

# 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 ...](http://arstechnica.com/tech-policy/20...) by @nathanmattise

## Social uses

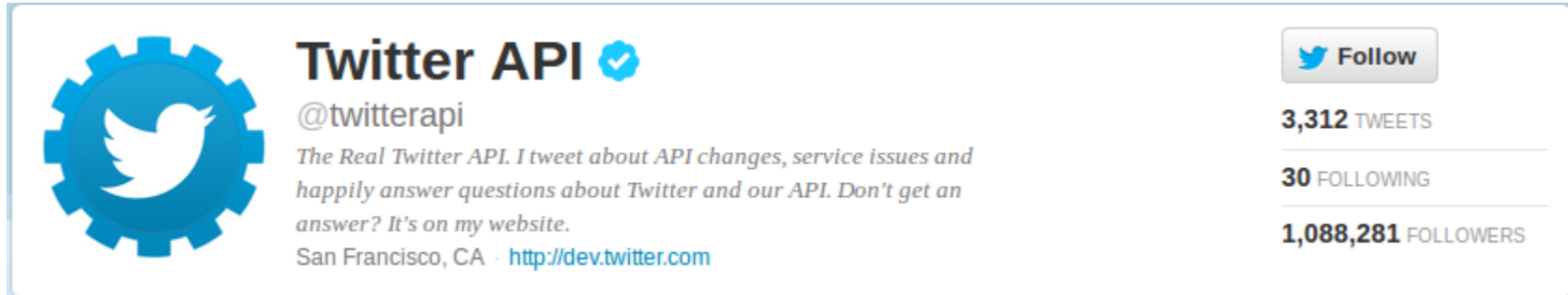
### 「ジャム」 Jam Gregory



@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



# Twitter User Profiles


A screenshot of a Twitter profile for the user @twitterapi. The profile includes a blue gear-shaped profile picture with a white bird icon inside. The name is "Twitter API" with a verified badge. The handle is "@twitterapi". The bio reads: "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." The location is "San Francisco, CA" and the website is "http://dev.twitter.com". On the right side, there is a "Follow" button, and statistics showing "3,312 TWEETS", "30 FOLLOWING", and "1,088,281 FOLLOWERS".

 **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>

 **Follow**

**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 #somanerrorsitsfunny",
  "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":"somanerrorsitsfunny","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



# 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

# 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)

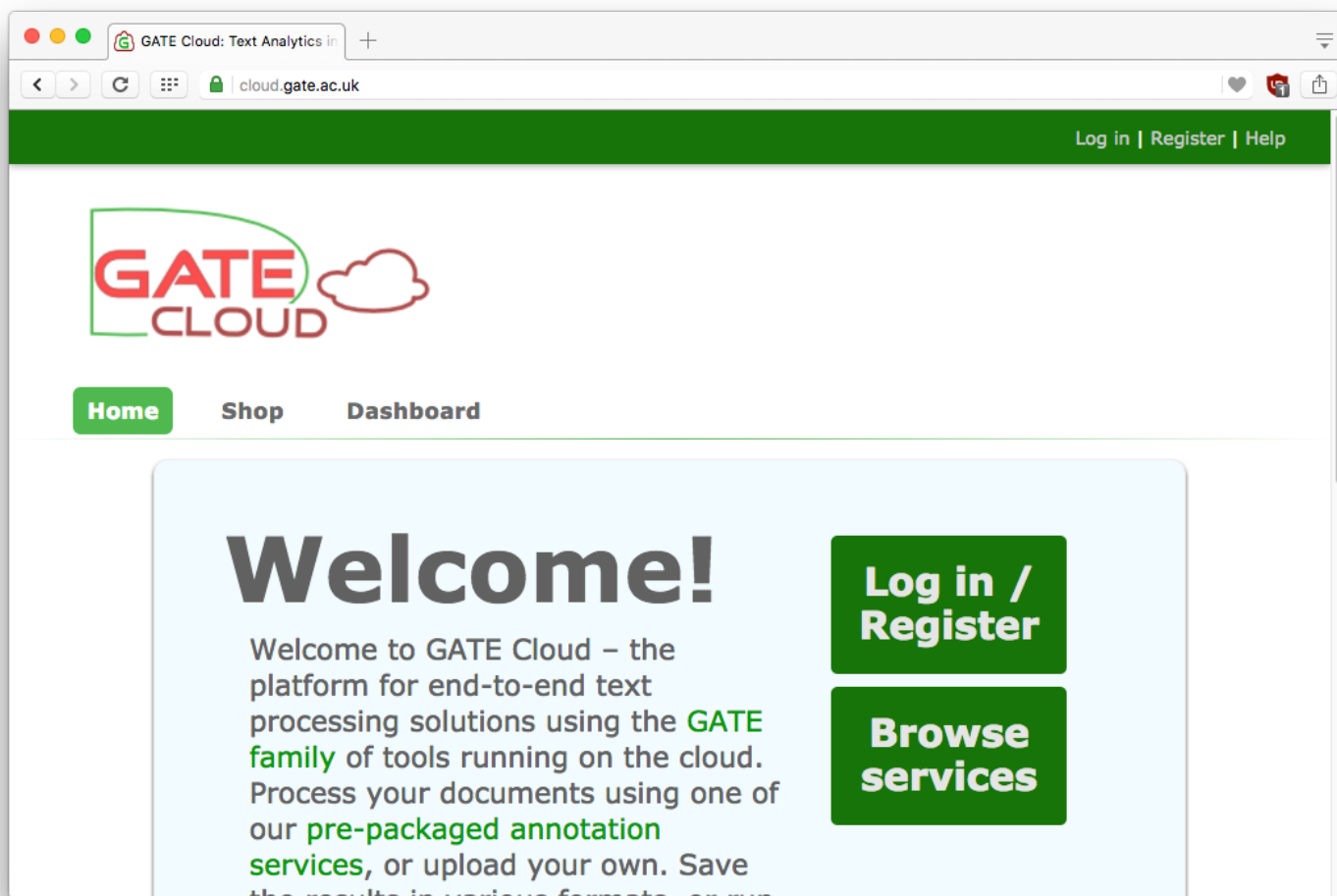
- First create an account for GateCloud
- Load some credit onto your account
- Order the service and wait for your reservation
- Start the machine and configure the collector!

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



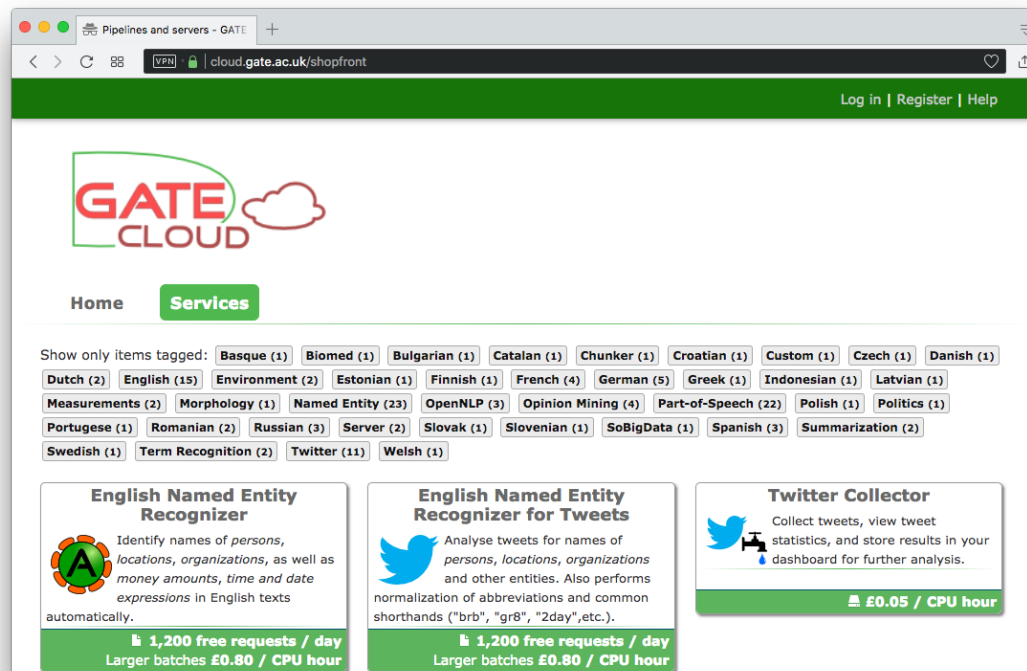
# GATE Cloud

<https://cloud.gate.ac.uk>



# 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

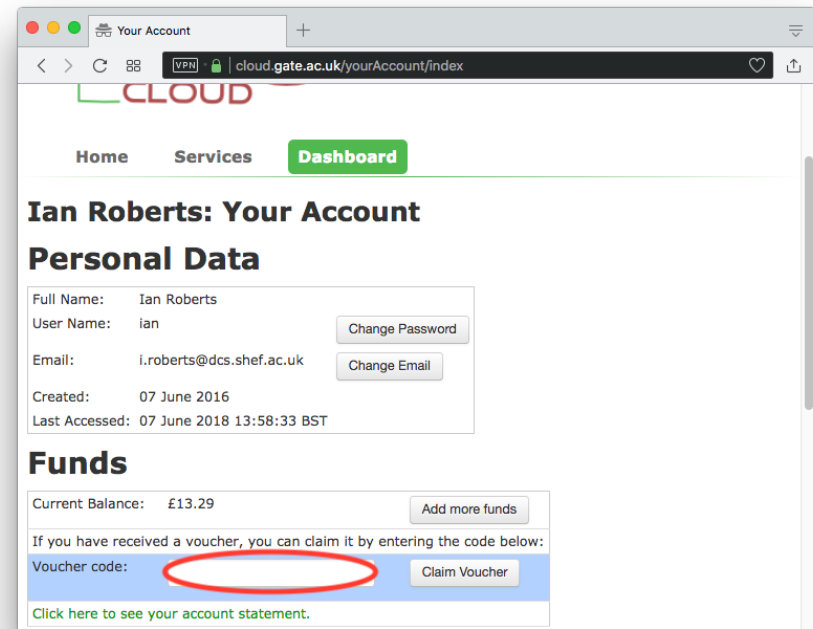
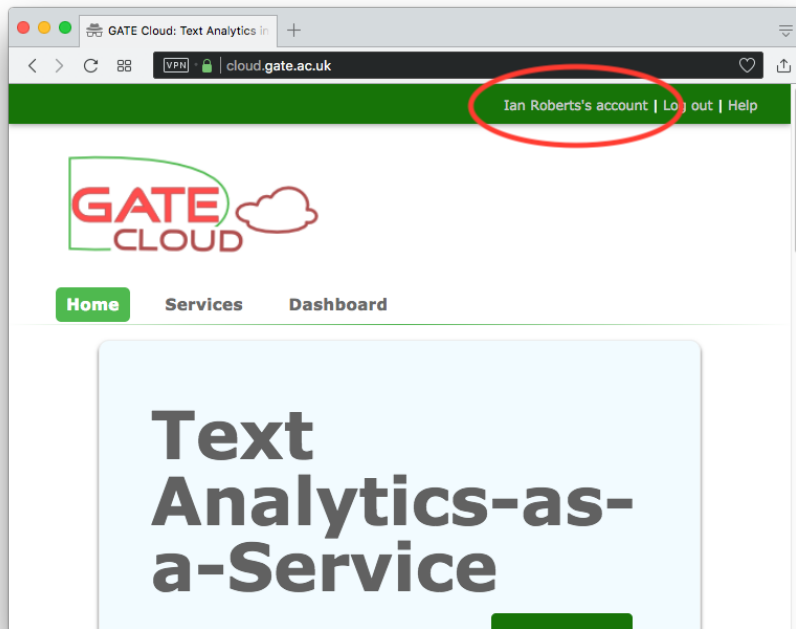


The screenshot shows the GATE Cloud website interface. At the top, there is a navigation bar with "Log in | Register | Help" links. Below the GATE Cloud logo, there are tabs for "Home" and "Services". A filter bar shows various tags for different languages and tasks, such as "English (15)", "Named Entity (23)", and "Twitter (11)". Three service cards are displayed:

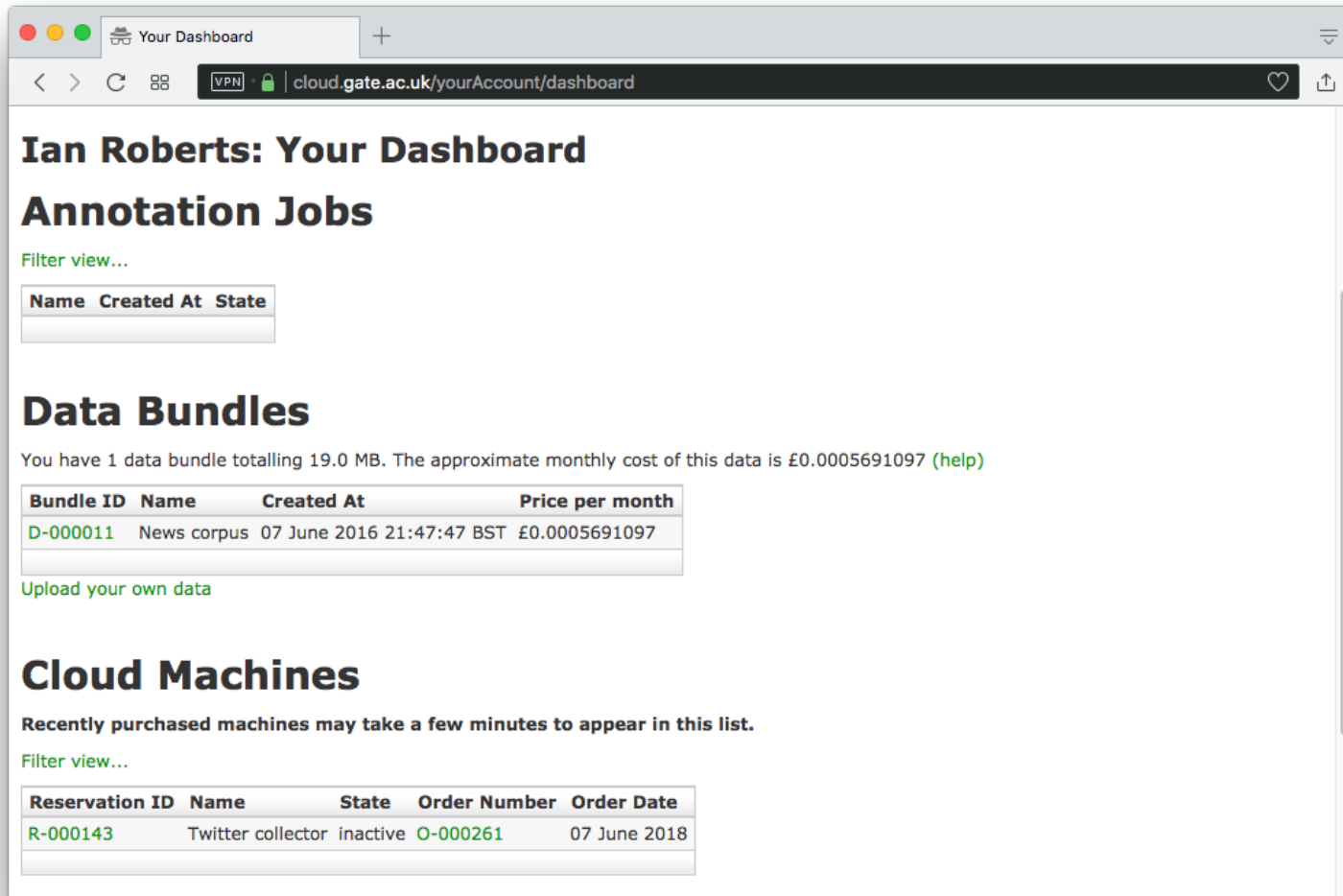
- English Named Entity Recognizer**: Identify names of persons, locations, organizations, as well as money amounts, time and date expressions in English texts automatically. **1,200 free requests / day**, Larger batches **£0.80 / CPU hour**.
- English Named Entity Recognizer for Tweets**: Analyse tweets 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



# Dashboard

A screenshot of a web browser displaying the GATE dashboard for user Ian Roberts. The browser's address bar shows the URL "cloud.gate.ac.uk/yourAccount/dashboard". The dashboard is titled "Ian Roberts: Your Dashboard" and features three main sections: "Annotation Jobs", "Data Bundles", and "Cloud Machines". Each section includes a "Filter view..." link and a table of data. The "Data Bundles" section shows one bundle with ID "D-000011", name "News corpus", and a price of £0.0005691097. The "Cloud Machines" section shows one reservation with ID "R-000143", name "Twitter collector", and state "inactive".

**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

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/yourAccount/machineReservationDetails/143`. The page header identifies the user as Ian Roberts and includes links for 'Log out' and 'Help'. The main content area features the GATE Cloud logo and a navigation menu with 'Home', 'Services', and 'Dashboard' (the active page). The main heading is 'Machine Reservation R-000143'. Below this, a table-like structure displays reservation details:

ID	R-000143	<a href="#">Destroy Reservation</a>
Name	Twitter collector	<a href="#">Rename</a>
Machine type	Twitter Collector	
Hourly price	£0.05	
State	inactive	<a href="#">Start Instance</a>
Instance ready	no	

Below the details table is a section for 'Backups':

Backups	
Slot 1	<empty> <a href="#">Create new backup</a>

The page concludes with the heading '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

The screenshot displays a LinkedIn profile page with the following sections:

- Header:** Search bar, navigation icons, and user profile picture.
- News Feed:**
  - Pulse recommendation:** "Europe's Tech Hubs: Let's Startup Somewhere Else" by Inge Geerdens on LinkedIn, 54m ago. Includes an image of a yellow mug.
  - Article:** "Bored at Work? Here's What To Do!" by Bernard Marr, 1h.
  - Article:** "Why Canada is where smart VC money is going in 2014" from venturebeat.com, 15h.
  - Article:** "Super Bowl 2014 Commercials: Watch Them All Here" from mashable.com, 12h.
- Jobs you may be interested in:**
  - Lead Lawyer** at Siemens.
  - UK Financial Controller** at Insight UK.
- Connections:** "18 people have new connections." Includes a row of profile pictures and a notification for Nick Jones connecting to Rajat Malhotra at Momentum Bioscience Ltd. 8m ago.
- Activity:** "Jose Maria Gomez Hidalgo" posted "CFP: 6th International Conference on Social Informatics (SocInfo 2014). DL: August 8".
- Right Sidebar:**
  - Connect:** Suggestions for Emili Christofidou (Architect), Robert Chorley (Director at Workspace Systems), and Ali Mehmet (Sales Advisor).
  - You Recently Visited:** AcEmpire co.uk (2nd), Founder, CEO.
  - Who's Viewed Your Profile:** 9 profile views in the past 30 days; 44 search result views in the past 15 days.
  - Who's Viewed Your Updates:** A post titled "Day1" on SlideShare (5d ago) has 33 views.

## 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

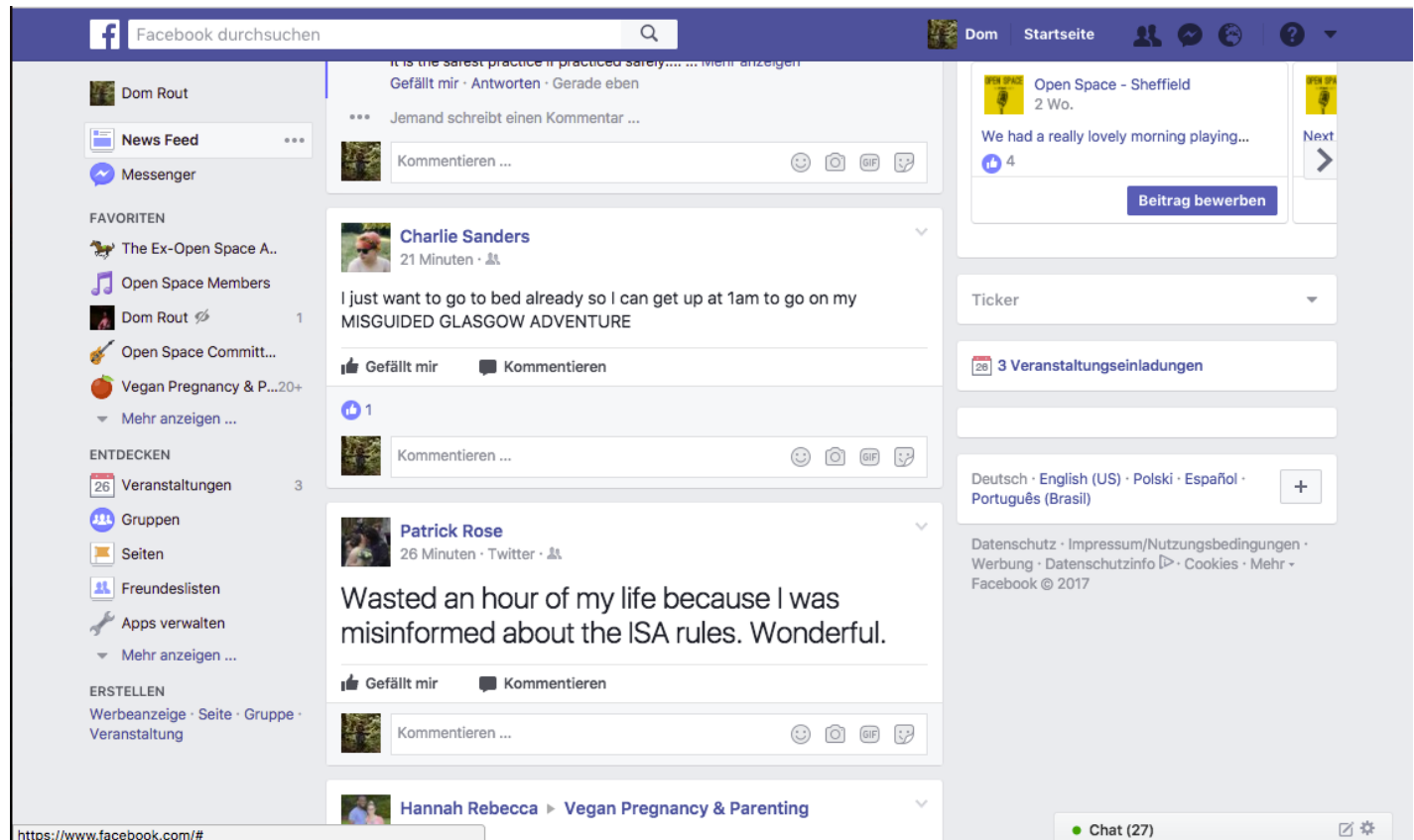
- Entities: people, stream, groups, mail, companies, job positions
- API is rate limited at application, user and developer level
- Limits quite high: e.g. 100k user profile queries per application per day

## 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



The screenshot displays the Facebook news feed interface. On the left, a navigation sidebar includes sections for 'FAVORITEN' (The Ex-Open Space A., Open Space Members, Dom Rout, Open Space Commit..., Vegan Pregnancy & P...20+), 'ENTDECKEN' (Veranstaltungen, Gruppen, Seiten, Freundeslisten, Apps verwalten), and 'ERSTELLEN' (Werbeanzeige, Seite, Gruppe, Veranstaltung). The main feed shows posts from Dom Rout, Charlie Sanders (commenting on a post about 'MISGUIDED GLASGOW ADVENTURE'), Patrick Rose (commenting 'Wasted an hour of my life because I was misinformed about the ISA rules. Wonderful.'), and Hannah Rebecca (commenting on a post from Vegan Pregnancy & Parenting). The right sidebar features a 'Beitrag bewerben' button, a 'Ticker', '3 Veranstaltungseinladungen', and language options (Deutsch, English (US), Polski, Español, Português (Brasil)). A footer at the bottom right shows a 'Chat (27)' notification.

# 3. Facebook

Main APIs for facebook data access: Graph, Public Feed (also others for web hosting, ads)

REST and JSON-based

- GET graph.facebook.com /{node-id}
- GET graph.facebook.com /{node-id}/{edge-name}
- Also POST, DELETE
- Example response; fields vary depending on entity type
- 
- 
- 

```
{
  "id": "4",
  "link": "https://www.facebook.com/zuck",
  "gender": "male",
  "username": "zuck",
  "picture": {
    "data": {
      "url": "https://fbcdn-profile-a.akamaihd.net/hprofile-ak-prn2/202896_4_1782288297_q.jpg",
      "is_silhouette": false
    }
  }
}
```

Many di

Optiona

One extra

API: Keyword insights

s, videos..)

- Access to demographic information given keywords, locations



# 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

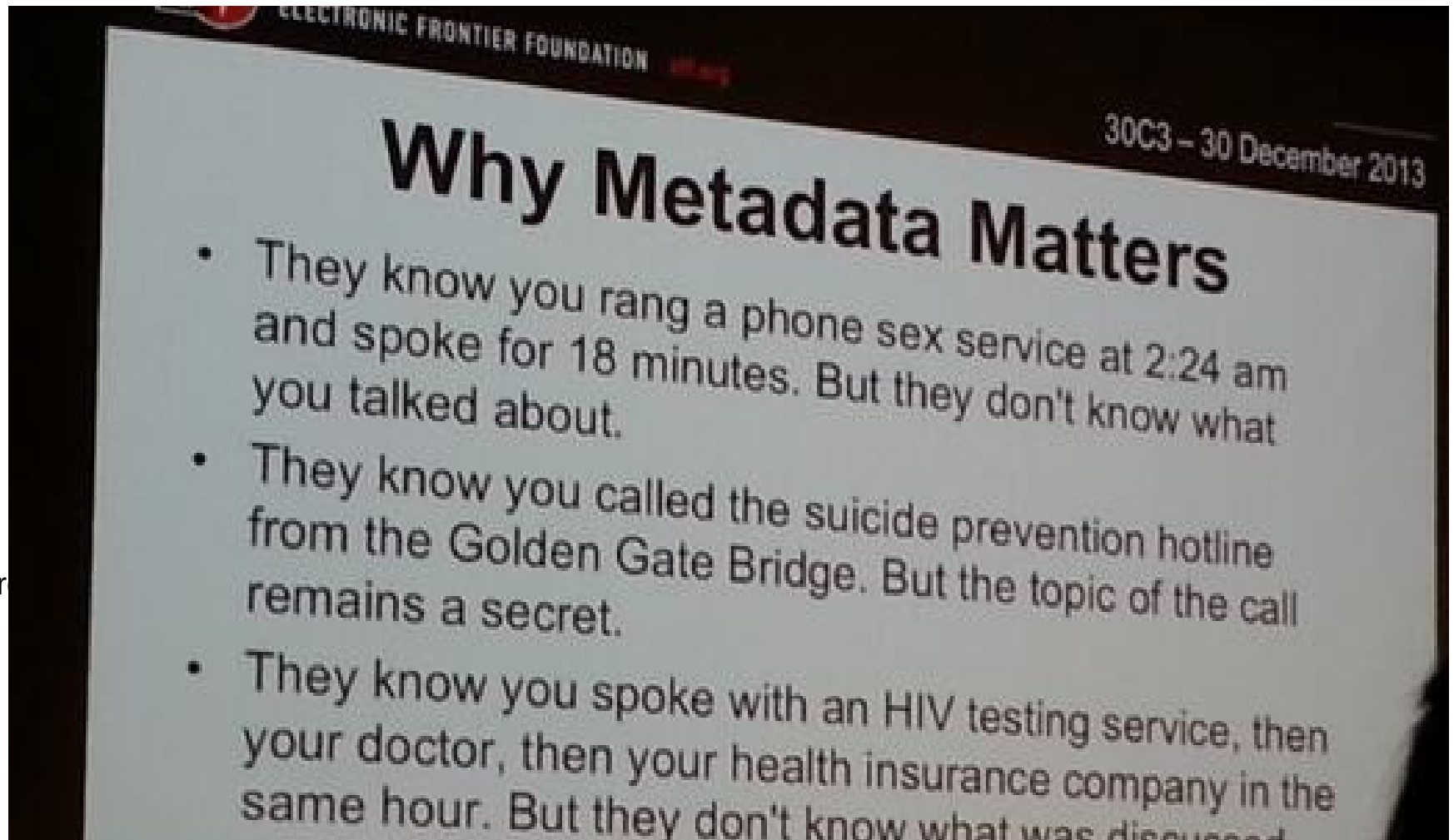
## What native data can we safely store?

- LinkedIn: Object IDs only
- Twitter: IDs and the freshest seen API call result
- Facebook: Anything that the user has given us access to

## 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

A photograph of a presentation slide. The slide is white with black text and is tilted slightly to the right. At the top left, the text "ELECTRONIC FRONTIER FOUNDATION" is visible. At the top right, the text "30C3 - 30 December 2013" is visible. The main title of the slide is "Why Metadata Matters". Below the title, there is a bulleted list of three points. The first point discusses a phone call to a sex service. The second point discusses a call to a suicide prevention hotline from the Golden Gate Bridge. The third point discusses a call to an HIV testing service and a doctor's office. The text is partially cut off at the bottom.

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 more relaxed, but data recipient must also have express permission from user

# 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
- Create a new GATE corpus called "Raw tweets" and save to DS
- Right-click on the corpus and choose "Populate from Twitter JSON files"
- See that you can choose which fields to import or ignore
- Select the JSON file used earlier, and make sure the "One document per tweet" box is checked, near the top
- Import with default fields for now
- Examine the different annotations in the document: text, username, date