Module 3: GATE and Social Media

Part 2: Gathering Social Media Data

Social media sites

Twitter, LinkedIn, Facebook

Twitter has varied uptake per country:

- Low in Denmark, Germany (Facebook is preferred)
- Medium in UK, though often complementary to Facebook
- High in USA

Networks have common themes:

- Individuals as nodes in a common graph
- Relations between people
- Sharing and privacy restrictions
- No curation of content
- Multimedia posting and re-posting

Other features: topics, closed groups, moderation, liking, media, groups, person discovery ..

1. Twitter

- Opened in 2006 as a short message blogging service
- Allows ‘subscription’ to interesting accounts
- Anyone can post, most messages are public
- Messages are <280 characters (used to be <140)
- Posts can come from PC, mobile, SMS, iPad etc
- Specialised markup: #hashtags and @mentions
- Has grown extremely popular
  - 100 million active users; over 230 million tweets a day http://www.guardian.co.uk/technology/pda/2011/sep/08/twitter-active-users

Example Uses

Public relations

Barack Obama

We just made history. All of this happened because you gave your time, talent and passion. All of this happened because of you. Thanks

Celebrity worship

Kidrauhl ♡

“One day you will forget me. You have a husband and be a mother. But I will never forget you, My Beliebers.” - Justin Bieber ♥

Broadcasting & Activism

Ars Technica

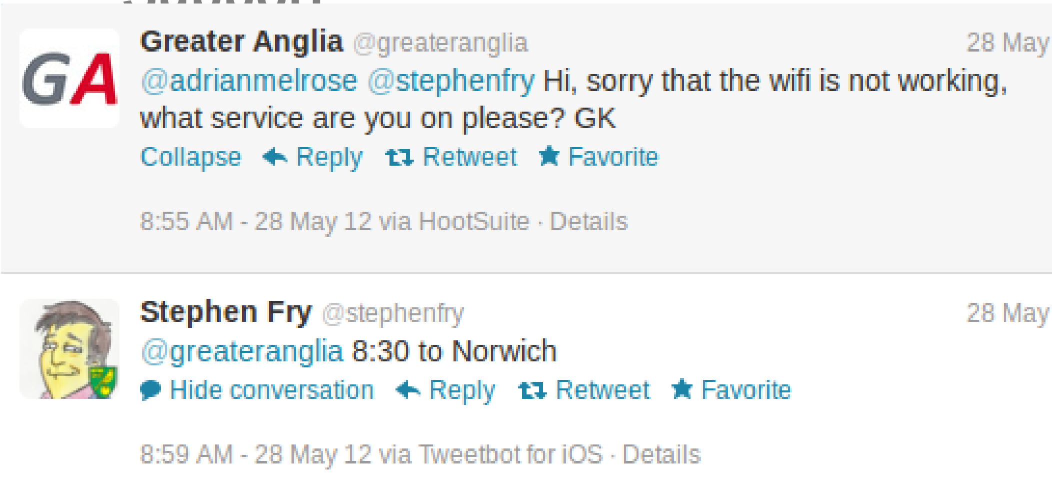
SOPA opponents unveil "Digital Bill of Rights" http://arstechnica.com/tech-policy/20 … by @nathanmattise

Social uses

「ジャム」 Jam Gregory

@RyanBibby: lots of people have been talking about it - need to make sure I watch it! Love @ninaconti, got a signed DVD at #EdFringe :D

Conversations/Customer Support



Greater Anglia @greateranglia 28 May

@adrianmelrose @stephenfry Hi, sorry that the wifi is not working, what service are you on please? GK

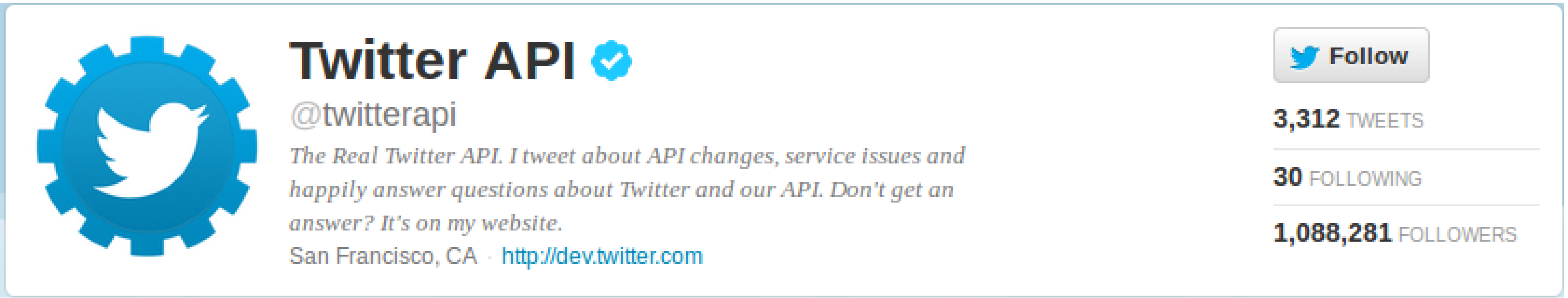
8:55 AM - 28 May 12 via HootSuite · Details

Stephen Fry @stephenfry 28 May

@greateranglia 8:30 to Norwich

8:59 AM - 28 May 12 via Tweetbot for iOS · Details

Twitter User Profiles



Twitter API

@twitterapi

The Real Twitter API. I tweet about API changes, service issues and happily answer questions about Twitter and our API. Don't get an answer? It's on my website.

San Francisco, CA · http://dev.twitter.com

3,312 TWEETS

30 FOLLOWING

1,088,281 FOLLOWERS

- Picture
- Name
- Location
- Website
- Bio (160 characters)

What is Twitter? (2)

- Interest-graph social media

Following/follower relationship is typically not bi-directional

- 77.6% of user connections are not reciprocated (Kwak 2010)

A large graph in which mutual follower/following relationships comprise the edges

Twitterers can 'retweet' one another, so information propagates via the graph quickly

- RTs typically contain links to interesting content

Users can be organised in lists, which introduces groupings

Example Tweet metadata in JSON

```
{    "contributors":null,
"text":"Automotive RDFa (a horribly researched SEO article on RDFa/Microformats):
http://ow.ly/5JSoS #somanyerrorsitsfunny“,
"geo":null,
"retweeted":false,
"in_reply_to_screen_name":null,
"truncated":false,  "entities":{"urls":[{"expanded_url":null,"indices":
[74,92],"url":"http://ow.ly/5JSoS"}],                              "hashtags":
[{"text":"somanyerrorsitsfunny","indices":[93,114]}],
"user_mentions":[]},
"in_reply_to_status_id_str":null,
"id":94029193863639040,
"source":"<a href=\"http://www.hootsuite.com\" rel=\"nofollow\">HootSuite<\/a>“,
"in_reply_to_user_id_str":null,
"favorited":false,
"in_reply_to_status_id":null,
"retweet_count":0,
 "created_at":"Thu Jul 21 13:01:21 +0000 2011",
```

Example Tweet metadata in JSON (2)

```
    "in_reply_to_user_id":null,
"id_str":"94029193863639040“,
"place":{"id":"c799e2d3a79f810e",
    "bounding_box":{"type":"Polygon",
            "coordinates":[[[6.6266397,35.4928765],
                            [18.5203619,35.4928765],
                            [18.5203619,47.0924248],
                            [6.6266397,47.0924248]]]},
        "place_type":"country",
        "name":"Italia",
            "attributes":{},
            "country_code":"IT“,
            "url":"http:/…/1/geo/id/c799e2d3a79f810e.json",
            "full_name":"Italia",
           "country":"Italia"
        },
```

Type of place, e.g. city

Country containing the place of origin

More: https://courses.ischool.berkeley.edu/i202/f11/sites/default/files/map-of-a-tweet.pdf

Example Tweet metadata in JSON (3)

```
"user":{"location":"Blacksburg, VA",
    …,
    "statuses_count":2404,
    "lang":"en",
    "id":20446311,
     …,
    "description":“Text from the user profile (max 160 chars)", …,
    "name":“User Name", …,
    "created_at":"Mon Feb 09 16:33:16 +0000 2009",
    "followers_count":1239,
    "geo_enabled":false, …,
    "url":“The author’s URL (optional)",
    "utc_offset":-21600,
    "time_zone":"Central Time (US & Canada)", ..,
    "friends_count":160, …,
    "screen_name":“twitter-user-name", …,
     "listed_count":189, …
}, …
```

Embedded user information can become out-of-sync, if the user changes it later

More: https://courses.ischool.berkeley.edu/i202/f11/sites/default/files/map-of-a-tweet.pdf

How to get tweets?

The REST API allows access to timelines, tweeting, following, etc.

- REST/JSON based
- Requires registration, and developer / app keys
- Contains access to what was previously the Search API
- Core entities: tweets, users, entities, places
- Heavily rate-limited

The Streaming API streams tweets in real time

- Various strengths available, from 1% to 100% sample (~$1M p.a.)
- May be filtered by language, location, user view, hashtag, search term
-

See https://dev.twitter.com/docs

Getting tweets in the cloud

Gate Cloud tools make getting tweets possible without any programming

- Makes use of the streaming twitter API
- Tweets are stored in real time
- Filter by keyword, username, location and language
- Tweets can be downloaded or stored in the cloud

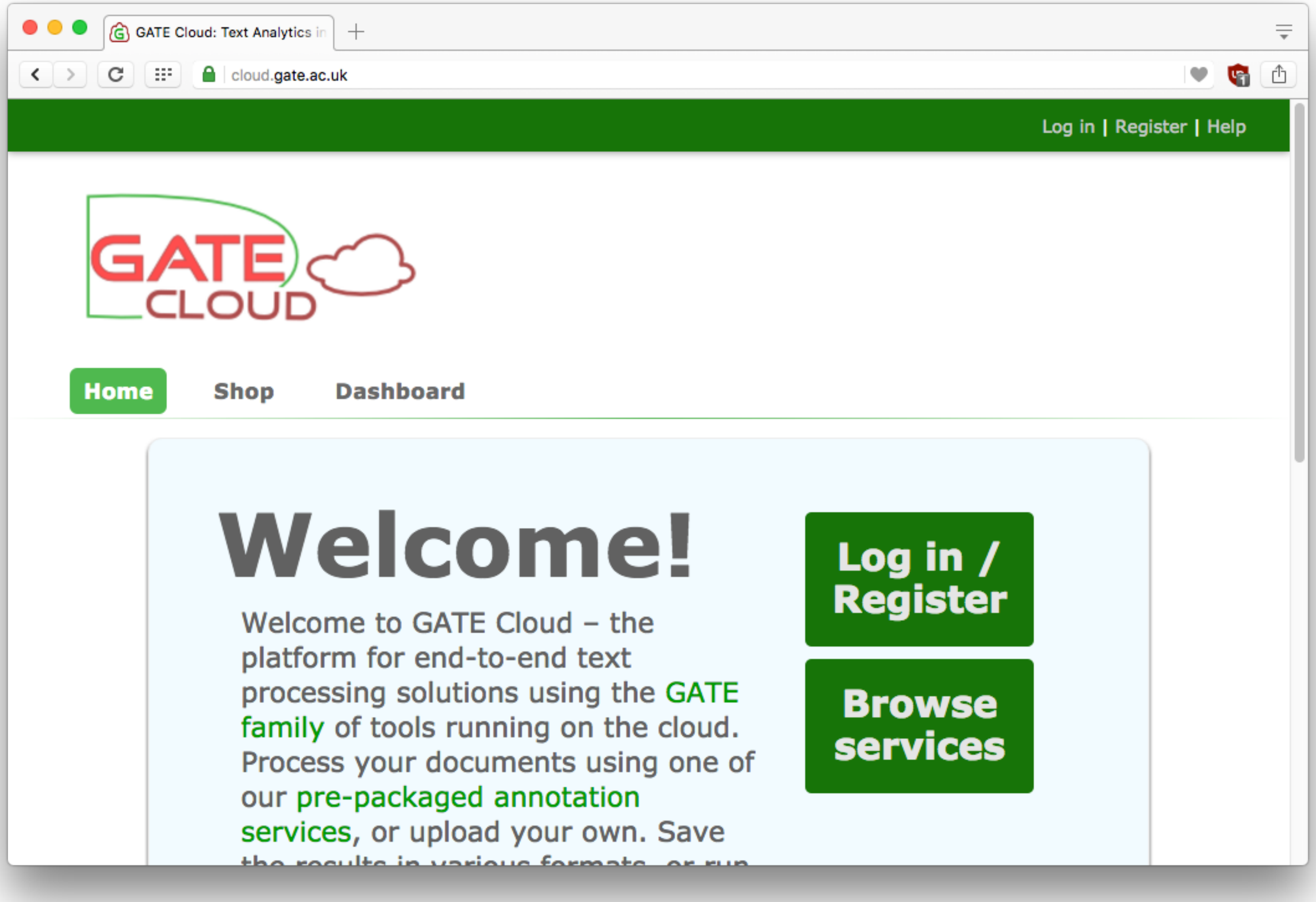
Pay hourly at a very reasonable rate (£0.05 an hour, or about £36 a month)

- First create an account for GateCloud
- Load some credit onto your account
- Order the service and wait for your reservation
- Start the machine and configure the collector!

It’s recommended to save tweets to S3 or GateCloud, as they’ll be deleted after a while if not downloaded.

GATE Cloud

https://cloud.gate.ac.uk

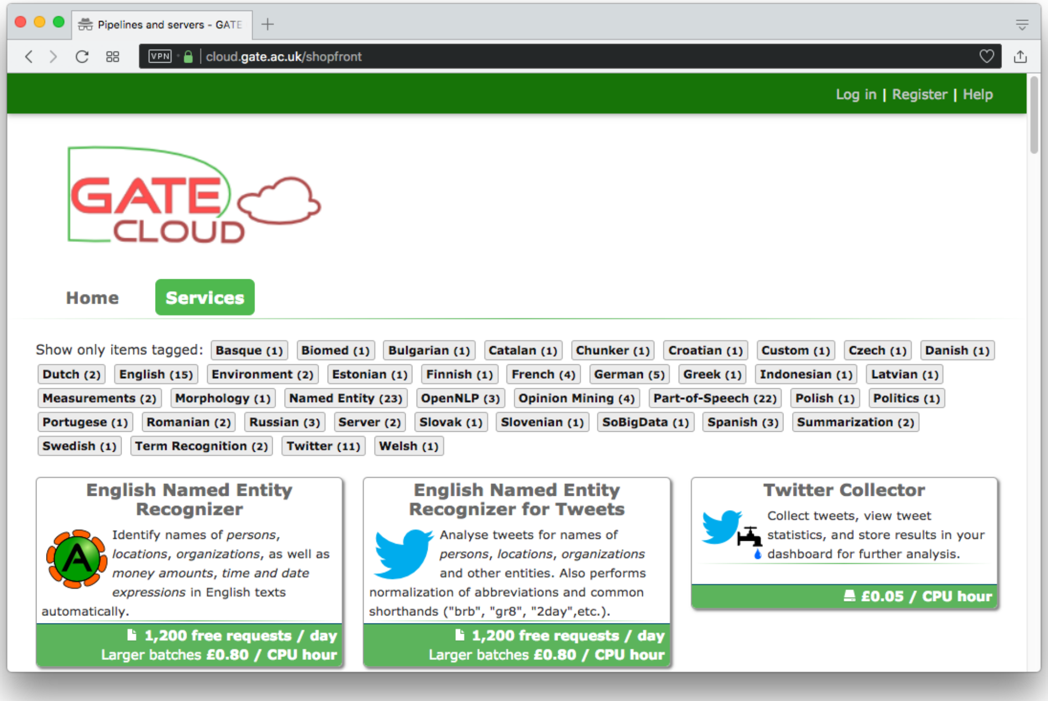


Welcome!

Welcome to GATE Cloud – the platform for end-to-end text processing solutions using the GATE family of tools running on the cloud. Process your documents using one of our pre-packaged annotation services, or upload your own. Save the results in various formats, or run

Dedicated servers

- Twitter collector is provided as a *dedicated server* – you rent a dedicated server for your private use
- Start and stop it as required
- Pay only for the hours it is running (though typically you would leave it running continuously)
- Backup and restore facility available



Show only items tagged: Basque (1) Biomed (1) Bulgarian (1) Catalan (1) Chunker (1) Croatian (1) Custom (1) Czech (1) Danish (1) Dutch (2) English (15) Environment (2) Estonian (1) Finnish (1) French (4) German (5) Greek (1) Indonesian (1) Latvian (1) Measurements (2) Morphology (1) Named Entity (23) OpenNLP (3) Opinion Mining (4) Part-of-Speech (22) Polish (1) Politics (1) Portugese (1) Romanian (2) Russian (3) Server (2) Slovak (1) Slovenian (1) SoBigData (1) Spanish (3) Summarization (2) Swedish (1) Term Recognition (2) Twitter (11) Welsh (1)

English Named Entity Recognizer

Identify names of *persons*, *locations*, *organizations*, as well as *money amounts*, *time and date expressions* in English texts automatically.

1,200 free requests / day
Larger batches **£0.80 / CPU hour**

English Named Entity Recognizer for Tweets

Analyse tweets for names of *persons*, *locations*, *organizations* and other entities. Also performs normalization of abbreviations and common shorthands ("brb", "gr8", "2day",etc.).

1,200 free requests / day
Larger batches **£0.80 / CPU hour**

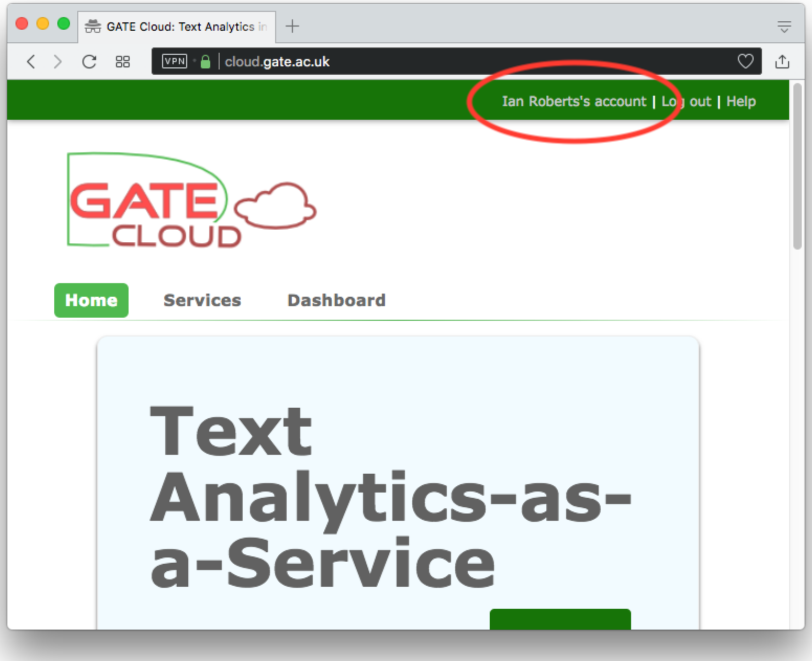
Twitter Collector

Collect tweets, view tweet statistics, and store results in your dashboard for further analysis.

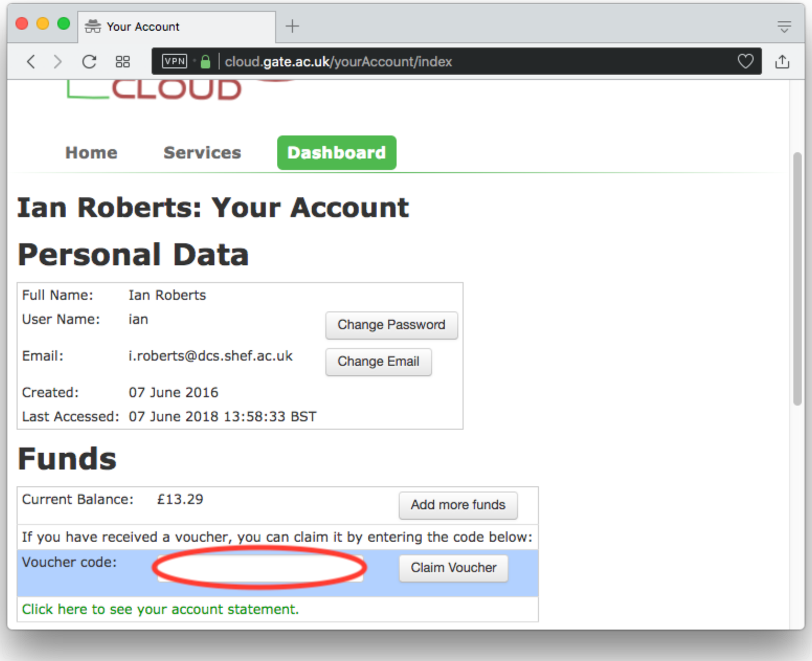
£0.05 / CPU hour

Reserving a server

- The usual e-commerce experience
  - Sign up for an account
  - Buy a top-up voucher (or use the free one we just gave you)
  - Find the server you want in the shop
  - Press “reserve this machine” and follow the instructions
- Server appears in your *dashboard*
- Behind the scenes, creates a persistent data *volume* for your data


GATE Cloud: Text Analytics in
cloud.gate.ac.uk
Ian Roberts's account | Log out | Help
GATE CLOUD
Home
Services
Dashboard

Text Analytics-as-a-Service



Ian Roberts: Your Account

Personal Data

Full Name:	Ian Roberts
User Name:	ian
Email:	i.roberts@dcs.shef.ac.uk
Created:	07 June 2016
Last Accessed:	07 June 2018 13:58:33 BST

Funds

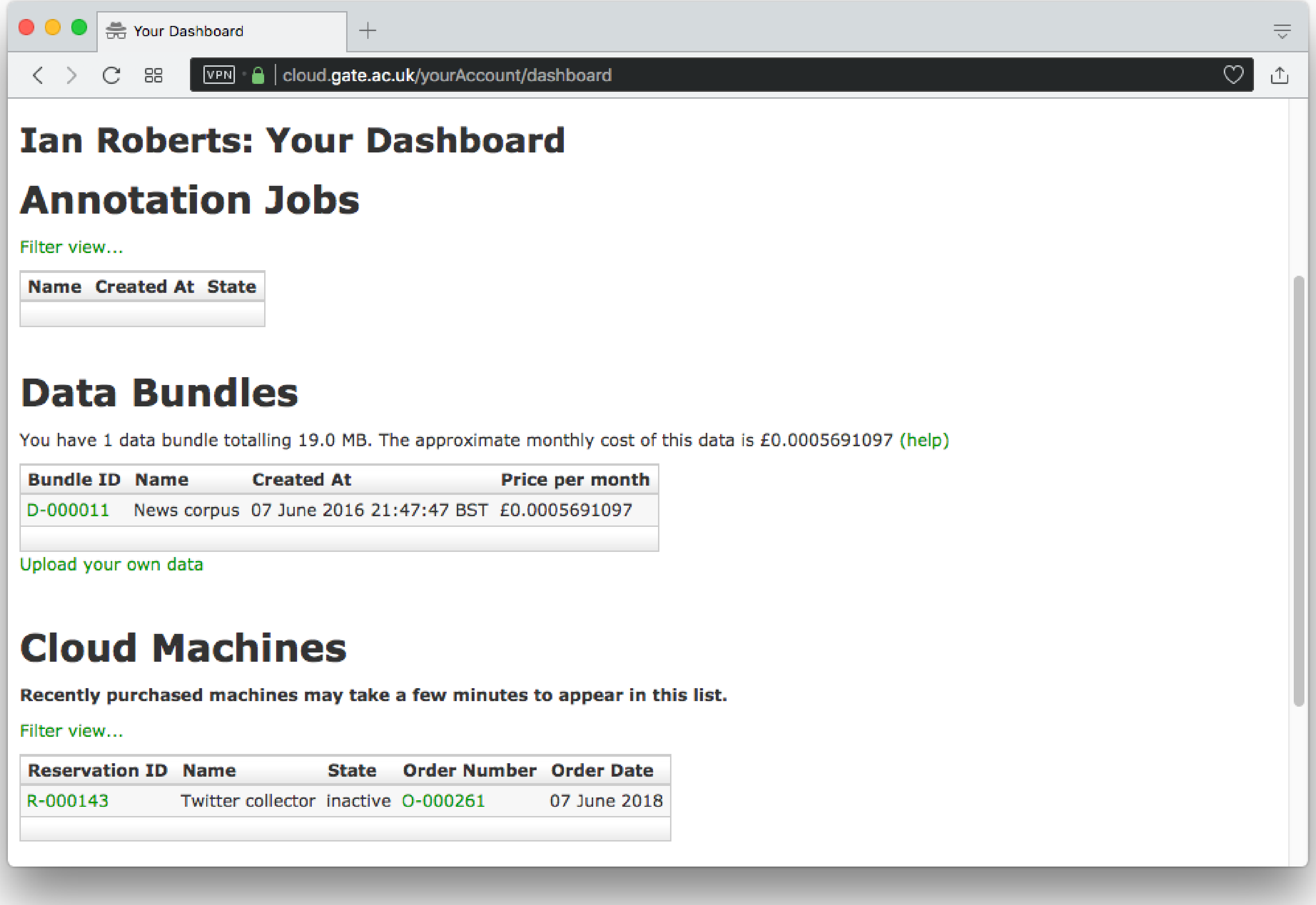
Current Balance: £13.29

If you have received a voucher, you can claim it by entering the code below:

Voucher code:

Click here to see your account statement.

Dashboard



Ian Roberts: Your Dashboard

Annotation Jobs

Filter view...

Name	Created At	State

Data Bundles

You have 1 data bundle totalling 19.0 MB. The approximate monthly cost of this data is £0.0005691097 (help)

Bundle ID	Name	Created At	Price per month
D-000011	News corpus	07 June 2016 21:47:47 BST	£0.0005691097

Upload your own data

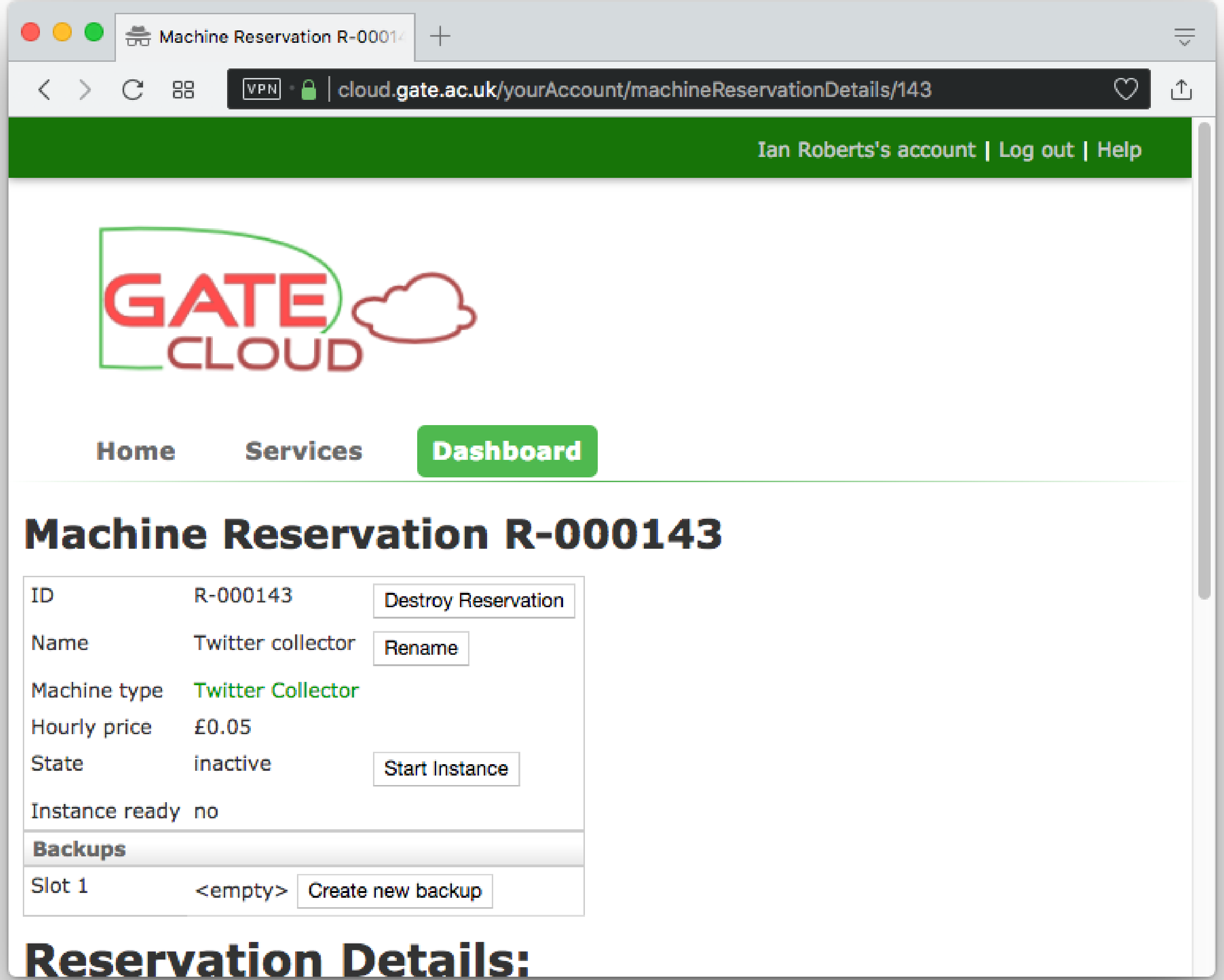
Cloud Machines

Recently purchased machines may take a few minutes to appear in this list.

Filter view...

Reservation ID	Name	State	Order Number	Order Date
R-000143	Twitter collector	inactive	O-000261	07 June 2018

Reservation control panel


Machine Reservation R-0001
cloud.gate.ac.uk/yourAccount/machineReservationDetails/143
Ian Roberts's account | Log out | Help
GATE CLOUD
Home
Services
Dashboard

Machine Reservation R-000143

ID	R-000143	Destroy Reservation
Name	Twitter collector	Rename
Machine type	Twitter Collector	
Hourly price	£0.05	
State	inactive	Start Instance
Instance ready	no	
Backups		
Slot 1	<empty>	Create new backup

Reservation Details:

Controlling the server

- Start and stop instance
  - Startup/shutdown takes a few minutes – system will email you when server is ready
  - You pay the hourly price whenever the instance is running

- Backup and restore
  - Save the state of your data volume so you can roll back later

- Destroy reservation
  - If you no longer need the server, destroy it to discard the data volume and all backups
  - *This cannot be undone*

Hands-on: GateCloud Collector

Start a Twitter collector

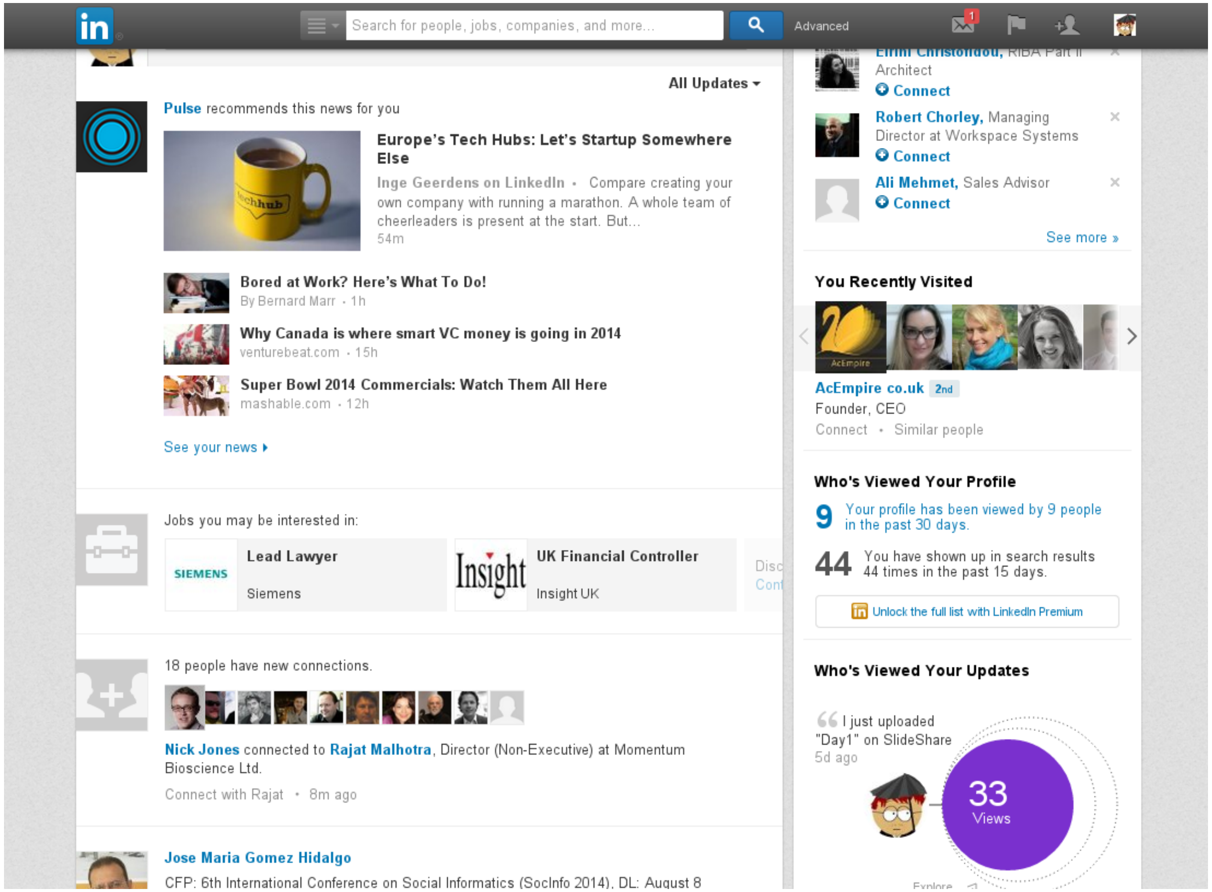
- Authenticate with your own Twitter account (create one if needed)
- Enter some search terms to track
- Check summary for search terms
- Try downloading tweet archive

2. LinkedIn

- Opened in 2003 as a professional networking portal
- Focus is on a CV-like profile
- Allows connection to your contacts
- Allows subscription and posting to forum-like groups
- Event-focused rather than message focused
- Posts can come from PC, mobile, SMS, iPad etc
- 260 million registered users

2. LinkedIn

- Feed-based output; information on new relations
- Focus on building networks: contact suggestions, contact history, people interested in you



All Updates

Pulse recommends this news for you

Europe's Tech Hubs: Let's Startup Somewhere Else
Inge Geerdens on LinkedIn • Compare creating your own company with running a marathon. A whole team of cheerleaders is present at the start. But...
54m

Bored at Work? Here's What To Do!
By Bernard Marr · 1h

Why Canada is where smart VC money is going in 2014
venturebeat.com · 15h

Super Bowl 2014 Commercials: Watch Them All Here
mashable.com · 12h

See your news ▸

Jobs you may be interested in:

SIEMENS	**Lead Lawyer** Siemens	Insight	**UK Financial Controller** Insight UK

Disc
Cont

18 people have new connections.

Nick Jones connected to **Rajat Malhotra**, Director (Non-Executive) at Momentum Bioscience Ltd.
Connect with Rajat • 8m ago

Jose Maria Gomez Hidalgo
CFP: 6th International Conference on Social Informatics (SocInfo 2014), DL: August 8

Eirini Christodoulou, RIBA Part II Architect
Connect

Robert Chorley, Managing Director at Workspace Systems
Connect

Ali Mehmet, Sales Advisor
Connect

See more »

You Recently Visited

AcEmpire co.uk 2nd
Founder, CEO
Connect • Similar people

Who's Viewed Your Profile

9 Your profile has been viewed by 9 people in the past 30 days.

44 You have shown up in search results 44 times in the past 15 days.

Unlock the full list with LinkedIn Premium

Who's Viewed Your Updates

“ I just uploaded "Day1" on SlideShare
5d ago

33 Views

Explore

2. LinkedIn

Data is available via API

No storage of data permitted: “**No LinkedIn data can be stored**”

- Except member ID
- User data can be stored only given explicit permission from that user
- Rationale: “LinkedIn users own their data. They need to have control over it. They might want to change it, change the visibility rules, or even delete it.”

Cross-referencing data is not permitted (via e.g. other networks)

- Creates problems for storing and communicating graph information
- Analysis must be live, but processing is not instantaneous – so no snapshots

API access is query driven: entities, items in streams

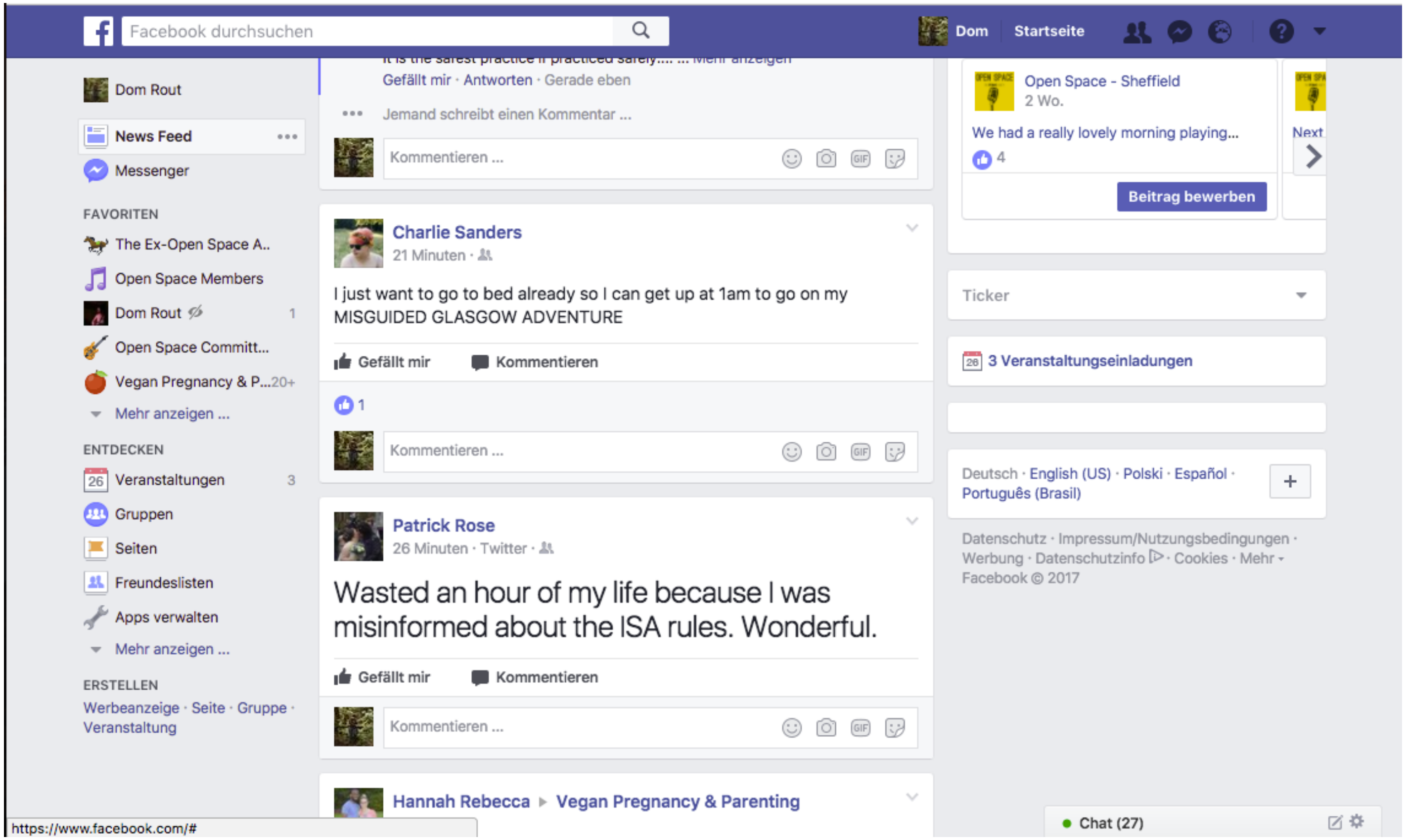
- Entities: people, stream, groups, mail, companies, job positions
- API is rate limited at application, user and developer level
- Limits quite high: e.g. 100k user profile queries per application per day

3. Facebook

- Opened in 2004 as a university student directory
- Communication is based on personal pages, to which messages are posted
- Allows connection to your contacts
- Allows subscription and posting to forum-like groups
- Message focused, with comments and voting systems (unidirectional)
- Posts can come from PC, mobile, SMS, iPad etc
- Millions of registered users
- Extensive privacy options for users

3. Facebook

- News items, with comments and likes
- Access network connections, events and private messaging



It is the safest practice if practiced safely.... ... Mehr anzeigen

Gefällt mir · Antworten · Gerade eben

Jemand schreibt einen Kommentar ...

Charlie Sanders
21 Minuten

I just want to go to bed already so I can get up at 1am to go on my MISGUIDED GLASGOW ADVENTURE

Patrick Rose
26 Minuten · Twitter

Wasted an hour of my life because I was misinformed about the ISA rules. Wonderful.

Hannah Rebecca ▸ **Vegan Pregnancy & Parenting**

Open Space - Sheffield
2 Wo.

We had a really lovely morning playing...

3. Facebook

Over the last few years Facebook has suffered from a number of privacy failures. As a result the access given to outside developers is now very limited.

Any application you develop has to be manually reviewed by Facebook and even then it can only access public pages, and you can’t store any information

Storing social media data

What would help us do our science?

- NLP and network analysis tools often data-driven, preferring “as much data as possible”
- Not only do the messages change over time – meta-information also
- A minimum: something that helps others reproduce your work
- Abstract annotations over the raw data != the raw data

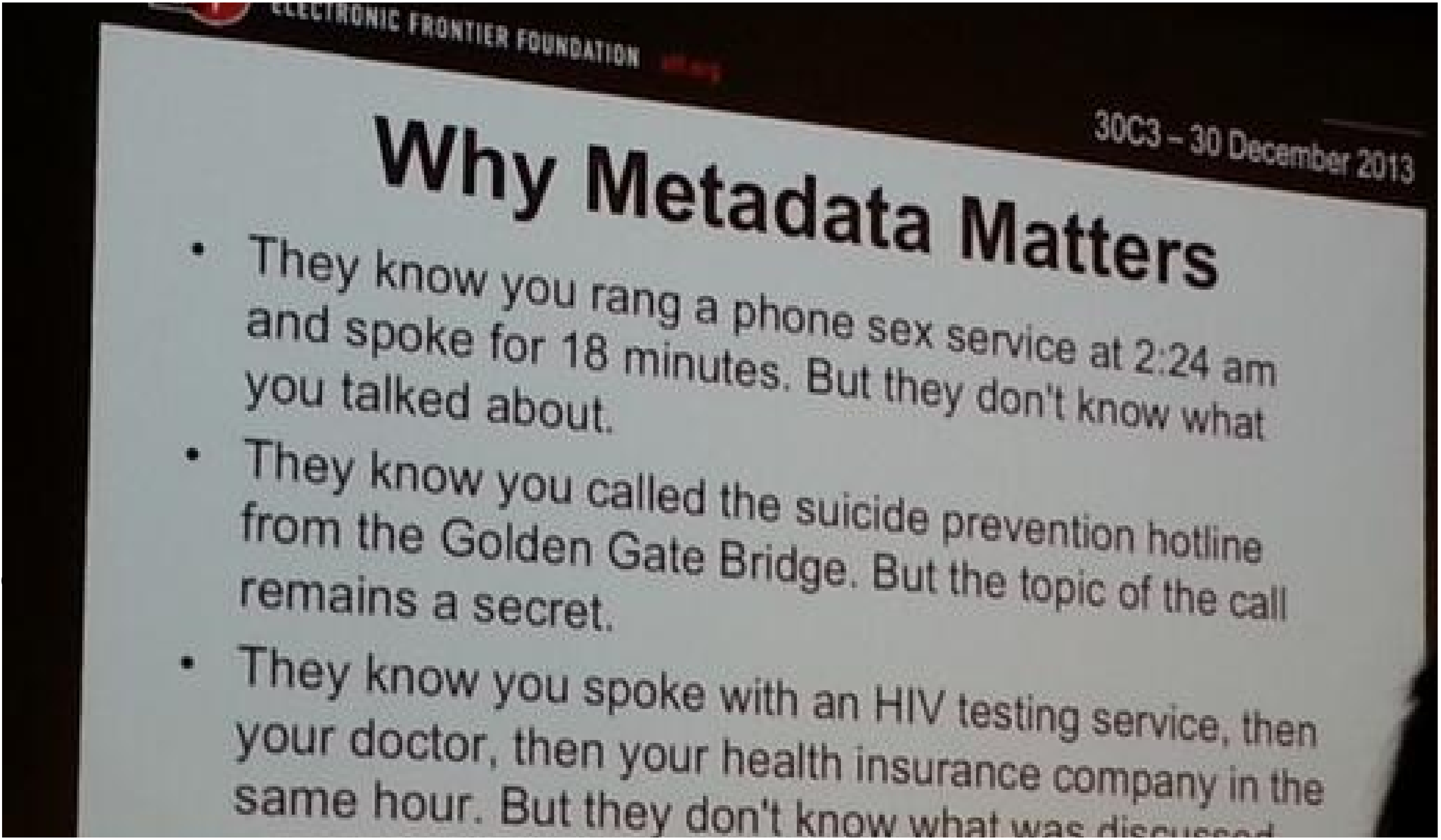
What native data can we safely store?

- LinkedIn: Object IDs only
- Twitter: IDs and the freshest seen API call result
- Facebook: Nothing

Ethical considerations

- We all have something to hide (e.g. from identity thieves)
- Important that personal data cannot proliferate once its owner removes / changes it
- How long to retain for? NSA's minimum 15-year seems excessive
-
- **Metadata just as powerful as text data**
- **Text data weaker without metadata**

Storing social media data



ELECTRONIC FRONTIER FOUNDATION

30C3 – 30 December 2013

Why Metadata Matters

- They know you rang a phone sex service at 2:24 am and spoke for 18 minutes. But they don't know what you talked about.
- They know you called the suicide prevention hotline from the Golden Gate Bridge. But the topic of the call remains a secret.
- They know you spoke with an HIV testing service, then your doctor, then your health insurance company in the same hour. But they don't know what was discussed

(fr

Social media corpora

Distribution concerns

- Social media corpora are difficult to distribute
- E.g. Twitter does not allow you to give other researchers / companies / anyone tweets you have collected and annotated in bulk
- Instead, distribute the tweet IDs and stand-off markup for the linguistic gold data
- The recipient re-collects all tweets himself, based on the IDs
- Necessary so user-deleted tweets are not propagated – privacy
-
- LinkedIn has even more stringent data sharing policy
- Facebook you can’t store so nothing to distribute

Social media corpora

Corpus completeness

- However, in some cases (e.g. misinformation, smear tweets) messages can be deleted
- Makes re-creating the corpus problematic
- Two classes of deletion:
    - Rapid deletions, usually within first few minutes (e.g. of spam, for editing the text)
    - Slower deletions (Petrovic et al. 2013)
- Our experience is that about 1 in 5 tweets are no longer available a year later.

Increased topic and entity drift: broader range of entities (Eisenstein 2013)

- Corpora age rapidly, and become less useful for some purposes (e.g. NEL)

Hands-on: Loading twitter data

- Open corpora/plain-tweets.json or your own corpus with a text viewer (such as notepad)
- Let's take a more useful view: find an online JSON viewer, and paste one line in. (e.g. “http://jsonviewer.stack.hu”)
- Note the hierarchical structure of the data, and embedded user profile
- Now, let's load some data into GATE. First, load the Twitter plugin and the Format: JSON plugin
- Create a new GATE corpus called “Raw tweets” and save to DS
- Right-click on the corpus and choose “Populate from JSON”
- Select the JSON file used earlier, and make sure the mime type is set to “text/x-json-twitter”
- Examine the different annotations in the document