IMPROVING MEDAL PROJECTIONS IN OLYMPIC BEST MARK SPORTS

UNITED STATES OLYMPIC & PARALYMPIC COMMITTEE

Suraj Bhuva

USOPC, Performance Data Scientist

September 23, 2023

What are best mark sports?

CGS sports

- Events that are measured in centimeters, grams and seconds

Defined by lack of interaction between competitors

- Results are less dependent on the competitive field

Examples:

Weightlifting

- Winners in each weight class are the athletes that record the greatest weight lifted [*Score = max(Clean & Jerk) + max(Snatch)*]

Athletics

- Track events
  - Winners are the top 3 athletes with the quickest times
- Field events
  - Winners are the top 3 athletes with the highest marks

What are medal projections and how are they determined?

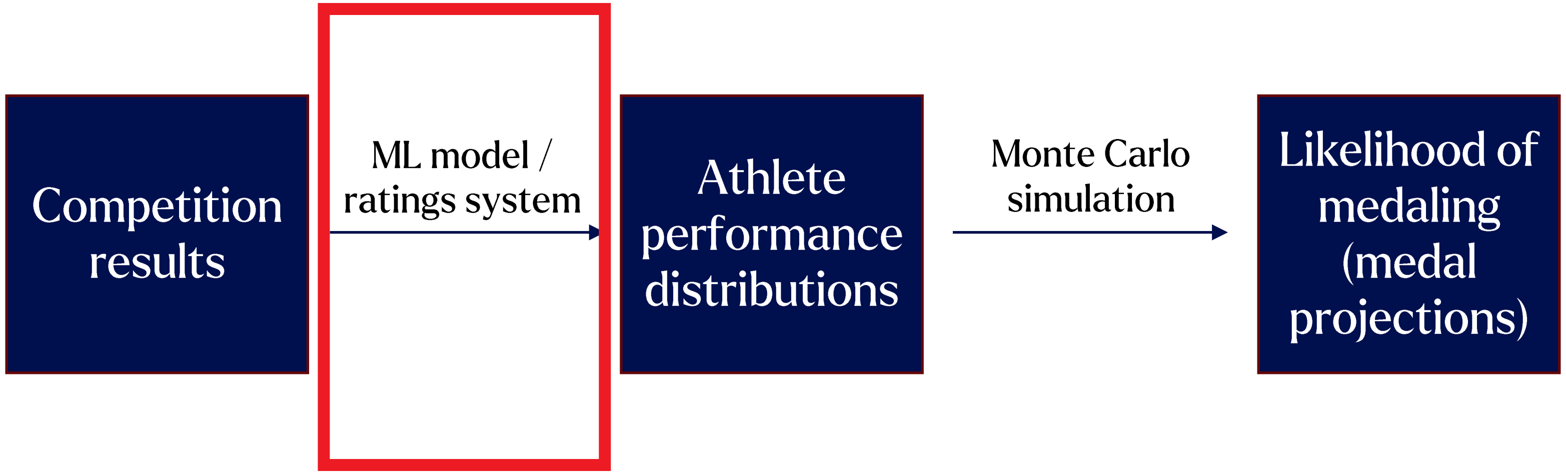
Medal Projections

Frequency with which we expect competitors to win a medal at the Olympic Games

What are medal projections and how are they determined?

Medal Projections

Frequency with which we expect competitors to win a medal at the Olympic Games



Sparse Competition Data

Competitor	NOC	EventGender	Event	final_mark	final_medal	Year
Lamont Marcell Jacobs	ITA	Men	100m	9.8	1	2021
Fred Kerley	USA	Men	100m	9.84	1	2021
Andre De Grasse	CAN	Men	100m	9.89	1	2021
Akani Simbine	RSA	Men	100m	9.93	0	2021
Ronnie Baker	USA	Men	100m	9.95	0	2021
Bingtian Su	CHN	Men	100m	9.98	0	2021
Enoch Adegoke	NGR	Men	100m	DNF	NULL	2021
Zharnel Hughes	GBR	Men	100m	DQ	NULL	2021

Imbalanced class data can lead to model bias

- Medalists account for 37.5% of data (3 medalists, 5 non-medalists) in the finals
- In practice, this number can be lower from DQ's after the competition, or for events with larger competition pools
- Proportion of medals from the entire Olympic field is much lower
- Tokyo 2021 saw 84 competitors in the men's 100m participant pool (3.6% of observations are medals)

Small quantity of data can lead to overfitting

- From Beijing 2008 – Tokyo 2021, only 28 "valid" observations for the men's 100m final

Predicting Medal Probabilities at Tokyo 2021

Initial Model

Predicting Medal Probabilities at Tokyo 2021

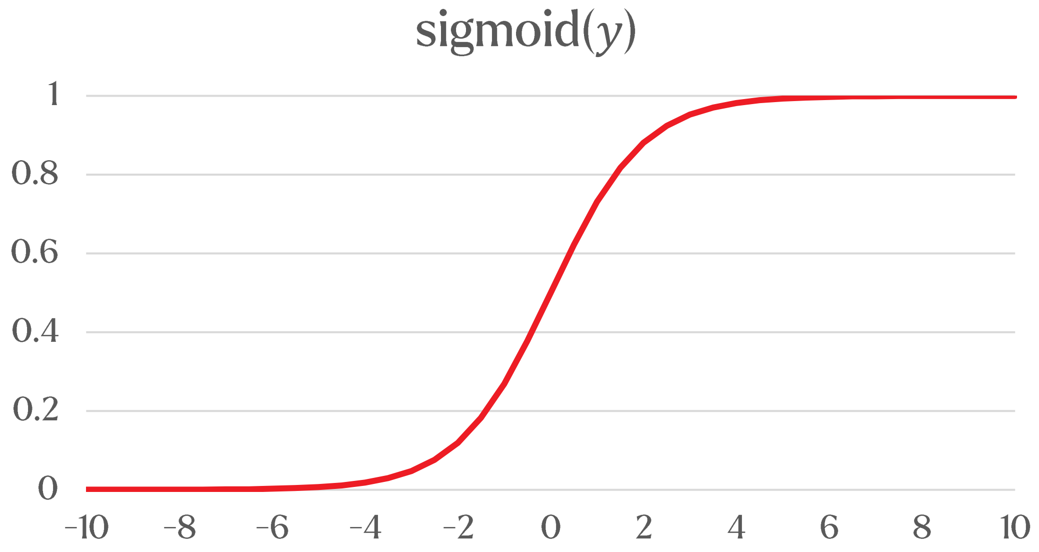
Initial Model

Logistic Regression Model:

$y = \dfrac{1}{1 + e^{-z}}$ where y is a probability between [0,1]

$z = b + w_1x_1 + w_2x_2 + \cdots + w_nx_n$

w_n = model weights, x_n = model features, b = bias term



Aggregated 1500m Data

Input Data

- Aggregated data on a 4-year window leading into each Olympic Games
- Build features and scale data where needed

Train Model

- Create logistic regression model using input features where dependent variable is the binary "final_medal" column

event	Competitor	NOC	feature_1	feature_2	feature_3	feature_4	final_medal	final_mark
1500m	Timothy Cheruiyot	KEN	214.0774	8721	31	208.41	1	209.26
1500m	Taoufik Makhloufi	ALG	214.535	11482	3	211.35	1	211.38
1500m	Marcin Lewandowski	POL	216.5827	11803	6	211.95	1	211.46
1500m	Jakob Ingebrigtsen	NOR	217.101	6956	13	210.16	0	211.7
1500m	Jake Wightman	GBR	217.7111	9218	9	213.96	0	211.87
1500m	Josh Kerr	GBR	218.6543	8033	2	213.6	0	212.52
1500m	Ronald Kwemoi	KEN	215.9082	8783	6	210.49	0	212.72
1500m	Matthew Centrowitz	USA	217.5324	10945	5	211.77	0	212.81
1500m	Asbel Kiprop	KEN	217.7408	6990	11	211.64	1	213.11
1500m	Kalle Berglund	SWE	218.7088	8609	4	214.89	0	213.7
1500m	Taoufik Makhloufi	ALG	215.4307	8866	9	210.8	1	214.08
1500m	Nick Willis	NZL	216.0306	9248	9	212.17	1	214.16
1500m	Mehdi Baala	FRA	215.3207	10960	14	210.8	1	214.21
1500m	Craig Engels	USA	217.7386	9289	5	214.04	0	214.24
1500m	Juan Carlos Higuero	ESP	216.1405	10974	10	211.57	0	214.44
1500m	Abdelaati Iguider	MAR	215.9478	7818	20	211.88	0	214.66
1500m	Juan Van Deventer	RSA	219.2071	9278	16	214.46	0	214.77
1500m	Leonel Manzano	USA	217.2348	10191	11	212.37	1	214.79
1500m	Abdelaati Iguider	MAR	215.6722	9267	14	211.47	1	215.13
1500m	Matthew Centrowitz	USA	218.8908	8329	6	214.46	0	215.17
1500m	Belal Mansoor Ali	BRN	217.0388	7246	18	211.49	0	215.23
1500m	Andrew Baddeley	GBR	218.5781	9557	14	214.36	0	215.37
1500m	Henrik Ingebrigtsen	NOR	219.6943	7835	2	216.39	0	215.43
1500m	Mekonnen Gebremedhin	ETH	214.6042	8701	12	211.45	0	215.44
1500m	Augustine Kiprono Choge	KEN	214.1607	7881	13	211.57	0	215.5
1500m	Silas Kiplagat	KEN	215.2673	8388	21	209.27	0	216.19
1500m	Ilham Tanui Ozbilen	KEN	215.5075	8191	10	211.37	0	216.72
1500m	Nick Willis	NZL	216.6353	10697	15	210.35	0	216.94
1500m	Neil Gourley	GBR	219.406	9007	4	215.95	0	217.3
1500m	Daham Najim Bashir	QAT	217.4733	10512	6	211.04	0	217.68
1500m	Youssouf Hiss Bachir	DJI	219.0333	11966	3	215.74	0	217.96
1500m	Belal Mansoor Ali	BRN	216.6066	8695	18	212.1	0	217.98
1500m	Nixon Kiplimo Chepseba	KEN	214.6823	7909	19	209.77	0	219.04
1500m	Christian Obrist	ITA	220.0057	10134	10	215.32	0	219.87

Medal Projections for 1500m

Evaluate Predictions

- Utilize brier score to measure accuracy of probabilistic predictions

$$\frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$$

p_i = probability of medaling for athlete i

y_i = binary value (medal/no medal) for athlete i

N = number of athletes in training set

event	Competitor	NOC	final_mark	final_medal	medal_projections
1500m	Jakob Ingebrigtsen	NOR	208.32	1	0.4724
1500m	Timothy Cheruiyot	KEN	209.01	1	0.0236
1500m	Josh Kerr	GBR	209.05	1	0.0472
1500m	Abel Kipsang	KEN	209.56	0	0.5734
1500m	Adel Mechaal	ESP	210.77	0	0.0866
1500m	Cole Hocker	USA	211.4	0	0.0102
1500m	Stewart Mcsweyn	AUS	211.91	0	0.6956
1500m	Michal Rozmys	POL	212.67	0	0.0038
1500m	Jake Heyward	GBR	214.43	0	0.216
1500m	Jake Wightman	GBR	215.09	0	0.1446
1500m	Oliver Hoare	AUS	215.79	0	0.1454
1500m	Charles Grethen	LUX	216.8	0	0.076
1500m	Ignacio Fontes	ESP	218.56	0	0.5052

Brier Score: 0.255

Equal chance Brier Score: 0.178

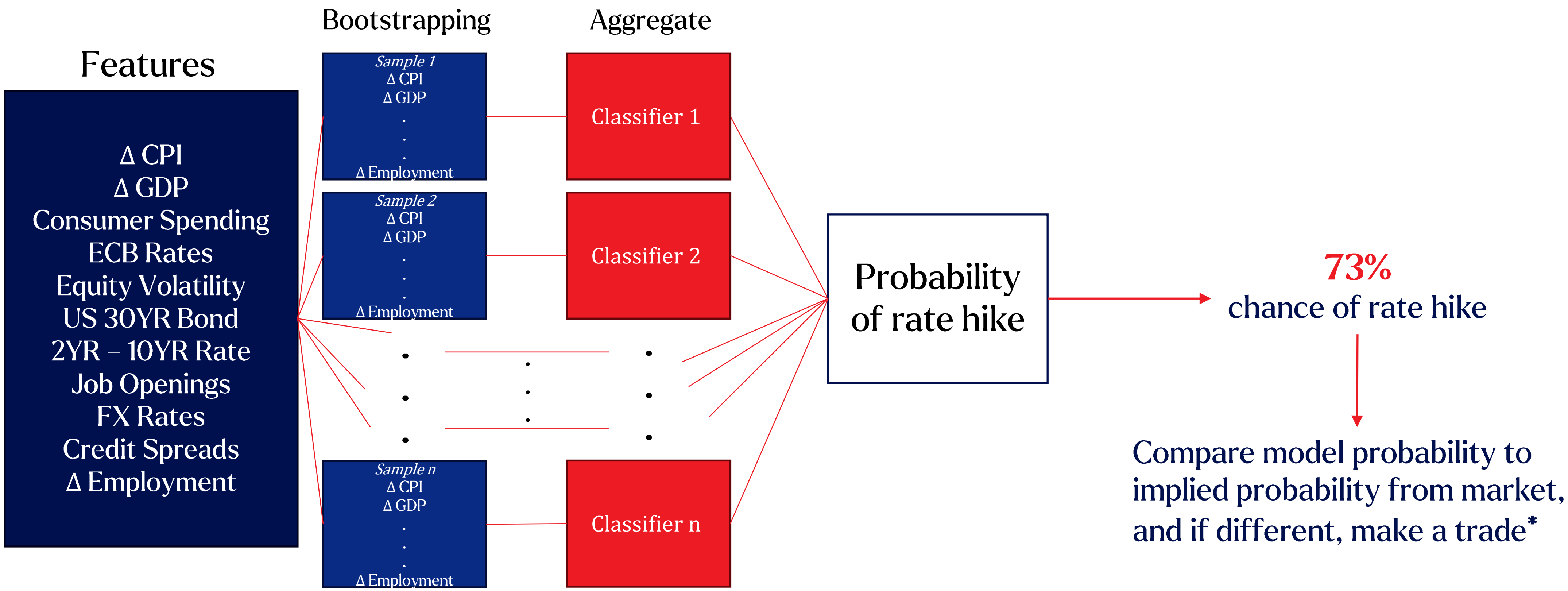
Medal Projections for 1500m

How can we get more robust predictions on this limited dataset?

Ensemble Learning

Predicting Probability of a Federal Reserve Rate Hike

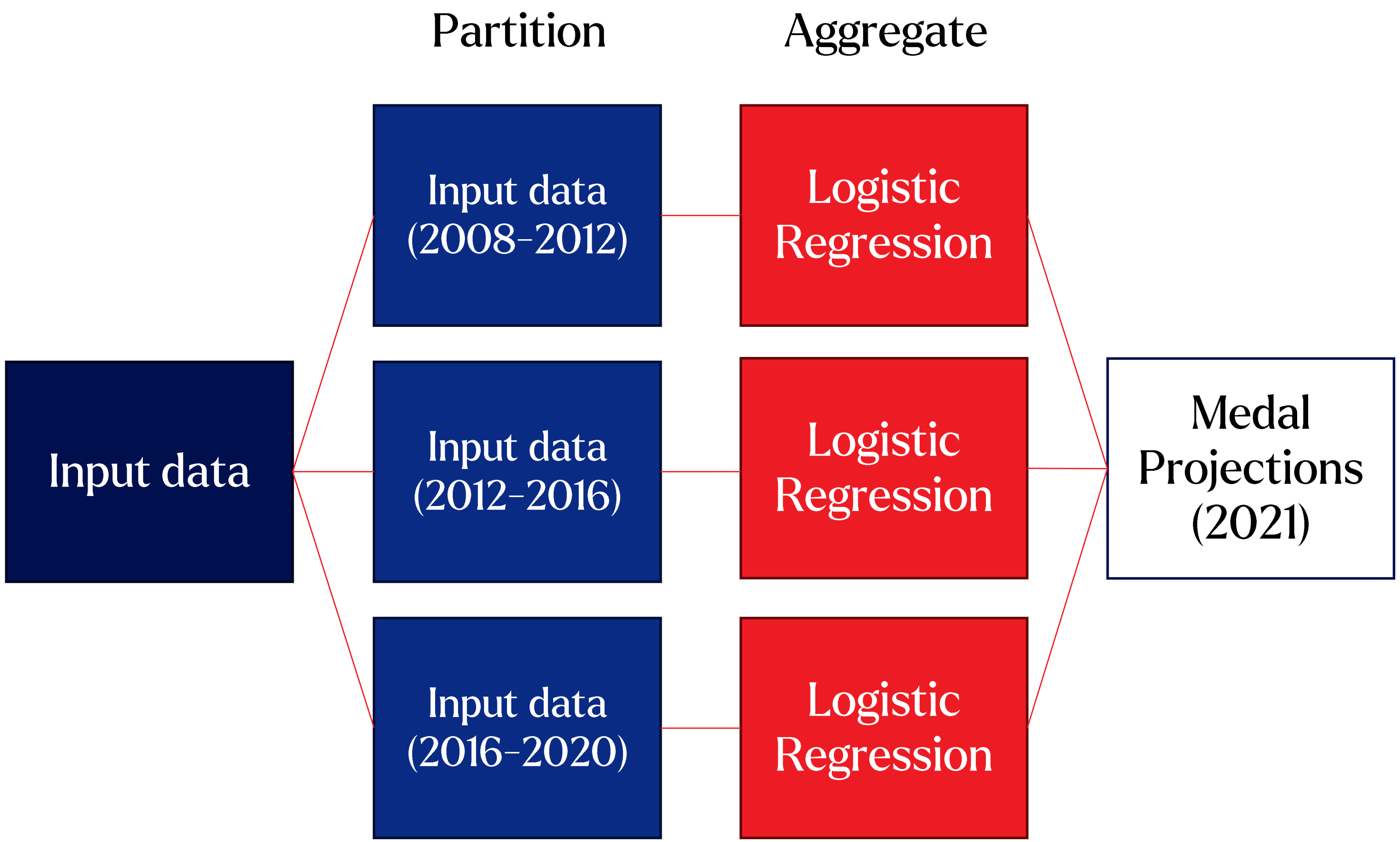
- Federal Reserve (Fed) conducts meetings most months to assess the strength of the economy and modify monetary policy
- The Fed can decide to modify the prime rate (the rate banks will lend to customers with good credit) higher or lower
- Theory suggests that by modifying this rate lower (higher), inflation will increase (decrease)



*The idea of mismatched odds has applications in sports betting. If you believe you have better information about the true odds of an event vs. what a sports book is showing, one can figure out an optimal bet size to maximize wealth (see *Fortune's Formula*)

Ensemble Learning

Proposed Model



Why does this approach work?

- Unique classifiers are trained on datasets which retain "local" characteristics
- Averaging predictions of different weak-learners helps create more accurate athlete performance distributions
- Diversification reduces variance
    - Modern Portfolio Theory

Ensemble Learning: Training Data

Partitioned training data by Olympic cycle results in greater consistency of data points

- There will not be any data points where non-medaling final marks are better than medaling final marks
- Each dataset represents the competitive landscape at that time
  - Advances in sport/tech do not have undue bias
- Process will now generate 3 different predictions for each athlete
  - Sparse data problem is improved, but not solved

event	Competitor	NOC	feature_1	feature_2	feature_3	feature_4	final_medal	final_mark
1500m	Asbel Kiprop	KEN	217.7408	6990	11	211.64	1	213.11
1500m	Nick Willis	NZL	216.0306	9248	9	212.17	1	214.16
1500m	Mehdi Baala	FRA	215.3207	10960	14	210.8	1	214.21
1500m	Juan Carlos Higuero	ESP	216.1405	10974	10	211.57	0	214.44
1500m	Abdelaati Iguider	MAR	215.9478	7818	20	211.88	0	214.66
1500m	Juan Van Deventer	RSA	219.2071	9278	16	214.46	0	214.77
1500m	Belal Mansoor Ali	BRN	217.0388	7246	18	211.49	0	215.23
1500m	Andrew Baddeley	GBR	218.5781	9557	14	214.36	0	215.37
1500m	Augustine Kiprono Choge	KEN	214.1607	7881	13	211.57	0	215.5
1500m	Daham Najim Bashir	QAT	217.4733	10512	6	211.04	0	217.68
1500m	Christian Obrist	ITA	220.0057	10134	10	215.32	0	219.87

event	Competitor	NOC	feature_1	feature_2	feature_3	feature_4	final_medal	final_mark
1500m	Taoufik Makhloufi	ALG	215.4307	8866	9	210.8	1	214.08
1500m	Leonel Manzano	USA	217.2348	10191	11	212.37	1	214.79
1500m	Abdelaati Iguider	MAR	215.6722	9267	14	211.47	1	215.13
1500m	Matthew Centrowitz	USA	218.8908	8329	6	214.46	0	215.17
1500m	Henrik Ingebrigtsen	NOR	219.6943	7835	2	216.39	0	215.43
1500m	Mekonnen Gebremedhin	ETH	214.6042	8701	12	211.45	0	215.44
1500m	Silas Kiplagat	KEN	215.2673	8388	21	209.27	0	216.19
1500m	Ilham Tanui Ozbilen	KEN	215.5075	8191	10	211.37	0	216.72
1500m	Nick Willis	NZL	216.6353	10697	15	210.35	0	216.94
1500m	Belal Mansoor Ali	BRN	216.6066	8695	18	212.1	0	217.98
1500m	Nixon Kiplimo Chepseba	KEN	214.6823	7909	19	209.77	0	219.04

event	Competitor	NOC	feature_1	feature_2	feature_3	feature_4	final_medal	final_mark
1500m	Timothy Cheruiyot	KEN	214.0774	8721	31	208.41	1	209.26
1500m	Taoufik Makhloufi	ALG	214.535	11482	3	211.35	1	211.38
1500m	Marcin Lewandowski	POL	216.5827	11803	6	211.95	1	211.46
1500m	Jakob Ingebrigtsen	NOR	217.101	6956	13	210.16	0	211.7
1500m	Jake Wightman	GBR	217.7111	9218	9	213.96	0	211.87
1500m	Josh Kerr	GBR	218.6543	8033	2	213.6	0	212.52
1500m	Ronald Kwemoi	KEN	215.9082	8783	6	210.49	0	212.72
1500m	Matthew Centrowitz	USA	217.5324	10945	5	211.77	0	212.81
1500m	Kalle Berglund	SWE	218.7088	8609	4	214.89	0	213.7
1500m	Craig Engels	USA	217.7386	9289	5	214.04	0	214.24
1500m	Neil Gourley	GBR	219.406	9007	4	215.95	0	217.3
1500m	Youssouf Hiss Bachir	DJI	219.0333	11966	3	215.74	0	217.96

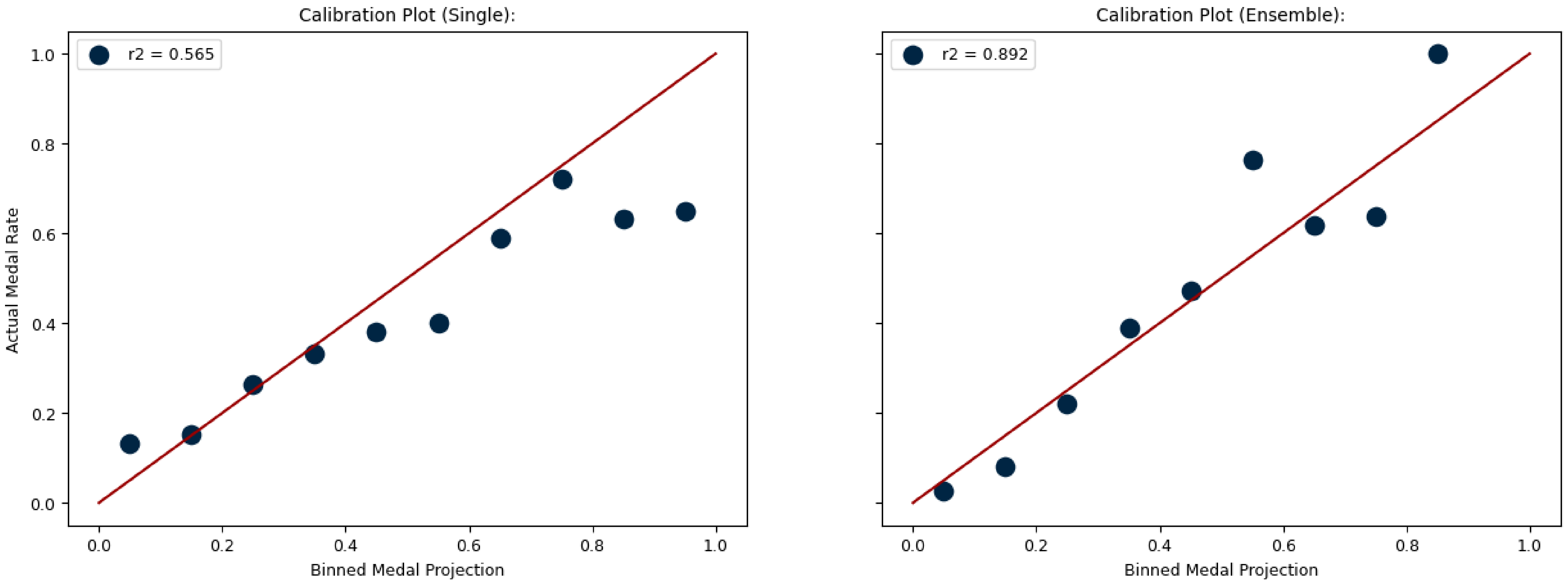
Ensemble Learning: Projections

event	Competitor	NOC	final_mark	final_medal	old_medal_projections	new_medal_projections
1500m	Jakob Ingebrigtsen	NOR	208.32	1	0.4724	0.4216
1500m	Timothy Cheruiyot	KEN	209.01	1	0.0236	0.4228
1500m	Josh Kerr	GBR	209.05	1	0.0472	0.2552
1500m	Abel Kipsang	KEN	209.56	0	0.5734	0.24
1500m	Adel Mechaal	ESP	210.77	0	0.0866	0.1648
1500m	Cole Hocker	USA	211.4	0	0.0102	0.1174
1500m	Stewart Mcsweyn	AUS	211.91	0	0.6956	0.4086
1500m	Michal Rozmys	POL	212.67	0	0.0038	0.1258
1500m	Jake Heyward	GBR	214.43	0	0.216	0.1424
1500m	Jake Wightman	GBR	215.09	0	0.1446	0.3016
1500m	Oliver Hoare	AUS	215.79	0	0.1454	0.1424
1500m	Charles Grethen	LUX	216.8	0	0.076	0.0702
1500m	Ignacio Fontes	ESP	218.56	0	0.5052	0.1872

Initial Model Brier Score: 0.255
Equal chance Brier Score: 0.178
New Brier Score: 0.129

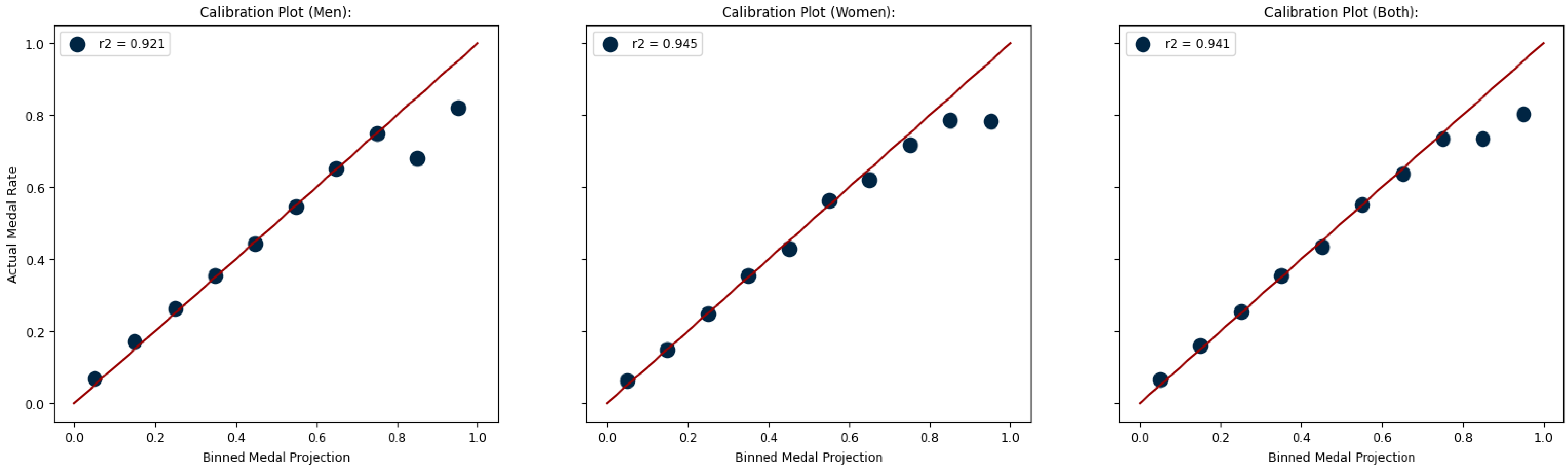
Ensemble Learning: Calibration Plots

Men's + Women's 1500m, Tokyo 2021



Ensemble Learning: Calibration Plots

All events, all Olympic games



How to improve accuracy of high probability predictions?

Ensemble Learning: Considerations/Takeaways

Overall brier score improved, but certain sports saw decreased accuracy

- Are there certain sports/characteristics of datasets that are better suited to an ensemble approach?

High medal projections can be better

- Bagging multiple predictions means medal projections (>0.8) tend to imply a competitor is a very strong winner
- Calibration plots indicate these projections are not well tuned

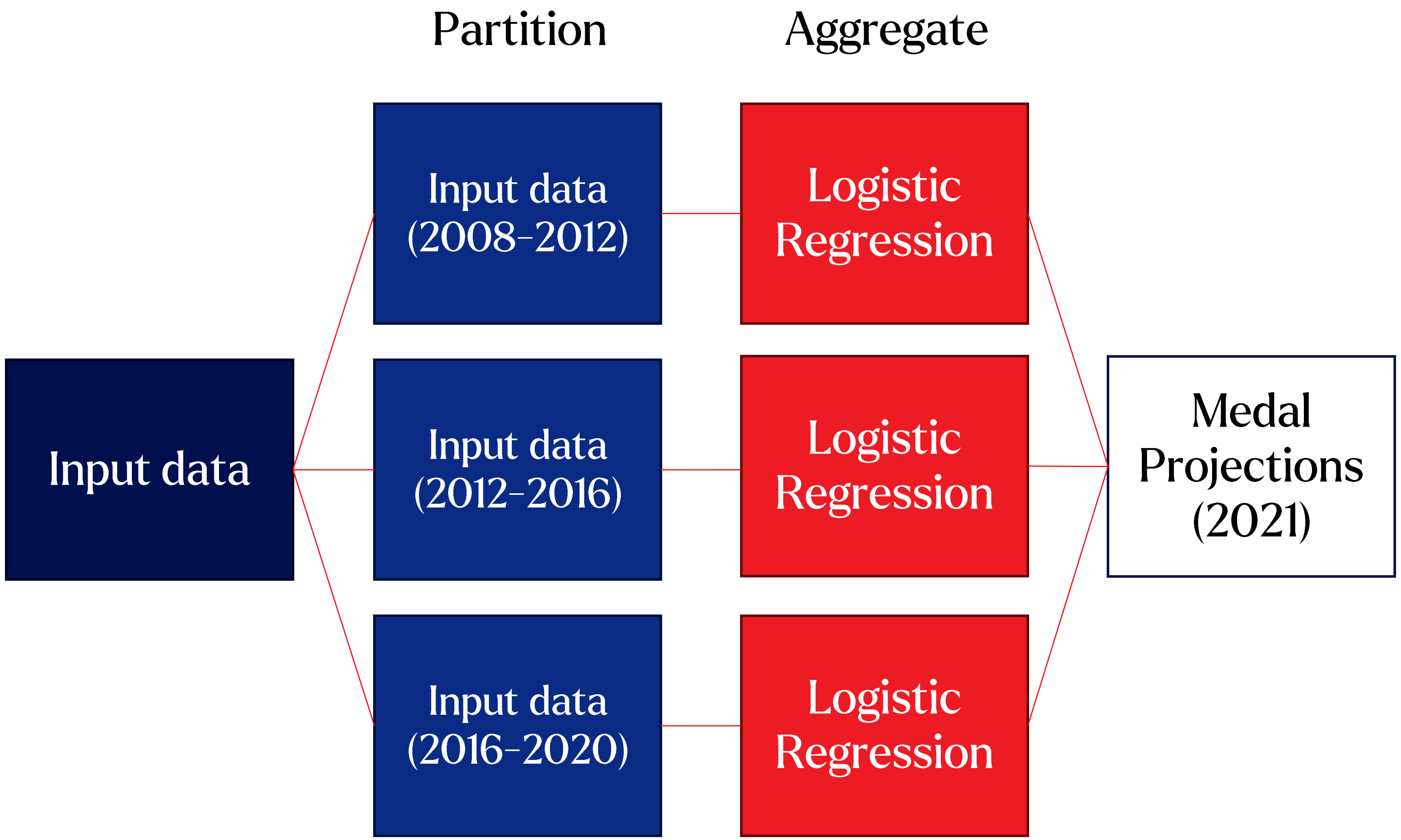
Can prediction aggregation be improved?

- Is averaging the best way to bag all predictions?

event	brier_single	brier_ensemble	Change
110m Hurdles	0.1237	0.1932	+0.070
200m	0.0888	0.1306	+0.042
800m	0.1567	0.1863	+0.030
3000m Steeplechase	0.0743	0.0925	+0.018
400m Hurdles	0.1516	0.1692	+0.018
Shot Put	0.0809	0.0907	+0.010
Javelin Throw	0.1965	0.203	+0.007
400m	0.1985	0.2022	+0.004
Marathon	0.0437	0.0458	+0.002
Pole Vault	0.1236	0.1255	+0.002
Hammer Throw	0.1459	0.1432	-0.003
Triple Jump	0.1511	0.1461	-0.005
Long Jump	0.1235	0.1176	-0.006
High Jump	0.1709	0.163	-0.008
Decathlon	0.0851	0.0712	-0.014
Discus Throw	0.1976	0.1641	-0.033
10000m	0.1019	0.066	-0.036
100m	0.3426	0.2523	-0.090
1500m	0.2546	0.1289	-0.126
5000m	0.3164	0.1263	-0.190

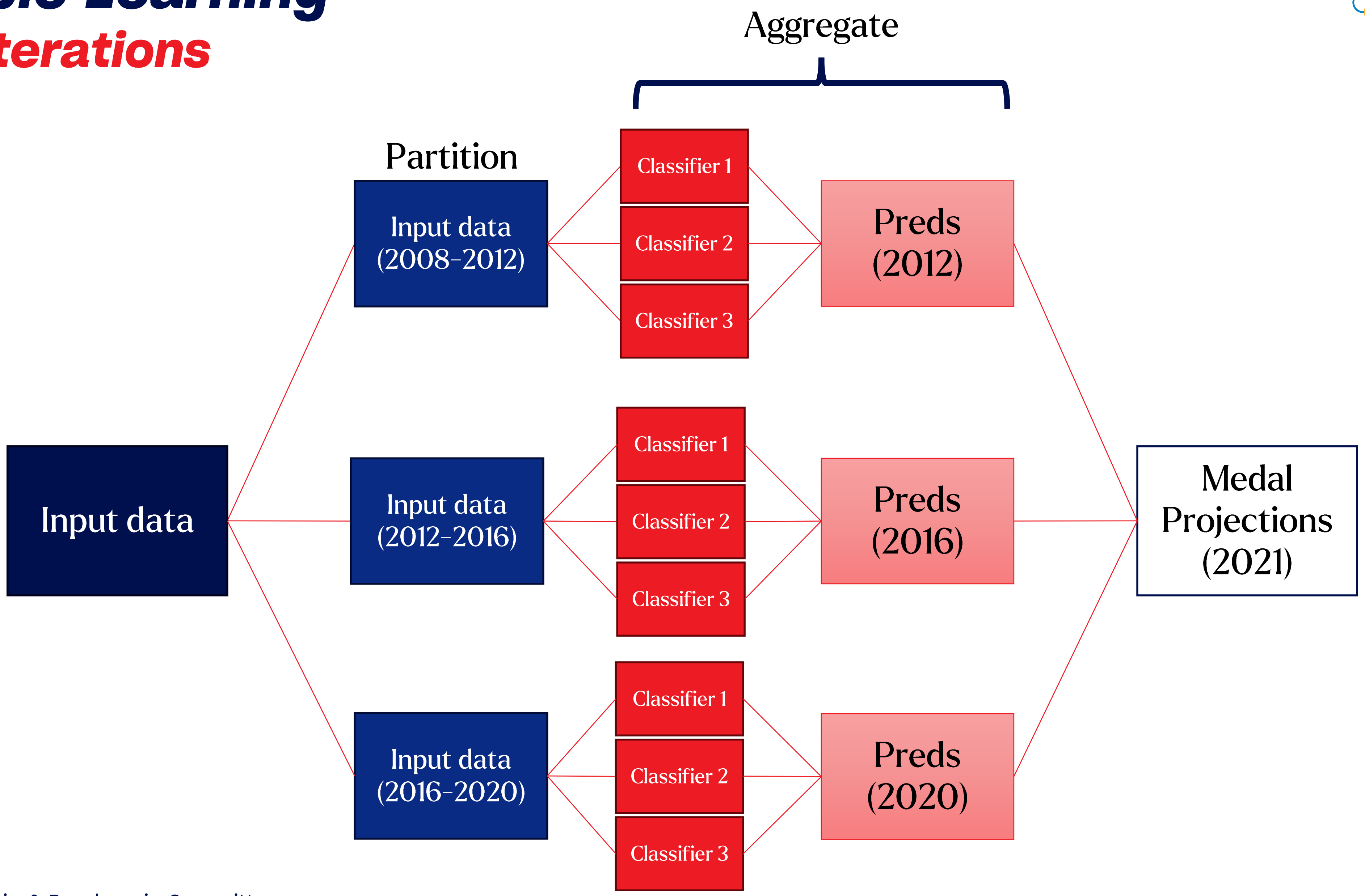
Ensemble Learning

First Iteration



Ensemble Learning

Future Iterations



QUESTIONS?