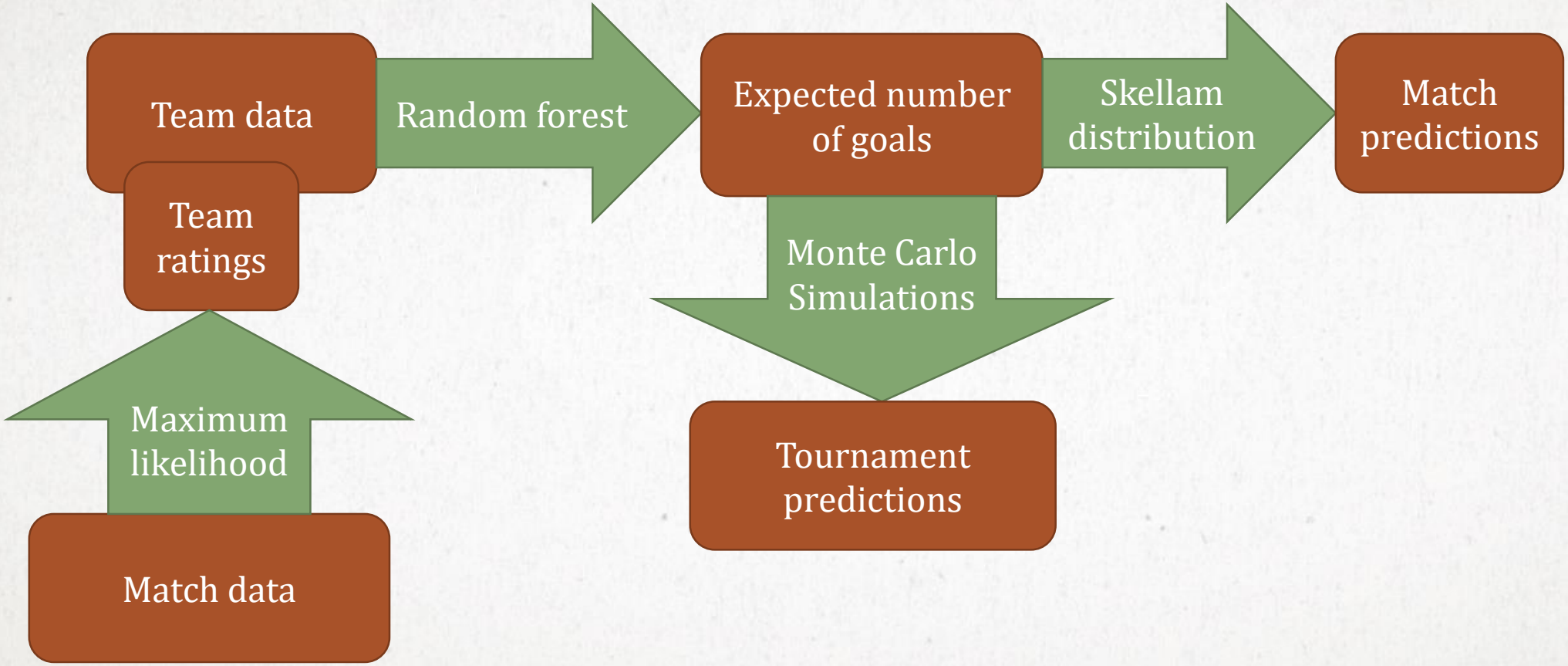
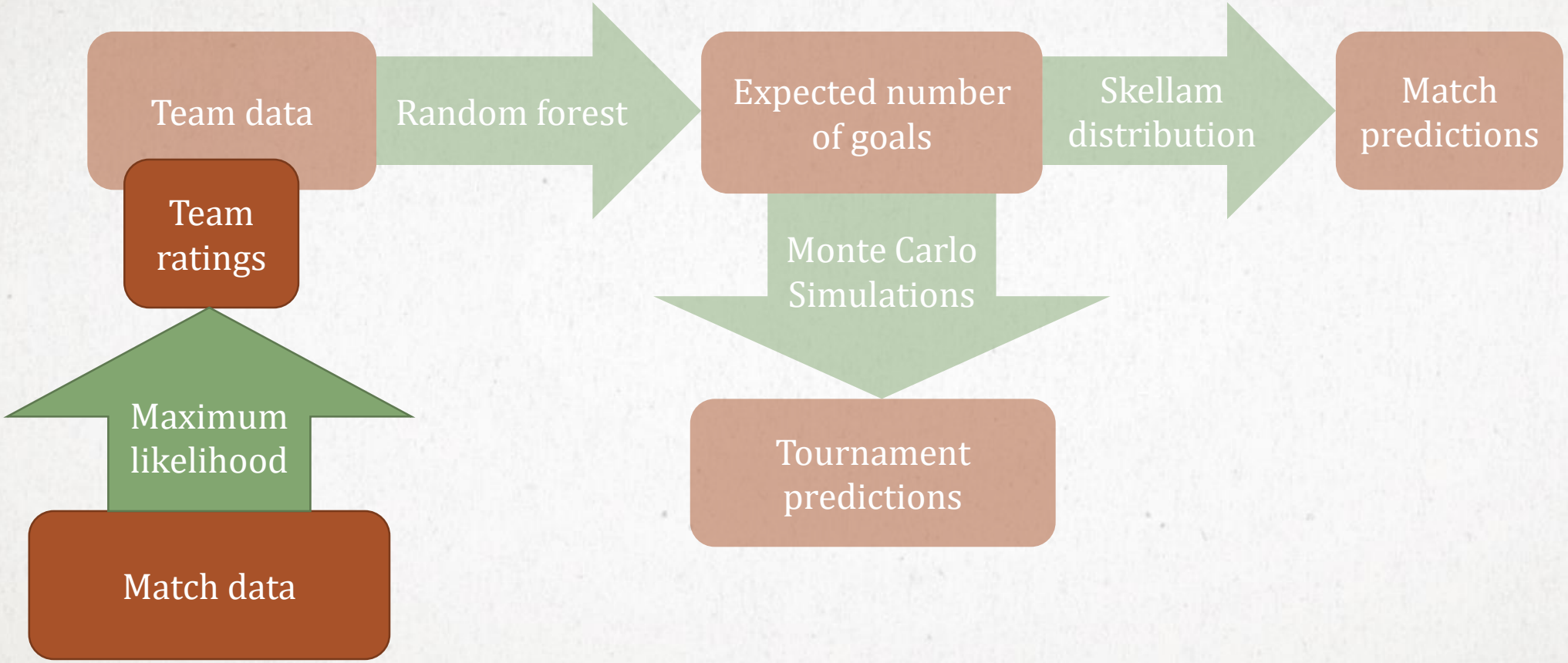
HYBRID PREDICTION MODEL FOR INTERNATIONAL SOCCER TOURNAMENTS

ANDREAS GROLL
CHRISTOPHE LEY
GUNTHER SCHAUBERGER
HANS VAN EETVELDE

THE HYBRID RANDOM FOREST



THE HYBRID RANDOM FOREST



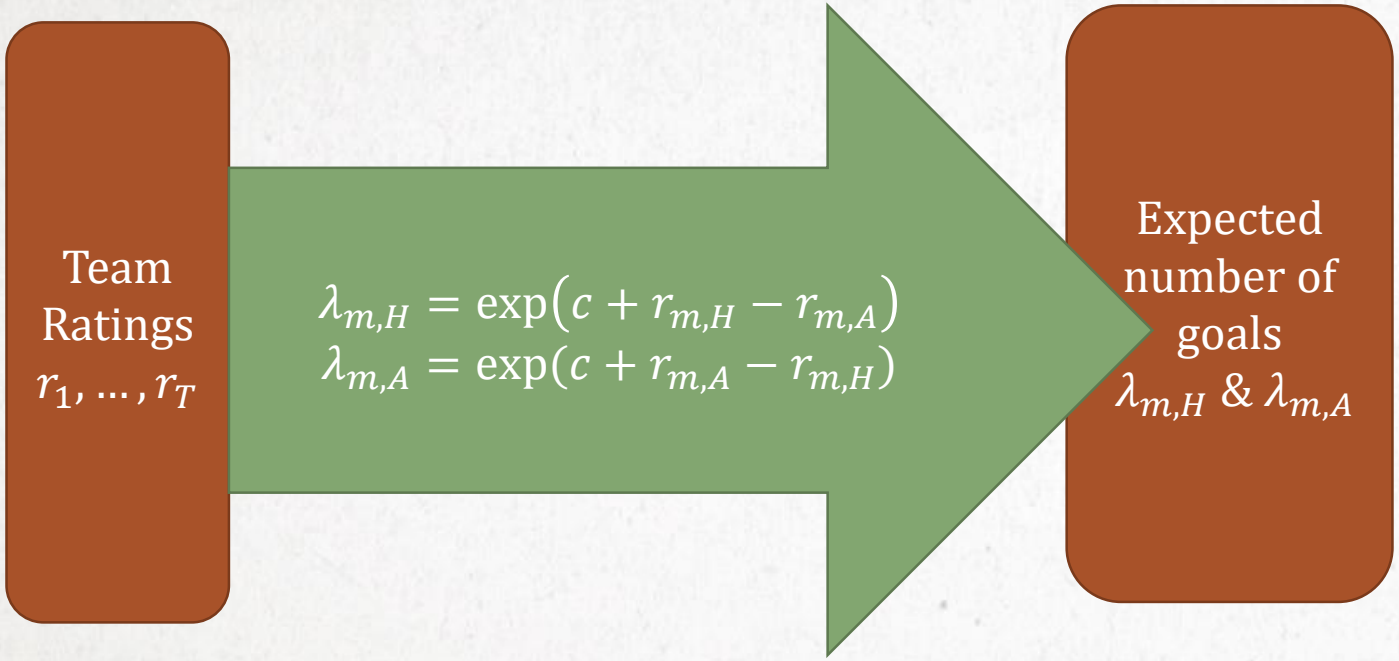
MATCH DATA

Date	Home team	Away team	Score	Country	Neutral
2018-06-12	Poland	Lithuania	4-0	Poland	FALSE
2018-06-12	Japan	Paraguay	4-2	Austria	TRUE
2018-06-11	Belgium	Costa Rica	4-1	Belgium	FALSE
2018-06-11	Korea Republic	Senegal	0-2	Austria	TRUE
2018-06-10	Austria	Brazil	0-3	Austria	FALSE
2018-06-09	France	USA	1-1	France	FALSE
2018-06-09	Tunisia	Spain	0-1	Russia	TRUE

TEAM RATINGS
BY MAXIMUM LIKELIHOOD ESTIMATION

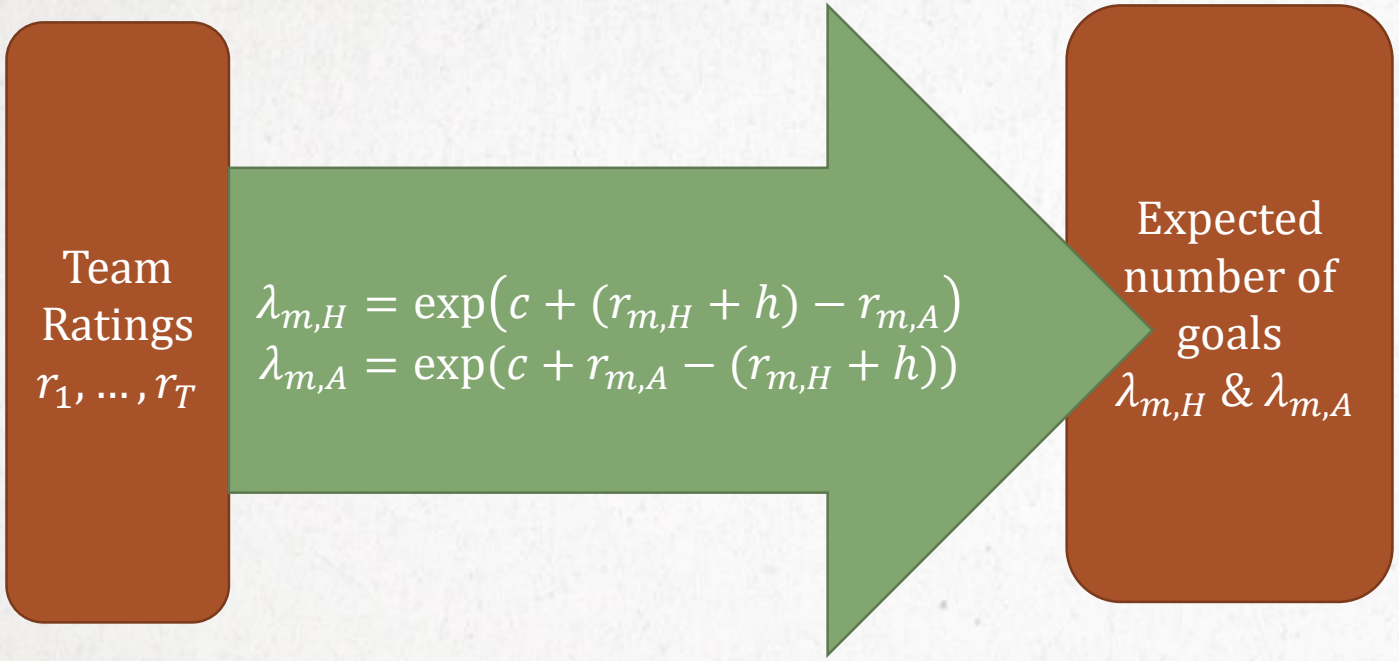
Team Ratings
$r_1, \ldots, r_T$

TEAM RATINGS
BY MAXIMUM LIKELIHOOD ESTIMATION



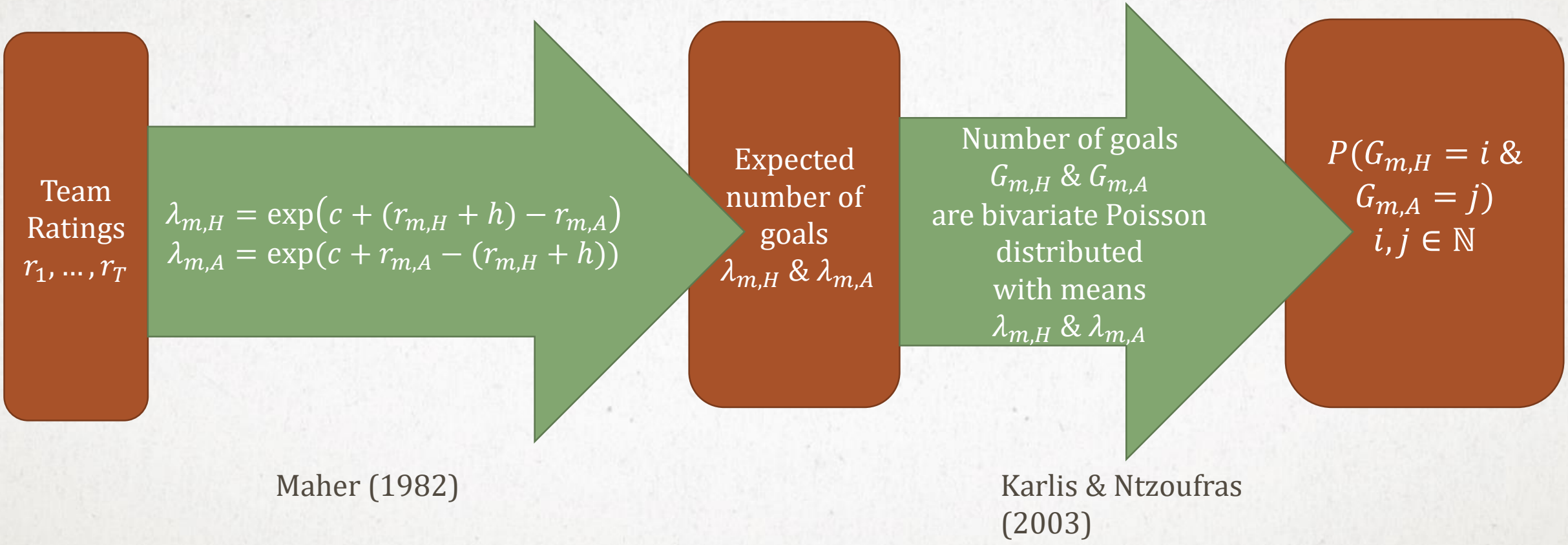
Maher (1982)

TEAM RATINGS
BY MAXIMUM LIKELIHOOD ESTIMATION



Maher (1982)

TEAM RATINGS
BY MAXIMUM LIKELIHOOD ESTIMATION



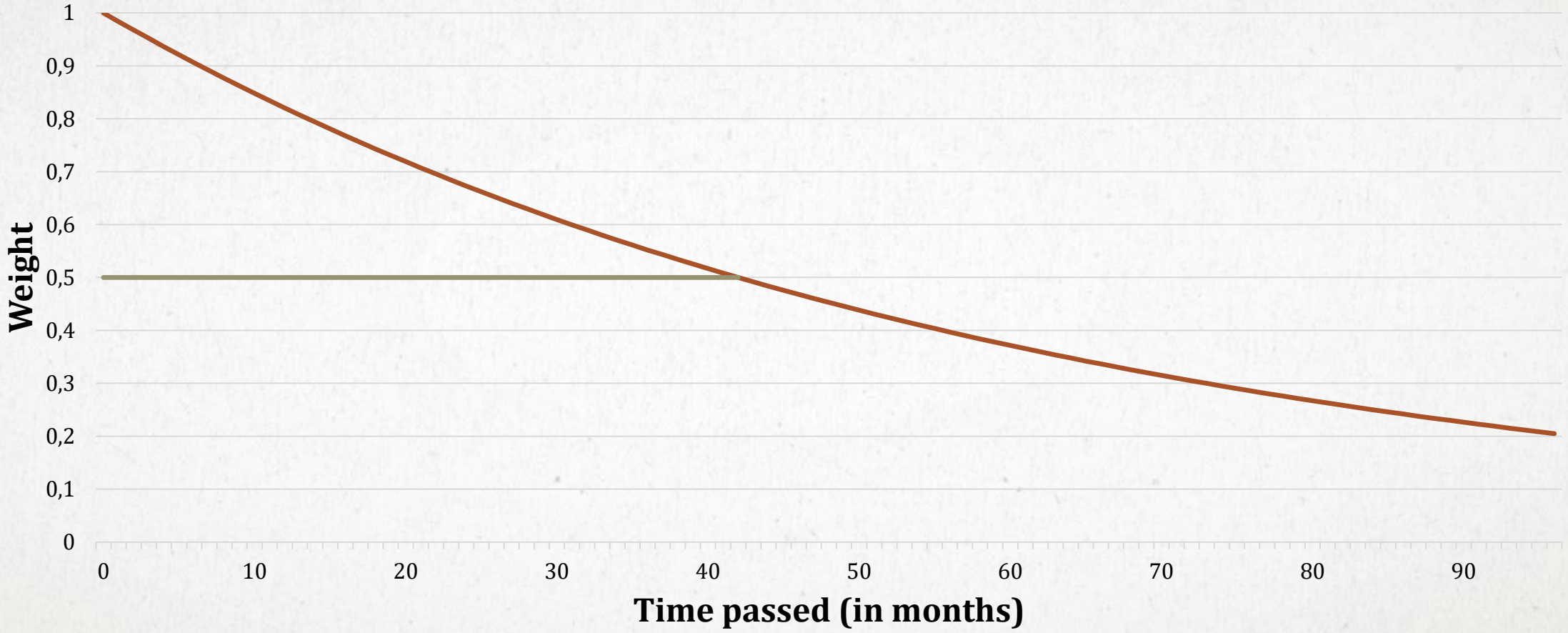
Maher (1982)

Karlis & Ntzoufras (2003)

TIME WEIGHT

$$w_m = \exp(-\alpha\, t_m)$$

with t_m the numbers of days ago that match m is played



WEIGHTED LIKELIHOOD FUNCTION

$$L = \prod_{m=1}^{M} P(G_{m,H} = x_{m,H}, G_{m,A} = x_{m,A})^{w_m}$$

WEIGHTED LIKELIHOOD FUNCTION

$$L = \prod_{m=1}^{M} P(G_{m,H} = x_{m,H}, G_{m,A} = x_{m,A})^{w_m}$$

$$P(G_{m,H} = x_{m,H}, G_{m,A} = x_{m,A}) =$$

$$\exp(-(\lambda_{m,H} + \lambda_{m,A} + \lambda_C)) \frac{\lambda_{m,H}^{x_{m,H}}}{x_{m,H}} \frac{\lambda_{m,A}^{x_{m,A}}}{x_{m,A}} \sum_{k=1}^{\min(x_{m,H}, x_{m,A})} \binom{x_{m,H}}{k} \binom{x_{m,A}}{k} k! \left(\frac{\lambda_C}{\lambda_1 \lambda_2}\right)^k$$

WEIGHTED LIKELIHOOD FUNCTION

$$L = \prod_{m=1}^{M} P(G_{m,H} = x_{m,H}, G_{m,A} = x_{m,A})^{w_m}$$

$$P(G_{m,H} = x_{m,H}, G_{m,A} = x_{m,A}) =$$

$$\exp(-(\lambda_{m,H} + \lambda_{m,A} + \lambda_C)) \frac{\lambda_{m,H}^{x_{m,H}}}{x_{m,H}} \frac{\lambda_{m,A}^{x_{m,A}}}{x_{m,A}} \sum_{k=1}^{\min(x_{m,H}, x_{m,A})} \binom{x_{m,H}}{k} \binom{x_{m,A}}{k} k! \left(\frac{\lambda_C}{\lambda_1 \lambda_2}\right)^k$$

$$\lambda_{m,H} = \exp(c + (r_{m,H} + h) - r_{m,A})$$

$$\lambda_{m,A} = \exp(c + r_{m,A} - (r_{m,H} + h))$$

WEIGHTED LIKELIHOOD FUNCTION

$$L = \prod_{m=1}^{M} P(G_{m,H} = x_{m,H}, G_{m,A} = x_{m,A})^{w_m}$$

$$P(G_{m,H} = x_{m,H}, G_{m,A} = x_{m,A}) =$$

$$\exp(-(\lambda_{m,H} + \lambda_{m,A} + \lambda_C)) \frac{\lambda_{m,H}^{x_{m,H}}}{x_{m,H}} \frac{\lambda_{m,A}^{x_{m,A}}}{x_{m,A}} \sum_{k=1}^{\min(x_{m,H}, x_{m,A})} \binom{x_{m,H}}{k} \binom{x_{m,A}}{k} k! \left(\frac{\lambda_C}{\lambda_1 \lambda_2}\right)^k$$

$$\lambda_{m,H} = \exp(c + (r_{m,H} + h) - r_{m,A}) \qquad \lambda_{m,A} = \exp(c + r_{m,A} - (r_{m,H} + h))$$

Parameters that will be estimated are: $r_1, \ldots, r_T$ and c, h, λ_C

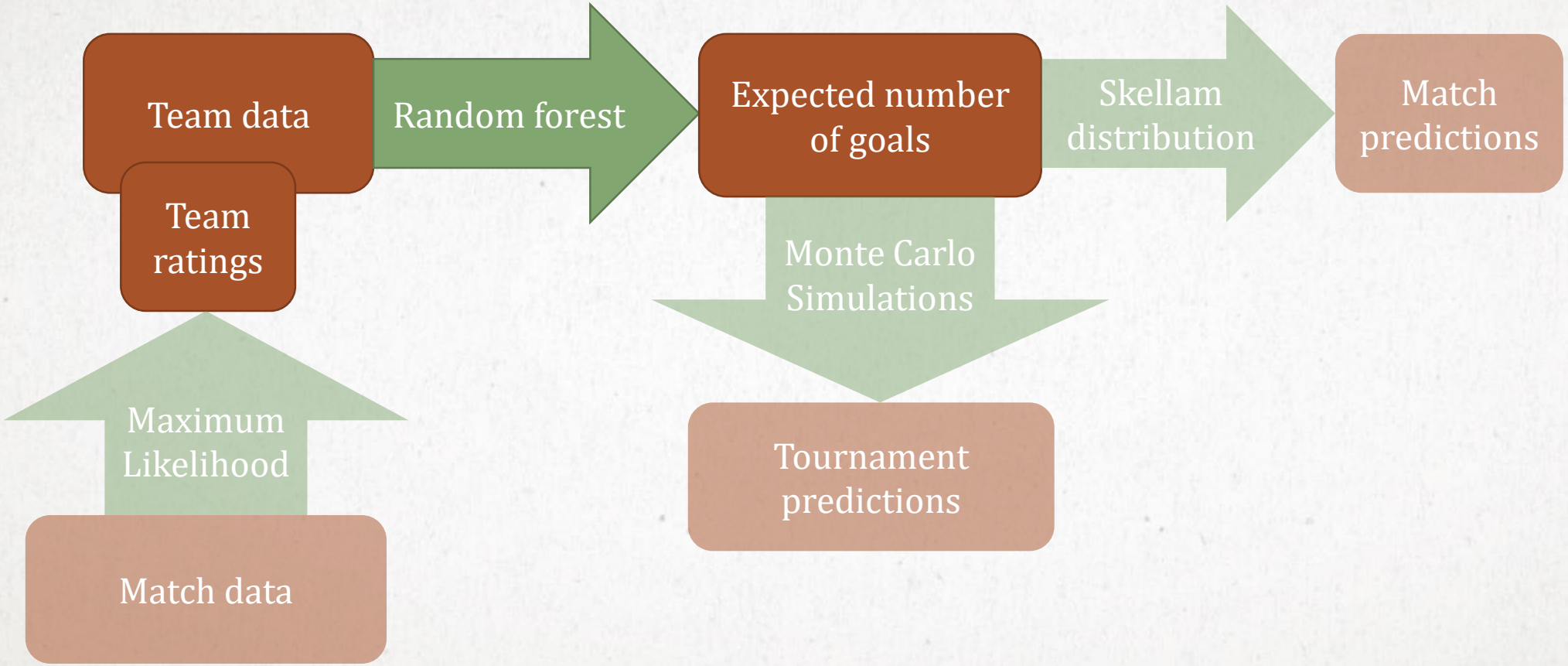
TEAM RATINGS IN JUNE 2018

Rank	Team	Rating r_i
1	Brazil	1,57
2	Germany	1,54
3	Argentina	1,53
4	Spain	1,53
5	Belgium	1,38
6	Colombia	1,38
7	France	1,36
8	Chile	1,35
9	Netherlands	1,31
10	Portugal	1,29

Rank	Team	Rating r_i
11	England	1,23
12	Uruguay	1,20
13	Peru	1,17
14	Croatia	1,15
15	Poland	1,14
16	Sweden	1,13
17	Denmark	1,12
18	Italy	1,12
19	Ecuador	1,07
20	Switzerland	1,06

Rank	Team	Rating r_i
21	Ukraine	1,02
22	Mexico	1,01
23	Serbia	0,98
24	Austria	0,97
25	Bosnia-Herzegovina	0,96
26	Russia	0,96
27	Wales	0,96
…	…	…
50	Greece	0,74
…	…	…

THE HYBRID RANDOM FOREST



TEAM DATA

- Economic factors
  - GDP Per capita
  - Population
- Home advantage
  - Host
  - Continent
  - Confederation
- Coach
  - Age
  - Tenure
  - Nationality (same as country or not)
- Team structure
  - Maximum number of teammates
  - Average age
  - Number of players in the semi-finals of the Champions League and Europe League
  - Number of players abroad
- Sportive factors
  - FIFA ranking
  - Elo-ratings (www.eloratings.net)
  - Max. Likelihood ratings
  - Bookmakers odds
- …

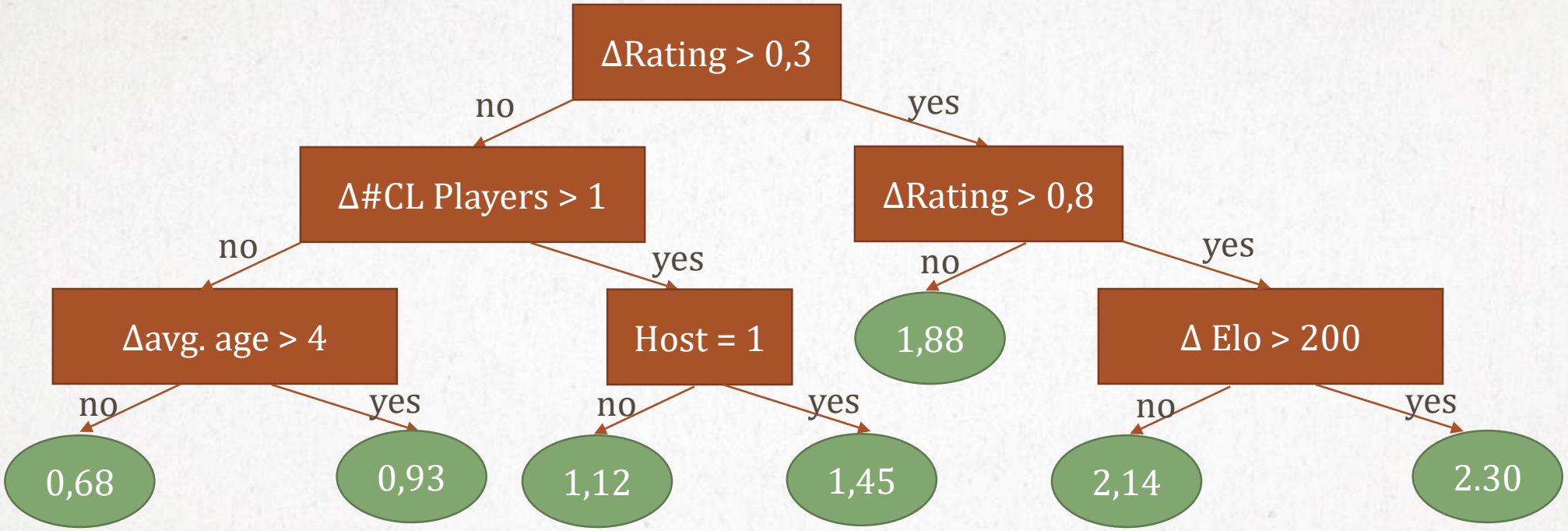
TEAM DATA WORLD CUP 2018

Team	Population	Host	Average age	#CL Players	Tenure coach	Rating	…
Argentina	0,005977	0	29,2	1	1	1,53	…
Australia	0,003313	0	28,1	0	0,4	0,79	…
Belgium	0,001538	0	27,6	1	2	1,38	…
Brazil	0,028203	0	28,6	4	2	1,57	…
Colombia	0,006616	0	28,1	1	6	1,38	…
Costa Rica	0,000662	0	29,5	1	3	0,83	…
Croatia	0,000557	0	27,9	3	1	1,15	…
Denmark	0,00077	0	27,1	0	2	1,12	…
Egypt	0,013291	0	29	1	3	0,80	…
England	0,007392	0	26	2	1,5	1,23	…

INPUT RANDOM FOREST

Team	Opponent	Δ Population	Host	Opp.Host	Δ avg. age	Δ#CL Players	Δ Rating	…
Russia	Saudi Arabia	0.0148	1	0	0.1	0	0.469	…
Saudi Arabia	Russia	-0.0148	0	1	-0.1	0	-0.469	…
Egypt	Uruguay	0.0128	0	0	0.9	1	-0.395	…
Uruguay	Egypt	-0.0128	0	0	-0.9	-1	0.395	…
Morocco	Iran	-0.0061	0	0	0.0	1	0.035	…
Iran	Morocco	0.0061	0	0	0.0	-1	-0.035	…
Portugal	Spain	-0.0048	0	0	-0.1	-6	-0.233	…
Spain	Portugal	0.0048	0	0	0.1	6	0.233	…

REGRESSION TREE



- The tree's are trained on the games of the 4 previous World Cups (2002, 2006, 2010, 2014)
- Splits chosen based on highest variance reduction

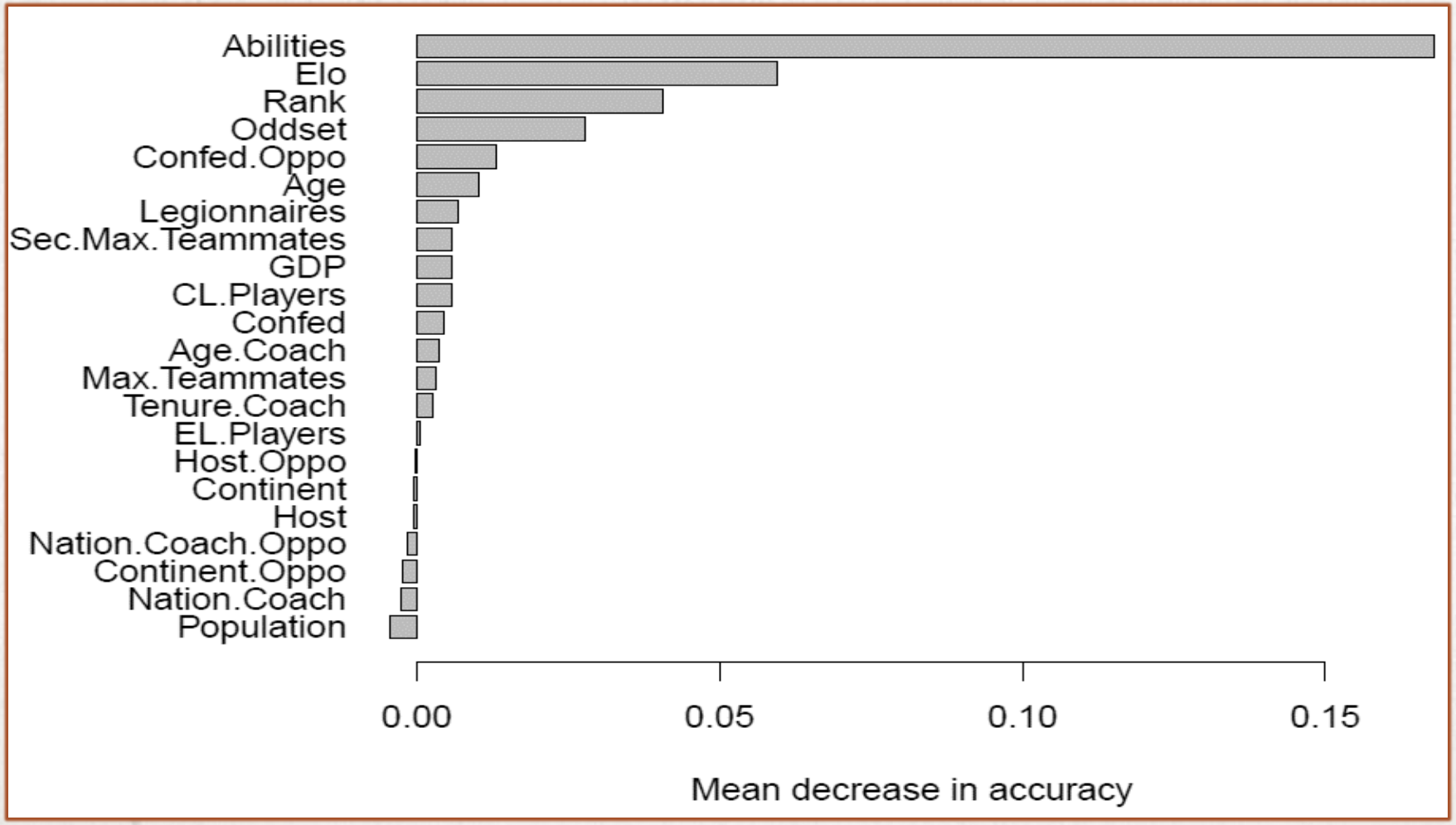
RANDOM FOREST

- “Forest”: Combination of many regression trees
- “Random”
  - For every tree, we take a random sample from our training data
  - At every node in each tree, we only consider a random subset of the covariates to make the best split.

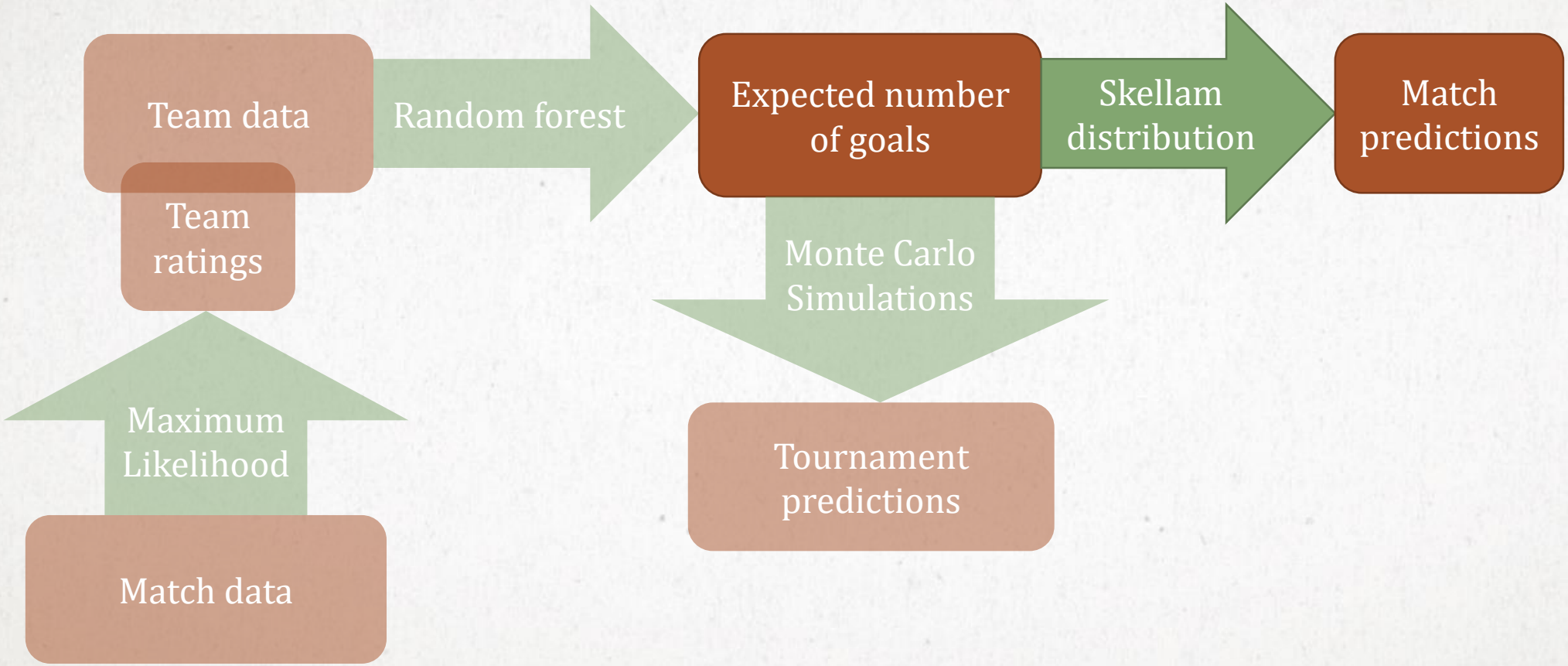
OUTPUT RANDOM FOREST

Team	Opponent	Δ Population	Host	Opp.Host	Δ avg. age	Δ#CL Players	Δ Rating	...	Exp. goals
Russia	Saudi Arabia	0.0148	1	0	0.1	0	0.469	...	1.89
Saudi Arabia	Russia	-0.0148	0	1	-0.1	0	-0.469	...	0.96
Egypt	Uruguay	0.0128	0	0	0.9	1	-0.395	...	0.79
Uruguay	Egypt	-0.0128	0	0	-0.9	-1	0.395	...	1.90
Morocco	Iran	-0.0061	0	0	0.0	1	0.035	...	1.30
Iran	Morocco	0.0061	0	0	0.0	-1	-0.035	...	1.12
Portugal	Spain	-0.0048	0	0	-0.1	-6	-0.233	...	0.69
Spain	Portugal	0.0048	0	0	0.1	6	0.233	...	1.51

VARIABLE IMPORTANCE



THE HYBRID RANDOM FOREST



PREDICTION OF SINGLE MATCHES

- Example: France vs Croatia
- The Random Forest gave us:

$$\lambda_{France\ vs\ Croatia} = 1.54$$

$$\lambda_{Croatia\ vs\ France} = 0.75$$

- We assume the goal difference is Skellam distributed with means 1.54 and 0.75, which gives us
  - $P(France\ wins) = P(Goal\ Difference > 0)$ =56%
  - $P(Draw) = P(Goal\ Difference = 0) = 26\%$
  - $P(Croatia\ wins) = P(Goal\ Difference < 0) = 18\%$

THE HYBRID RANDOM FOREST

TOURNAMENT SIMULATIONS

- For each team in each game in the group stage: sample the number of goals from the Poisson distribution, given the expected number of goals.

TOURNAMENT SIMULATIONS

- For each team in each game in the group stage: sample the number of goals from the Poisson distribution, given the expected number of goals.
- Deduce the outcome of each game and the standings in each group

TOURNAMENT SIMULATIONS

- For each team in each game in the group stage: sample the number of goals from the Poisson distribution, given the expected number of goals.
- Deduce the outcome of each game and the standings in each group
- Deduce which teams will face each other in the round of 16

TOURNAMENT SIMULATIONS

- For each team in each game in the group stage: sample the number of goals from the Poisson distribution, given the expected number of goals.
- Deduce the outcome of each game and the standings in each group
- Deduce which teams will face each other in the round of 16
- Calculate the expected goals for the round of 16 and simulate these games in the same way as in the group stage

TOURNAMENT SIMULATIONS

- For each team in each game in the group stage: sample the number of goals from the Poisson distribution, given the expected number of goals.
- Deduce the outcome of each game and the standings in each group
- Deduce which teams will face each other in the round of 16
- Calculate the expected goals for the round of 16 and simulate these games in the same way as in the group stage
- Do the same for the quarter-finals, semi-finals & final

TOURNAMENT SIMULATIONS

- For each team in each game in the group stage: sample the number of goals from the Poisson distribution, given the expected number of goals.
- Deduce the outcome of each game and the standings in each group
- Deduce which teams will face each other in the round of 16
- Calculate the expected goals for the round of 16 and simulate these games in the same way as in the group stage
- Do the same for the quarter-finals, semi-finals & final
- Repeat 100 000 times

RESULTS

	Team	Winning prob.
1	Spain	13,7%
2	Germany	11,5%
3	France	10,8%
4	Brazil	10,3%
5	Belgium	9,9%
6	England	7,5%
7	Argentina	5,4%
8	Croatia	3,8%
9	Portugal	3,2%
10	Colombia	3,2%

11	Switzerland	2,9%
12	Uruguay	2,8%
13	Denmark	2,6%
14	Sweden	1,9%
15	Serbia	1,6%
16	Poland	1,3%
17	Peru	1,3%
18	Iceland	1,0%
19	Senegal	1,0%
20	Morocco	0,8%
21	Mexico	0,7%

22	Tunisia	0,7%
23	Australia	0,3%
24	Nigeria	0,3%
25	Costa Rica	0,3%
26	Egypt	0,3%
27	Russia	0,3%
28	Japan	0,2%
29	South-Korea	0,2%
30	Iran	0,1%
31	Panama	0,1%
32	Saoudi Arabia	0,0%

EVALUATION

We “predicted” Spain as the world champion

But…


6
2
18
19


THE METHOD IS WRONG

EVALUATION

Method	Rank Probability Score (RPS)
Hybrid Random Forest	0.190
Random Forest	0.193
Bookmakers	0.194
Ranking	0.194
Hybrid Lasso	0.197
Lasso	0.207

$$RPS = \frac{1}{2M} \sum_{m=1}^{M} \left(P_{m,H} - y_{m,H}\right)^2 + \left(P_{m,A} - y_{m,A}\right)^2$$

EVALUATION

On website fifaexperts.com

- Probabilities (1X2) for each game of the world cup
- Evaluation based on the Brier Score
- More then 500 participating teams

EVALUATION

On website fifaexperts.com

- Probabilities (1X2) for each game of the world cup
- Evaluation based on the Brier Score
- More then 500 participating teams

1. Esportes em Números: 4650 points
2. Andreas Groll: 4644 points
3. Danilo Lopes: 4634 points
4. Natanael Prata: 4634 points
5. Chance de Gol: 4611 points
6. Wilson Chaves: 4597 points
7. Sigma Benedek: 4589 points
8. Márcio Diniz: 4587 points
9. Francesco Beatrice: 4574 points
10. Alun Owen: 4565 points

EVALUATION

On website fifaexperts.com

- Probabilities (1X2) for each game of the world cup
- Evaluation based on the Brier Score
- More then 500 participating teams

"it's not the winning but the taking part that counts"

1. Esportes em Números: 4650 points
2. Andreas Groll: 4644 points
3. Danilo Lopes: 4634 points
4. Natanael Prata: 4634 points
5. Chance de Gol: 4611 points
6. Wilson Chaves: 4597 points
7. Sigma Benedek: 4589 points
8. Márcio Diniz: 4587 points
9. Francesco Beatrice: 4574 points
10. Alun Owen: 4565 points

CONCLUSIONS

- Combine statistical methods (Maximum Likelihood Estimation) with machine learning methods (Random Forest)

Hans.VanEetvelde@UGent.be

CONCLUSIONS

- Combine statistical methods (Maximum Likelihood Estimation) with machine learning methods (Random Forest)
- Work together!
  - Christophe Ley and myself: ranking methods
  - Andreas Groll and Gunther Schauberger: regression methods

Hans.VanEetvelde@UGent.be

CONCLUSIONS

- Combine statistical methods (Maximum Likelihood Estimation) with machine learning methods (Random Forest)
- Work together!
  - Christophe Ley and myself: ranking methods
  - Andreas Groll and Gunther Schauberger: regression methods
  - (Women World Cup) Achim Zeileis: ratings based on bookmakers odds and inverse tournament simulation

Hans.VanEetvelde@UGent.be