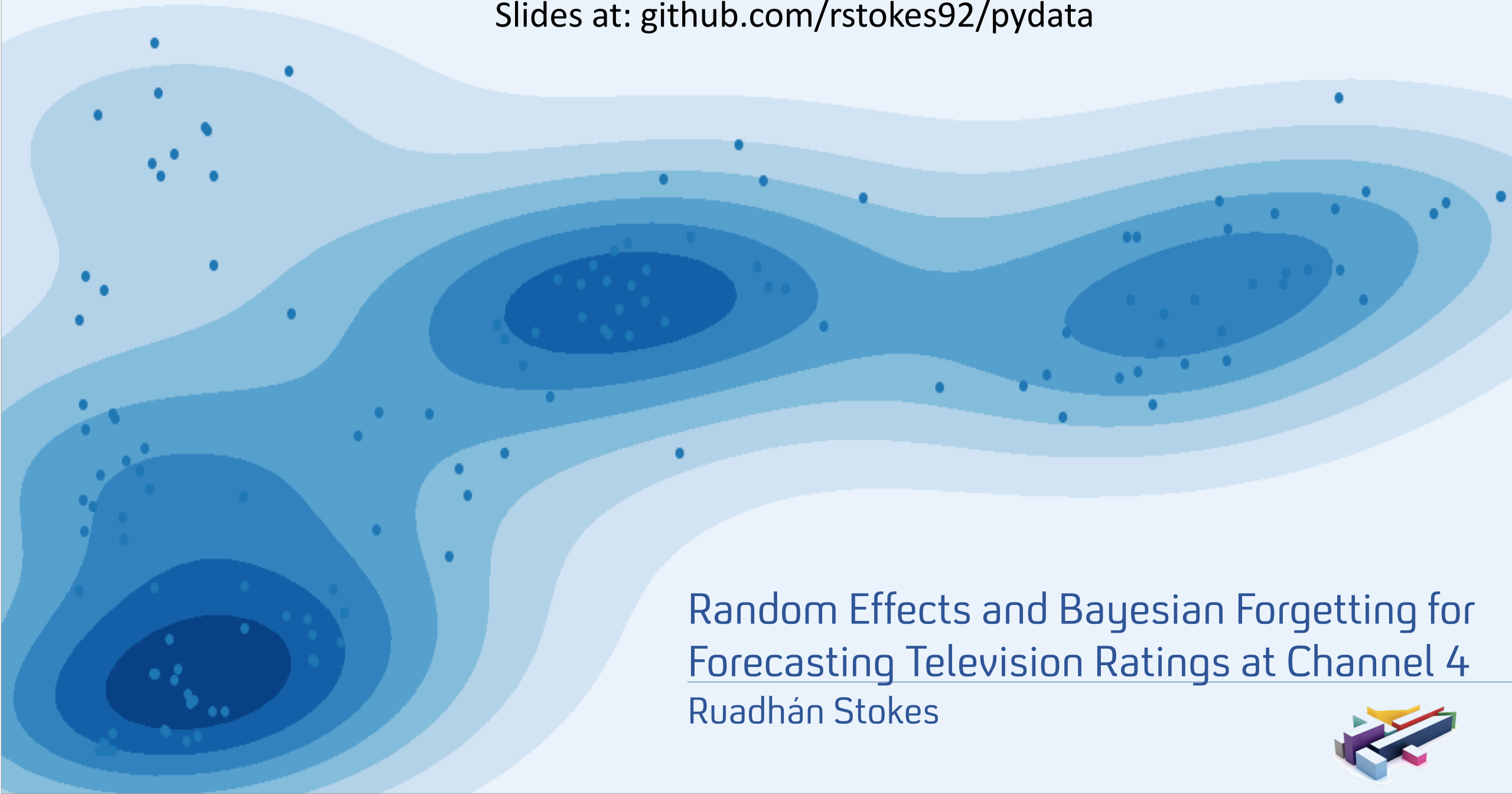
Slides at: github.com/rstokes92/pydata



Random Effects and Bayesian Forgetting for Forecasting Television Ratings at Channel 4

Ruadhán Stokes

TV Ratings & The Ad Market

- The TV Rating (TVR) is the currency of the TV ad market. 1 TVR represents 1% of viewing from a whole demographic.

- British Audience Research Board (BARB) is jointly funded by advertisers and broadcasters to be an independent body trusted with measuring TV viewership.

- BARB measurements are based on data collected from 5100 homes, chosen to be representative of the UK populace.

TV Ratings & The Ad Market

- BARB measure exactly *who* is watching and *what* they are watching.

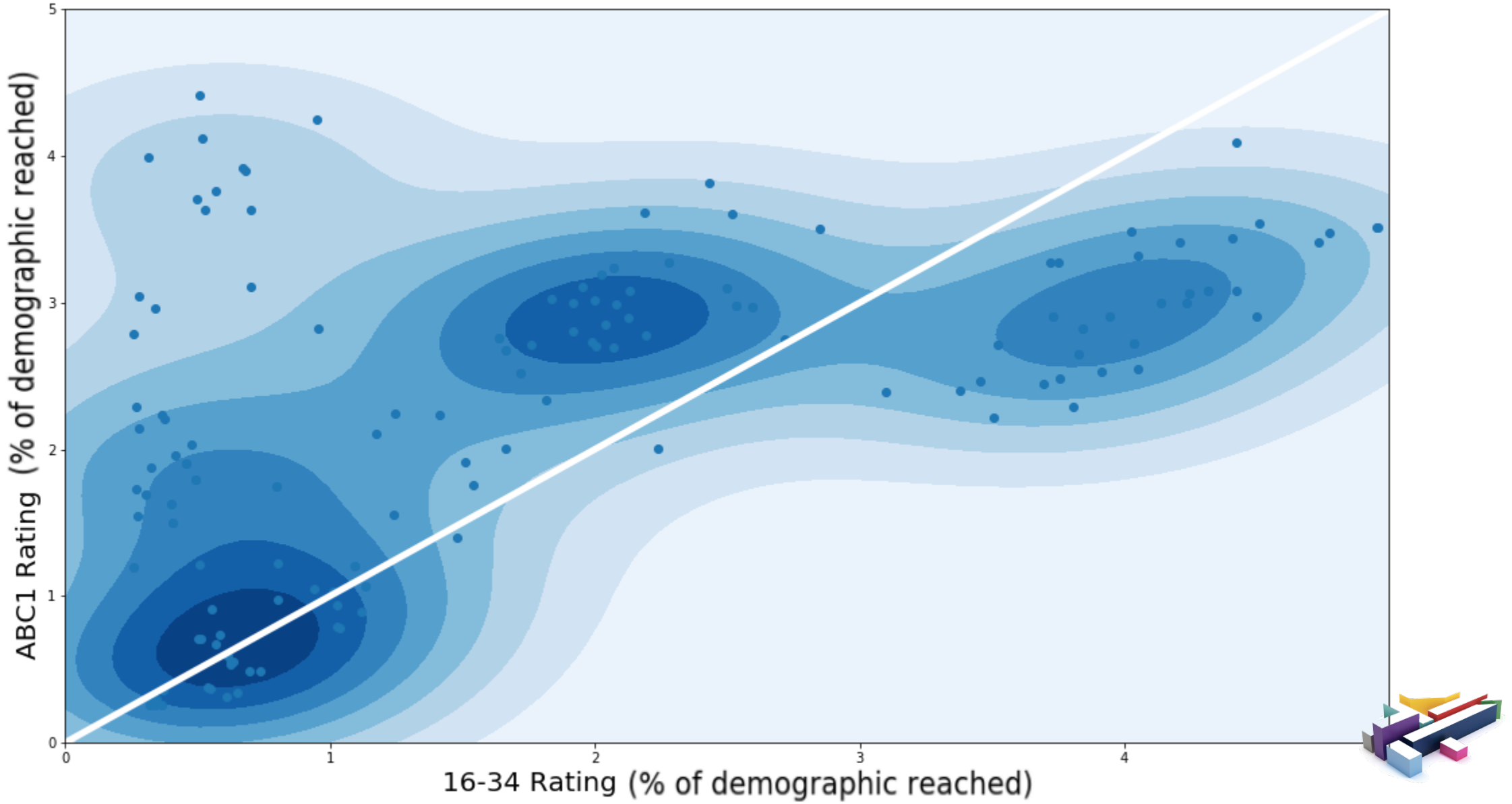
- Advertisers are concerned with the *who!* Typically only wanting to reach a specific subset of people.

- The *who* measured by BARB is the key to demographic targeting in television. It is also the main thing we wish to forecast.

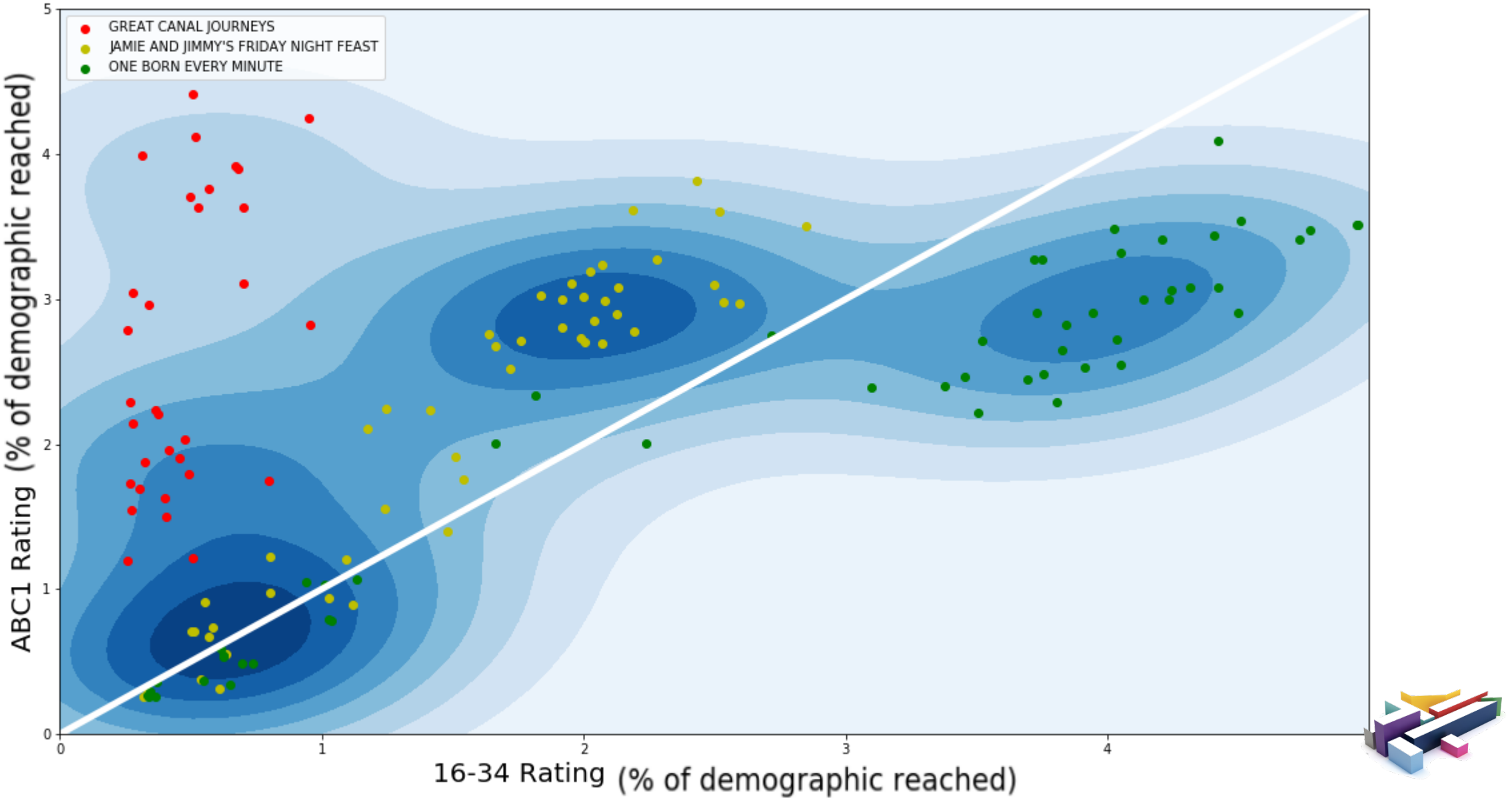
TV Ratings & The Ad Market

- Advertisers can choose from 16 audiences to buy.
- ***They only pay for viewers from the chosen audience!***
- Audiences are segmented along social class, age and gender.

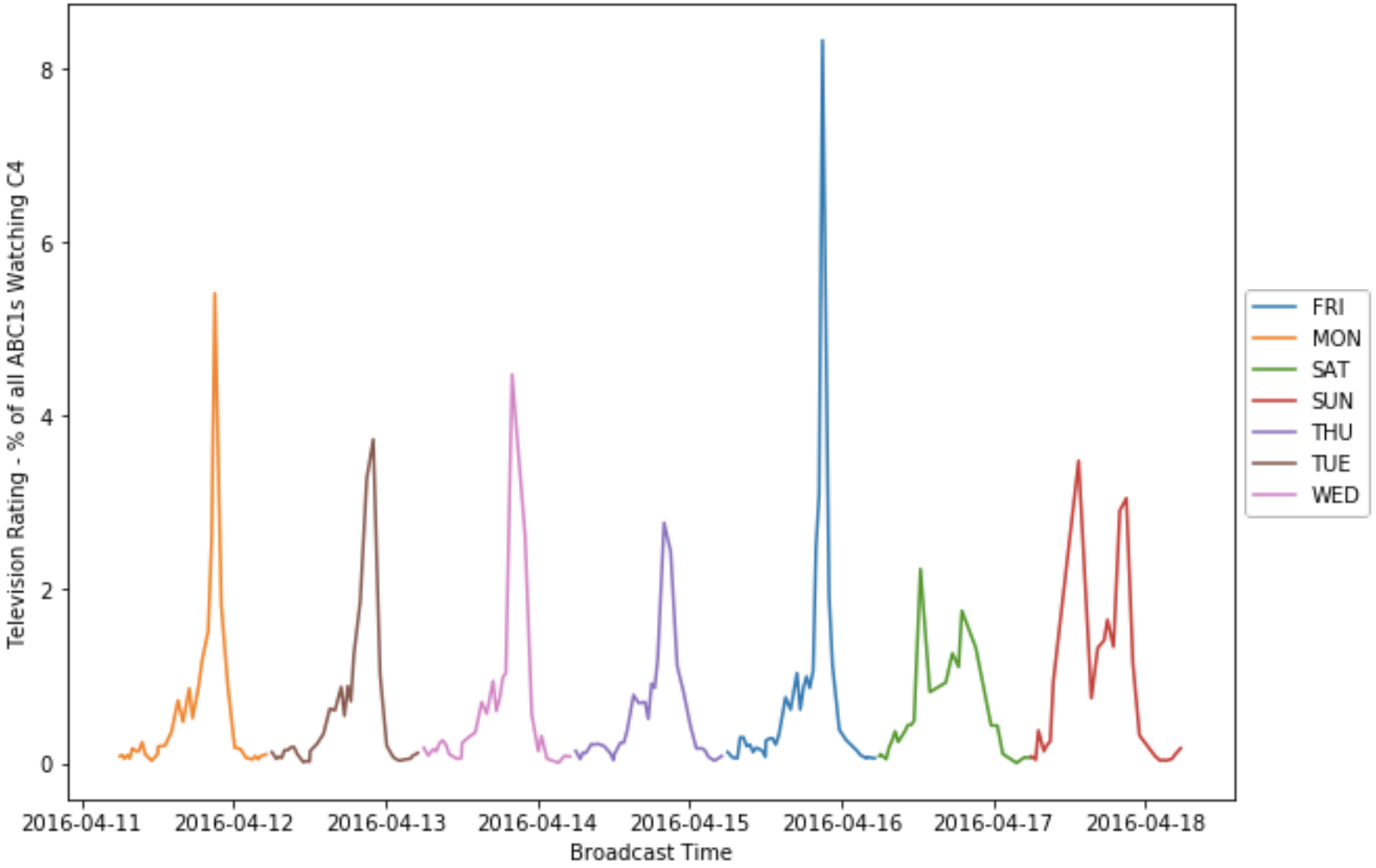
TV Ratings: “Youngs” Vs. “ABC1s”



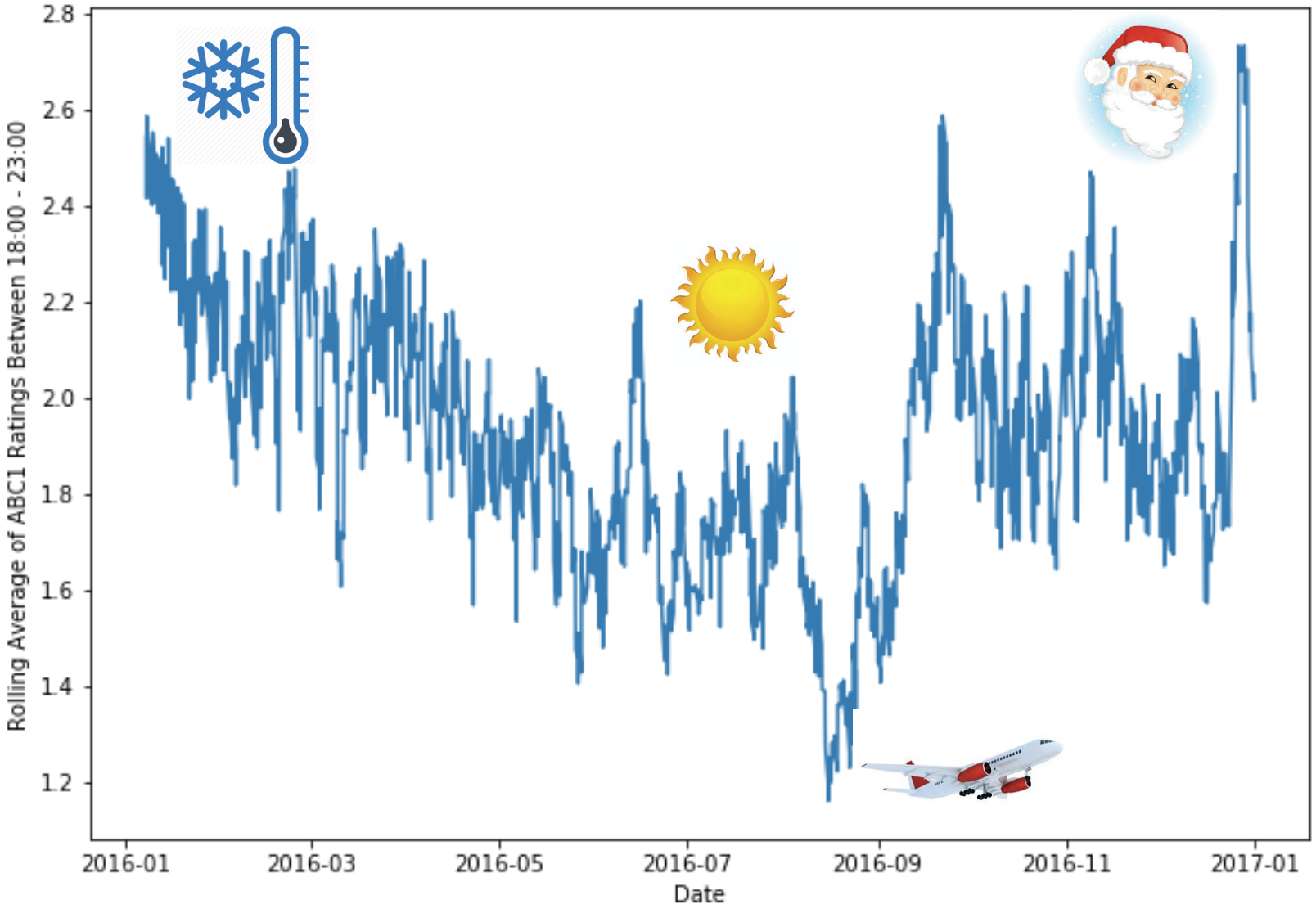
TV Ratings: “Youngs” Vs. “ABC1s”



TV Ratings: Weekly Patterns in Time



TV Ratings: Yearly Patterns in Time



Modelling

- Clearly time covariates and past programme ratings are strongly predictive of the TVR.
- We also have other predictors such as Genre and Repeat Status.
- We must also be able to make good ‘cold start’ predictions for broadcasts of a new programme.

Random Effects model

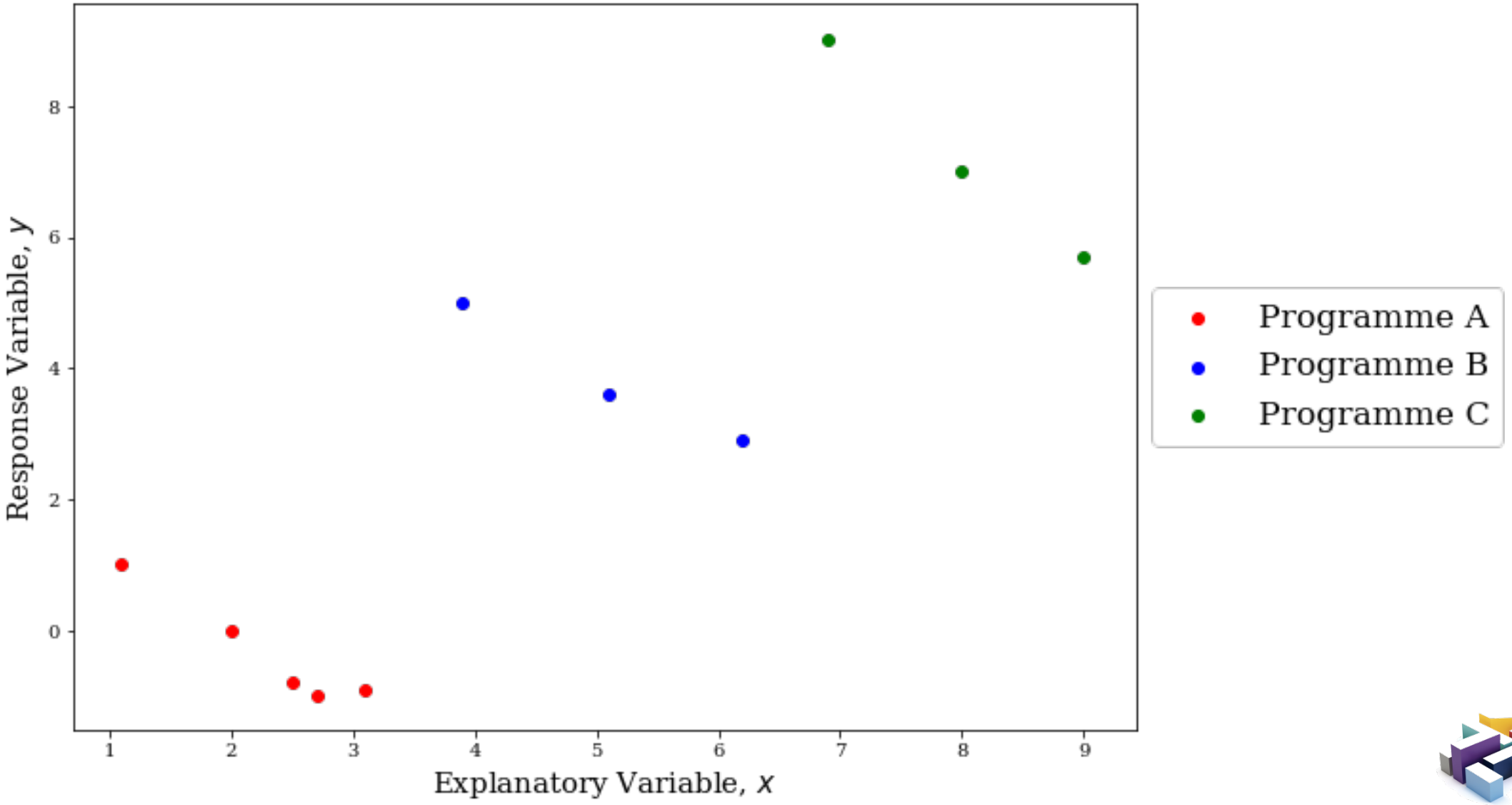
- Rating is a linear combination of ‘fixed effects’ predictors X, (genre, day of year, day of week, time of day, duration, repeat status) with a programme specific ‘random effect’ δ .

$$y_i = \boldsymbol{\beta}_{RE} X_i + \delta_p + e_i$$

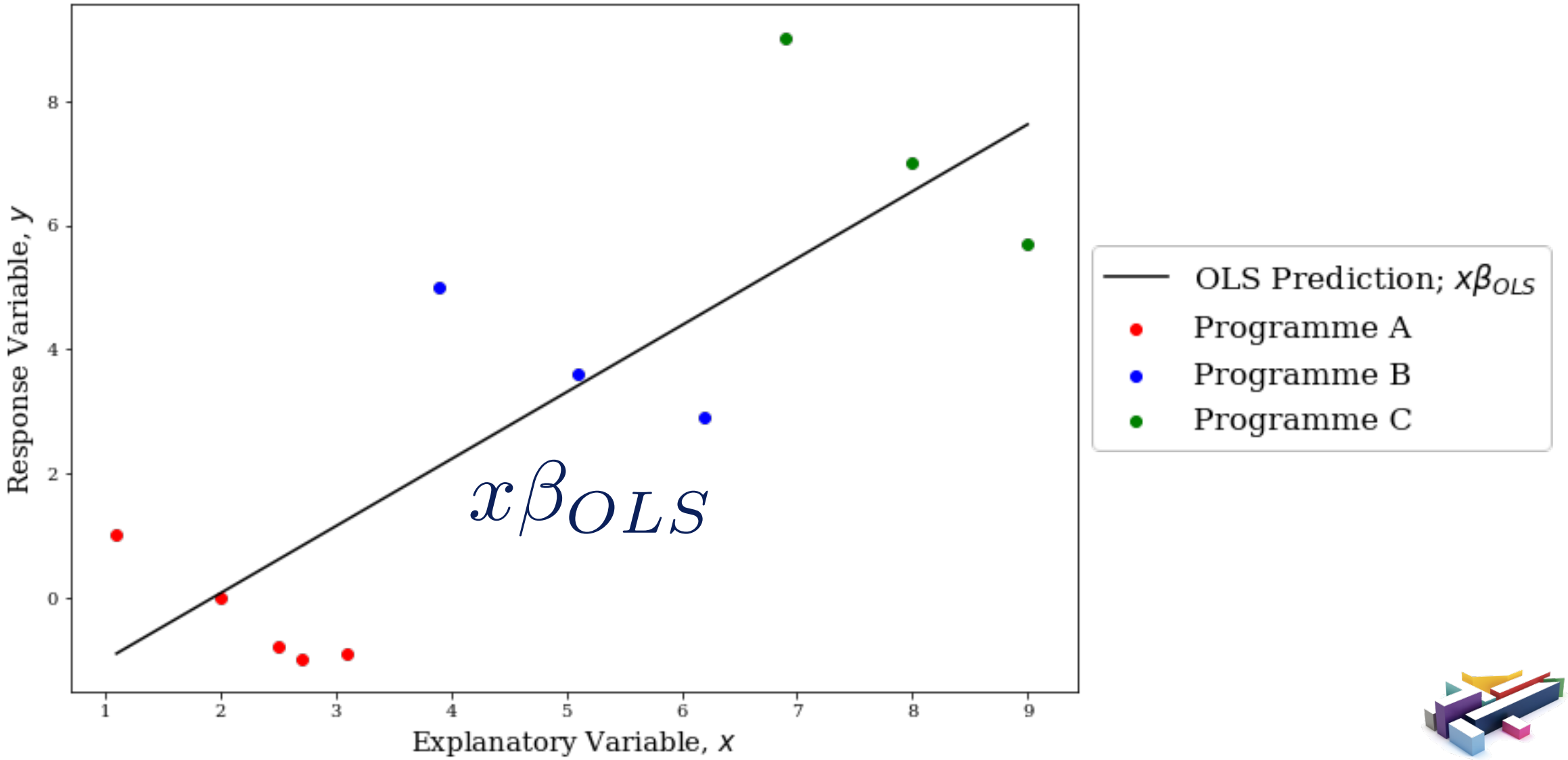
$$\delta_p = \frac{1}{|P|} \sum_{i \in P} y_i - \boldsymbol{\beta}_{RE} X_i$$

- Each i is a broadcast and P is the set of broadcasts of programme p

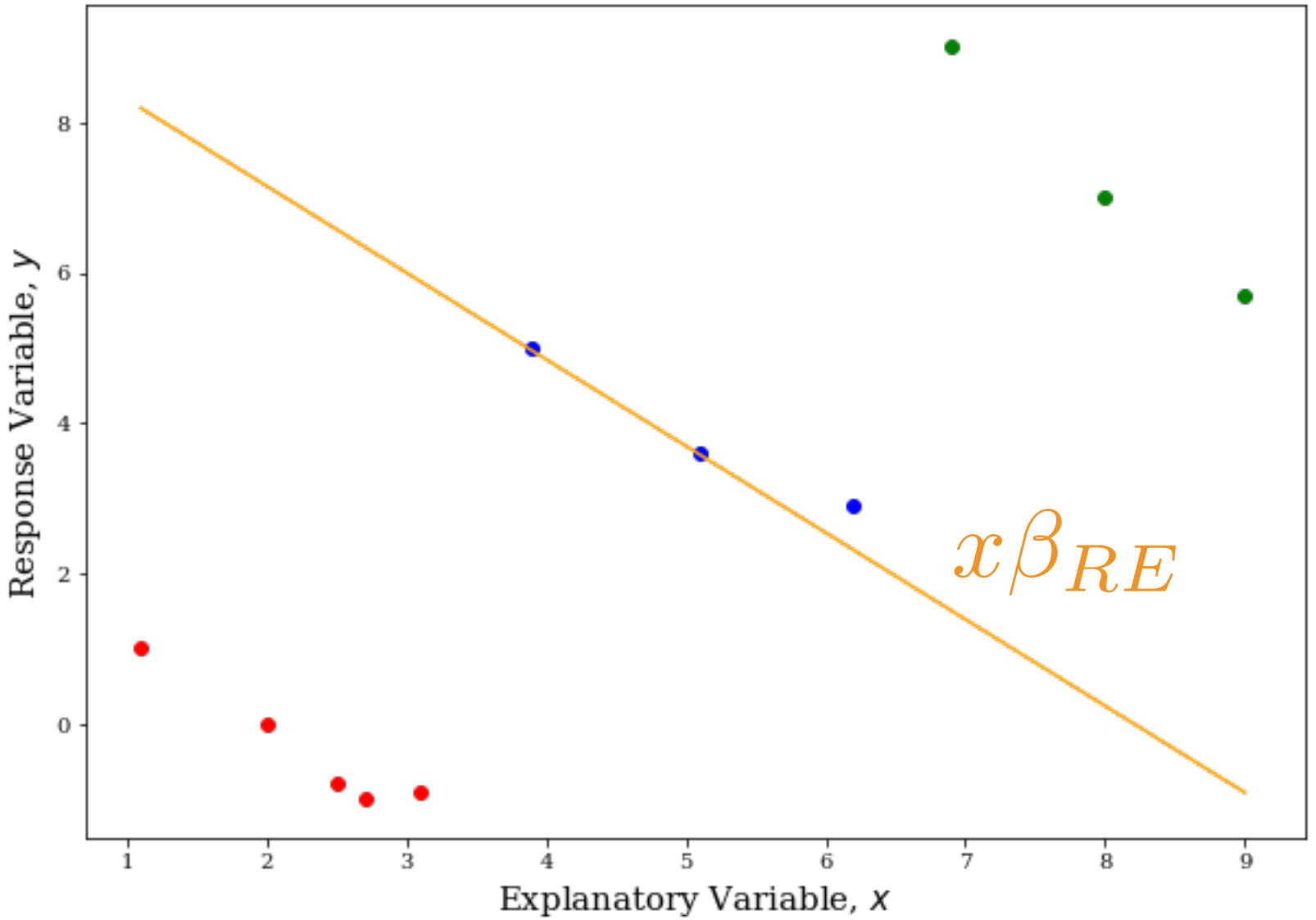
OLS Vs. Random Effects, Toy Data

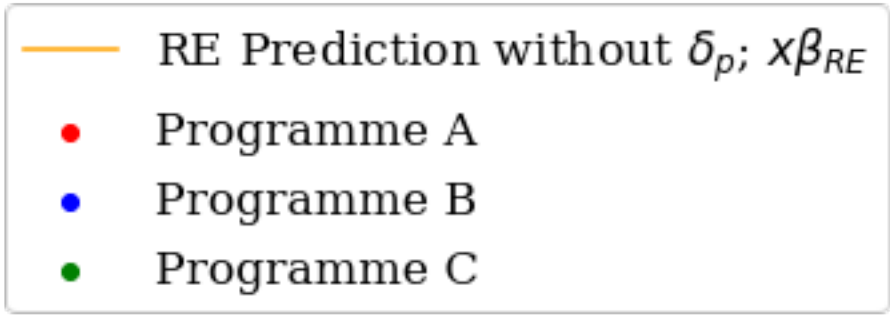


Ordinary Least Squares Fit, β_{OLS}

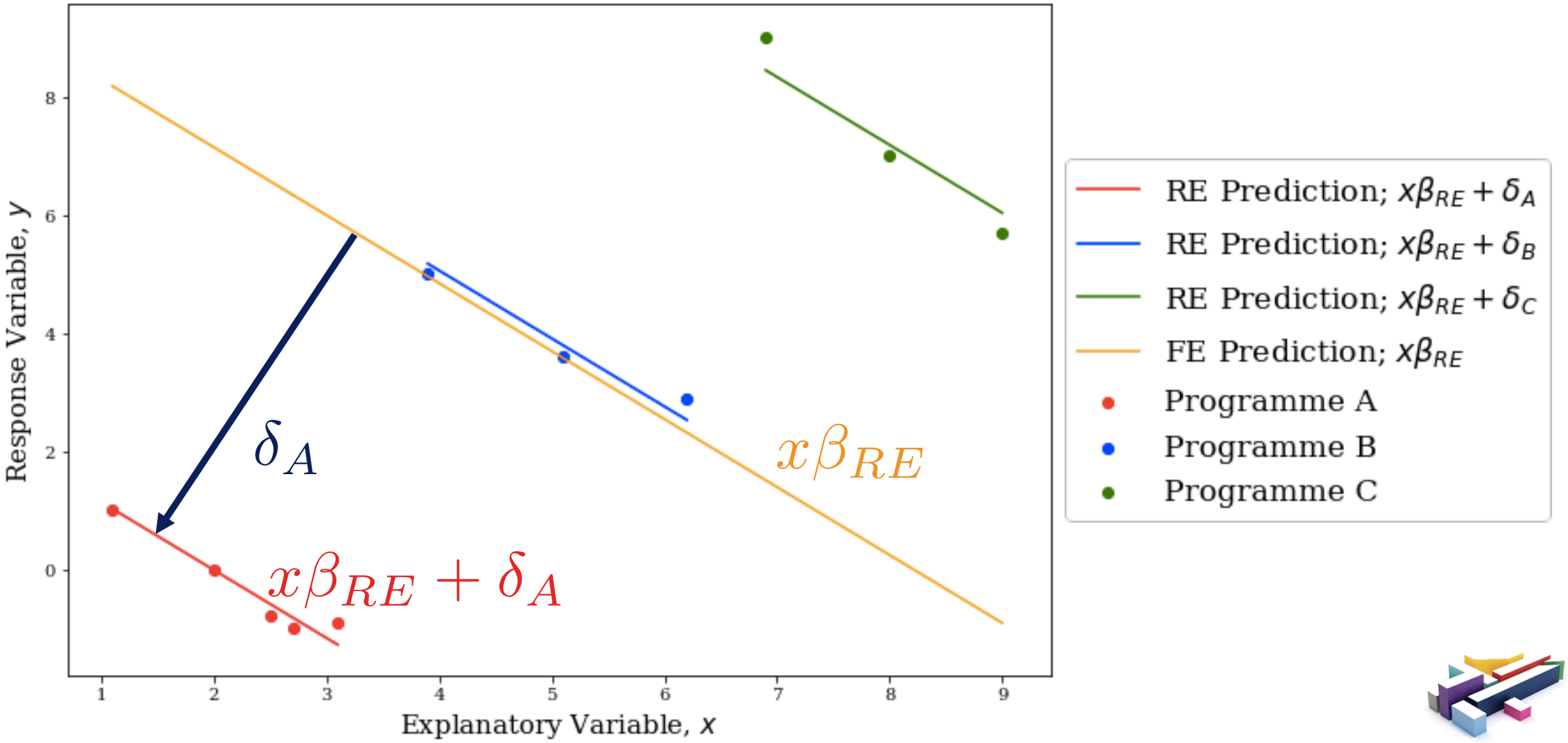


Random Effects Model, Fixed Effects Only; β_{RE}



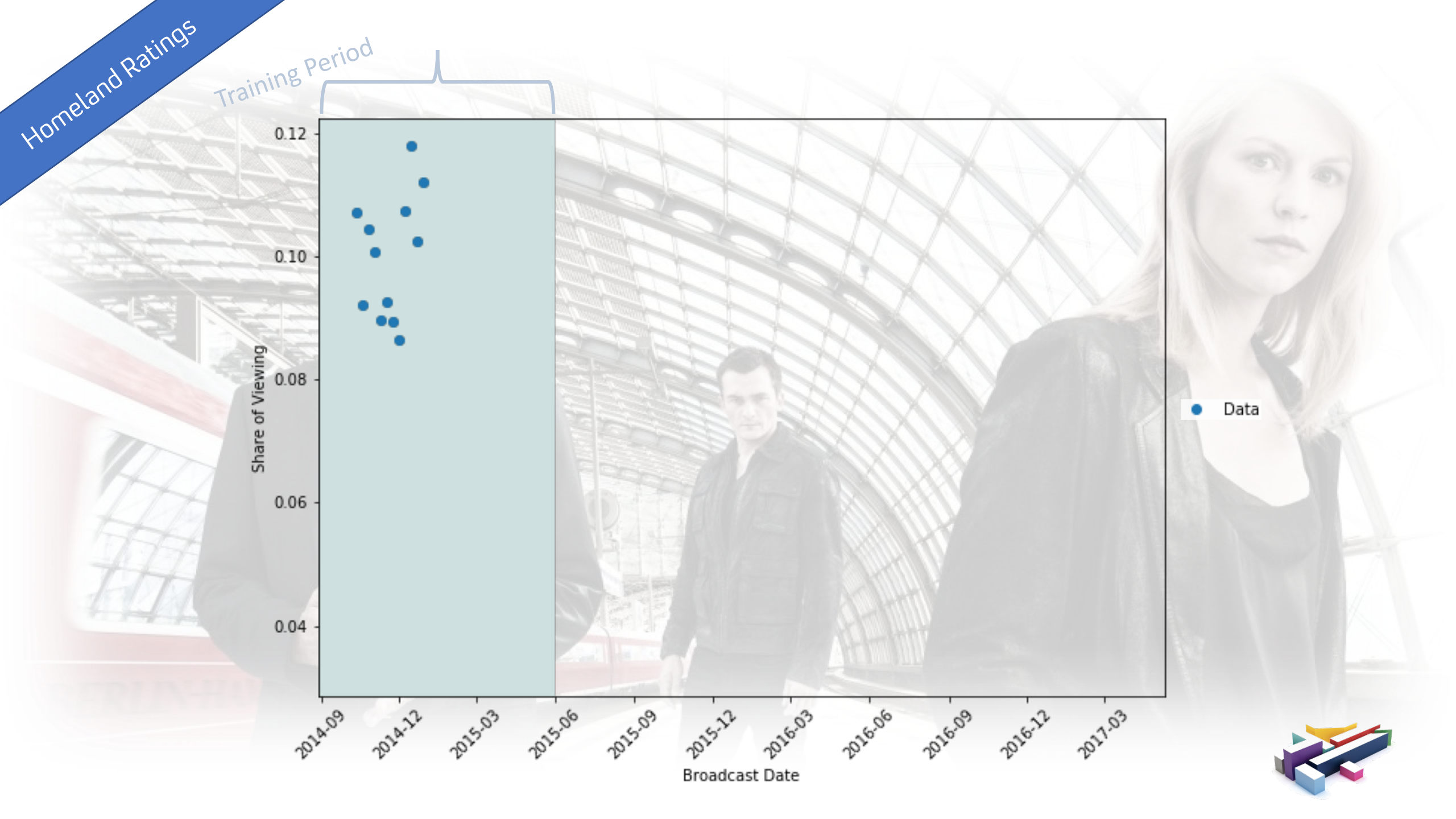


Fixed and Random Effects; $\beta_{RE} + \delta_p$

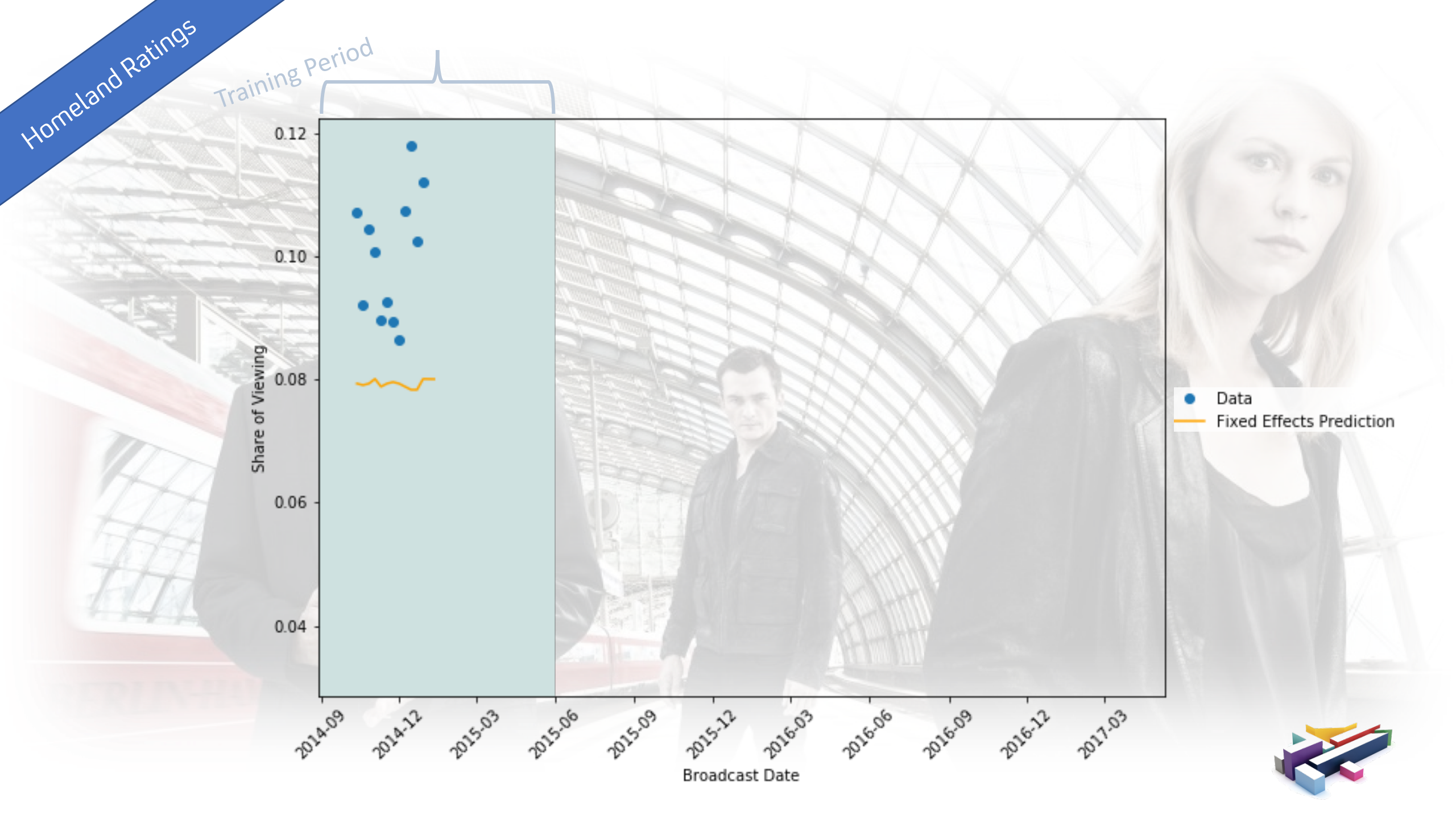



BERLIN-HA

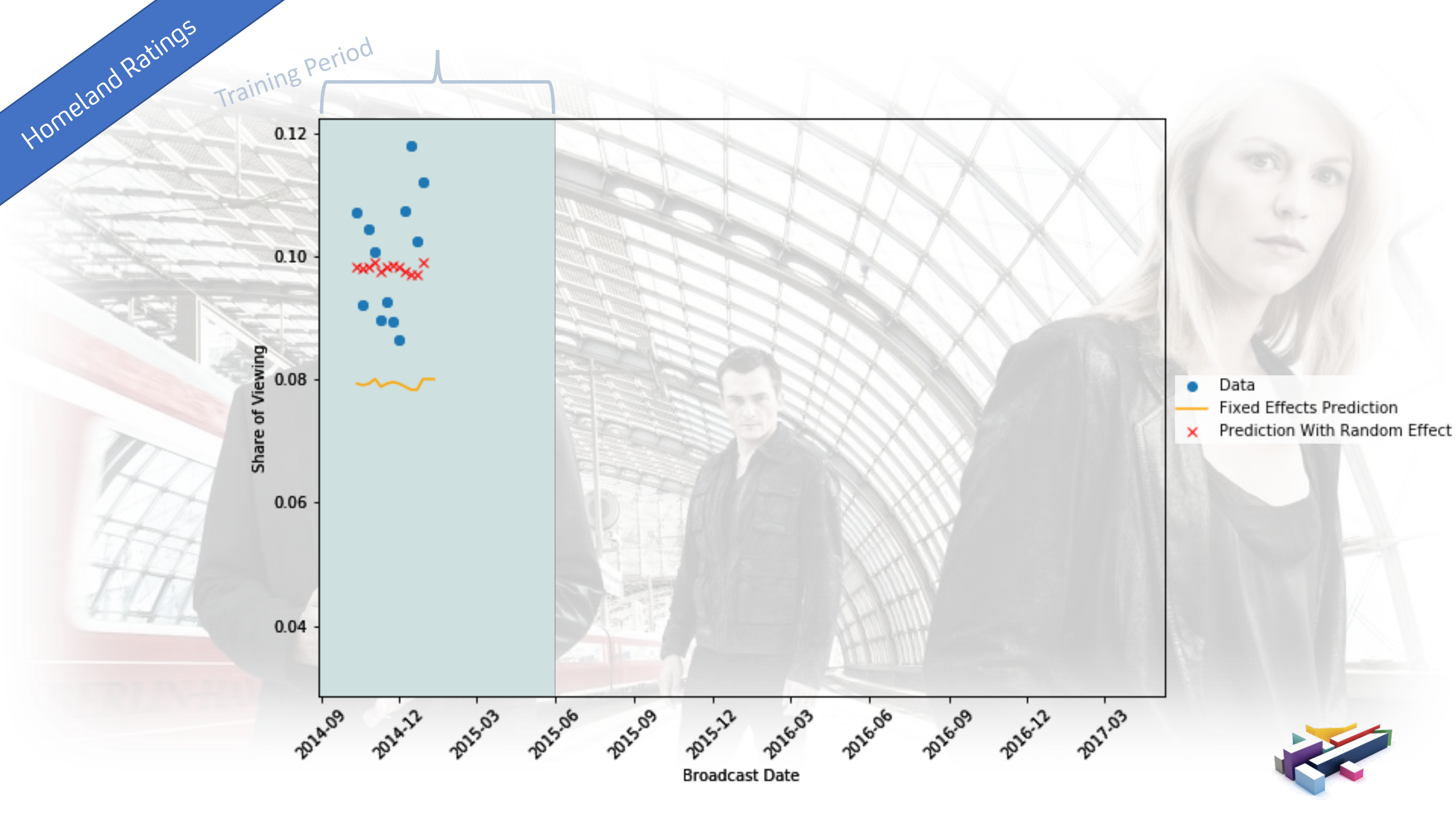
Homeland Ratings



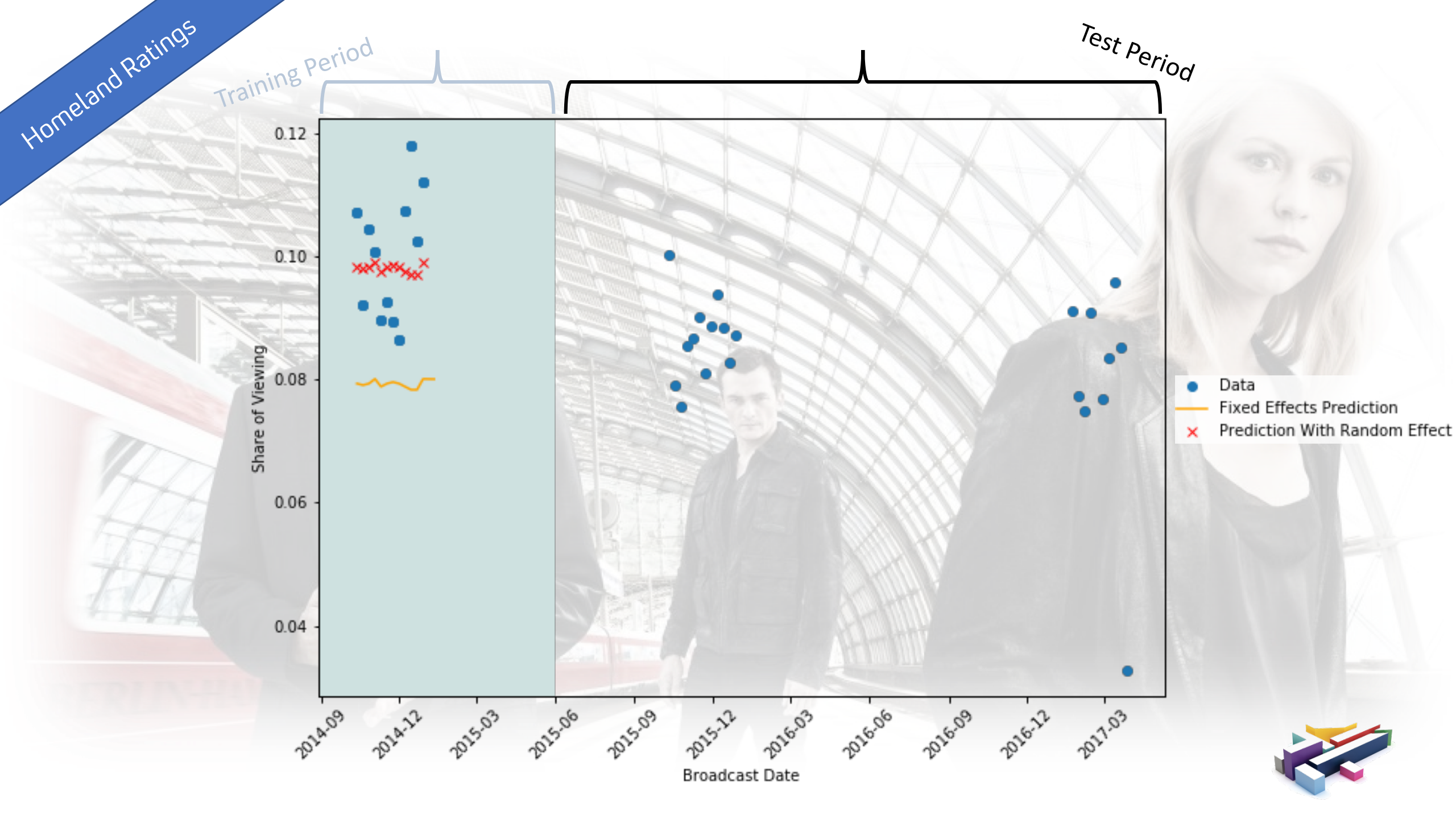
Homeland Ratings



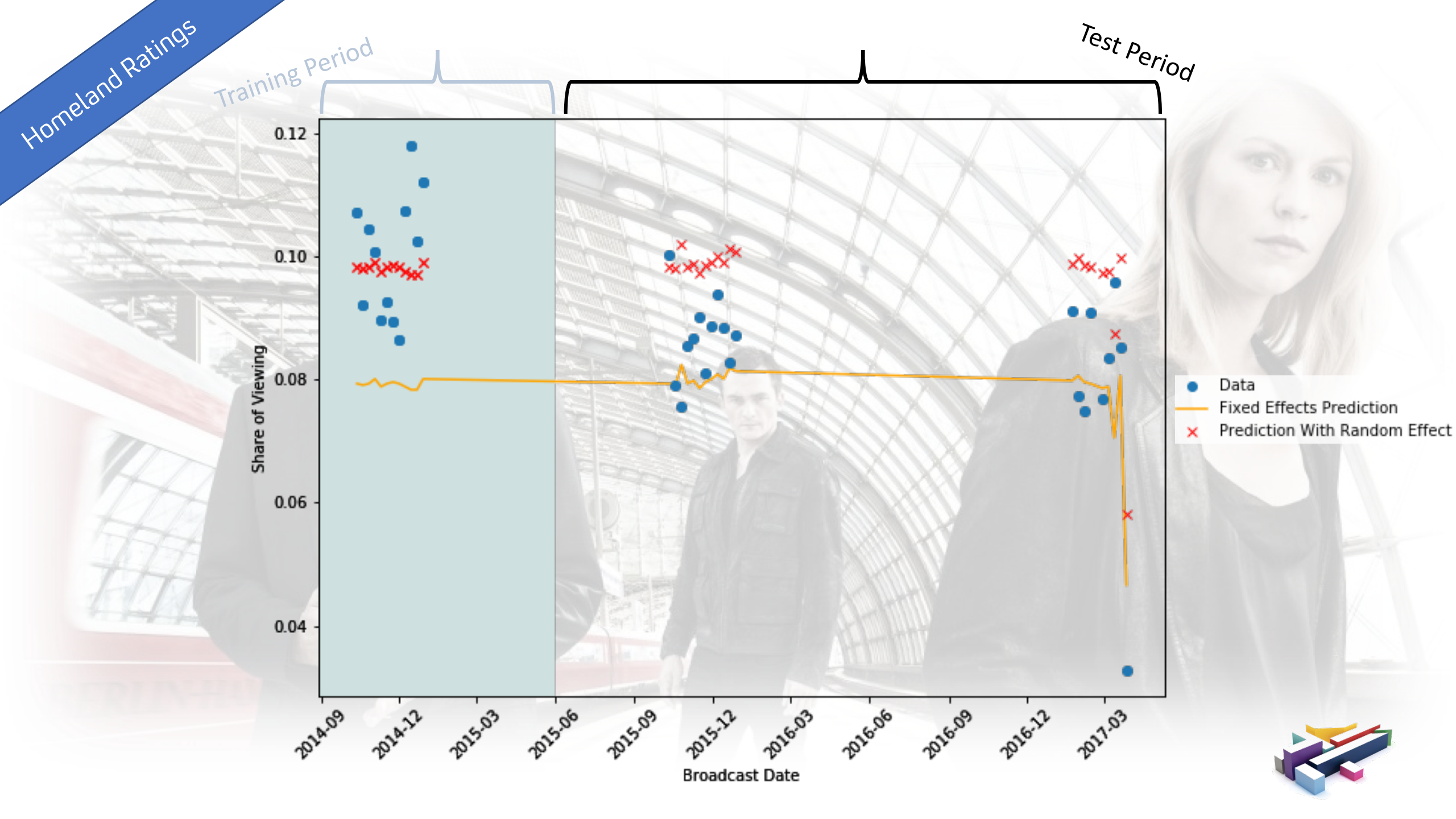
Homeland Ratings



Homeland Ratings



Homeland Ratings



Estimating a 'drifting' mean

- Simple method: Down-weight or **forget** old data, for data $z_{1:n}$

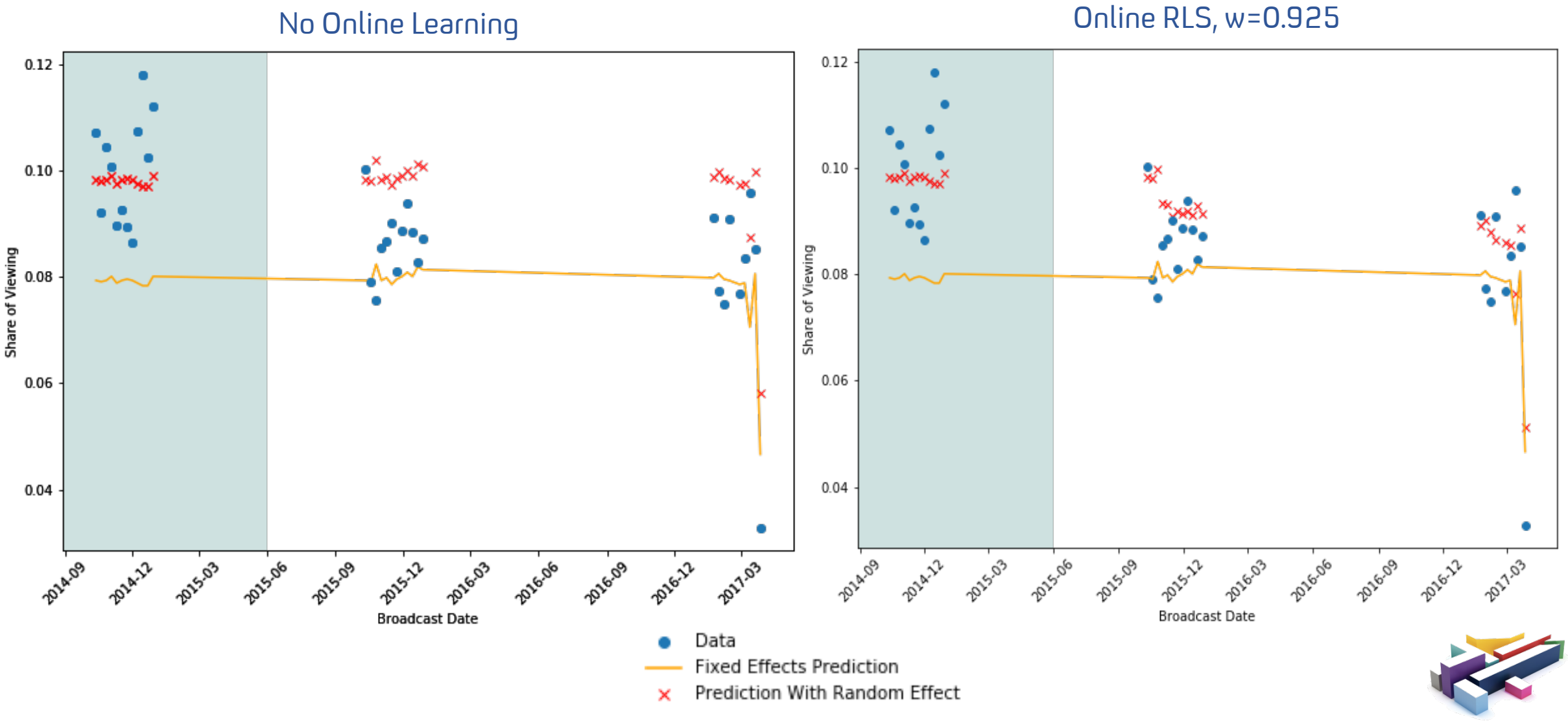
$$\mu_n = \frac{1}{\sum_i^n w_i} \sum_i^n w_i z_i, \quad w_n = 1, \quad w_i < w_{i+1}$$

- Same results can be achieved by setting a fixed **forgetting factor** $w < 1$ and recursively counting and accumulating the data.

$$s_n = w s_{n-1} + z_n, \quad c_n = w c_{n-1} + 1, \quad \mu_n = \frac{s_n}{c_n}$$

- These updates define the **recursive least squares** algorithm for estimating a time varying average.

Recursive Least Squares for Homeland



Bayesian Inference of Normal

- For n observations $z_{1:n} \sim \mathcal{N}(\mu, \sigma^2)$ the Bayesian posterior can be fully defined in closed form (Yay conjugacy!) as a function of the statistics:

$$V_n = \begin{bmatrix} \sum_i z_i^2 & \sum_i z_i \\ \sum_i z_i & n \end{bmatrix}, \quad \nu_n = n, \quad \{V_0, \nu_0\} = \text{Shaping Parameters of Prior}$$

- The posterior is Normal inverse Gamma

$$f(\mu, \sigma^2 | V_n, \nu_n) \propto \left(\frac{1}{\sqrt{\sigma^2}}\right)^{\nu_n} \exp\left[-\frac{1}{2\sigma^2} \operatorname{tr}\left(\begin{bmatrix} -1 & \mu \end{bmatrix} V_n \begin{bmatrix} -1 \\ \mu \end{bmatrix} \right) \right]$$

Bayesian Inference of Normal

- Importantly $V,\ \nu$ can be updated recursively with a forgetting factor in the same way as RLS.

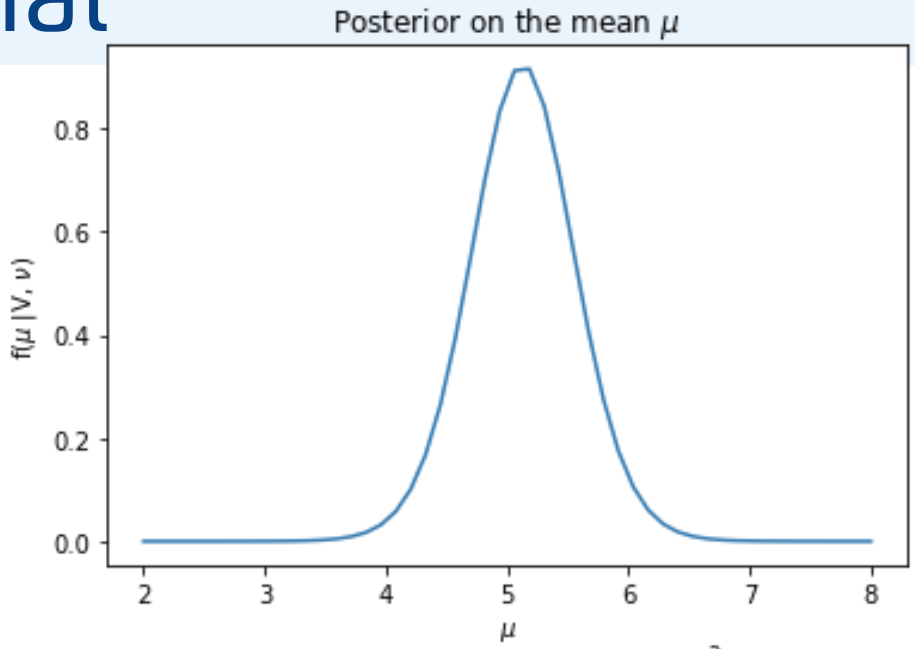
$$V_n = wV_{n-1} + (1-w)V_0 + \begin{bmatrix} z_n^2 & z_n \\ z_n & 1 \end{bmatrix}, \quad w\nu_{n-1} + (1-w)\nu_0 + 1$$

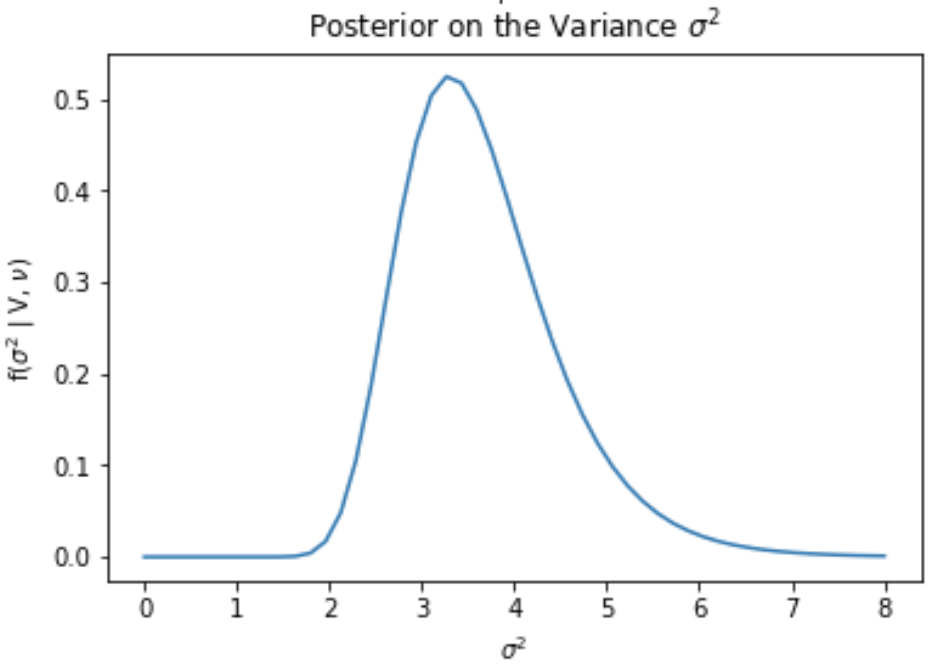
- For a non informative prior (i.e. small$\{V_0, \nu_0\}$) the posterior is more affected by the latest observations, just as the maximum likelihood estimate is in recursive least squares.

Bayesian Inference of Normal

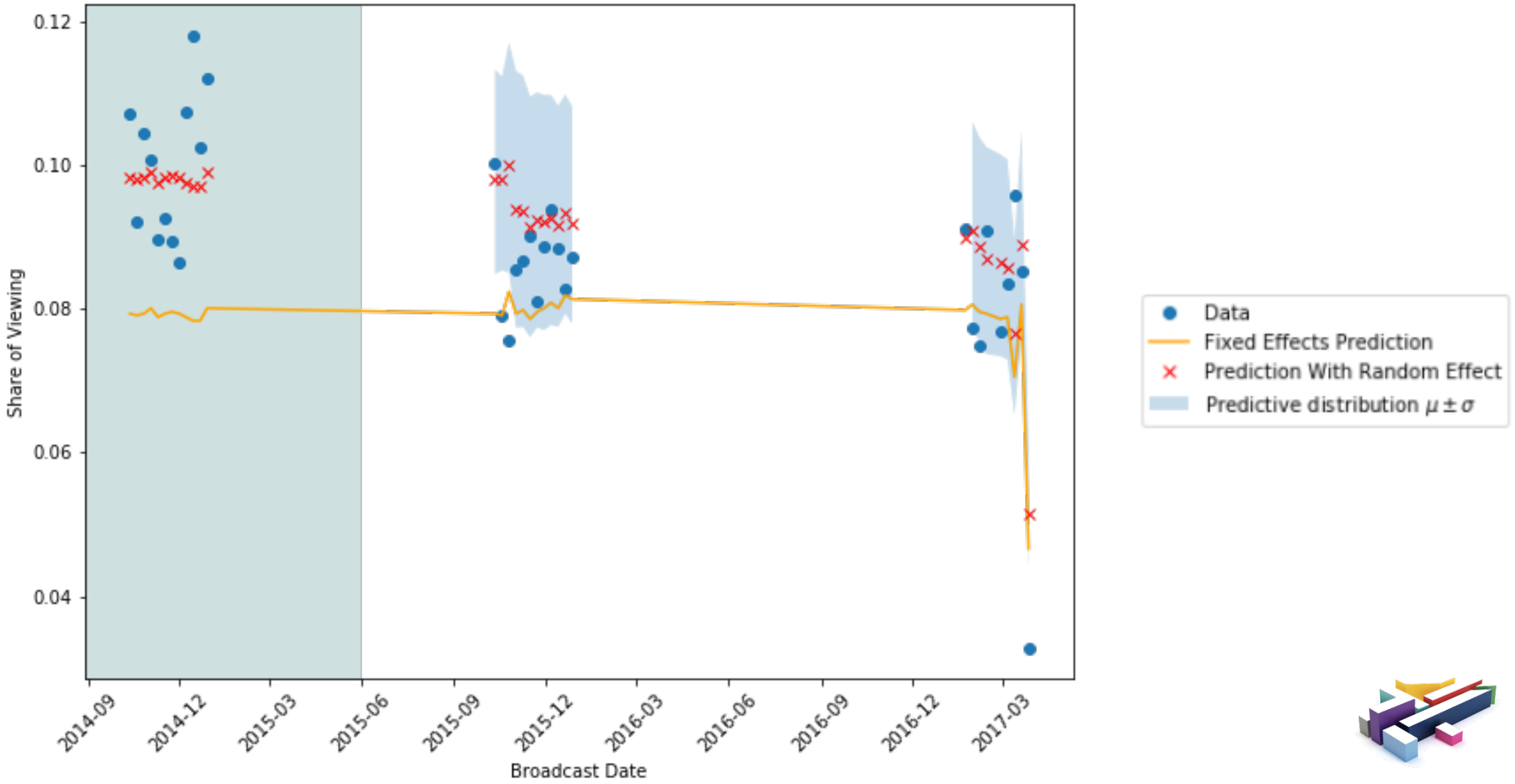
```python
import numpy as np
# Generate Normal data
data = np.random.normal(loc=5, scale = 2, size=2000)

## Uninformative Prior ##
V = np.zeros((2,2))
nu = 0

forgetting = 0.95 # Set Forgetting Factor

for latest_x in data:
    vec = np.array([latest_x, 1])

    # Recursive update with forgetting
    V = forgetting*V + np.outer(vec, vec)
    nu = forgetting*nu + 1

mean_map = V[1,0]/V[1,1]
variance_map = (V[0,0] - (V[1,0]**2)/V[1,1])/(nu)
print "Mean Estimate: {}".format(mean_map)
print "Variance Estimate: {}".format(variance_map)
```

```
Mean Estimate: 5.1277776728
Variance Estimate: 3.63982180795
```

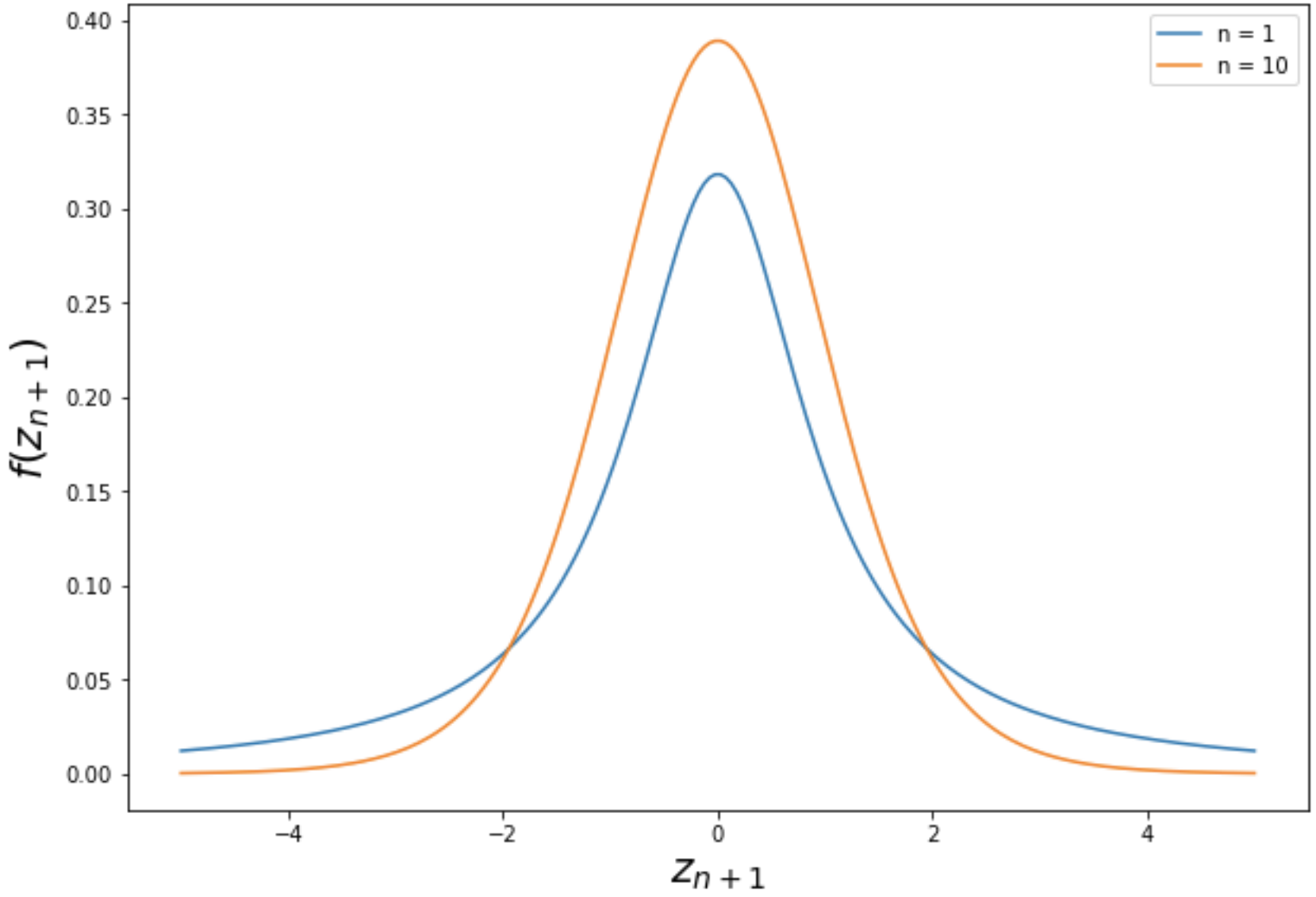




Predictive Distribution With Bayesian Forgetting

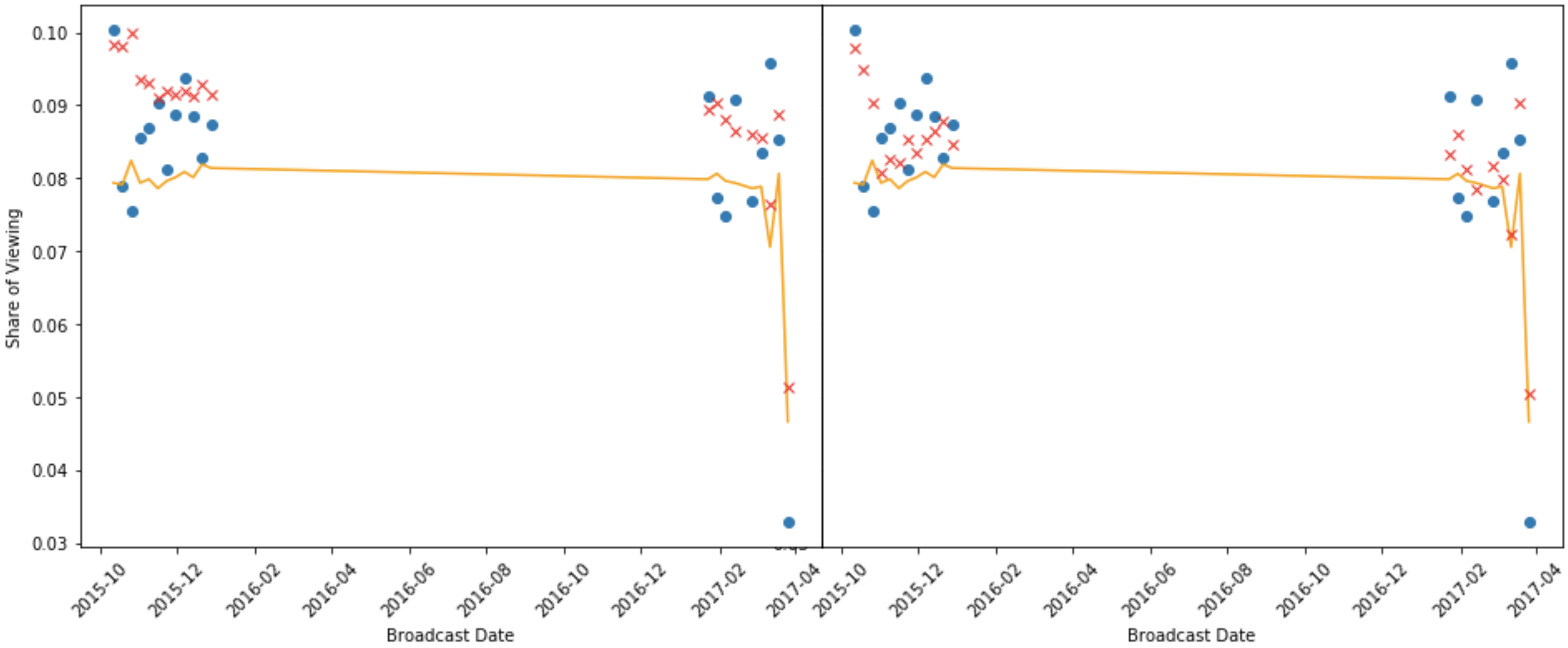


Predictive Distribution With Forgetting



Choosing the forgetting factor, w

$w = 0.925$ $w = 0.5$



Choosing the forgetting factor, w

- Forgetting factor is a nuisance parameter.
- Optimal forgetting factor depends on how quickly the ratings are changing.
- This means it is different for each programme.
- A fixed forgetting factor cannot adapt to sudden changes in the mean so the forgetting factor itself should ideally be time-varying.
- ***We want the forgetting factor at each time step to be estimated from the data!***

Choosing the forgetting factor, w: Forgetting as approximate inference

- The formal Bayesian solution to estimating time variant parameters is:

$$f(\mu_{n+1}|z_{1:n}) = \int f(\mu_{n+1}|\mu_n) f(\mu_n|z_{1:n}) d\mu_n$$

Exact Posterior at time n+1 — Parameter Evolution Model — Posterior at time n

Which reduces to the Kalman filter under some stringent conditions.

- In our case the Parameter Evolution Model is **unknown**, but....
  The forgetting factor w defines a bound κ on the KL Divergence

$$\mathcal{D}[f(\mu_{n+1}|z_{1:n})||f(\mu_n|z_{1:n})] \leq \kappa$$

making it a measure on the Parameter Evolution Model:

$$f(\mu_{n+1}|\mu_n, w_n)$$

Choosing the forgetting factor, w: Forgetting as approximate inference

- It turns out that the posterior under forgetting

$$f(\mu, \sigma^2 | V_n, \nu_n) \propto (\frac{1}{\sqrt{\sigma^2}})^{\nu_n} \exp\left[-\frac{1}{2\sigma^2} \operatorname{tr}\left(\begin{bmatrix} -1 & \mu \end{bmatrix} V_n \begin{bmatrix} -1 \\ \mu \end{bmatrix} \right) \right]$$

Where

$$V_n = wV_{n-1} + (1-w)V_0 + \begin{bmatrix} z_n^2 & z_n \\ z_n & 1 \end{bmatrix}, \quad w\nu_{n-1} + (1-w)\nu_0 + 1$$

Is the distribution with **maximum entropy** (or least informative) of all distributions **satisfying the KLD bound**, making it an appropriate choice to approximate the 'true' distribution.

Choosing the forgetting factor, w: Forgetting as approximate inference

- Now formal joint estimation of the parameters and forgetting factor can be written:

$$f(\mu_n, w_n | z_{1:n}) \propto f(z_n|\mu_n) f(\mu_n|\mu_{n-1}, w_n) f(\mu_{n-1}|z_{1:n-1}) f(w_n)$$

Joint Posterior — Likelihood Z_n — Parameter Evolution Model — Posterior at time n-1 — Prior on w_n

- And marginalising out μ we get the posterior on the forgetting factor:

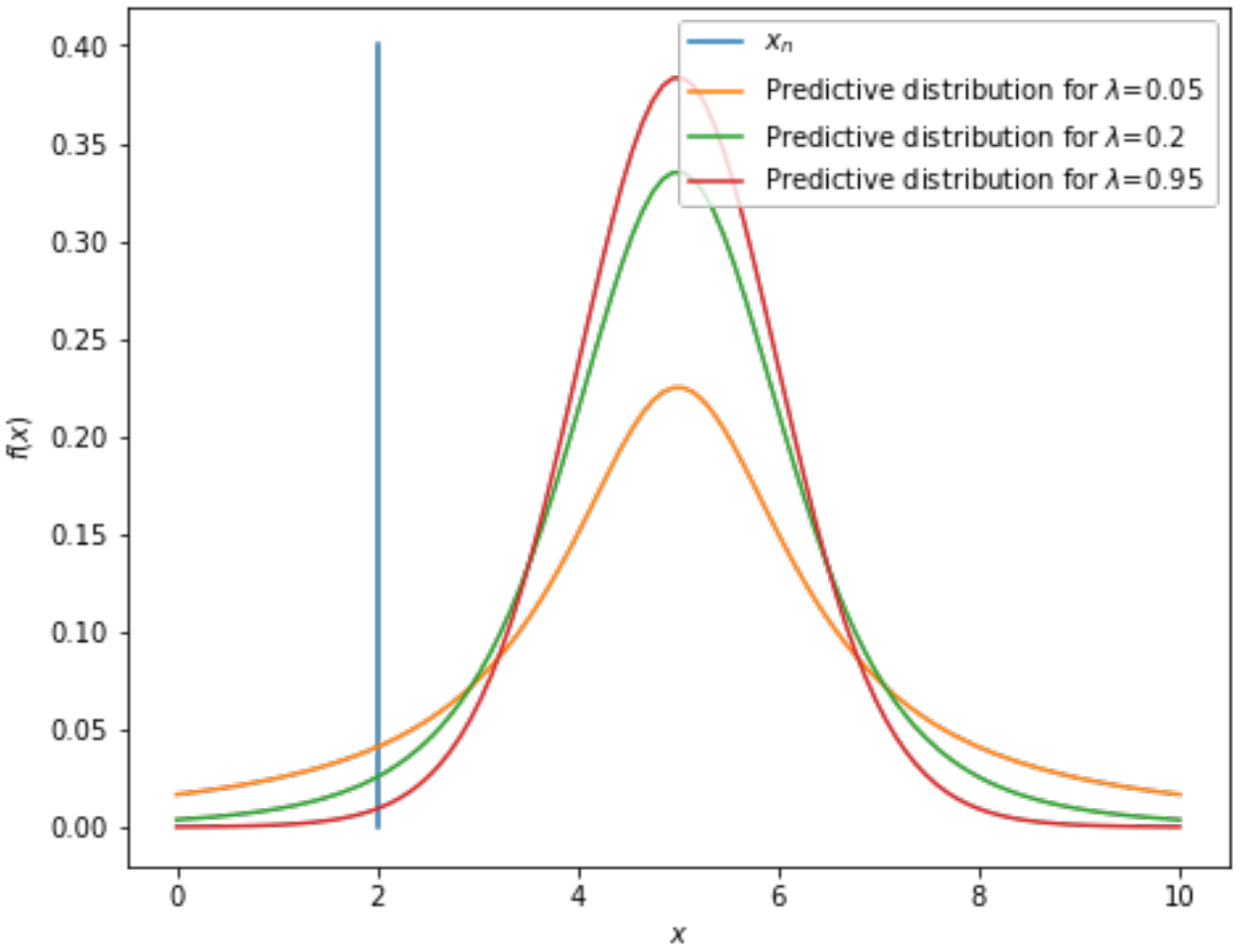
$$f(w_n|z_{1:n}) \propto f(z_n|z_{1:n-1}, w_n) f(w_n)$$

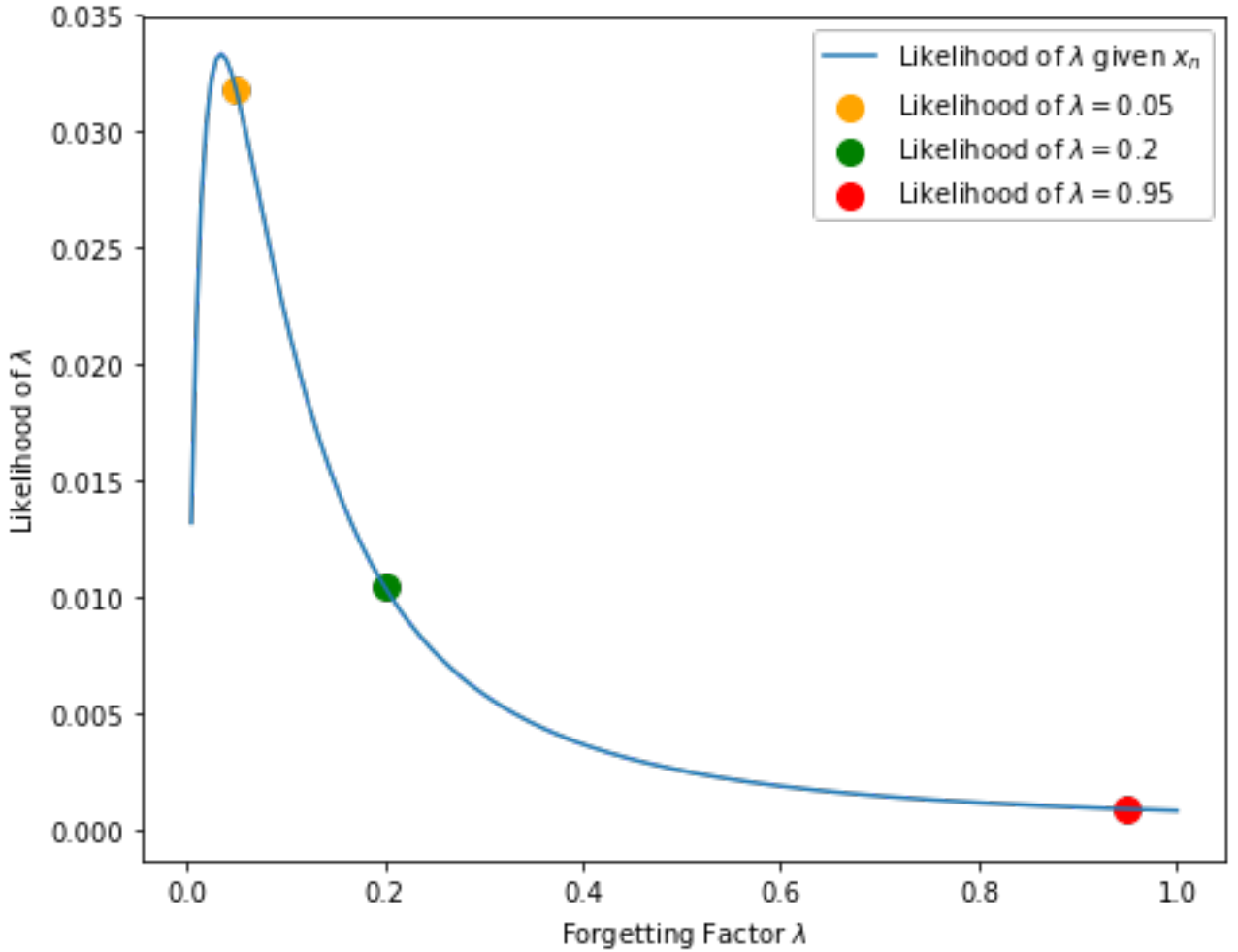
Marginal Posterior

Likelihood of w_n
Notice it is also the Predictive distribution at time n-1

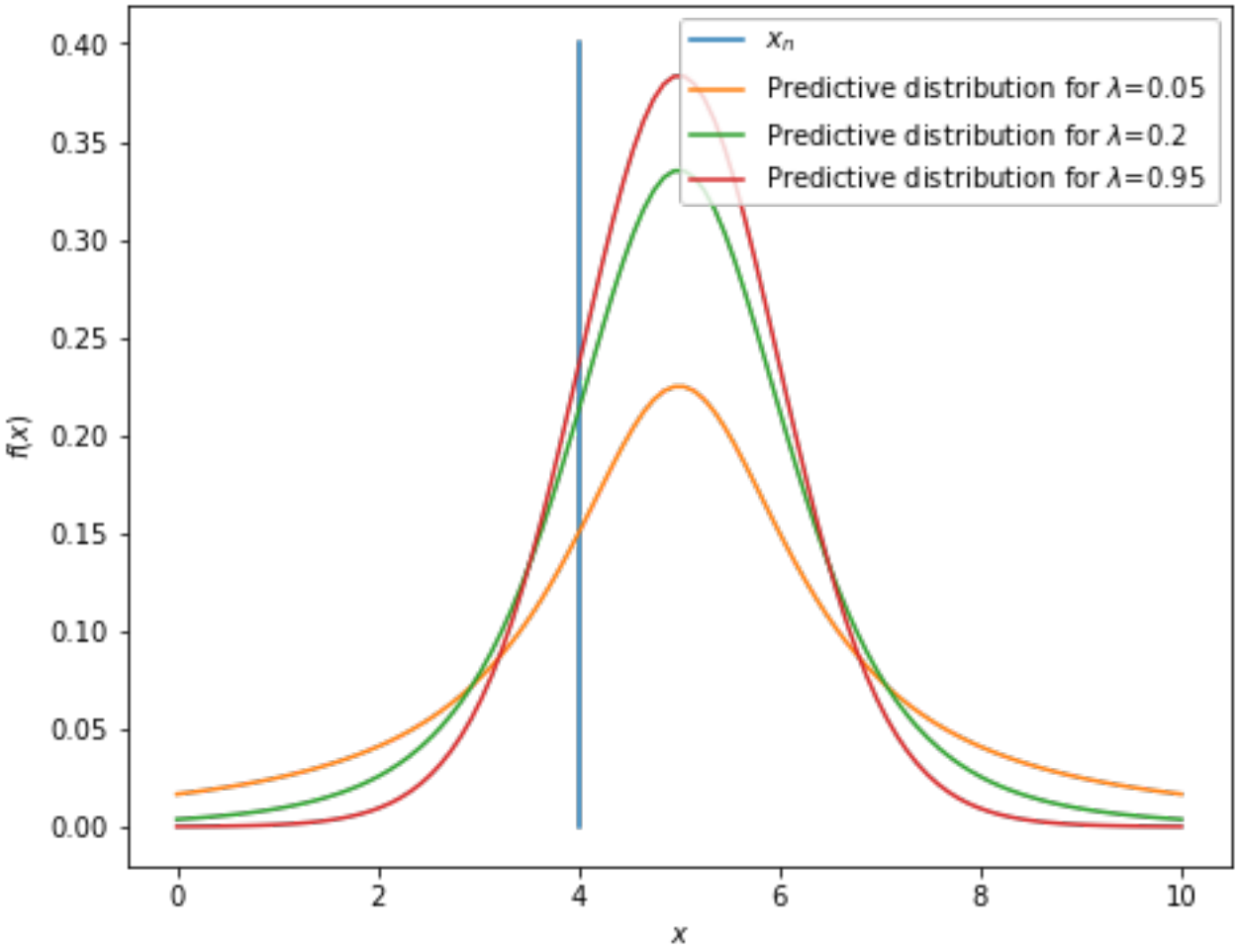
Prior on w_n

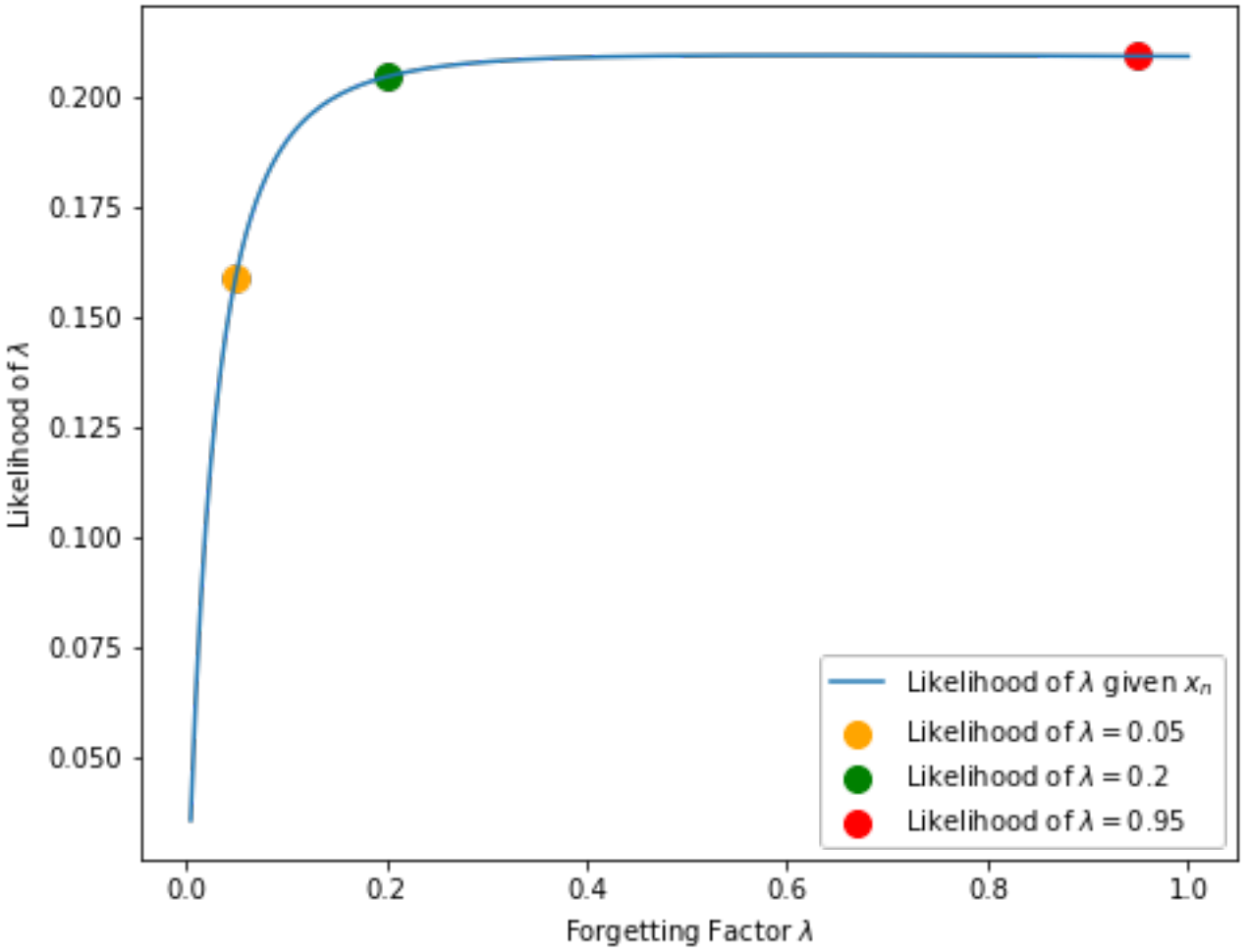
Forgetting Factor Likelihoods



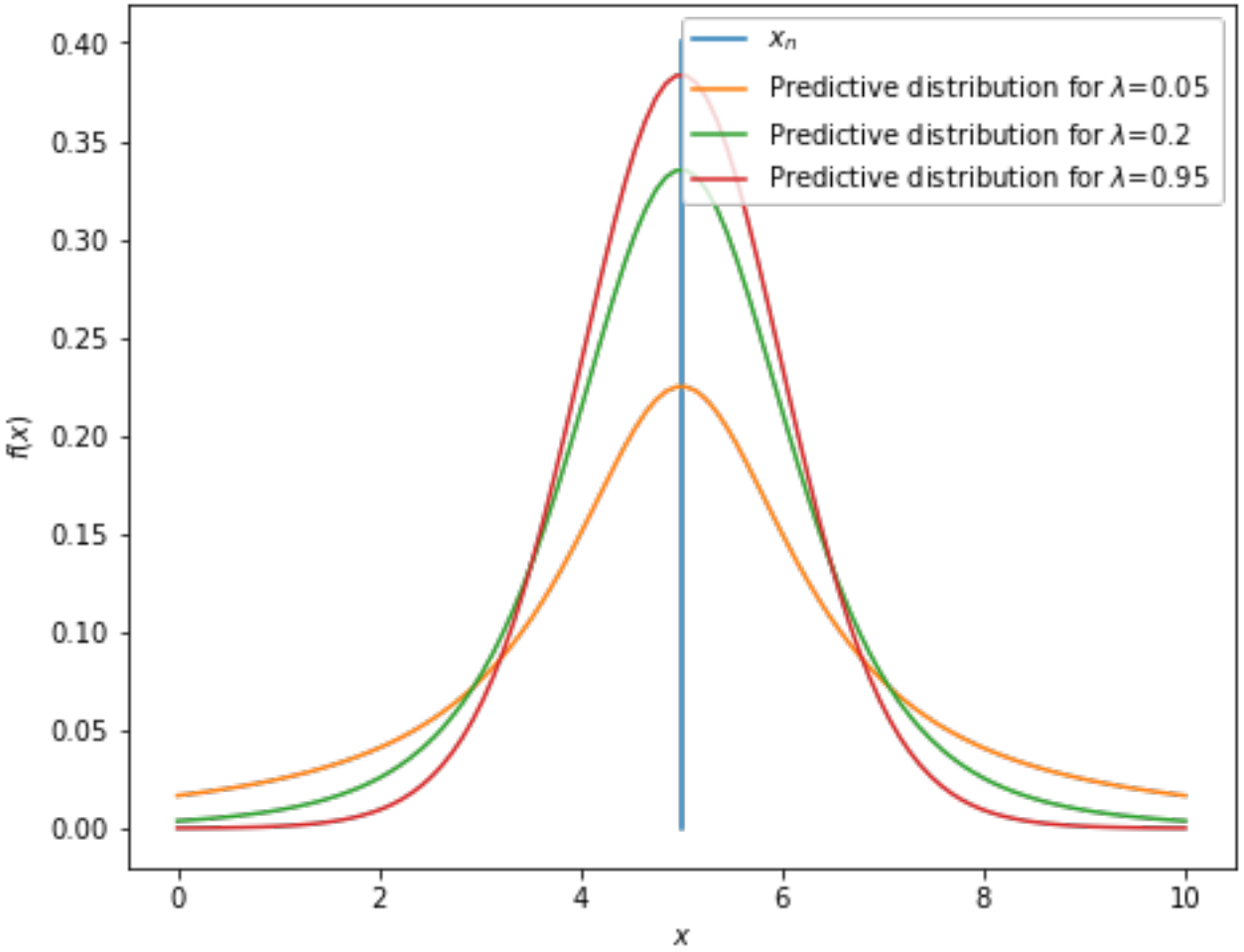


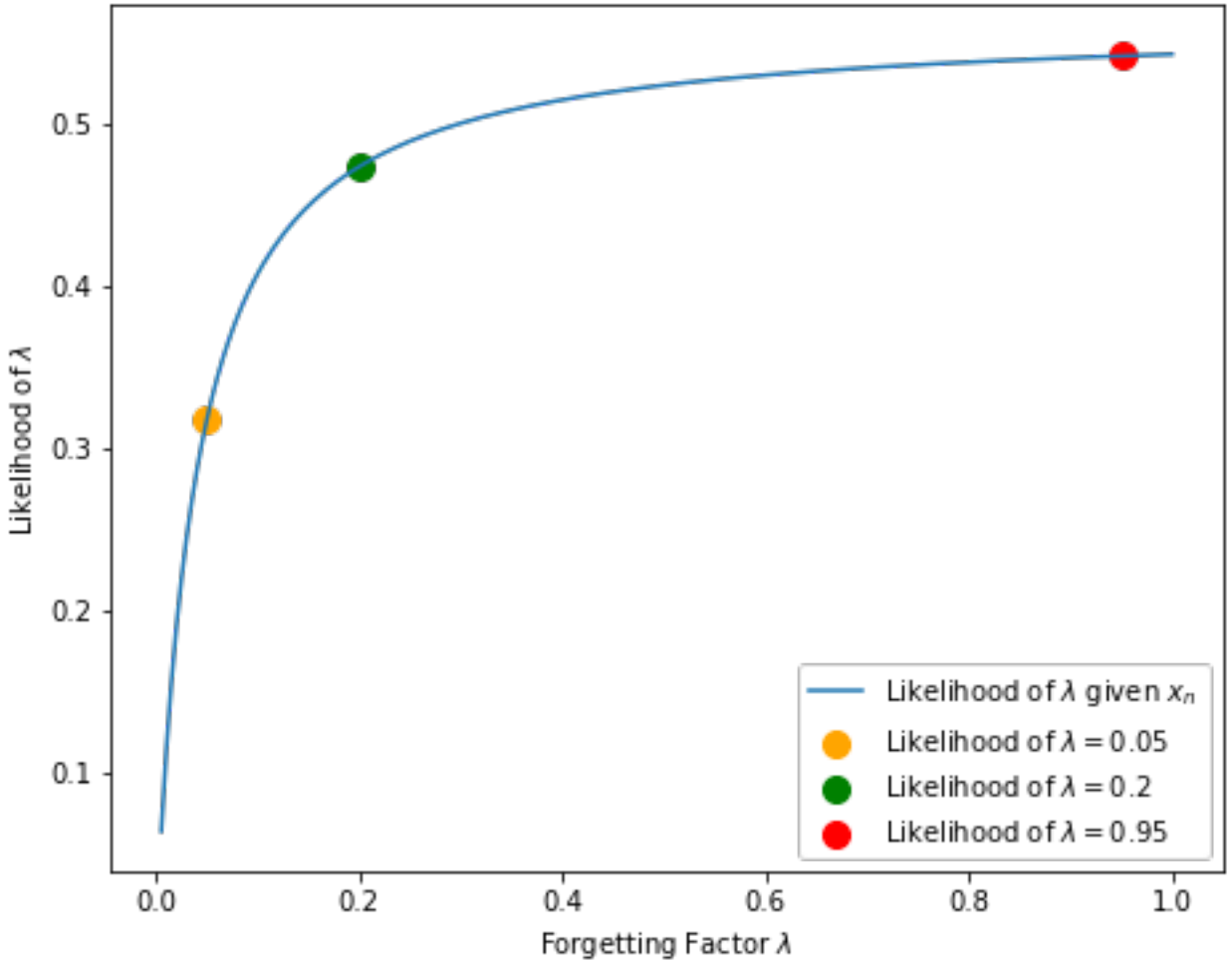
Forgetting Factor Likelihoods





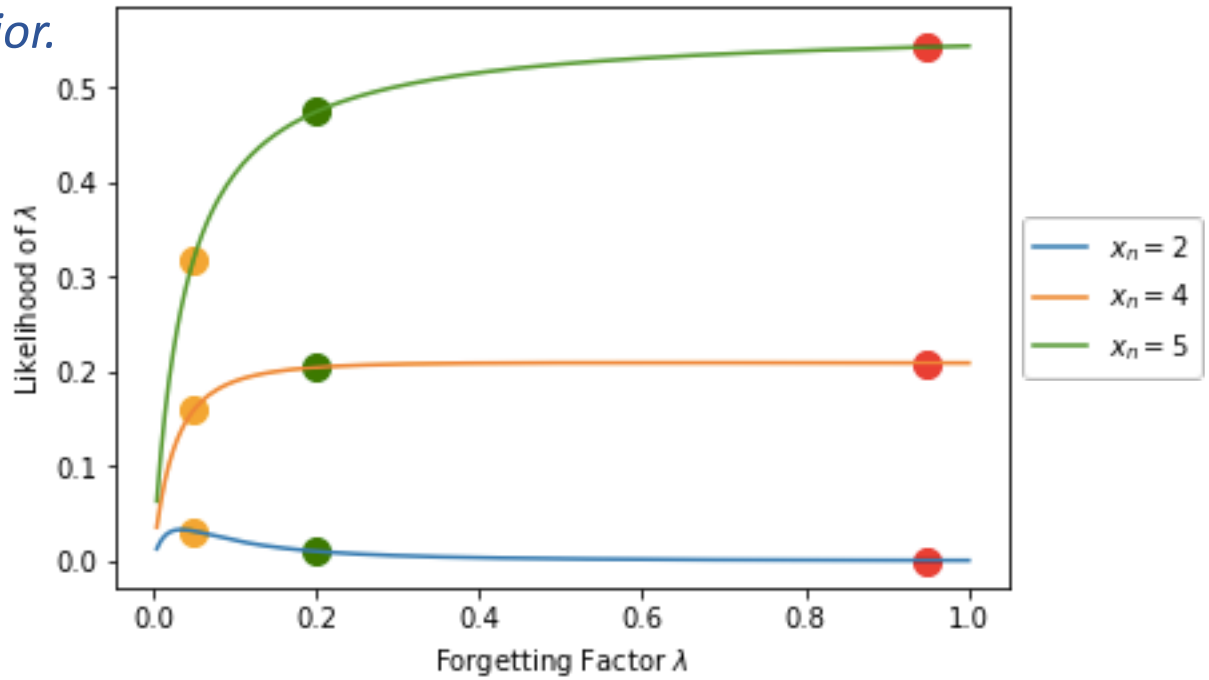
Forgetting Factor Likelihoods



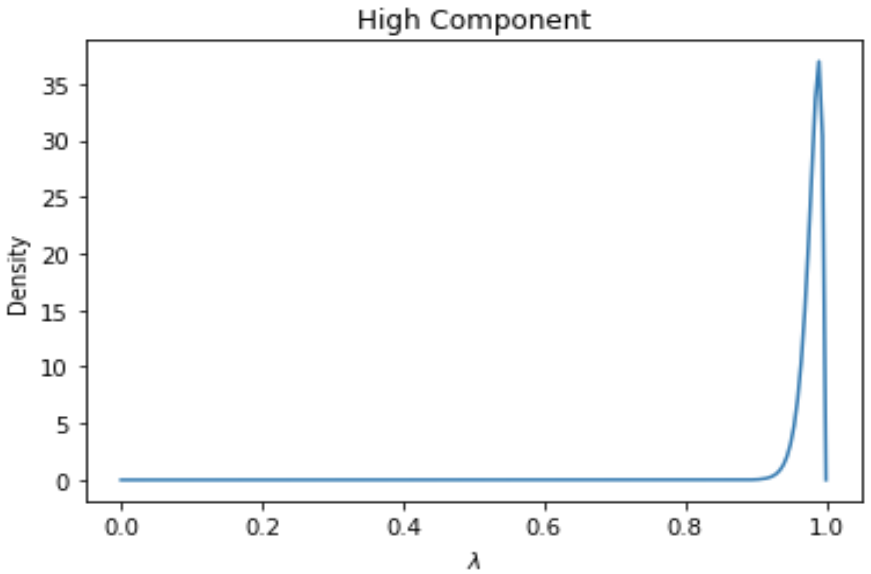


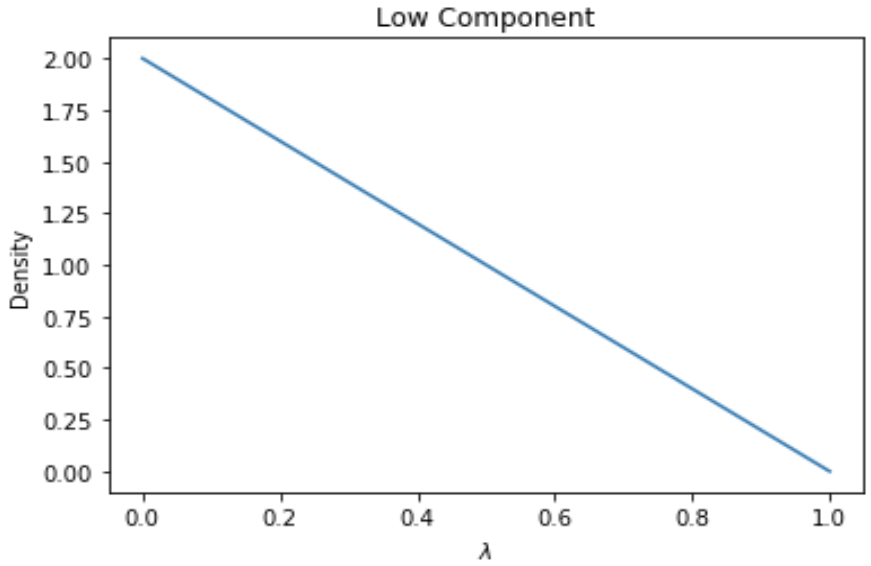
Flatness of Likelihoods

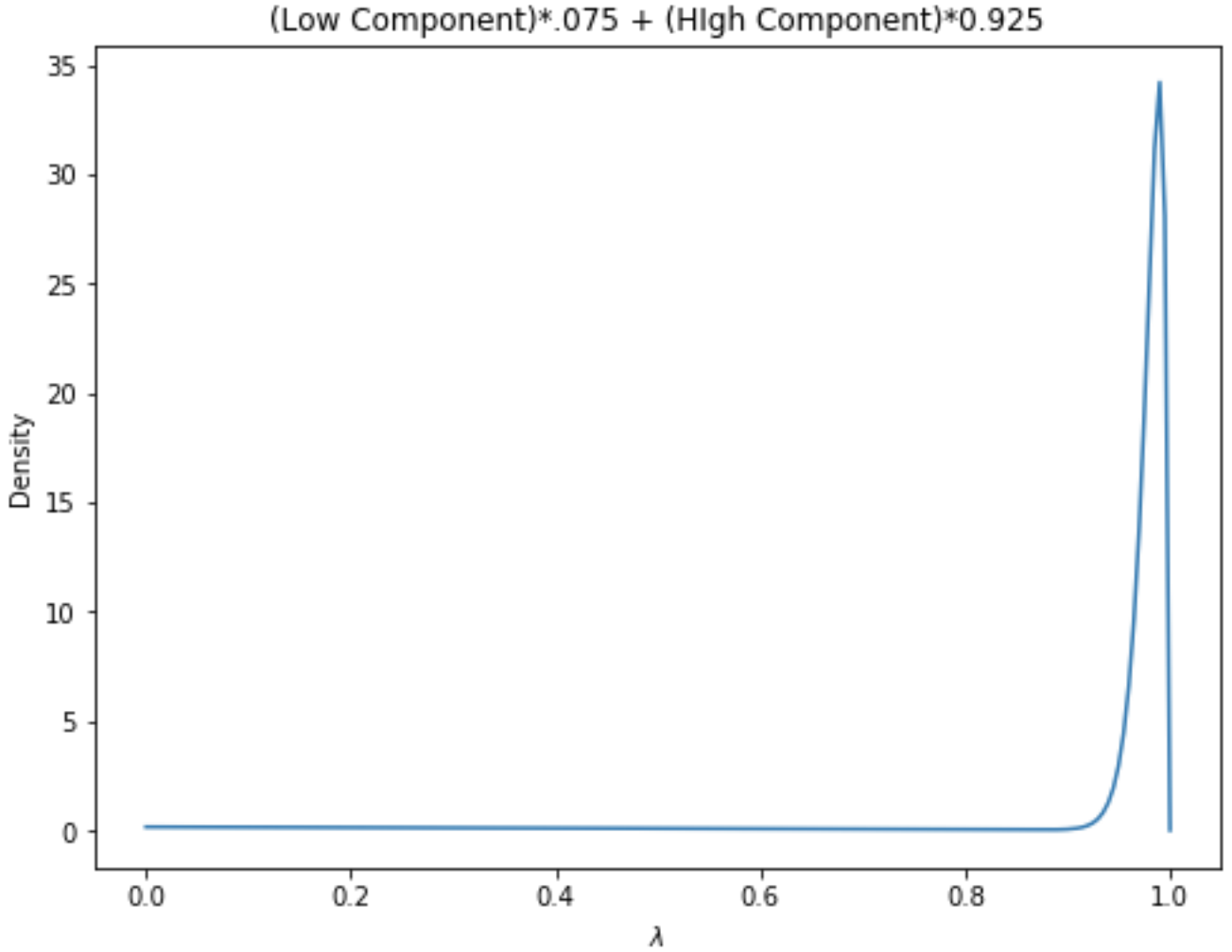
- Our Bayesian analysis is telling us that the data is often not informative of the most likely forgetting factor.
- Upon reflection, this makes sense intuitively.
- If the most recent observation is in an area of high predictive density, we cannot know whether it was drawn from this distribution or any other distribution with high density at that point.
- Because of this, we need to impose an *informative prior.*
- Because a shows ratings are typically relatively stable, at least over a period of several weeks, we choose a prior with most of its density towards 1.



Choice of Beta Mixture Prior

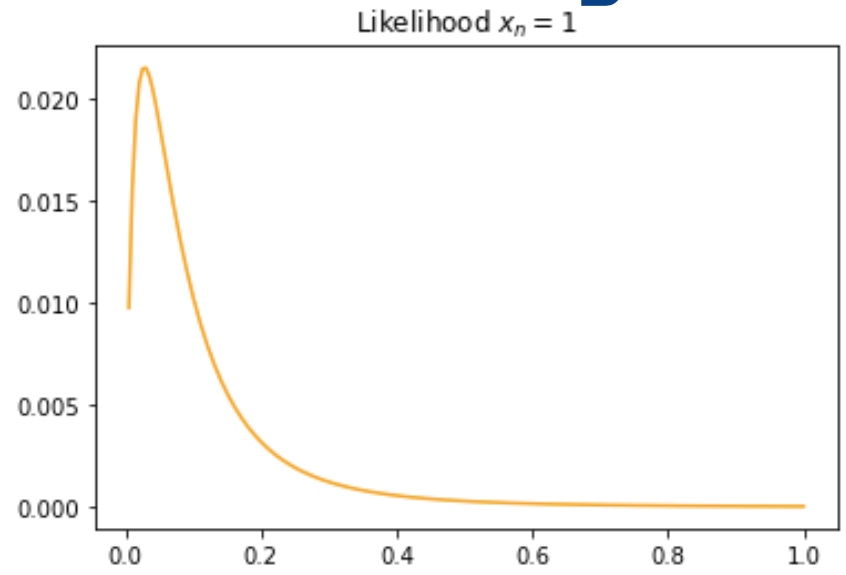


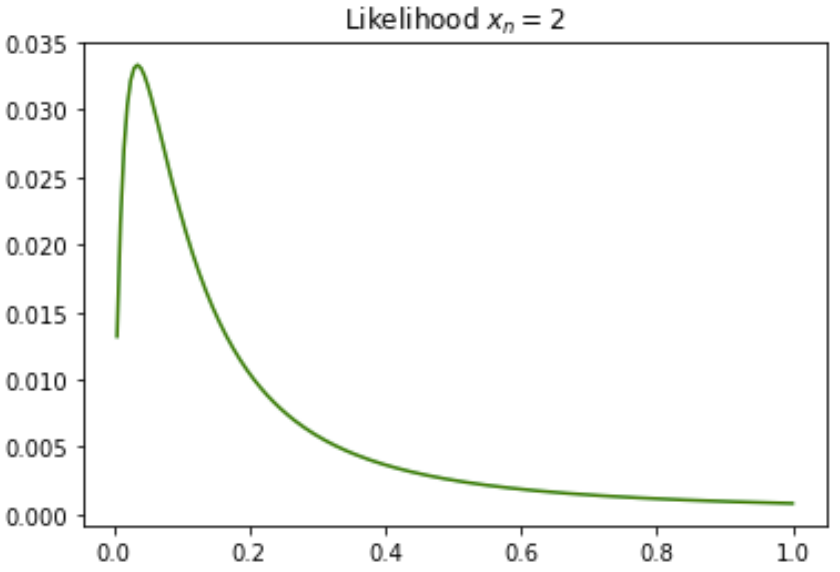


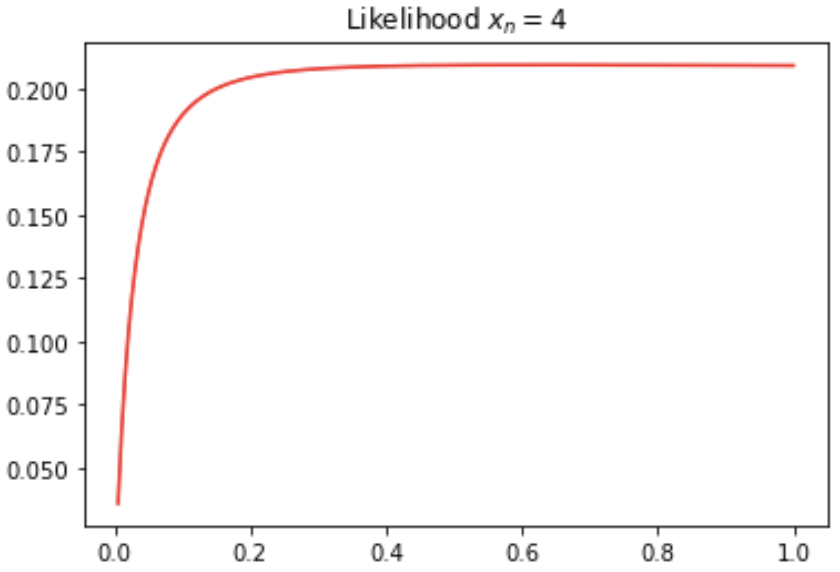


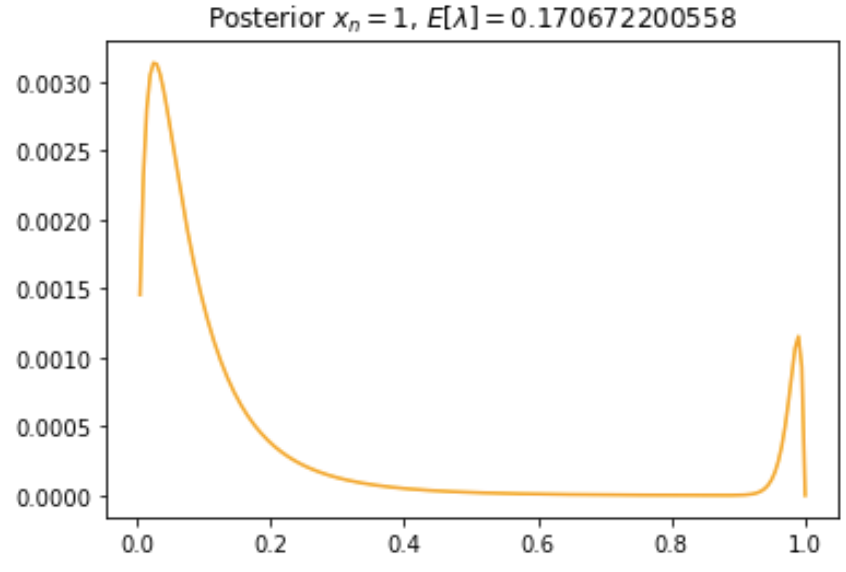
$$E[\lambda] = 0.93$$

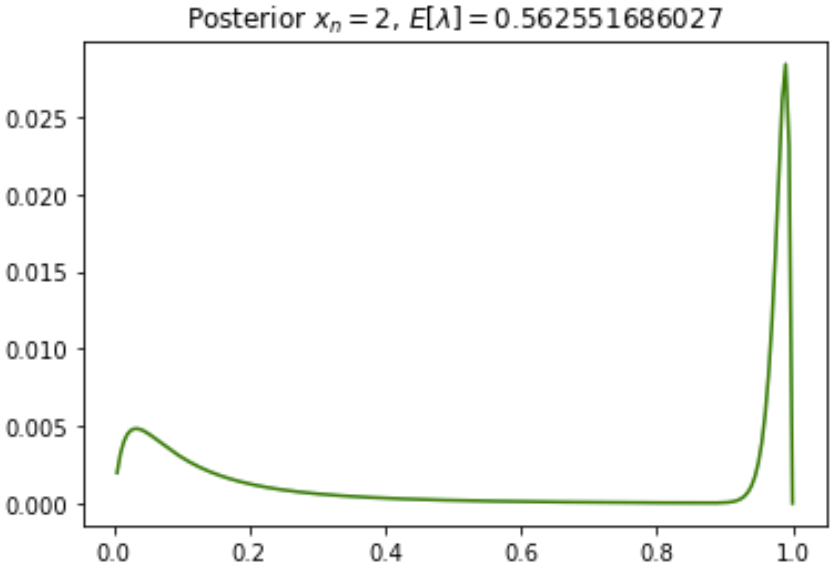
Resulting Posterior Distributions

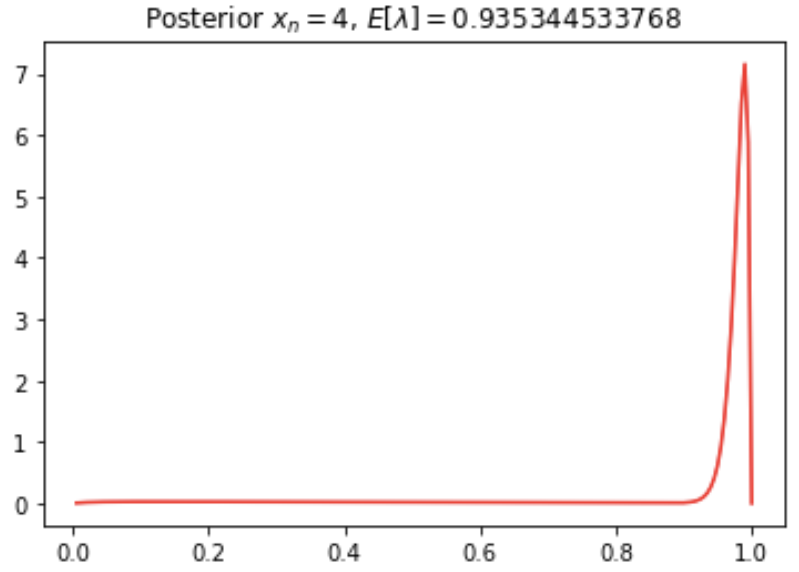




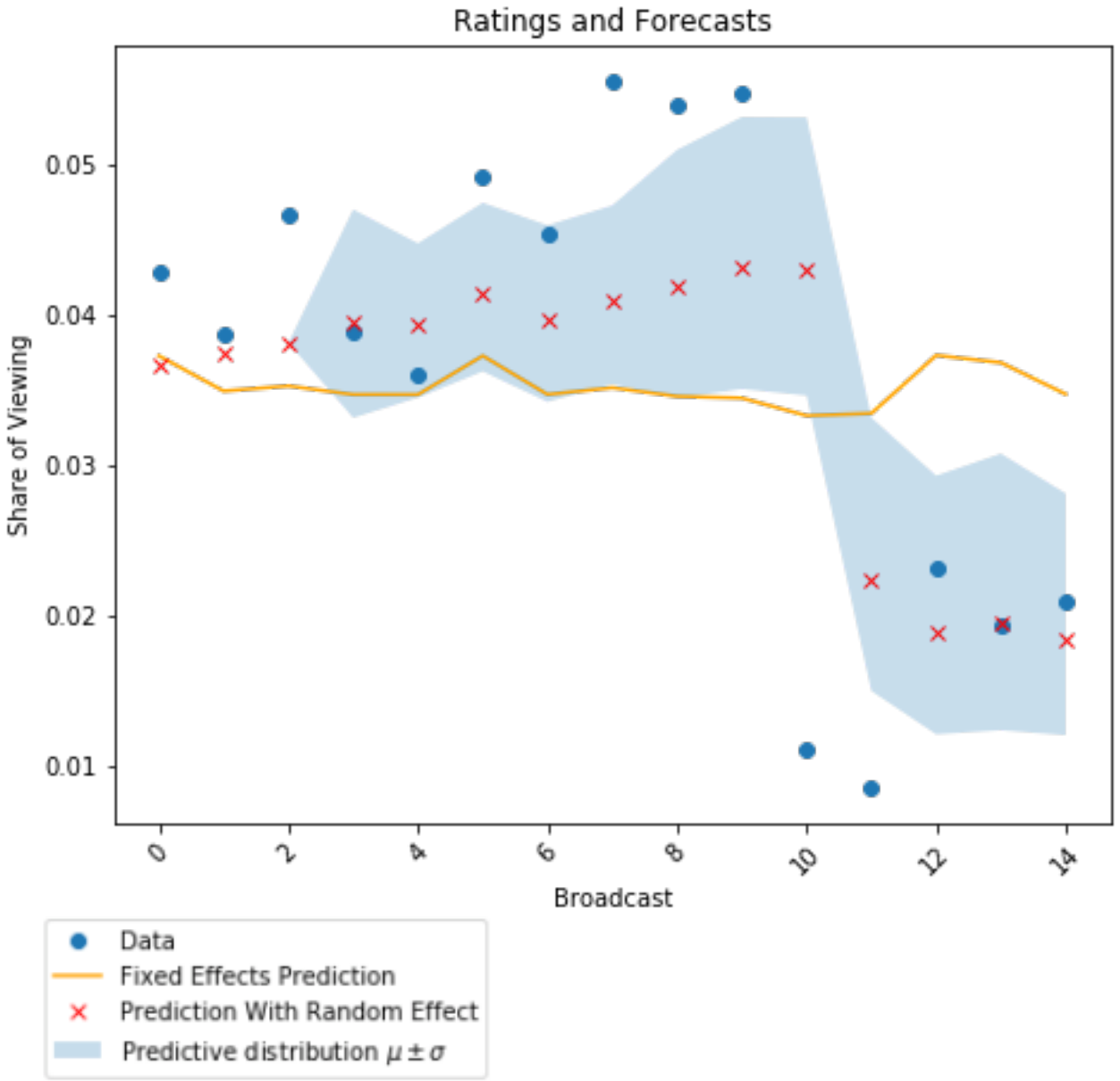


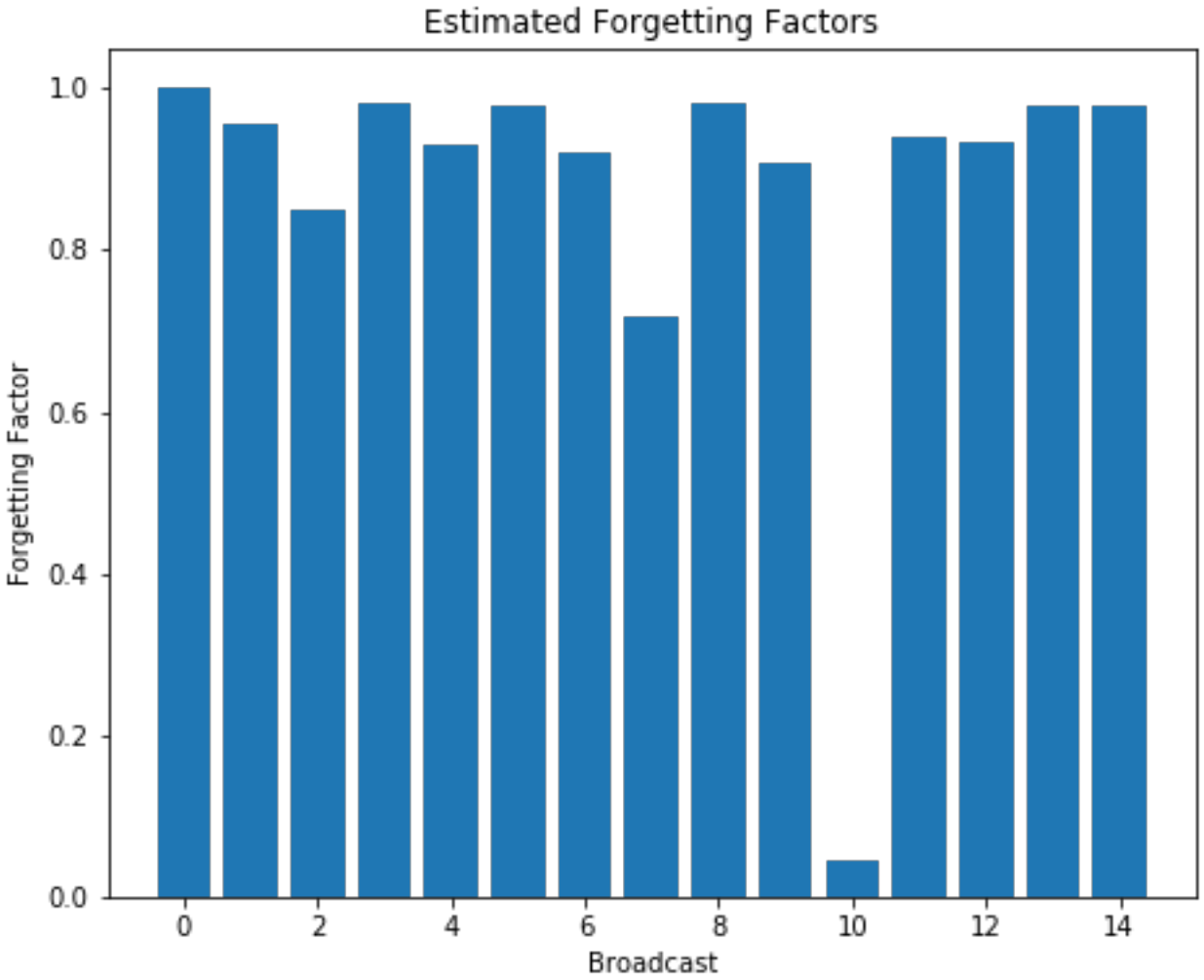




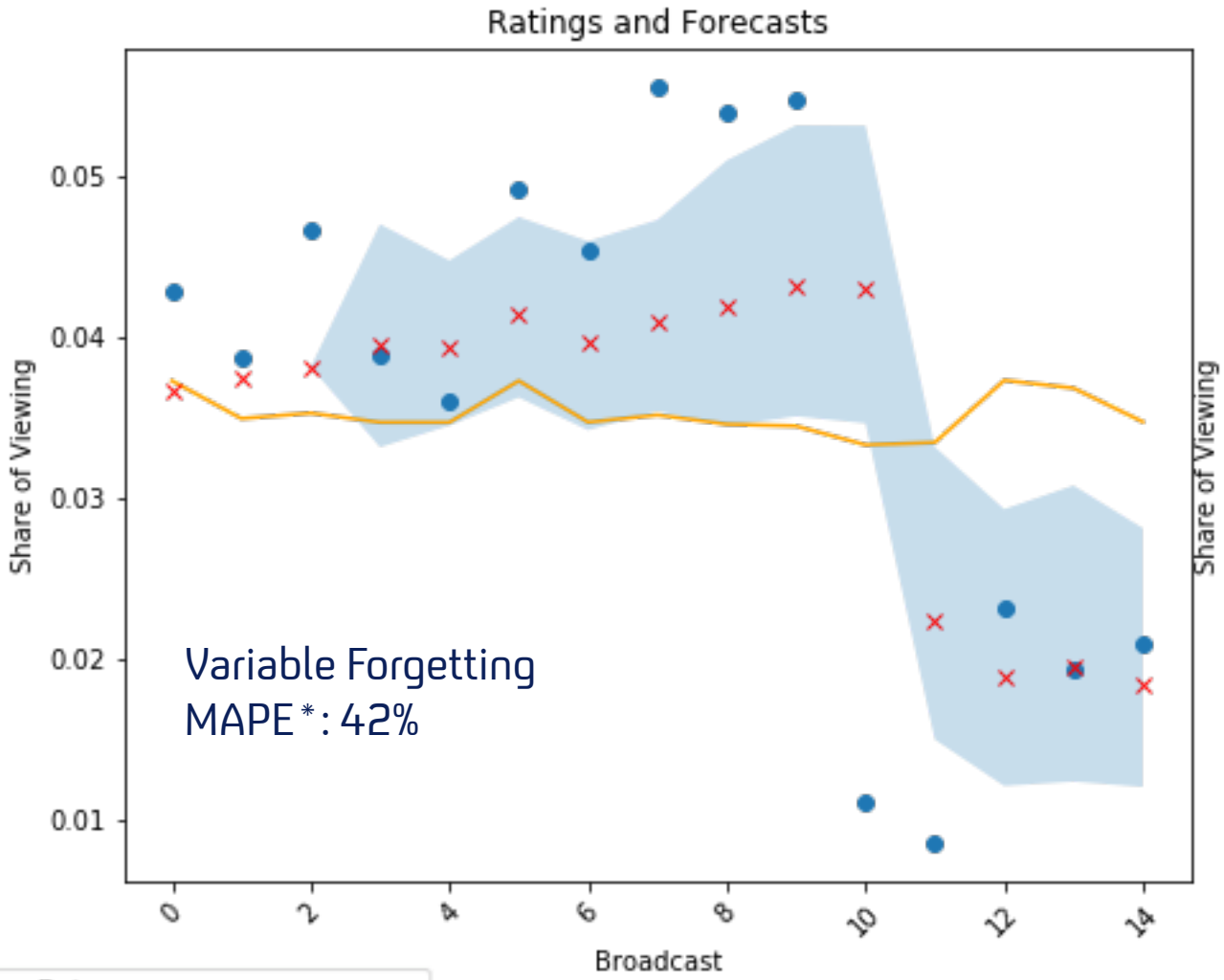


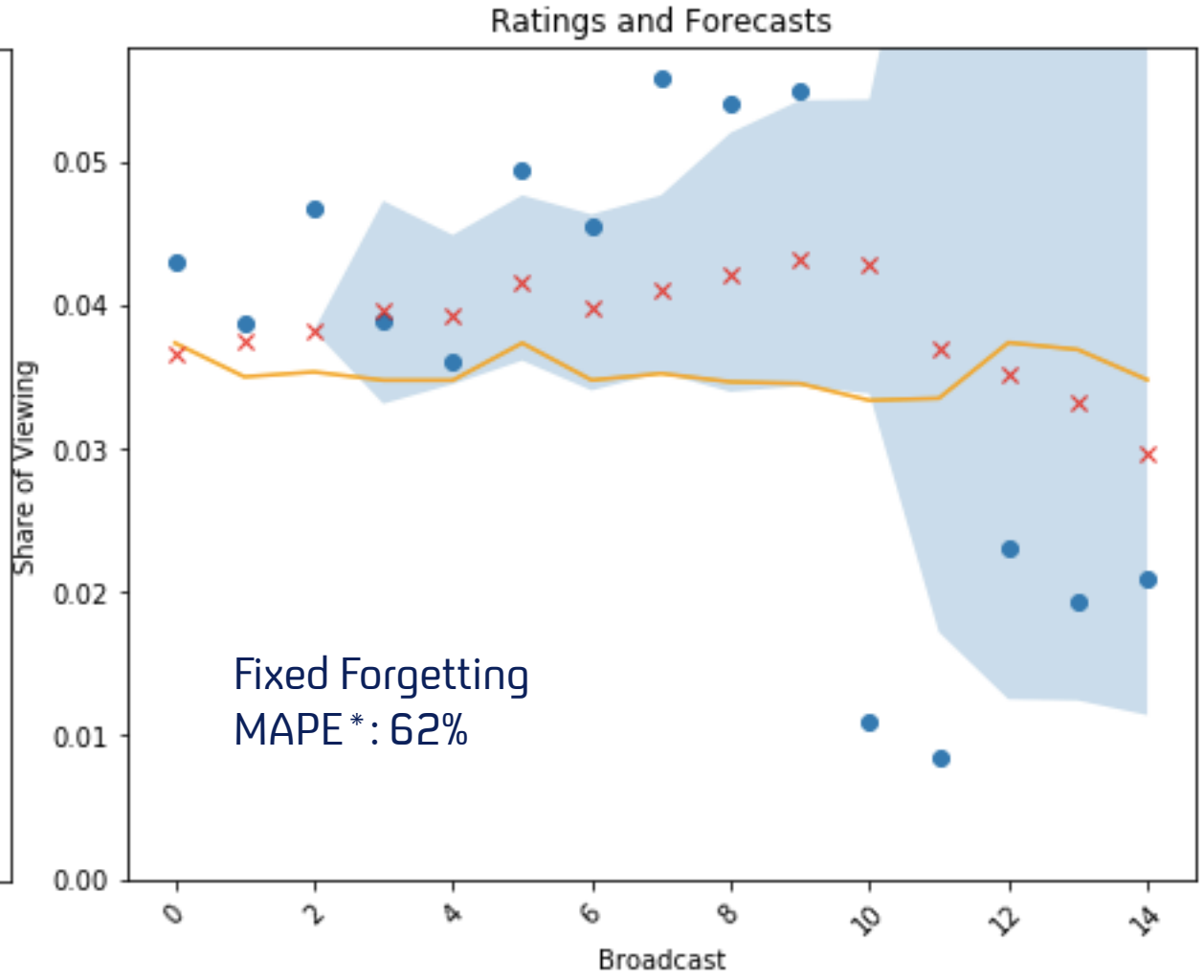
Detecting Changes in Ratings: Shipping Wars

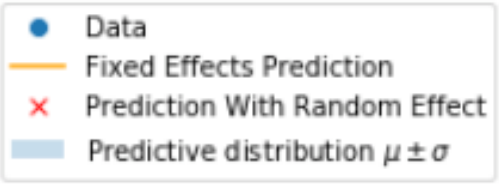




Shipping Wars, Variable Vs. Fixed Forgetting

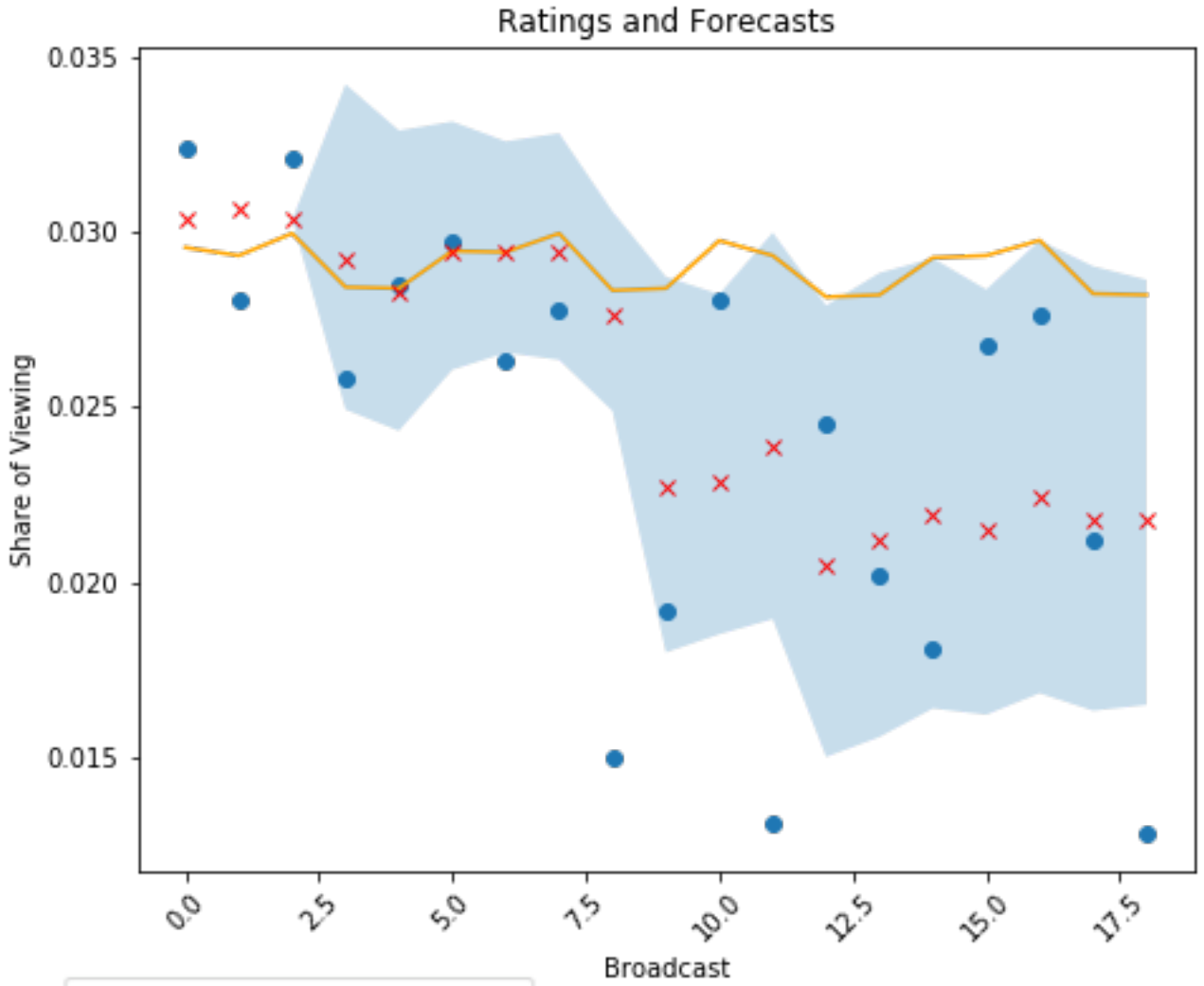


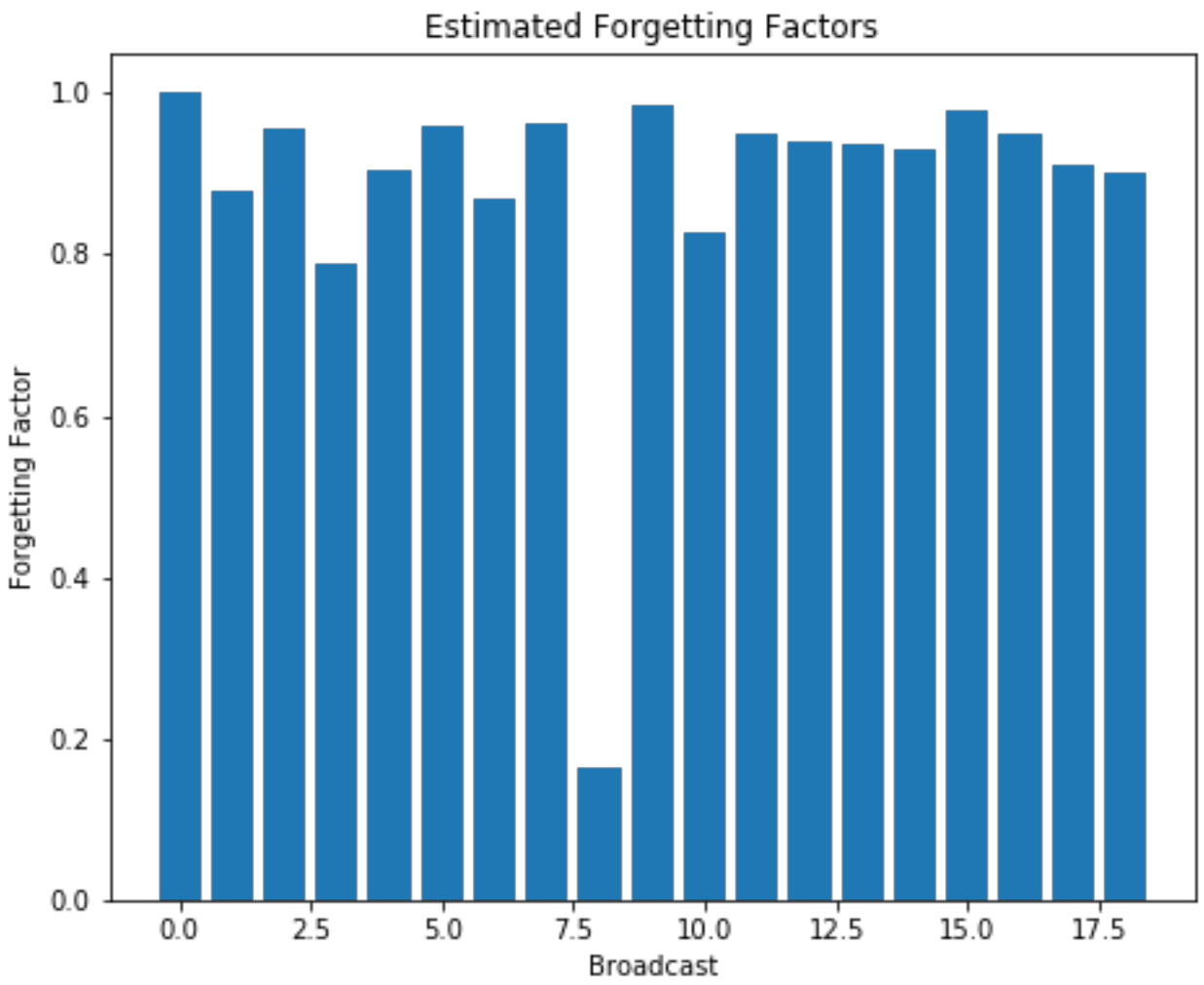




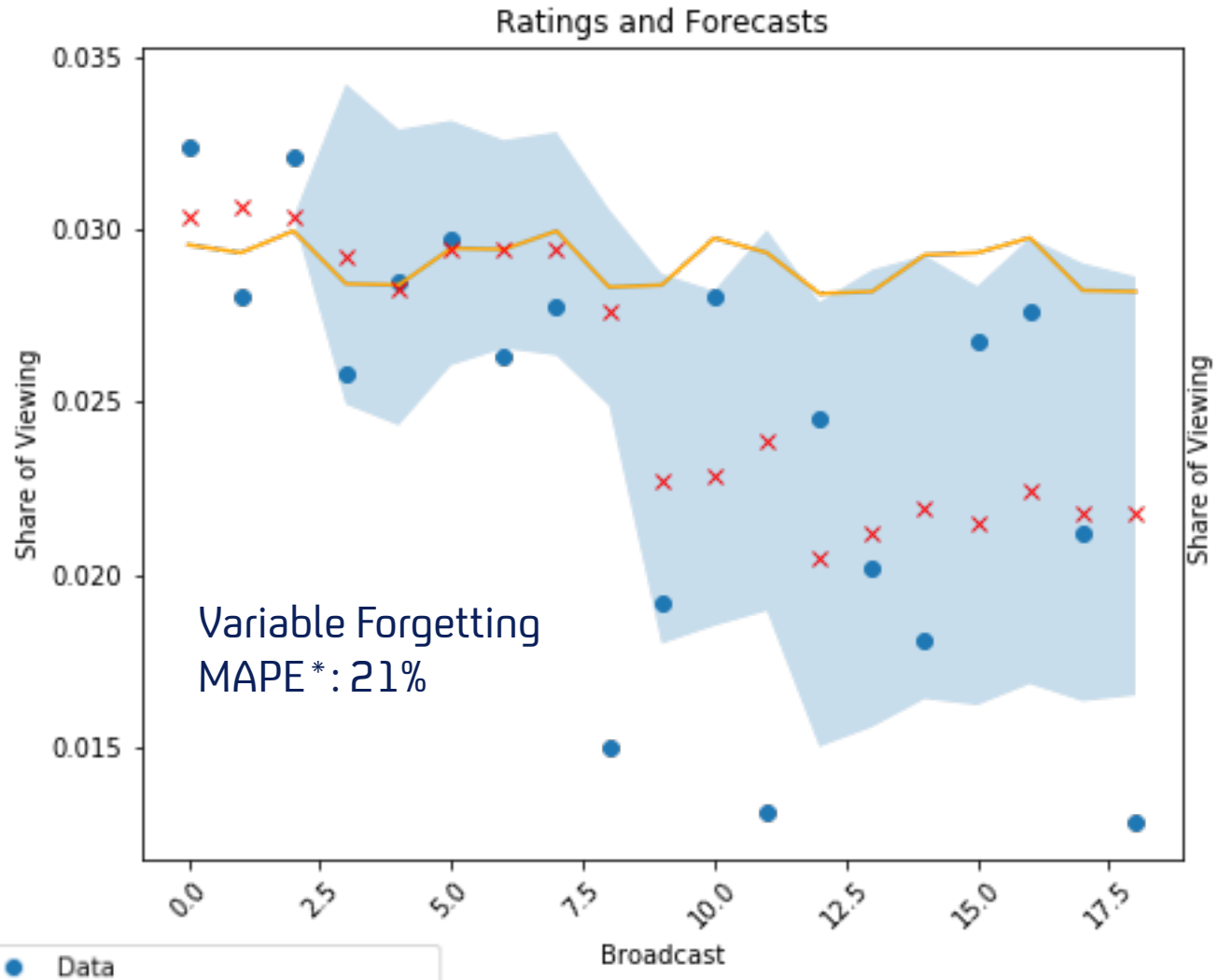
* MAPE = Mean Absolute Percentage Error

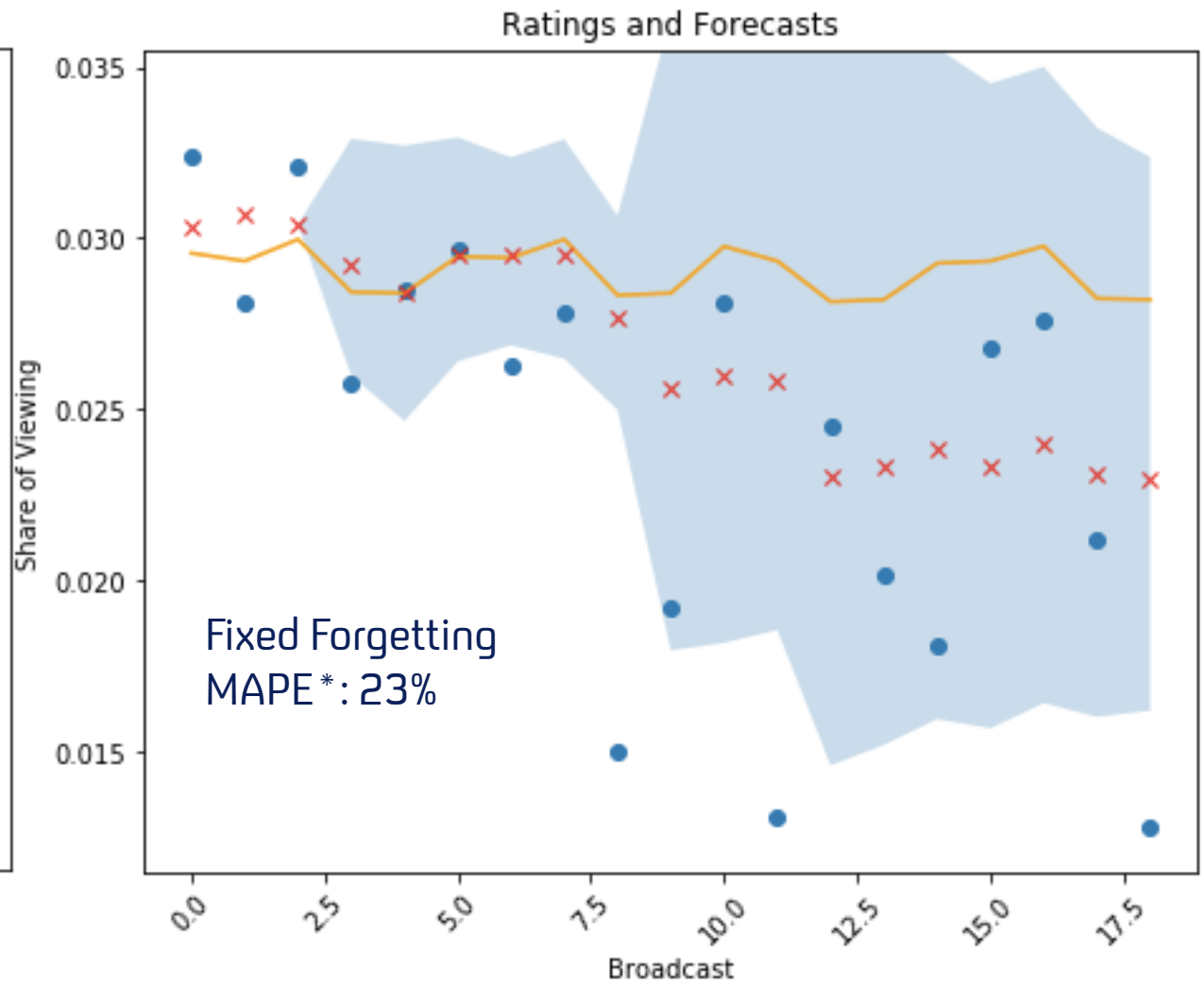
Detecting Changes in Ratings: Gordon Ramsay





Gordon Ramsay, Variable Vs. Fixed Forgetting





* MAPE = Mean Absolute Percentage Error

References

Gustaffson, Smidl; BAYESIAN ESTIMATION OF FORGETTING FACTOR IN ADAPTIVE FILTERING AND CHANGE DETECTION

R. Kulhavy; ON DUALITY OF EXPONENTIAL AND LINEAR FORGETTING

M. Karny; **Approximate** Bayesian **recursive** estimation

P. Danaher; Forecasting television **ratings**

TV Ratings & The Ad Market

Audi orders 5 ABC1 TVRs at £50k/TVR

C4 Creates the programming schedule

C4 Chooses what programmes to put the Audi ad into

Monday:
Location, Location, Location
ABC1 TVRs: 2.3 = £115k
Young TVRs: 0.6 = £30k

Tuesday:
The Inbetweeners
ABC1 TVRs: 0.4 = £20k
Young TVRs: 1.5= £75k

Wednesday:
Peep Show
ABC1 TVRs: 2.4 = £120k
Young TVRs: 1.1 = £55k