March Madness ML

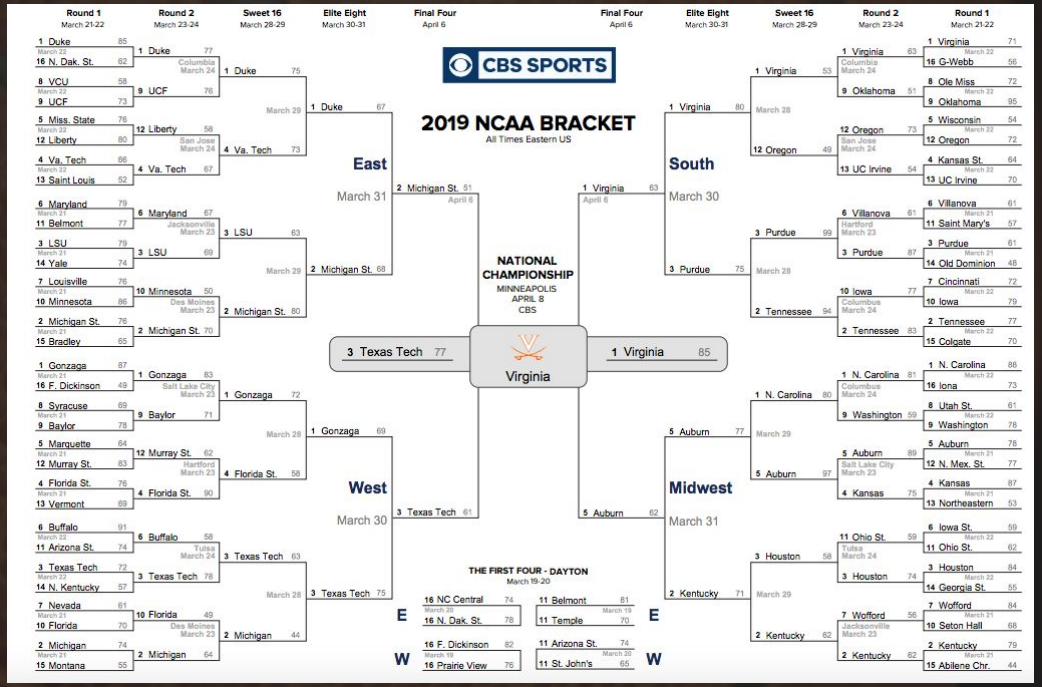
Chris Rabasa, Darshan Gove, Liam Peng

The Problem

Bracket competitions have become more popular in recent years

Famously hard to predict

Famous problem that attracts many statisticians



Whos cares? It's just basketball!

CNN World » US | Africa | Americas | Asia | Australia | China | Europe | Middle East | India | UK International Edition +

The odds of a perfect March Madness bracket are off-the-charts crazy

Tyler Greenawalt | NCAA.com

Updated 2011 GMT (0411 HKT) March 7, 2018

EDITOR'S PICK | 71,503 views | Mar 14, 2019, 08:00am

2019 March Madness Bracket Contest Prize: $1 Million A Year For Life From Warren Buffett

Kurt Badenhausen Forbes Staff

SportsMoney

I cover sports business with rare dips into b-schools, local economies

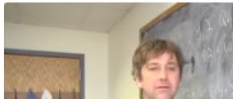
NCAA March Madness gambling total to hit $8.5B

By Thomas Barrabi | Published March 18, 2019 | Spor

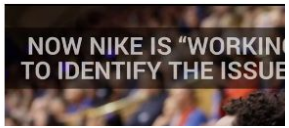
TIME

Duke Math Professor Says Odds of a Perfect Bracket are One in 2.4 Trillion

Bracket math isn't an exact science, but for years mathematicians have told us that the odds of picking a perfect NCAA tournament bracket are a staggering 1 in 9,223,372,036,854,775,808 (that's *9.2 quintillion*).



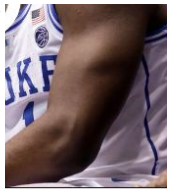
SPORTS • BASKETBALL



We Made 10 Million Computer-Generated March Madness Brackets. Will Any of Them Be Perfect?

College Sports

Gregg Nigl's perfect NCAA tournament bracket is busted — all the way to 348th place.



Unproductive workers cost employers $4 billion during March Madness NCAA Tournament

Ben Tobin | USA TODAY

Published 4:16 PM EDT Mar 22, 2019

on sprains knee when sho

Williamson had his shoe burst apart mid-game

AA "March Madness" men's basketball

How Brackets are Formed

Most people use methods without much logic

Bias to teams near them

Bias to favorite/popular players

Unscientific prediction methods

MARCH MADNESS | COLLEGE BASKETBALL | COLLEGE FOOTBALL

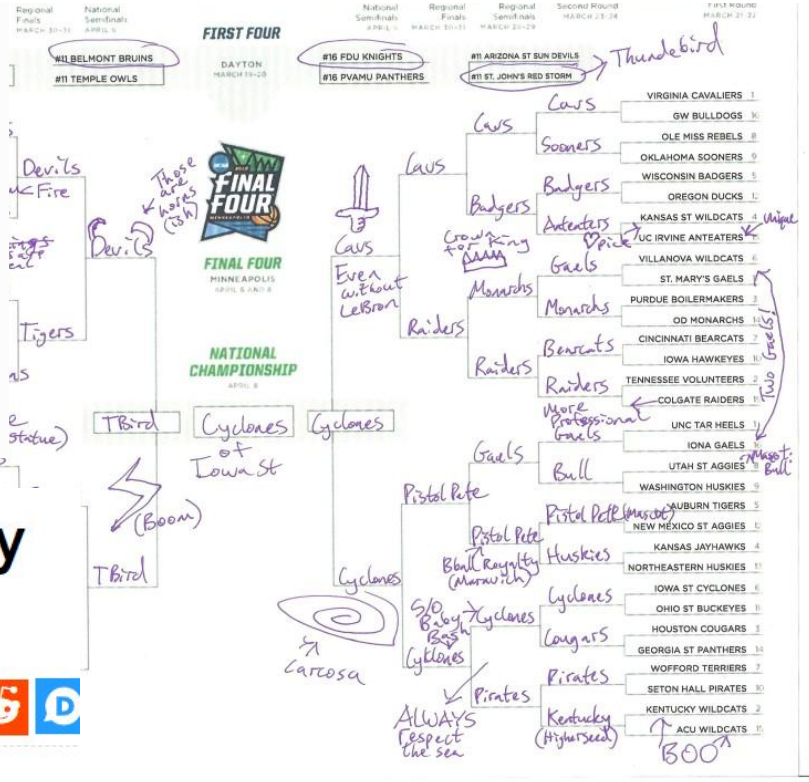
2019 March Madness predictions based entirely on football

Time for the smartest predictions method of all: using the wrong sport.

By Jason Kirk | Mar 20, 2019, 9:51am EDT

Method 3: Team Colors

If you are more a visual person than a numbers person, this method is for you. Pick the team who wears blue. But be warned: this is best employed after the first two rounds. Since 1939 blue has been the dominant team color of schools who have gone to the Final Four (44%), the Final Two (55%) and won the championship (55%). Betting on blue gives better odds than a coin toss.

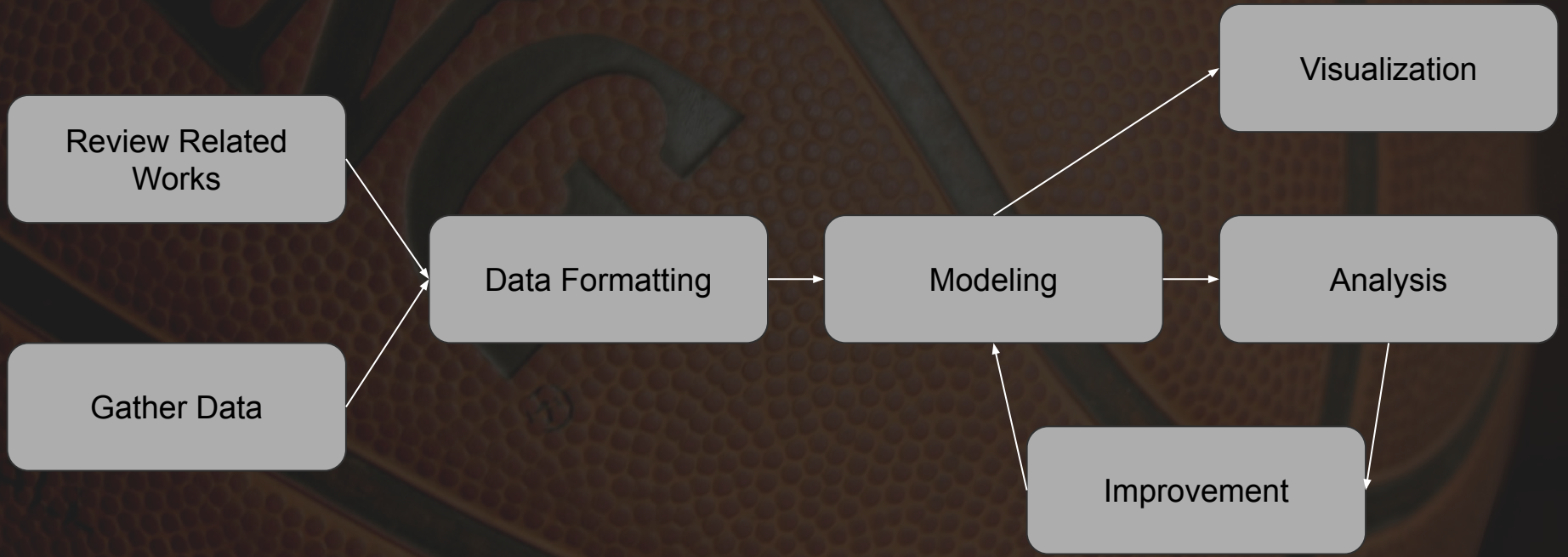


Re-seeding the NCAA Tournament by each school's most famous alum

BY **MATT MULLIN**

PhillyVoice Staff

Hey, you, whatcha gonna do?



The Data

Primary Source: Google Cloud & NCAA® ML Competition 2019-Men's on Kaggle

- Contains NCAA game results and play-by-play information from 2003 to 2018
- Mostly used detailed game results

Secondary Source: sportsreference.com

- Contains player, team stats for every team that has ever played in the NCAA
- Contains some derived metrics like Strength of Schedule and Simple Ranking System

In [4]:
```
# This one contains stats for every single regular season game played between 1985 and 2018. It mainly
# contains info on the score of the game, the IDs for each team, and where the game was played.
reg_season_compact_pd = pd.read_csv(path + 'RegularSeasonCompactResults.csv')
reg_season_compact_pd.head()
```

Out[4]:

	Season	DayNum	WTeamID	WScore	LTeamID	LScore	WLoc	NumOT
0	1985	20	1228	81	1328	64	N	0
1	1985	25	1106	77	1354	70	H	0
2	1985	25	1112	63	1223	56	H	0
3	1985	25	1165	70	1432	54	H	0
4	1985	25	1192	86	1447	74	H	0

In [8]:
```
# This one expands on the previous data frame by going into more in depth stats like 3 point field goals,
# free throws, steals, blocks, etc.
#NOTE: Whereas the previous dataframe has data from 1985, this one only has data from 2003 on!
reg_season_detailed_pd = pd.read_csv(path+'RegularSeasonDetailedResults.csv')
reg_season_detailed_pd.columns
```

Out[8]:
```
Index(['Season', 'DayNum', 'WTeamID', 'WScore', 'LTeamID', 'LScore', 'WLoc',
       'NumOT', 'WFGM', 'WFGA', 'WFGM3', 'WFGA3', 'WFTM', 'WFTA', 'WOR', 'WDR',
       'WAst', 'WTO', 'WStl', 'WBlk', 'WPF', 'LFGM', 'LFGA', 'LFGM3', 'LFGA3',
       'LFTM', 'LFTA', 'LOR', 'LDR', 'LAst', 'LTO', 'LStl', 'LBlk', 'LPF'],
      dtype='object')
```

Data Formatting

Had team stats for each game

Easy to compute stats for whole season

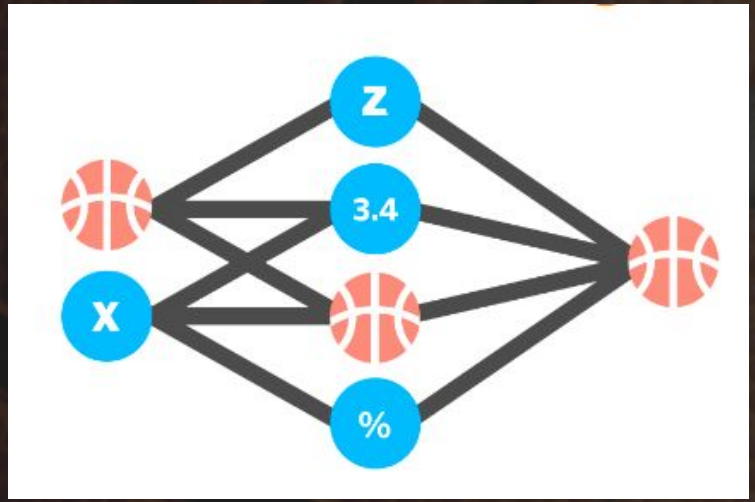
Better to compute stats up to each game (Only for the current season)

A problem! Winner is always first

Solutions: Data doubling, then randomization

Formatting: Concatenation, then subtraction

Modeling

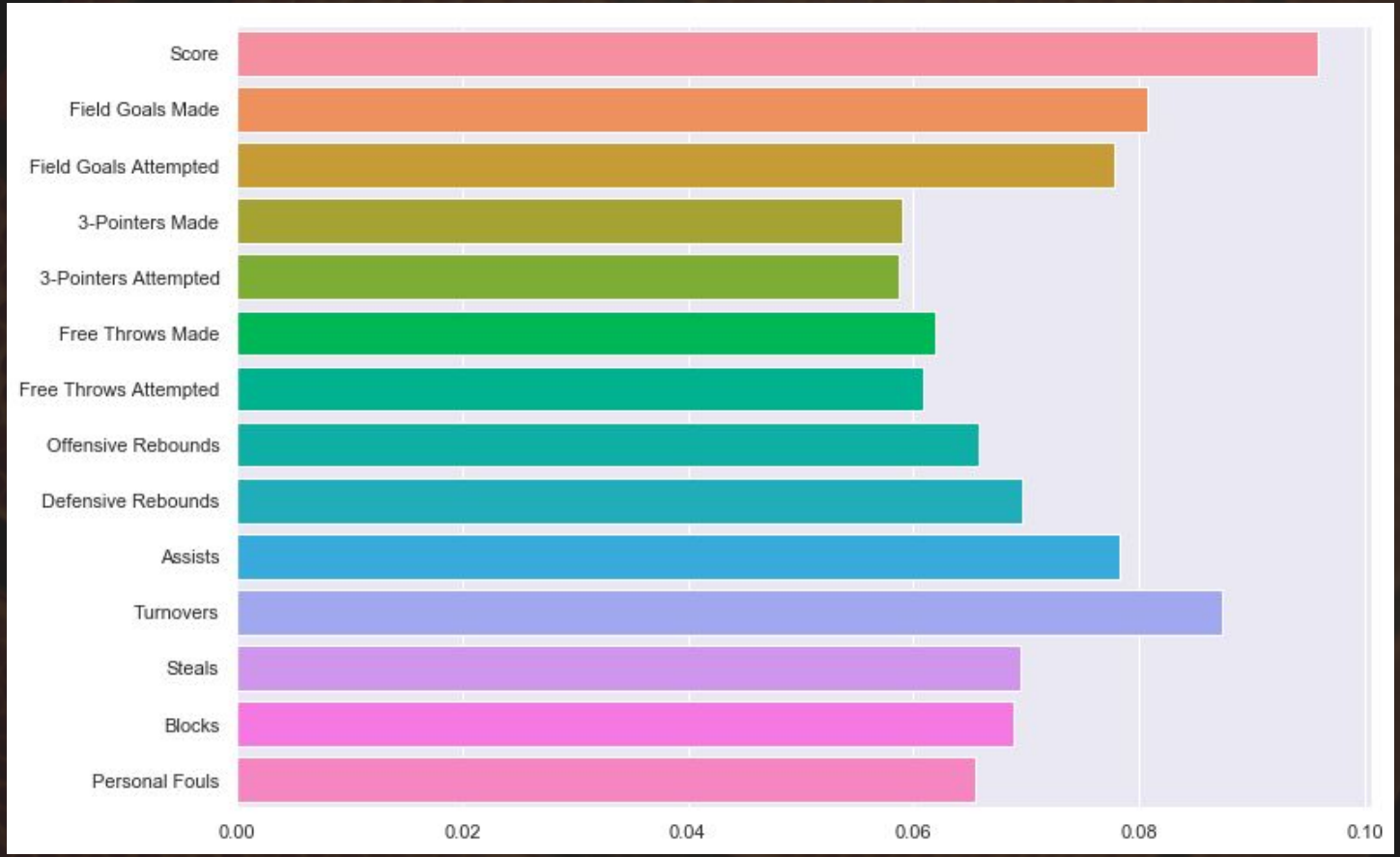


Tried the following models:

- Random Forest: 64%
- Multi-Layer Perceptron: 67%
- Logistic Regression: 67%
- KNN: 60%
- SVC: 51%
- Naive Bayes: 63%

```
LogisticRegression(C=1, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='12', random_state=0, solver='liblinear',
          tol=0.001, verbose=0, warm_start=False)
```

Feature Importances




NCAA men's basketball championship
EAST REGIONAL
Washington, D.C.
WEST REGIONAL
Anaheim
MIDWEST REGIONAL
Kansas City
SOUTH REGIONAL
Louisville
FINAL FOUR
NATIONAL CHAMPIONSHIP
April 8
Minneapolis
Mich St
vs.
Villanova
Villanova
Winner
1 Duke
16 NC Cen/N. Dak. St
8 VCU
9 UCF
Duke
Columbia
VCU
5 Mississippi St.
12 Liberty
4 Virginia Tech
13 Saint Louis
Liberty
Jacksonville
VA Tech
Duke
VA Tech
6 Maryland
11 Belmont/Temple
3 LSU
14 Yale
Belmont
Jacksonville
LSU
Belmont
Mich St
7 Louisville
10 Minnesota
2 Michigan St.
15 Bradley
Louisville
Des Moines
Mich St
Mich St
1 Gonzaga
16 F. Dickinson/Prairie View
8 Syracuse
9 Baylor
Gonzaga
Salt Lake City
Syracuse
Gonzaga
Gonzaga
5 Marquette
12 Murray State
4 Florida State
13 Vermont
Murray St
Vermont
Murray St
6 Buffalo
11 Ariz. St./St. John's
3 Texas Tech
14 Northern Kentucky
Ariz St
N KY
Ariz St
Nevada
7 Nevada
10 Florida
2 Michigan
15 Montana
Nevada
Michigan
Nevada
Gonzaga
11 Belmont
11 Temple
16 Fairview Dickinson
16 Prairie View
16 North Dakota St.
16 N.C. Central
11 Arizona St.
11 St. John's
Northeastern
UNC
UNC
Utah St
1 North Carolina
16 Iona
8 Utah State
9 Washington
Northeastern
Auburn
Salt Lake City
Northeastern
8 Auburn
12 New Mexico St.
4 Kansas
13 Northeastern
Houston
Ohio St
Tulsa
Houston
6 Iowa State
11 Ohio State
3 Houston
14 Georgia State
Houston
Wofford
Wofford
Kentucky
7 Wofford
10 Seton Hall
2 Kentucky
15 Abilene Christian
Villanova
Virginia
Virginia
Oklahoma
1 Virginia
16 Gardner-Webb
8 Mississippi
9 Oklahoma
Virginia
Oregon
Kansas St
Kansas St
5 Wisconsin
12 Oregon
4 Kansas St.
13 UC Irvine
Villanova
Villanova
Villanova
Purdue
6 Villanova
11 St. Mary's (Calif.)
3 Purdue
14 Old Dominion
Villanova
Cincinnati
Cincinnati
Tennessee
7 Cincinnati
10 Iowa
2 Tennessee
15 Colgate

Analysis

Those were… interesting results.

Why would the model struggle in an NCAA tournament setting?

What could be done to improve performance?

Challenges

Predicting a sport with a lot of randomness

Using data from multiple sources

Putting data into a suitable format for models

Working with a massive amount of data

Future work

Use expert rating and ranking systems (seed, SOS, SRS, KenPom, Massey Ordinals)

Include individual player statistics and importance metrics to deal with injuries and player matchups

Add more metrics that are indicative of play style (time on possession, shot quality metrics)

Thank You