GATE and Social Media: Gathering Social Media Data

Dom Rout
Kalina Bontcheva
Leon Derczynski (slides)

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 <140 characters

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

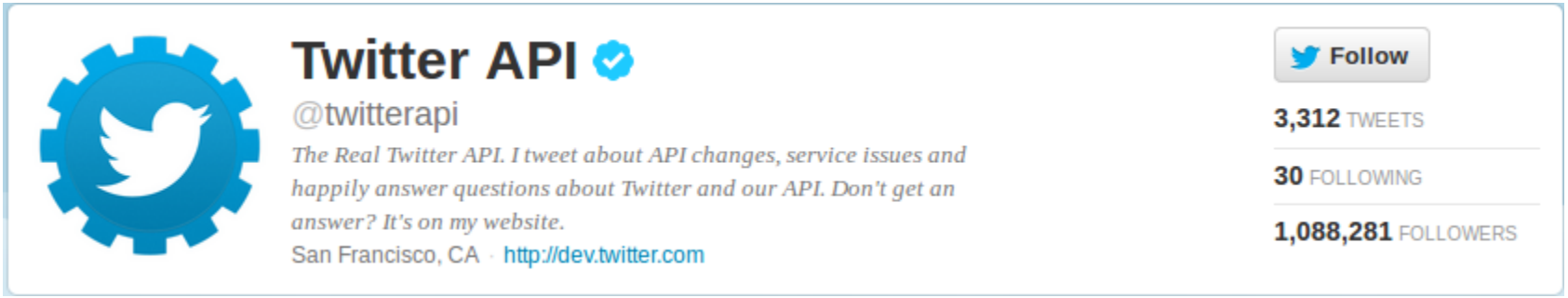
Social uses

「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

Twitter User Profiles



- 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":"9402919386363904O“,
"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 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

GateCloud 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 (£0.02 an hour, or about £14.60 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.

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

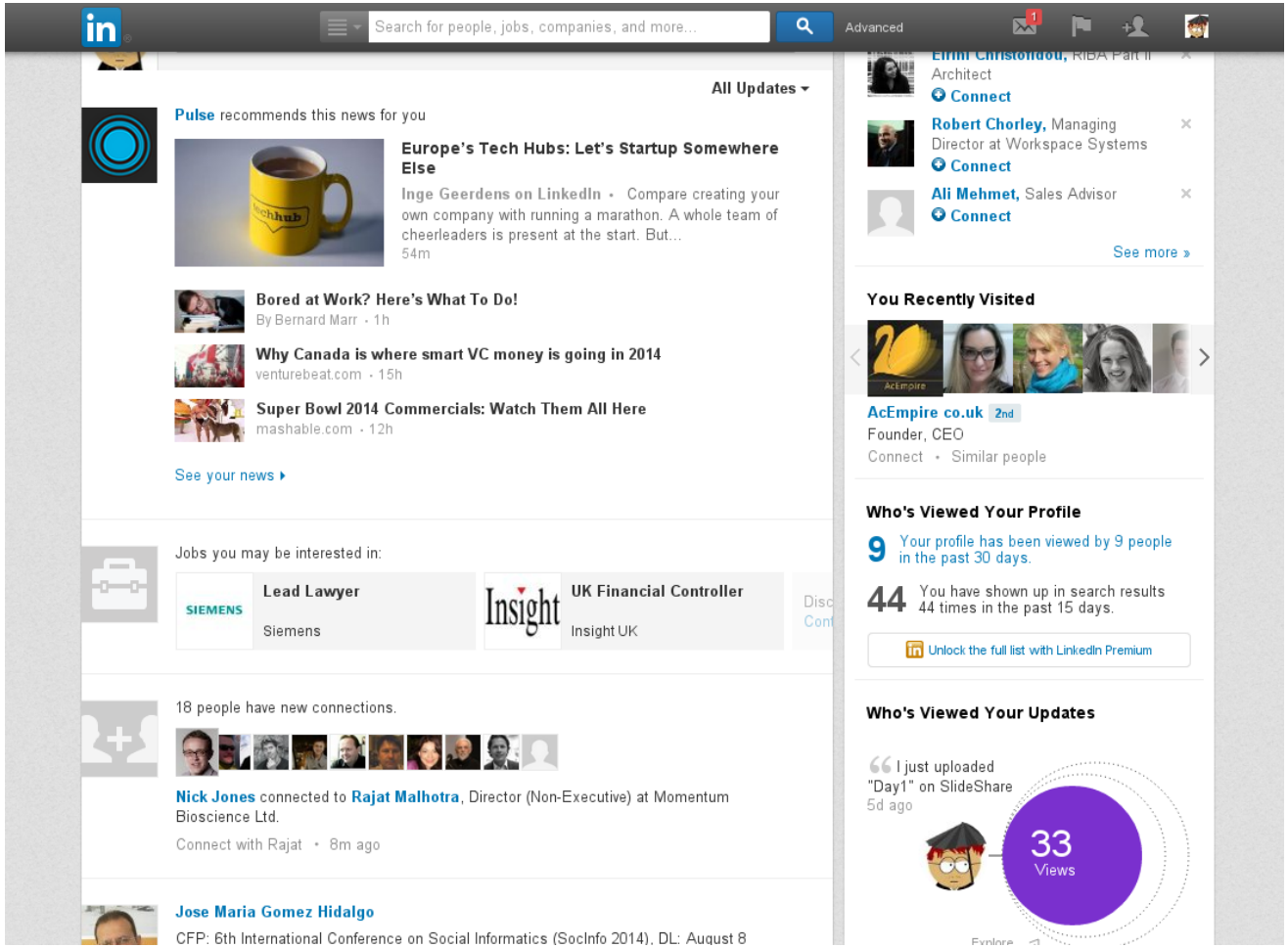
Posts can come from PC, mobile, SMS, iPad etc

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



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

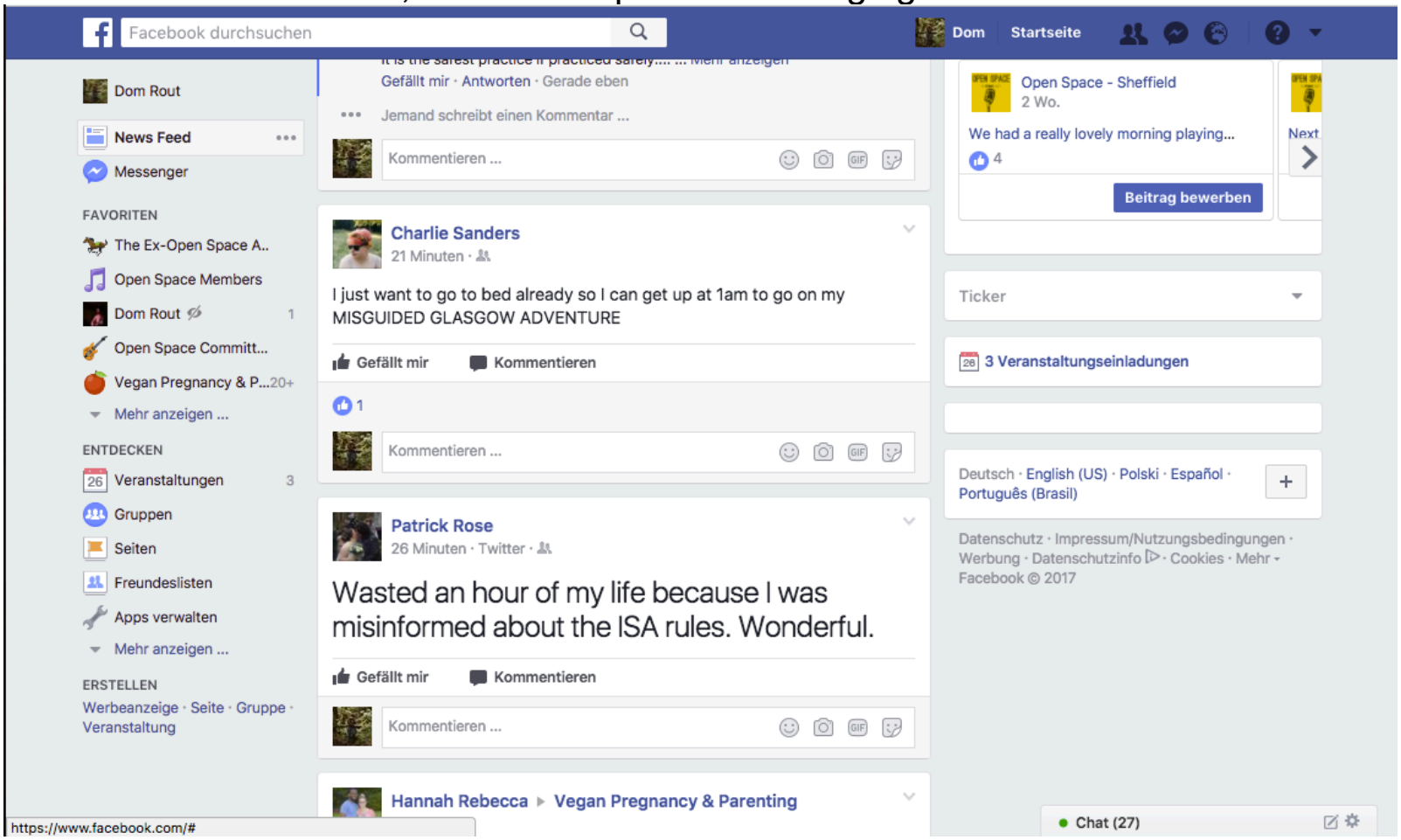
Millions of registered users

Extensive privacy options for users

3. Facebook

News items, with comments and likes

Access network connections, events and private messaging



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.jp
      "is_silhouette": false
    }
  }
}
```


Many different entity types (messages, links, photos, events, posts, payments, videos..)

Optional FQL access – Facebook Query Language

One extra API: Keyword Insights

- 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?

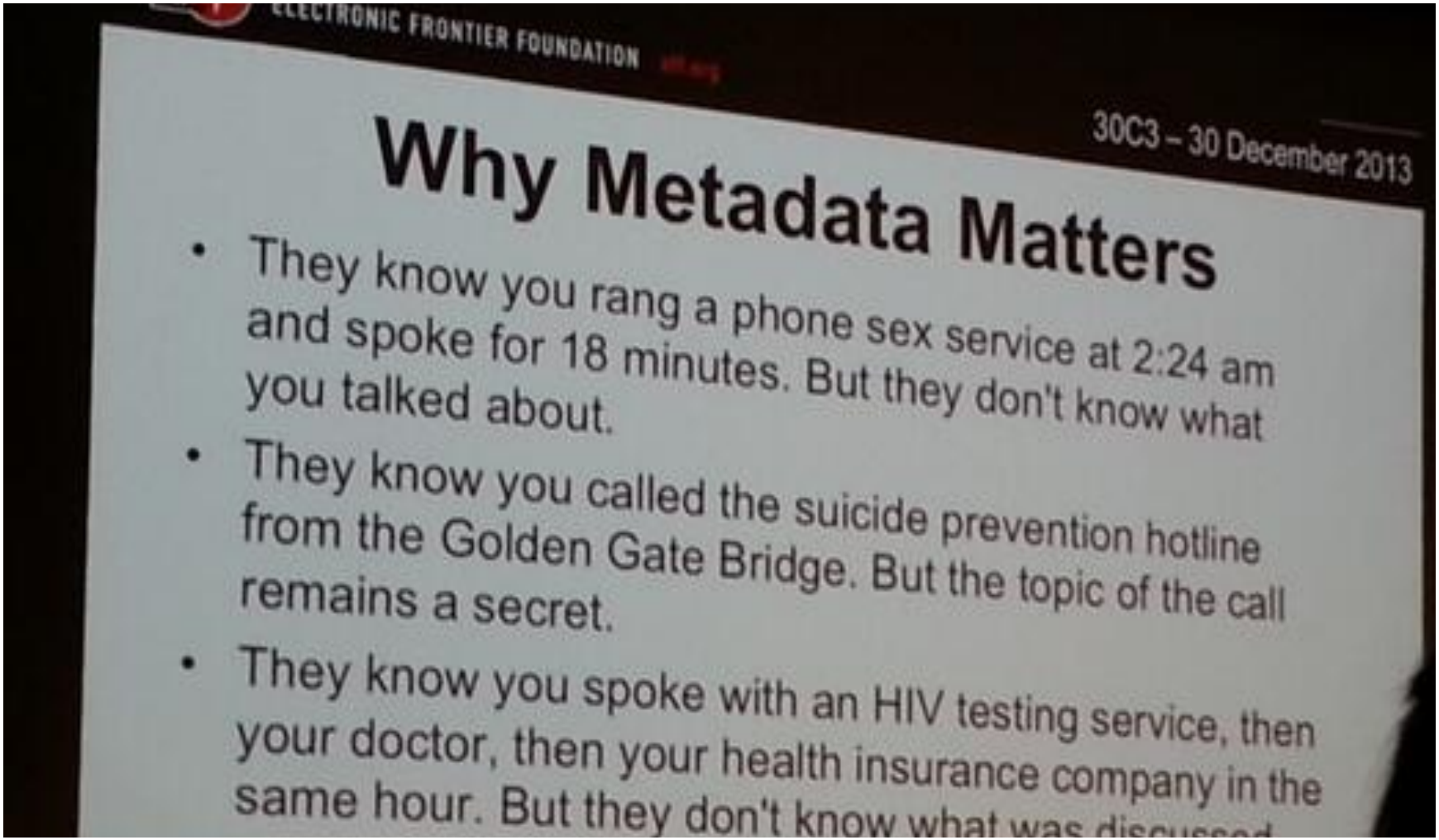
- LinkedIn: Object IDs only
- Twitter: IDs and the freshest seen API call result
- Facebook: Anything that the user has given us access to

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



(from Kurt Opshal's slides at the Chaos Communication Congress, photo by Marion Marschalek)

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 is 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 checker, near the top

Import with default fields for now

Examine the different annotations in the document: text, username, date