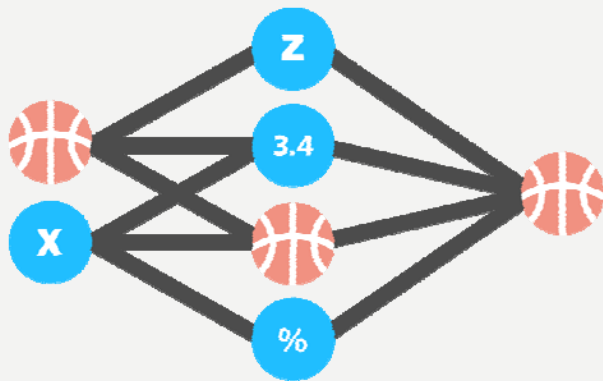
*noun*

the activity of predicting the participants in and outcomes of the games in a sports tournament, especially the NCAA college basketball tournament.

# HISTORICAL RESULTS

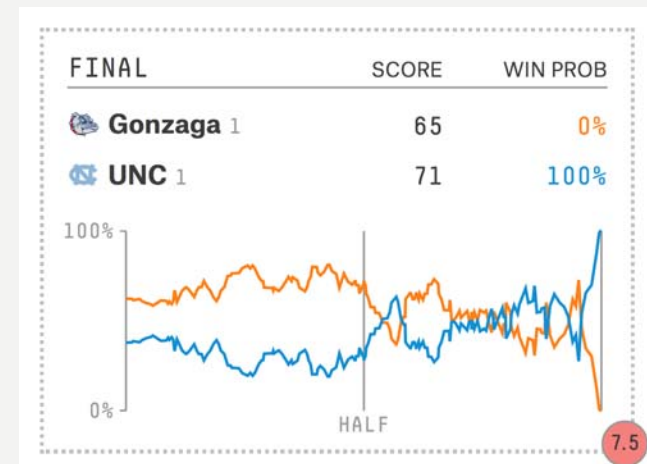
Matchup	Higher Seed Win %
1 vs 16	100.0%
2 vs 15	93.9%
3 vs 14	84.1%
4 vs 13	80.3%
5 vs 12	64.4%
6 vs 11	62.9%
7 vs 10	61.4%
8 vs 9	50.8%

# USING ML TO BUILD BRACKETS



Kaggle's March Machine  
Learning Mania 2017

FiveThirtyEight's  
Prediction Engine



R64	R32	S16	E8	F4	NCG	CHAMPION	TOTAL	PCT
250	200	160	160	160	320	UNC	1250	96.6
260	220	80	80	160	320	UNC	1120	90.5
270	140	120	80	160	320	UNC	1090	88.4
240	140	80	80	160	320	UNC	1020	84.6
220	180	200	160	160	0	Villanova	920	80.0
250	120	120	80	160	0	Gonzaga	730	59.2
270	120	160	160	0	0	Villanova	710	56.5
290	220	120	0	0	0	Kentucky	630	43.4
280	180	160	0	0	0	Kentucky	620	41.4
230	220	160	0	0	0	Villanova	610	39.4
240	200	160	0	0	0	Villanova	600	37.4
260	200	120	0	0	0	UCLA	580	33.1
230	180	40	0	0	0	Kentucky	450	10.6
210	60	0	0	0	0	Saint Mary's	270	2.6

← Jonathan  
← Andrew 😞

**PICKING A  
BRACKET IS  
REALLY  
HARD!**





# MODELS

# PREVIOUS WORK

- Mostly statistical modeling
- Poor choice of features
  - Seeding
  - Vegas point spread

## HOW WE FIX THIS:

- Pick only relevant features
- Analyze a wide range of ML techniques
- Build our own model using random forests



The Ringer/Getty Images

# THE DATA

- Points
- Field Goals Made
- Field Goals Attempted
- 3-Pointers Made
- 3-Pointers Attempted
- Free Throws Made
- Free Throws Attempted
- Offensive Rebounds
- Defensive Rebounds
- Assists
- Turnovers
- Steals
- Blocks
- Personal Fouls
- Game Location  
(Home or Away/Neutral)

# CONSTRUCTING TRAINING DATA

- 14 chosen statistics computed based on rolling averages of previous games
  - For teams and their opponents
- Only include data from 15 most recent games
- Game Location represented as binary flag
- No feature for wins or seed



- Adaptive Boosting
- K-Nearest Neighbors
- Naïve Bayes
- Neural Network
- Logistic Regression
- Support Vector Machine
- **Random Forests**

## LEARNING TECHNIQUES

Trained for *classification* (picking a winner and loser for each matchup)

AND

determining the *likelihood* of each outcome

# RANDOM FORESTS

- The combination of learning models increases classification accuracy
- **Bagging**: average noisy and unbiased models to create a model with low variance

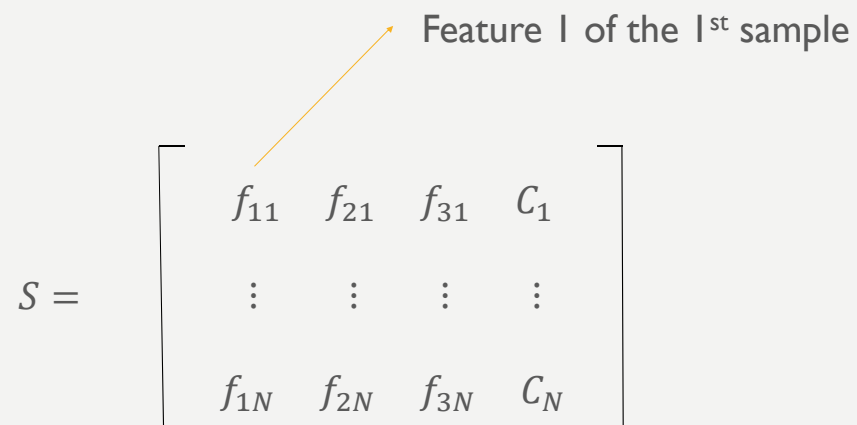
The random forest algorithm acts as a large collection of decorrelated **decision trees**



# DECISION TREE

$$S = \begin{bmatrix} f_{11} & f_{21} & f_{31} & C_1 \\ \vdots & \vdots & \vdots & \vdots \\ f_{1N} & f_{2N} & f_{3N} & C_N \end{bmatrix}$$

Feature 1 of the 1<sup>st</sup> sample



Feature 2 of the N<sup>th</sup> sample

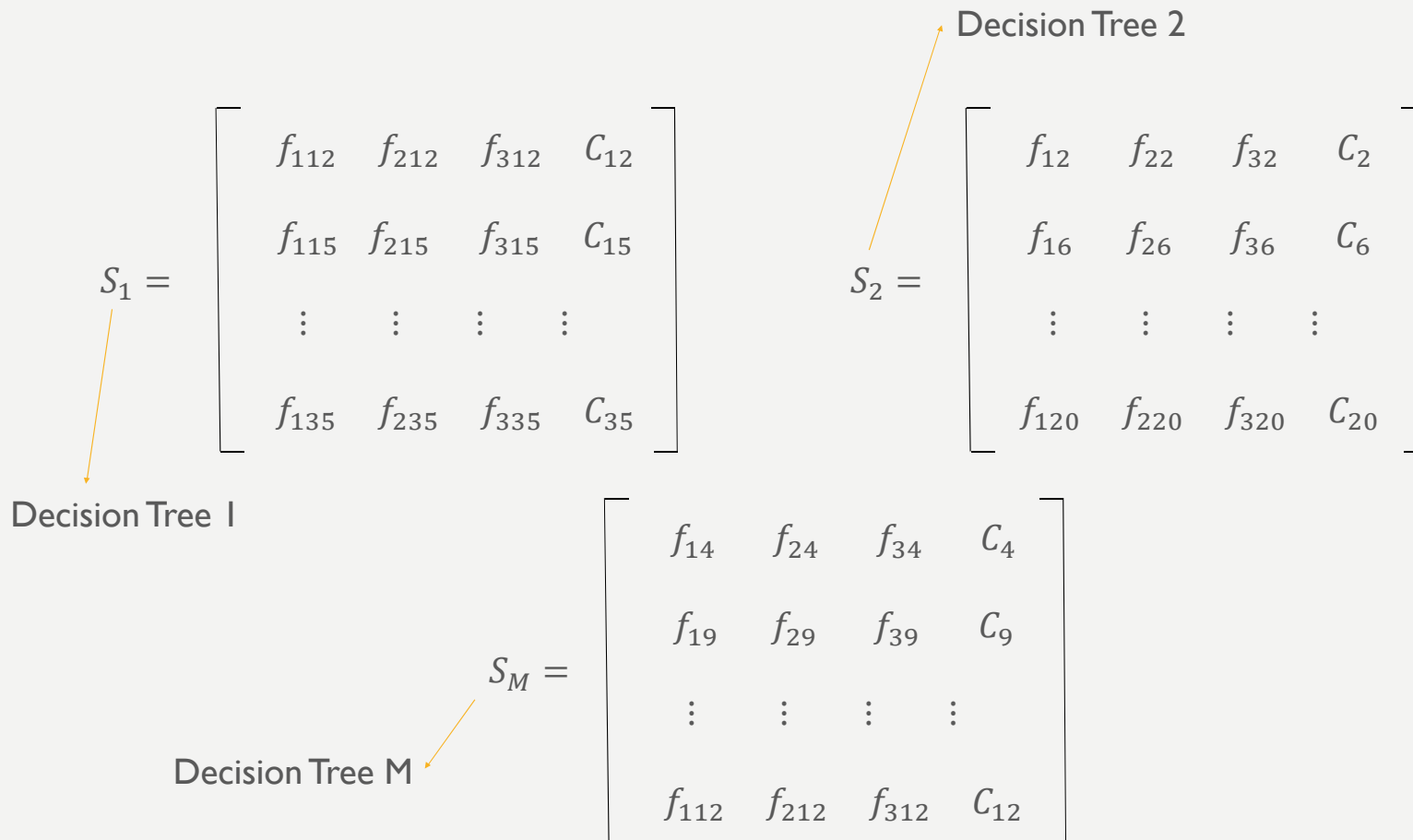
# RANDOM SUBSETS

$$S_1 = \begin{bmatrix} f_{112} & f_{212} & f_{312} & C_{12} \\ f_{115} & f_{215} & f_{315} & C_{15} \\ \vdots & \vdots & \vdots & \vdots \\ f_{135} & f_{235} & f_{335} & C_{35} \end{bmatrix}$$

$$S_2 = \begin{bmatrix} f_{12} & f_{22} & f_{32} & C_2 \\ f_{16} & f_{26} & f_{36} & C_6 \\ \vdots & \vdots & \vdots & \vdots \\ f_{120} & f_{220} & f_{320} & C_{20} \end{bmatrix}$$

$$S_M = \begin{bmatrix} f_{14} & f_{24} & f_{34} & C_4 \\ f_{19} & f_{29} & f_{39} & C_9 \\ \vdots & \vdots & \vdots & \vdots \\ f_{112} & f_{212} & f_{312} & C_{12} \end{bmatrix}$$

# RANDOM SUBSETS



# VOTING

$S_1 \rightarrow 1$

$S_2 \rightarrow 0$

$S_3 \rightarrow 1$

$\vdots$

$S_N \rightarrow 1$

Label as majority vote



# METHOD EVALUATION

- Accuracy
- Bracket score
- Log Loss (for probabilities)

$$\text{LogLoss} = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

where  $n$  is the number of games played,  
 $y_i$  is the predicted probability of team 1 beating team 2,  
and  $y_i$  is 1 if team 1 wins, 0 if team 2 wins.



**RESULTS**



# CLASSIFICATION ACCURACY

Algorithm	Accuracy
KNN	.619
AdaBoost	.667
SVM	.683
Bayes	.698
Random Forests	.698
Regression	.762
Neural Net	.794



# BRACKET SCORES

Algorithm	Score
AdaBoost	400
SVM	570
KNN	600
Bayes	610
Neural Net	650
Regression	670
Random Forests	900



The Ringer/Getty Images



# LOG LOSS

Algorithm	Score
AdaBoost	1.261
KNN	.687
SVM	.657
Bayes	.579
Random Forests	.578
Neural Net	.545
Regression	.529



The Ringer/Getty Images

Algorithm	Log Loss	Accuracy	Bracket Score
AdaBoost	1.261	.667	400
KNN	.687	.619	600
SVM	.657	.683	570
Bayes	.579	.698	610
Random Forests	.578	.698	900
Neural Net	.545	.794	650
Regression	.529	.762	670

# REFERENCES

Image Credits: Getty Images, The Ringer

B.J. Coleman, J.M. DuMond, and A.K. Lynch, “Evidence of Bias in NCAA Tournament Selection and Seeding”, *Managerial and Decision Economics*, vol. 31, 2010.

Kaggle.com, March Machine Learning Mania 2017”, “<https://www.kaggle.com/c/march-machine-learning-mania-2017>”, 2017.

L.H.Yuan et al, “A mixture-of-modelers approach to forecasting NCAA tournament outcomes”, *Journal of Quantitative Analysis of Sports*, vol. 11, 2014.

M.J. Lopez and G.J. Matthews, “Building an NCAA men’s basketball predictive model and quantifying its success”, *Journal of Quantitative Analysis of Sports*, vol. 11, 2014.

Z. Shi, S. Moorthy, and A. Zimmermann, “Predicting NCAAAB match outcomes using ML techniques - some results and lessons learned”, *ArXiv e-prints*, “<https://arxiv.org/abs/1310.3607>”, 2013.