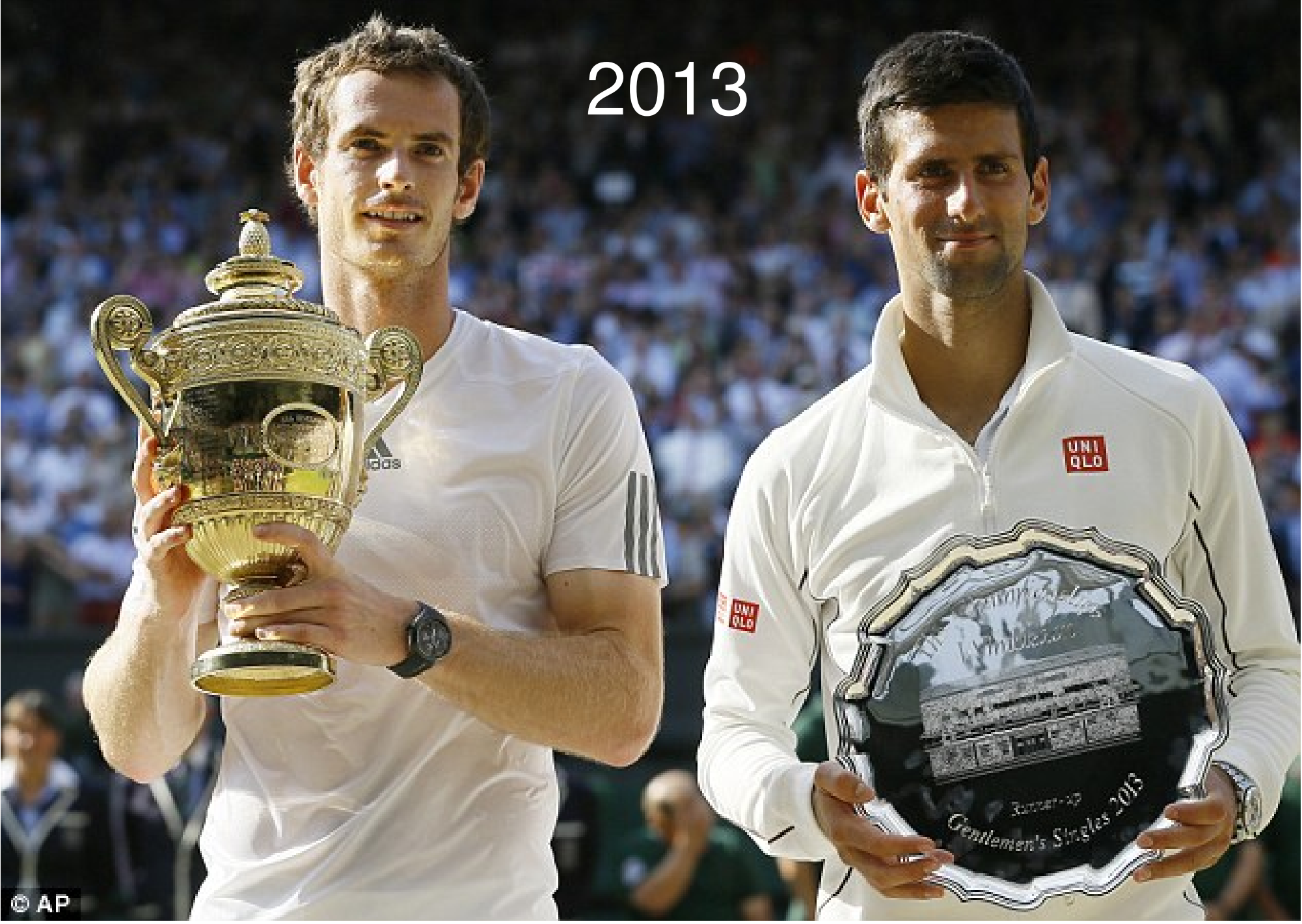
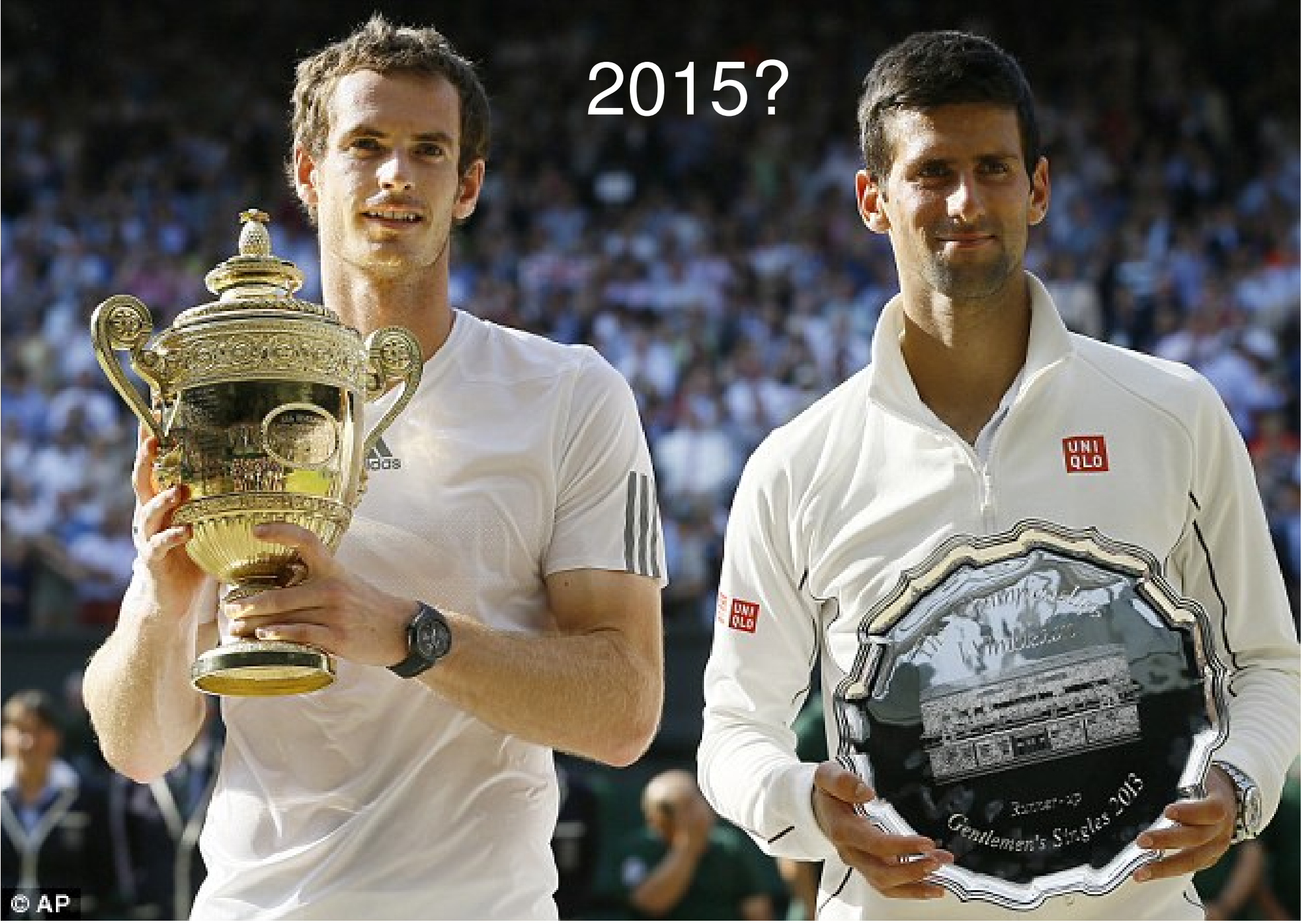
Will Andy Murray win Wimbledon 2015?

Robert Johnson, Sporting Advantage Limited

PyDataLondon, 7th July 2015


2013
UNI QLO
Gentlemen's Singles 2013
© AP


2015?
UNI QLO
Gentlemen's Singles 2015
© AP

Wimbledon Questions

- Murray won in 2013. Will he win again this year?
- Is Murray on better form this year than in 2013?
- How do the top players compare in the mens' tournament?

Outline

Data sources.

- Tennis results are freely available: `www.tennis-data.co.uk`.
- I used all ATP matches from 2010 until now.
- Start estimating player abilities from January 2012.

Deriving the model
Maximum likelihood estimation

- Model inspired by Dixon and Coles (1997) and McHale and Morton (2011).
- Assume each player has an (unknown) ability parameter θ.
- Model P(Player *i* defeats player *j*) using a logistic function of $\theta_i - \theta_j$

$$\text{P}(i \text{ defeats } j) = \frac{1}{1+\exp\left(-(\theta_i - \theta_j)\right)}$$

Likelihood function

- $L(\theta, k = 1, \ldots, n) = \prod_{k=1}^{n} \frac{1}{1+\exp(-(\theta_i - \theta_j))}$
- Player indices i and j in the equation above depend on match k.

Dynamic model

- Let's follow the method of Dixon and Coles (1997) and exponentially downweight older matches
- Gives us the pseudo-likelihood for each time point t:

$$L_t(\theta;\phi) = \prod_{k \in A_t} \left\{ \frac{1}{1+\exp\left(-(\theta_i - \theta_j)\right)} \right\}^{\phi(t-t_k)}$$

- where t_k is the time match k is played, $A_t = \{k : t_k < t\}$ and ϕ is an exponential downweighting function.

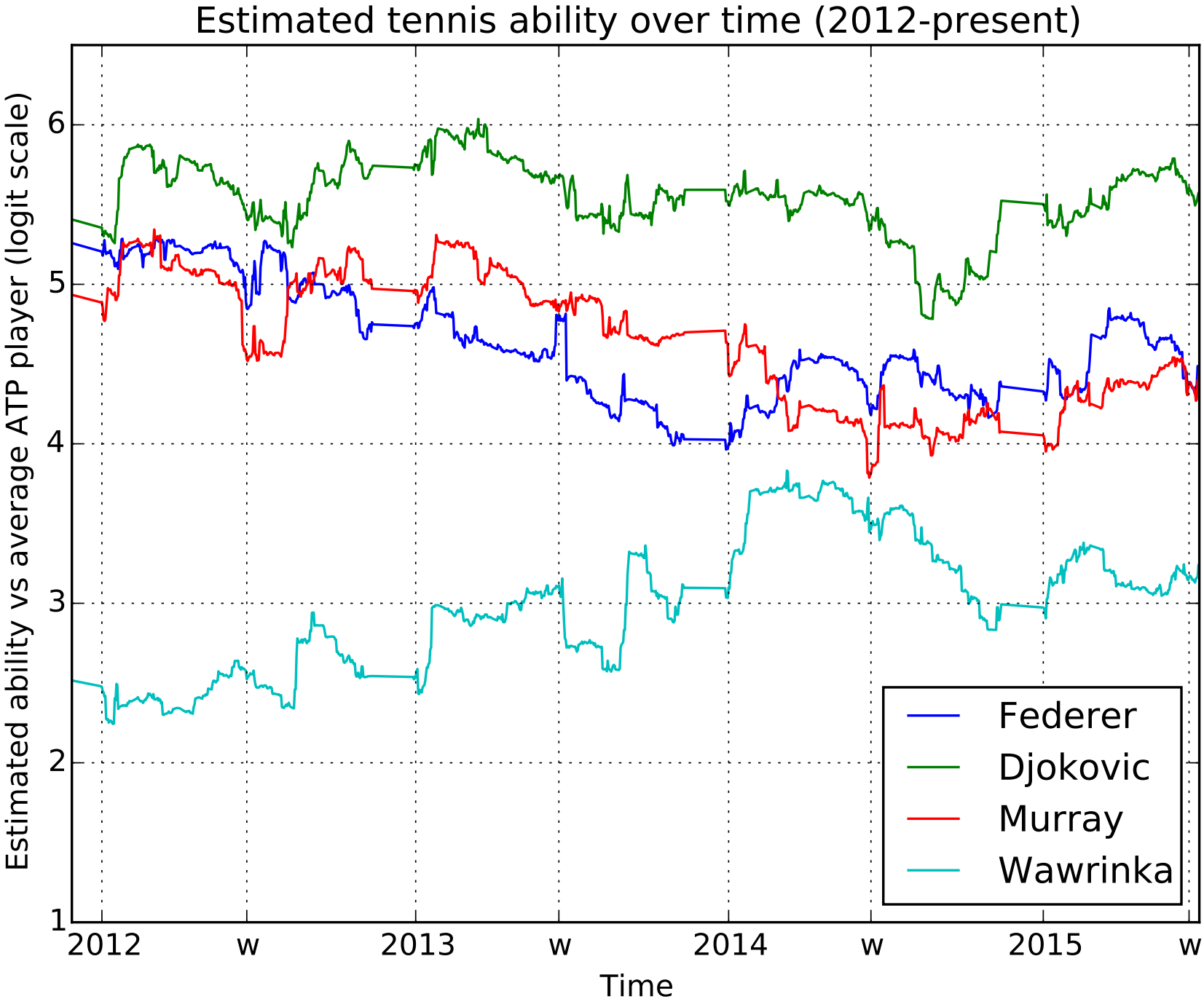
Implementation

- Coded that all up in Python last Sunday
- Used numpy, pandas, scipy
- Numerically maximised the pseudo-likelihood using minimize function in scipy.optimize
- Fitted on a day by day rolling basis
- Plotted using matplotlib

Implementation

Care needs to be taken to use the numpy vectorised operations for fitting the model

- $\approx$ 4500ms to evaluate a single likelihood using simple loops
- Iterating over rows of a pandas DataFrame is a bad idea...
- $\approx$ 2ms to calculate using numpy vectorised operations


Estimated tennis ability over time (2012-present)
Estimated ability vs average ATP player (logit scale)
6
5
4
3
2
1
2012
w
2013
w
2014
w
2015
w
Time
Federer
Djokovic
Murray
Wawrinka

If you trust these estimates...

2015 Mens Winner

Going In-Play | Cash Out | Rules | Matched: **GBP 3,415,822** | Refresh

9 selections			Back all	Lay all		
Novak Djokovic » £0.00	**2.66** £1955	**2.68** £2081	**2.7** £225	**2.72** £1359	**2.74** £418	**2.8** £27
Andy Murray » -£33.00	**3.15** £233	**3.2** £262	**3.25** £1622	**3.3** £1700	**3.35** £1375	**3.4** £500
Roger Federer » £58.00	**5.6** £1608	**5.7** £438	**5.8** £618	**5.9** £74	**6** £714	**6.2** £542
Stanislas Wawrinka » £0.00	**9** £110	**9.2** £157	**9.4** £362	**9.8** £443	**10** £1106	**10.5** £253
Marin Cilic » £0.00	**42** £14	**50** £45	**65** £83	**75** £71	**80** £118	**90** £78
Richard Gasquet » £0.00	**60** £30	**65** £629	**70** £63	**75** £153	**80** £183	**85** £110
Kevin Anderson » £0.00	**75** £55	**80** £55	**85** £242	**95** £99	**100** £70	**110** £10
Gilles Simon » £0.00	**100** £96	**110** £102	**120** £21	**150** £10	**200** £115	**300** £50
Vasek Pospisil » £0.00	**250** £34	**260** £171	**270** £57	**310** £12	**330** £12	**380** £18

- There appears to be value going long Federer, short Murray.
- Screenshot taken from Betfair yesterday evening (6th July).

Wimbledon this year...

- Don't bet on this!
- Betting markets are extremely efficient
  - This discrepancy is due to factors not included in my model
- However model suggests Murray and Federer have $\approx$ equal probabilities of winning Wimbledon
- Djokovic is the clear favourite to win

Summary

- Many interesting applications of Data Science to the world of sport
- We provide consultancy on sports modelling: see `www.sporting-advantage.co.uk`

Bibliography

M. Dixon and S. Coles.
Modelling Association Football Scores and Inefficiencies in the Football Betting Market.
Applied Statistics 46, No. 2, pp. 265–280, 1997.

I. McHale and A. Morton.
A Bradley-Terry type model for forecasting tennis match results.
International Journal of Forecasting, 27:619–630, 2011.