University of BRISTOL

The Alan Turing Institute

Epicosm: A framework for linking social media data in large-scale birth cohorts

Dr Oliver Davis

Associate Professor and Turing Fellow

University of Bristol and the Alan Turing Institute

bristol.ac.uk

MRC Integrative Epidemiology Unit


Plus
08:30
94%
Music
Remote
Photos
Camera
iTunes Store
App Store
Maps
Settings
Foursquare
Calendar
Instacast
Google Maps
Evernote
UP
Messenger
Facebook
ReadWrite
Netbot
Tweetbot
Phone
Messages
Mail
Safari

- Engage participants and cohort leaders to assess acceptable approaches
- Develop easy-to-use open-source software for securely linking Twitter data
- Proof-of-principle linking of Twitter data in the ALSPAC cohort
- Dissemination of the software and findings across CLOSER cohorts, funders, participants and policy makers

- Roll out the software to other cohorts
- Develop software to link other social media platforms

- Build the framework for algorithm developers to validate their approaches against “ground truth” in cohorts

What do participants think?

What do participants think?

“Children of the Nineties is fine because I know you’re not going to sell it”

“We’re always sharing our data with loads of people all the time who are using our data for advertising and selling it on. At least with this we would have given our consent and knew it was for a good cause.”

What do participants think?

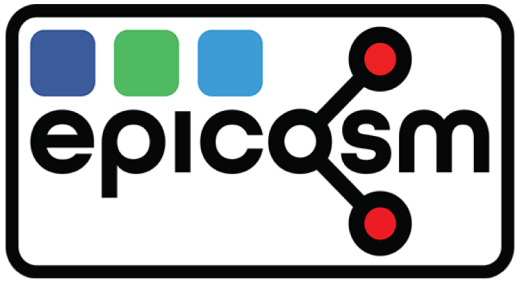
“Children of the Nineties is fine because I know you’re not going to sell it”

“We’re always sharing our data with loads of people all the time who are using our data for advertising and selling it on. At least with this we would have given our consent and knew it was for a good cause.”

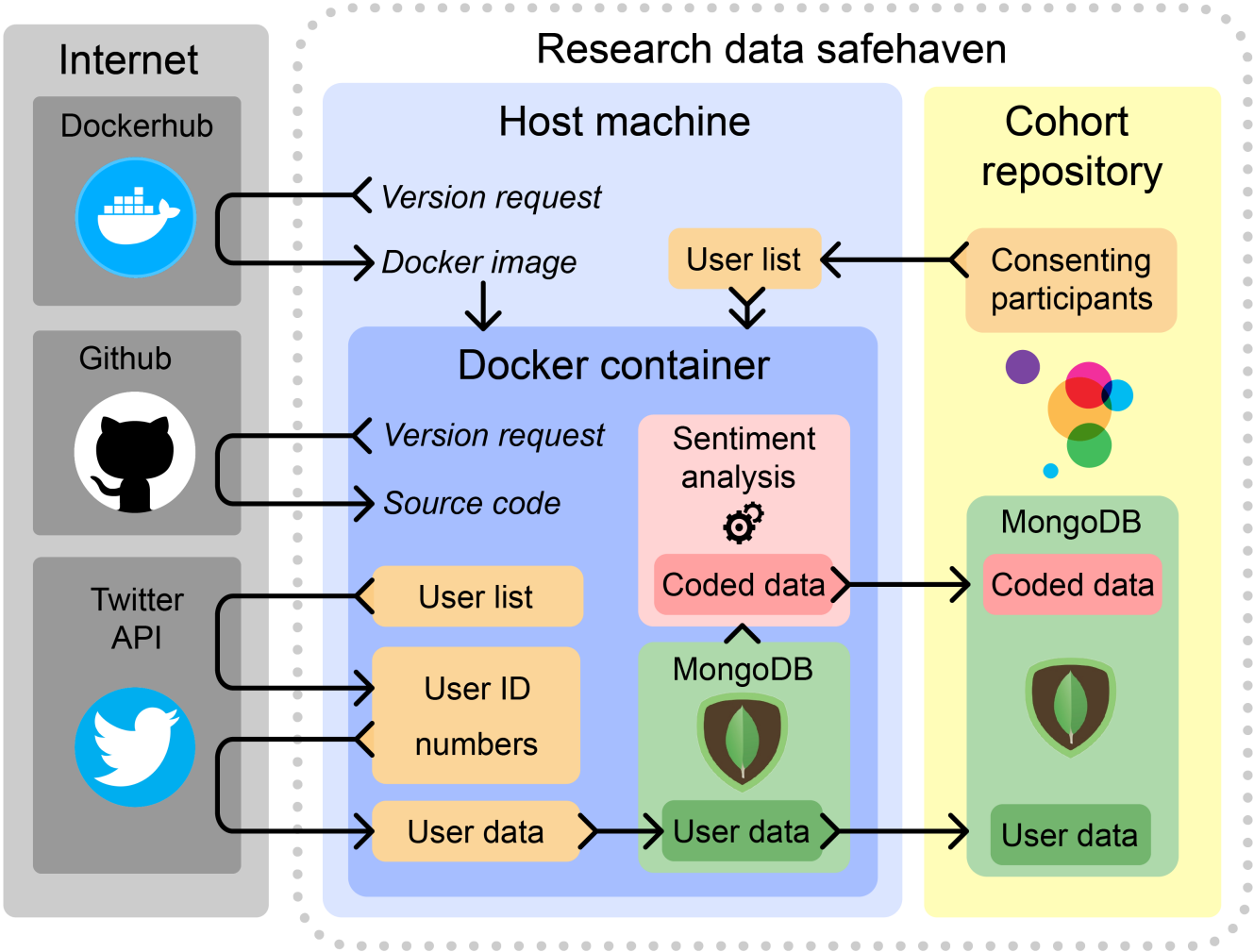
What do participants think?

“Children of the Nineties is fine because I know you’re not going to sell it”

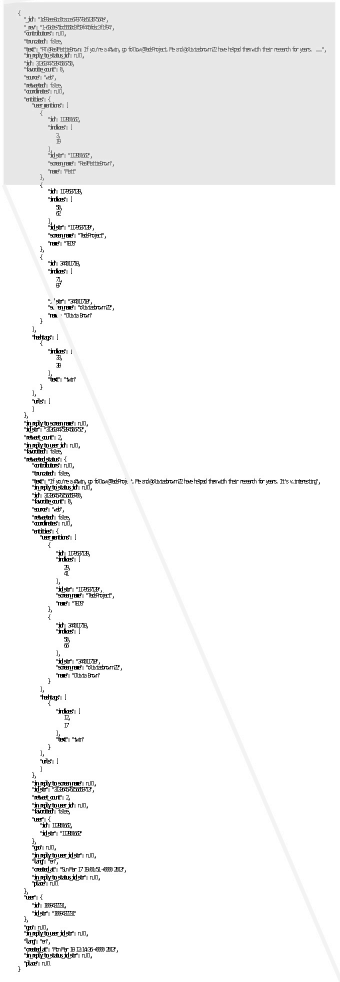
“I think it’s important. Because to get the fullest roundest picture you need to do that anyway, don’t you.”



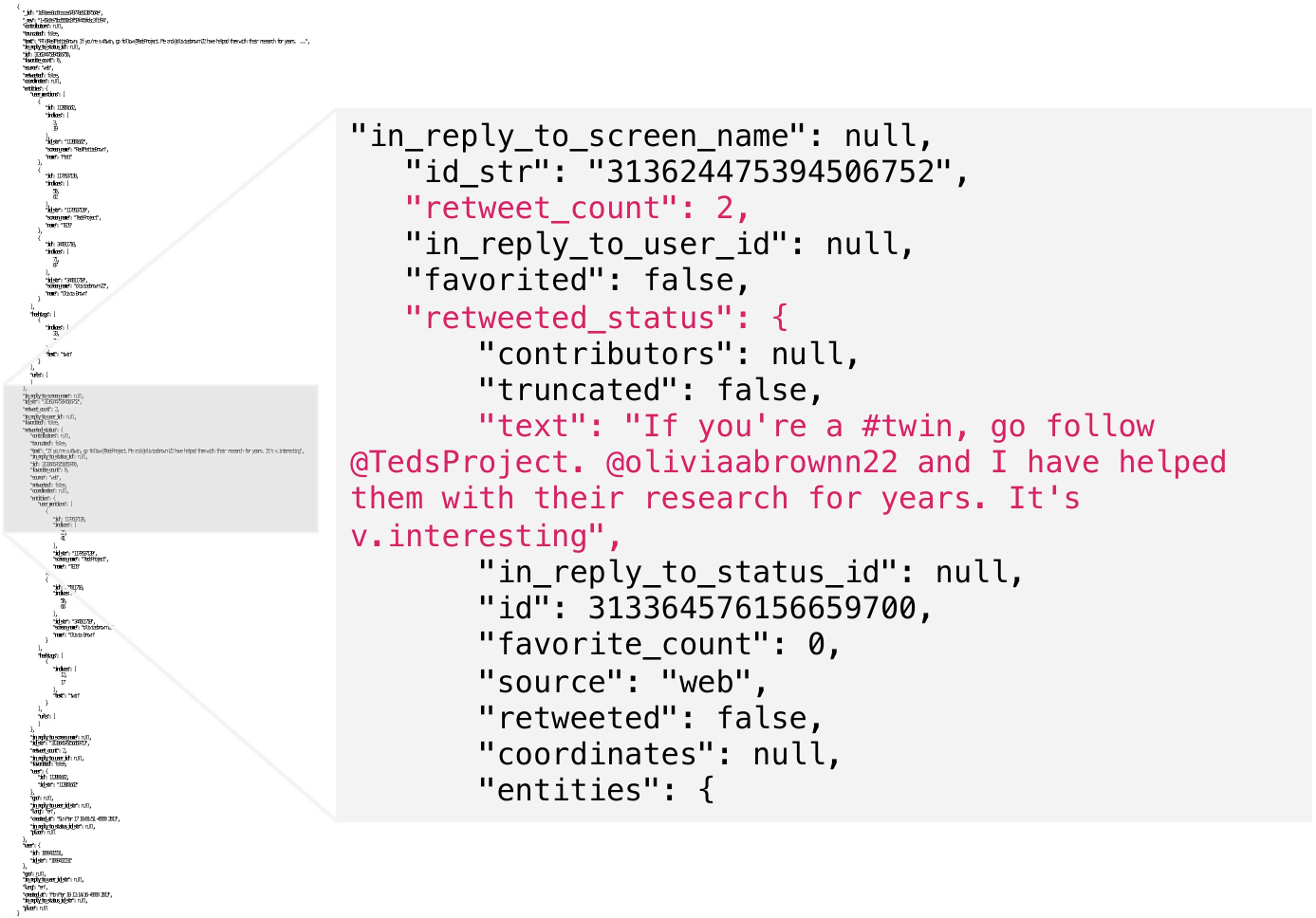
Epidemiological Cohort Online Social Media


Internet
Dockerhub
Github
Twitter API
Research data safehaven
Host machine
Cohort repository
Version request
Docker image
User list
Consenting participants
Docker container
Version request
Source code
Sentiment analysis
Coded data
MongoDB
Coded data
User list
User ID numbers
MongoDB
User data
User data
User data

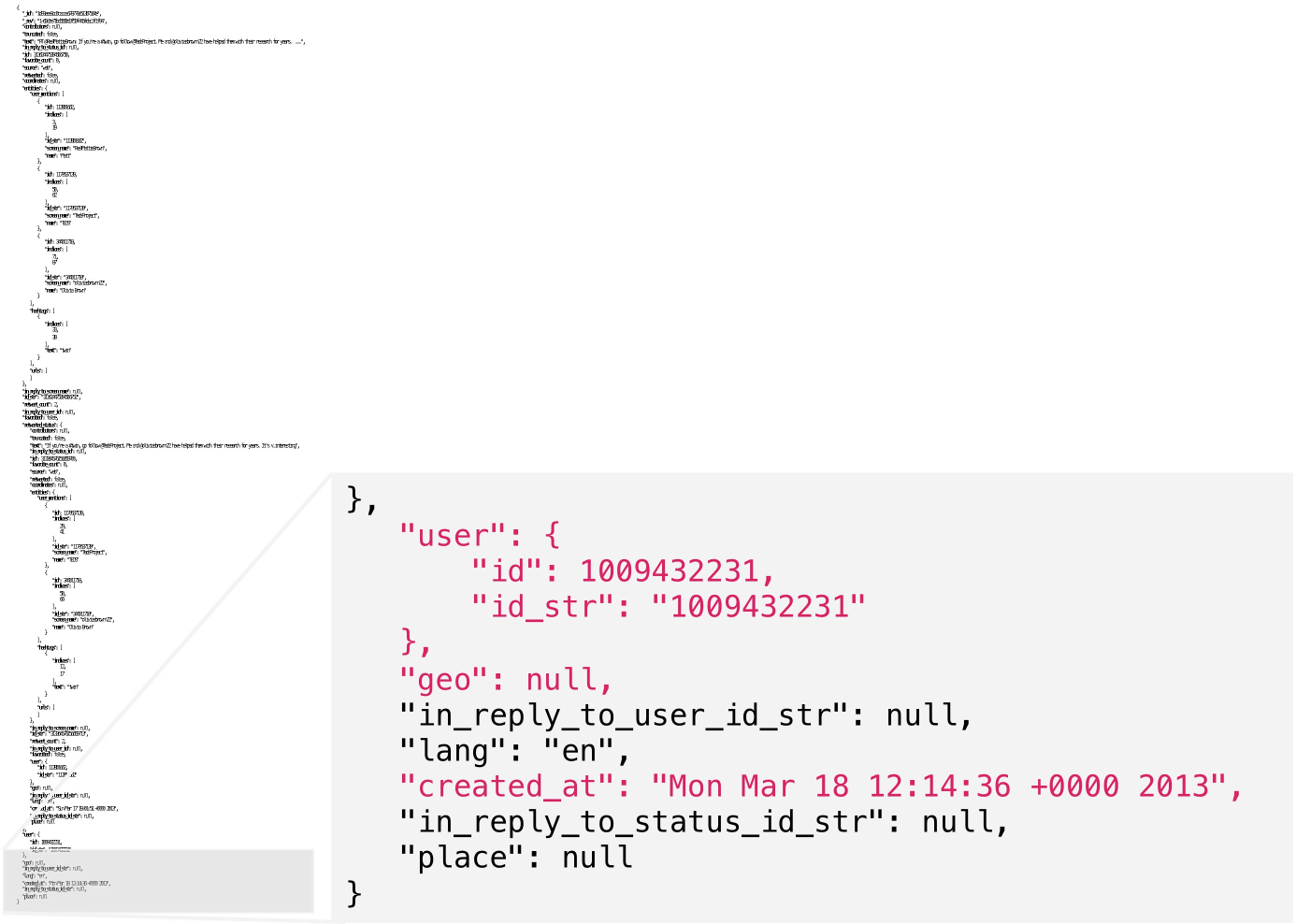
If you're a #twin, go follow @TedsProject. @oliviaabrownn22 and I have helped them with their research for years. It's v. interesting




```
{
    "_id": "1d93eaa8cc0cacce678778a51307564e",
    "_rev": "1-69c0e71bd5550d3f59f445fdac3f1f94",
    "contributors": null,
    "truncated": false,
    "text": "RT @RealMattieBrown: If you're a #twin, go follow @TedsProject. @oliviaabrownn22 and I have helped them with their research for years.  ...",
    "in_reply_to_status_id": null,
    "id": 313624475394506750,
    "favorite_count": 0,
    "source": "web",
    "retweeted": false,
    "coordinates": null,
    "entities": {
        "user_mentions": [
            {
                "id": 112801662,
                "indices": [
                    3,
                    19
                ],
                "id_str": "112801662",
                "screen_name": "RealMattieBrown",
                "name": "Matt"
            },
```




```
"in_reply_to_screen_name": null,
   "id_str": "313624475394506752",
   "retweet_count": 2,
   "in_reply_to_user_id": null,
   "favorited": false,
   "retweeted_status": {
       "contributors": null,
       "truncated": false,
       "text": "If you're a #twin, go follow
@TedsProject. @oliviaabrownn22 and I have helped
them with their research for years. It's
v.interesting",
       "in_reply_to_status_id": null,
       "id": 313364576156659700,
       "favorite_count": 0,
       "source": "web",
       "retweeted": false,
       "coordinates": null,
       "entities": {
```



```
},
    "user": {
        "id": 1009432231,
        "id_str": "1009432231"
    },
    "geo": null,
    "in_reply_to_user_id_str": null,
    "lang": "en",
    "created_at": "Mon Mar 18 12:14:36 +0000 2013",
    "in_reply_to_status_id_str": null,
    "place": null
}
```

participants

researcher

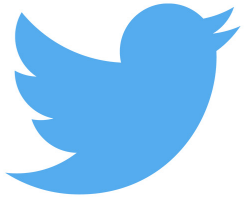
harvester code

Twitter API

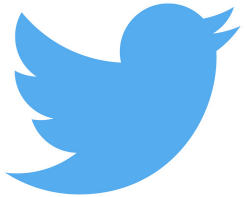
Tweet-scoring code

mongoDB.

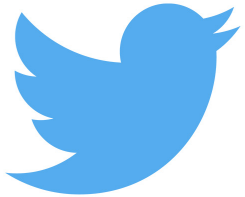
The researcher asks participants for consent and user names



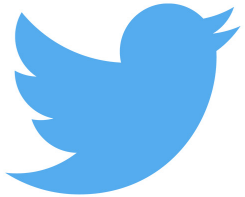
The researcher provides the user names to the Python Tweet-harvesting code



The harvester contacts the Twitter API to convert the participants' user names to persistent identifiers



Every week, the harvester contacts the database to find the ID of the most recent Tweet collected for each participant



weekly

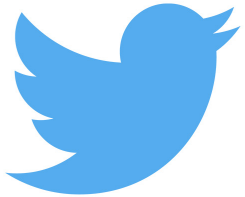
mongoDB.

Then the harvester contacts the Twitter API to download all the Tweets from the past week for each participant (as JSON objects)

weekly

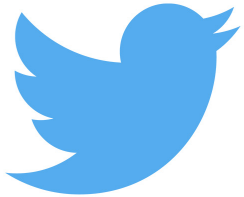
mongoDB.

The harvester stores the Tweets in the database natively as JSON documents



weekly

When requested, the database exports the Tweet data to a text file and sends the file to the Tweet-scoring code

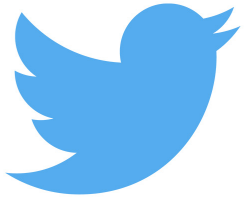


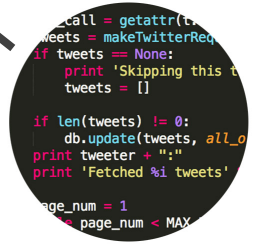
```
if tweets == None:
    print 'Skipping this
    tweets = []

if len(tweets) != 0:
    db.update(tweets, all_o
print tweeter + ":"
print 'Fetched %i tweets'
page_num = 1
page_num < MAX
```

on demand

mongoDB.

The Tweet-scoring code analyses the text of every Tweet and provides scores to the researcher





If you're a #twin, go follow @TedsProject. @oliviaabrownn22 and I have helped them with their research for years. It's v. interesting

if	4.66	
you're	5.30	
a	5.24	
twin	6.14	6.14
go	5.54	
follow	5.66	
and	5.22	
i	5.92	
have	5.82	
helped	7.28	7.28
them	4.92	
with	5.72	
their	5.16	
research	6.46	6.46
for	5.22	
years	5.28	
it's	4.88	
very	6.12	6.12
interesting	7.52	7.52

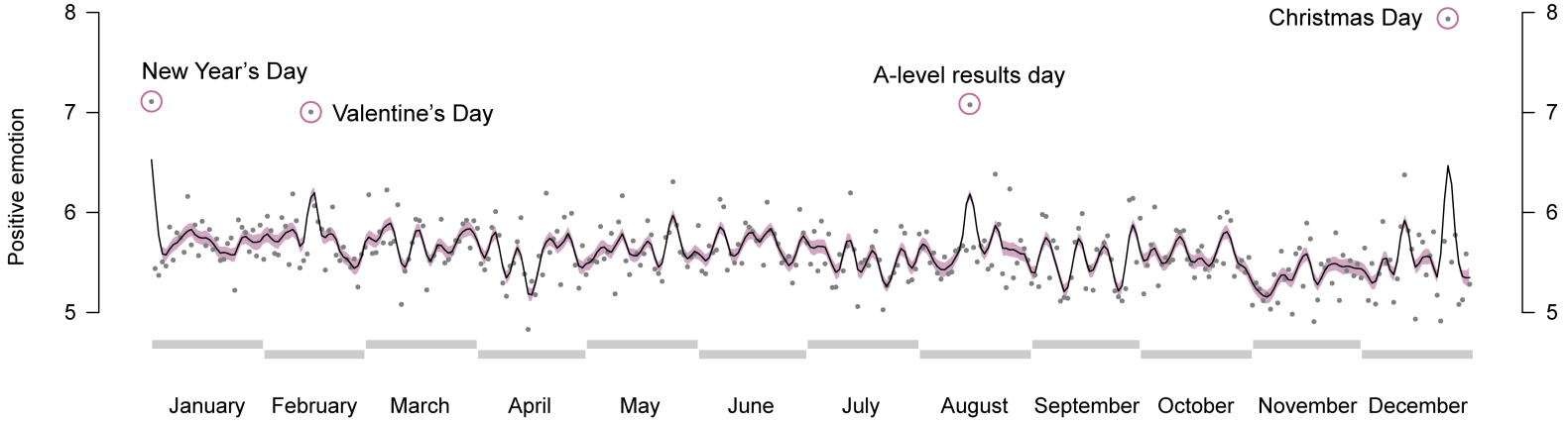
6.14 + 7.28 + 6.46 + 6.12 + 7.52

= 33.52

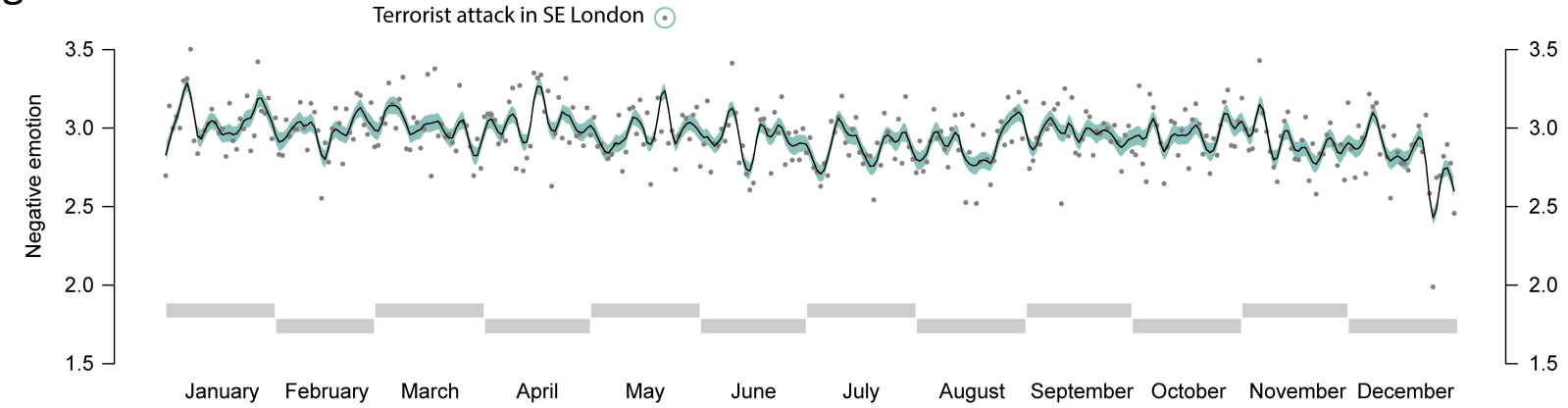
33.52 / 5

= 6.704

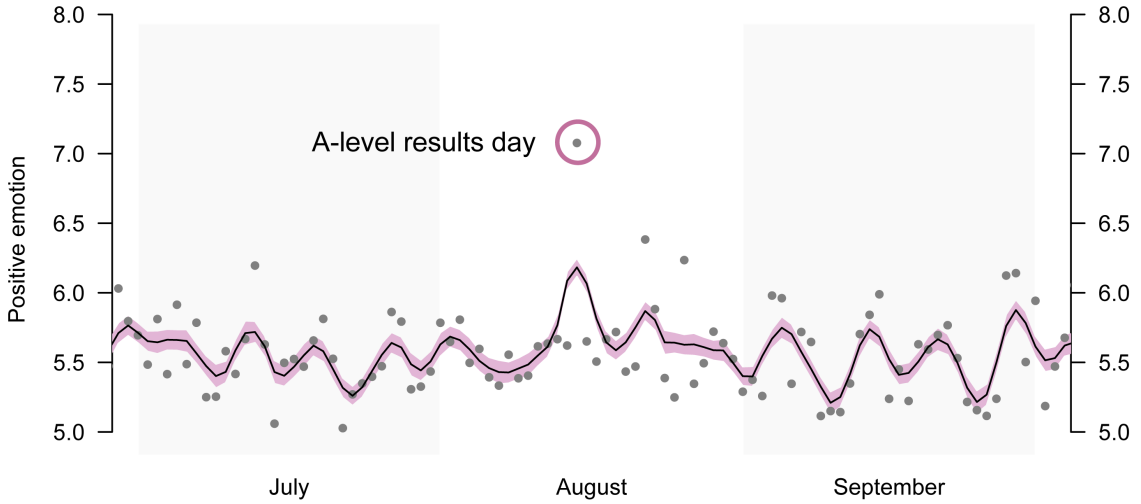
Positive emotion



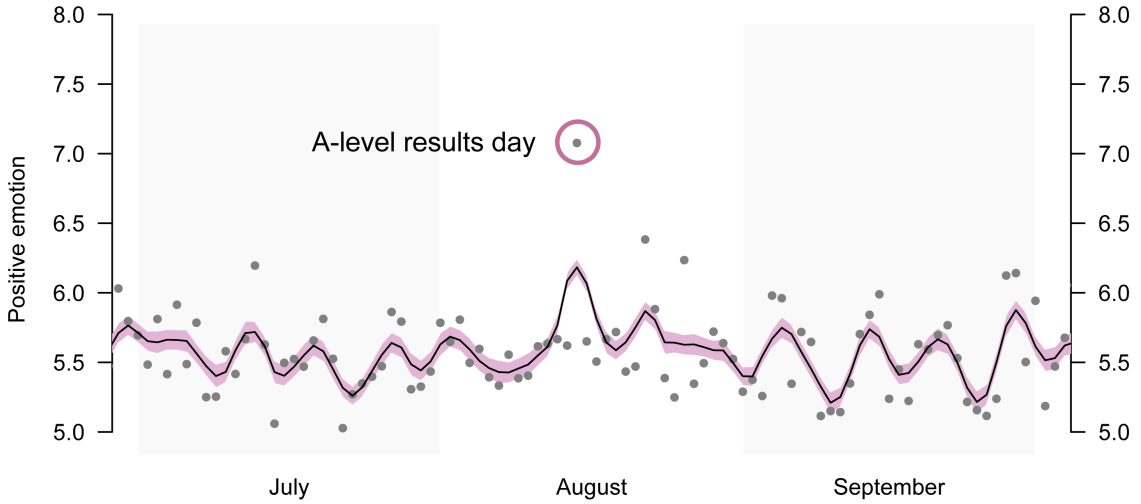
Negative emotion



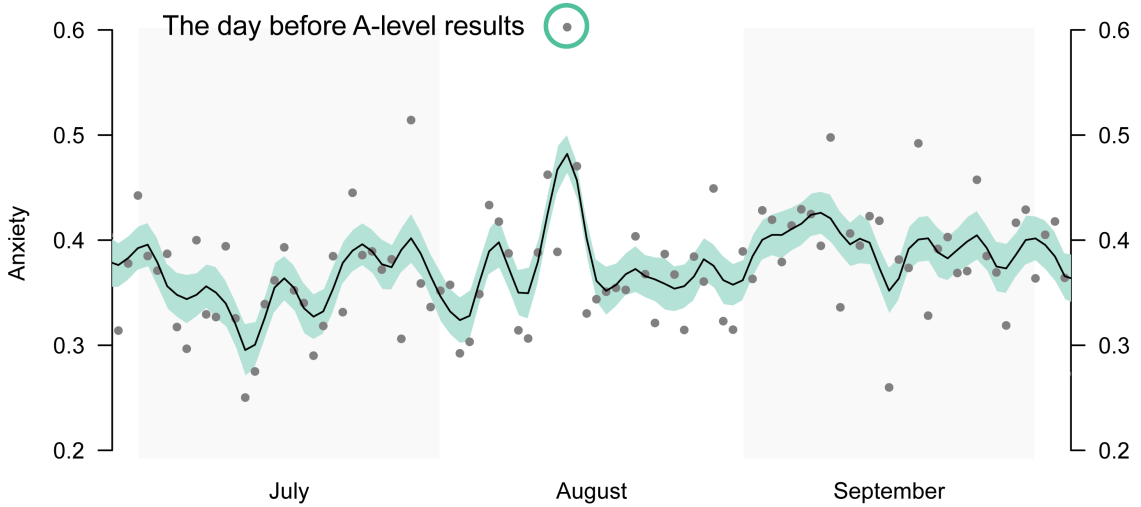
Positive emotion



Positive emotion



Anxiety



Acknowledgements

DYNAMIC GENETICS

University of Bristol

Claire Haworth

Al Tanner

Nina Di Cara

Andy Boyd

Claire Bowring

Richard Thomas

Lynn Molloy

King's College London

Robert Plomin

Stephanie Kliskey

Andy McMillan

UCL

Lisa Calderwood

Cardiff University

Luke Sloan

University of Essex

Tarek Al Baghal

University of BRISTOL

The Alan Turing Institute

LEVERHULME TRUST

MRC Integrative Epidemiology Unit