

CHALLENGES AND OPPORTUNITIES PRESENTED BY (BIG) DIGITAL DATA

Yin Yin Lu, Oxford Internet Institute

EAST GENESyS Fall Methods Workshop Training Day

Friday 13 October 2017



THE TRAINERS



- **Yin Yin Lu** is a doctoral candidate at the Oxford Internet Institute (OII) and Balliol College, Oxford.
- **Martin Wynne** is a digital research specialist in the Bodleian Libraries at the University of Oxford, and the National Coordinator for CLARIN UK.
- **Chico Camargo** is a research assistant at the OII and a doctoral candidate at the University of Oxford's Department of Physics.
- **Folger Karsdorp** is a post-doctoral researcher at the Meertens Institute of the Royal Netherlands Academy of Arts and Sciences.
- **Mike Thelwall** is a Professor of Information Science and leads the Statistical Cybermetrics Research Group at the University of Wolverhampton.

THE SCHEDULE (MORNING)

- **9-9.30am:** Challenges and Opportunities Presented by (Big) Digital Data [[Yin Yin Lu](#)]
- **9.35-10.10am:** CLARIN Tools and Resources for Digital Data and East Asian Languages [[Martin Wynne](#)]
- **10.15-10.45am:** Workshop preparation / troubleshooting session
- **10.45am-12pm:** Introduction to Python for Digital Text Analysis (Part I) [[Chico Camargo](#)]

THE SCHEDULE (AFTERNOON)

- 12-1pm: Much-needed lunch break ☺
- 1-2pm: Introduction to Python for Digital Text Analysis (Part II) [[Yin Yin Lu](#)]
- 2-3.15pm: Topic Modelling—An Empirical Approach to Theme Detection [[Folger Karsdorp](#)]
- 3.30-4.45pm: Sentiment Analysis—The Emotionality of Discourse [[Mike Thelwall](#)]
- 4.45-5pm: Closing reflections (we survived!)

THE GOAL(S)

- This is **not** a programming in Python workshop!
- Rather, the goal is to illuminate the **potential of Python for large-scale text analysis**, and to make you a) excited about the possibilities and b) cautious because of the limitations (and issues with internet data).
- Will teach a **variety of techniques**, so everyone should learn something relevant for their project.
- ***Disclaimer.*** if you are not familiar with the basics of Python, some parts of the workshops will be overwhelming. **Please do not give up!**

THE DATASET

- YouTube comments on [202 Korean pop \(Kpop\) music videos](#).
- **Four prominent groups:** BTS (58 videos), EXO (47), Twice (46), Black Pink (51)
- [Internationally viral](#). Fascinating contemporary pop phenomenon, especially in the context of East Asian youth culture.
- **Two sets of text files:** comments only, comments with metadata.
- (Most) non-English comments have been removed.
- **GitHub repository:** tinyurl.com/genesysnlp

MEASUREMENT

- Traditional instruments of social science: interviews, surveys, focus groups. **Indirect measurement** of underlying phenomena.
- Social media (and web data in general) allow for a **direct lens** on social behaviour and phenomena.
- Two high-level problems:
 - Representativeness
 - Online sample vs. offline population?
- Hybrid approach: **trace interviewing** (Dubois & Ford, 2015)

POINTS OF ACCESS TO WEB DATA

- **Application Programming Interfaces (APIs)**
 - **Structured** form of data collection. Need to know programming.
 - **Versatile**—you have full control over the parameters.
 - APIs allow web services to control who has access to their data and how much is distributed → severe **rate limits**.
- **Web scraping**
 - **Unstructured** form of data collection. Need to know programming.
 - Full control over parameters; no rate limits.

POINTS OF ACCESS TO WEB DATA

■ Collection tools

- Little or no programming knowledge is necessary.
- Less control over parameters and output format.
- Same rate limit constraint as APIs.
- Many have built-in analysis and visualisation tools.
- Many are free.

■ Preexisting public datasets

- Many are large and free.
- Least control over parameters.

TWITTER

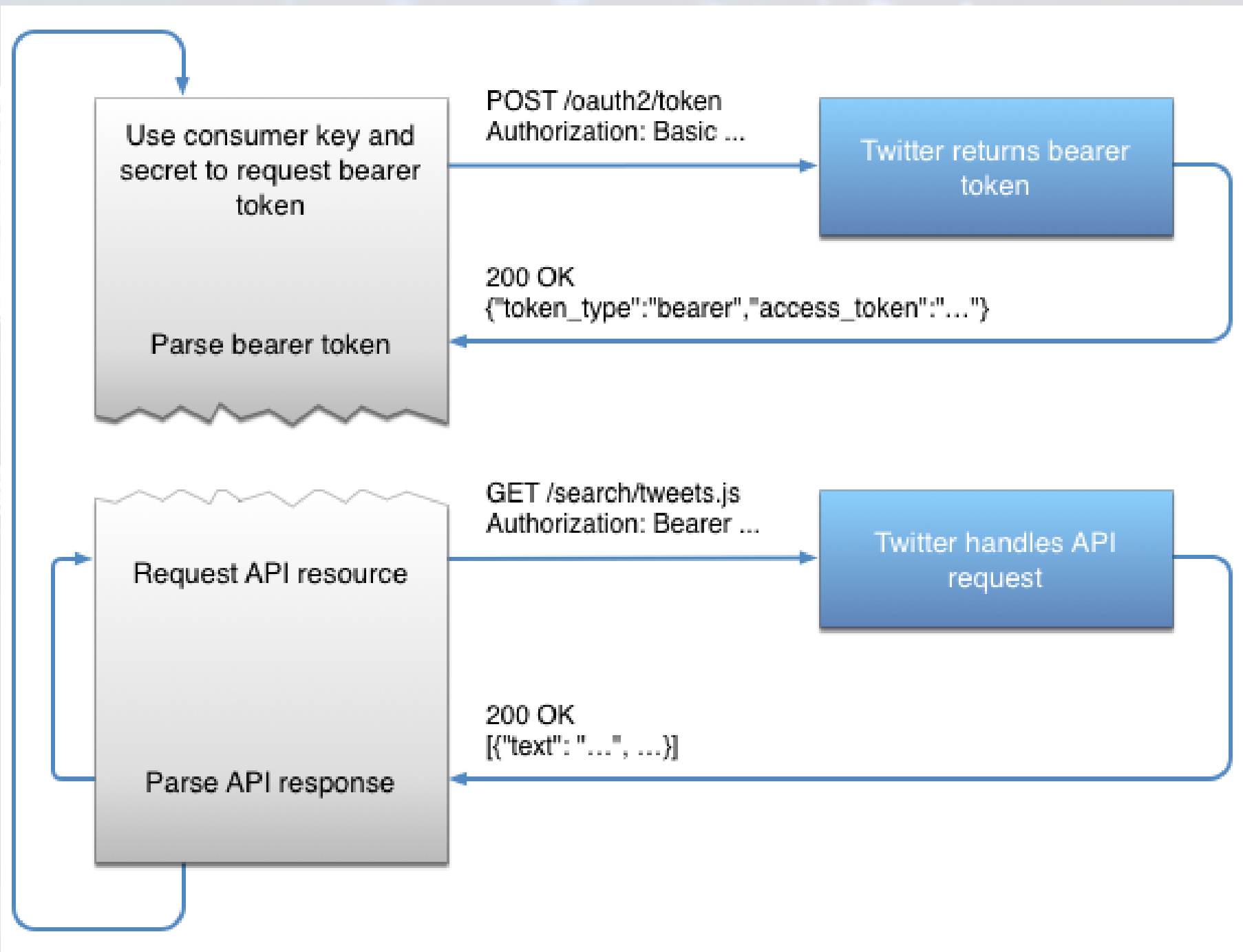
▪ Search/REST API

- Keyword or username queries of [historical tweets](#).
- Rate limit: 180 calls every 15 minutes or 18,000 tweets over the last 7-day period, whichever is reached first.
- ‘Focused on [relevance](#) and not [completeness](#).’

▪ Streaming API

- Track up to 400 keywords, 5,000 users, or 25 locations in [real time](#).
- Rate limit: 1% of total tweet volume (~500M/day → 5M).
- A more [complete](#) sample (if the query is not for popular keywords).
- ‘A small [random sample](#) of all [public statuses](#).’

▪ Documentation: developer.twitter.com/en/docs




```
{
  "statuses": [
    {
      "metadata": {
        "iso_language_code": "en",
        "result_type": "recent"
      },
      "created_at": "Sat Jan 31 13:58:42 +0000 2015",
      "id": 561523971448123392,
      "id_str": "561523971448123392",
      "text": "@Gerrarrdus you really have to fear for the state of Oxford CMD after reading that article in the Church Times.",
      "source": "\u003ca href=\"http://twitter.com/#!/download/ipad\" rel=\"nofollow\" \u003eTwitter for iPad\u003c/a\u003e",
      "truncated": false,
      "in_reply_to_status_id": 561521412973035520,
      "in_reply_to_status_id_str": "561521412973035520",
      "in_reply_to_user_id": 206849269,
      "in_reply_to_user_id_str": "206849269",
      "in_reply_to_screen_name": "Gerrarrdus",
      "user": {
        "id": 21115092,
        "id_str": "21115092",
        "name": "Phil Ritchie \u00646",
        "screen_name": "philritchie",
        "location": "Essex",
        "profile_location": null,
        "description": "dad, husband, revd canon, team rector great baddow team ministry, drummer, chicken keeper & man utd fan",
        "url": "http://t.co/ZbsY3csM8Z",
        "entities": {
          "url": {
            "urls": [
              {
                "url": "http://t.co/ZbsY3csM8Z",
                "expanded_url": "http://www.philipstreehouse.blogspot.com",
                "display_url": "philipstreehouse.blogspot.com",
                "indices": [0, 22]
              }
            ]
          },
          "protected": false,
          "followers_count": 1510,
          "friends_count": 875,
          "listed_count": 87,
          "created_at": "Tue Feb 17 18:35:54 +0000 2009",
          "favourites_count": 308,
          "utc_offset": 0,
          "time_zone": "London",
          "geo_enabled": false,
          "verified": false,
          "statuses_count": 19338,
          "lang": "en",
          "contributors_enabled": false,
          "is_translator": false,
          "is_translation_enabled": false,
          "profile_background_color": "C6E2EE",
          "profile_background_image_url": "http://pbs.twimg.com/profile_background_images/27682398/mellitus_icon.jpg",
          "profile_background_image_url_https": "https://pbs.twimg.com/profile_background_images/27682398/mellitus_icon.jpg",
          "profile_background_tile": true,
          "profile_image_url": "http://pbs.twimg.com/profile_images/505047850502606849/FjixsyQV_normal.jpeg",
          "profile_image_url_https": "https://pbs.twimg.com/profile_images/505047850502606849/FjixsyQV_normal.jpeg",
          "profile_banner_url": "https://pbs.twimg.com/profile_banners/21115092/1404765200",
          "profile_link_color": "1F98C7",
          "profile_sidebar_border_color": "C6E2EE",
          "profile_sidebar_fill_color": "DAECF4",
          "profile_text_color": "663B12",
          "profile_use_background_image": true,
          "default_profile": false,
          "default_profile_image": false,
          "following": false,
          "follow_request_sent": false,
          "notifications": false
        },
        "geo": null,
        "coordinates": null,
        "place": null,
        "contributors": null,
        "retweet_count": 0,
        "favorite_count": 0,
        "entities": {
          "hashtags": [],
          "symbols": [],
          "user_mentions": [
            {
              "screen_name": "Gerrarrdus",
              "name": "Gerrarrdus",
              "id": 206849269,
              "id_str": "206849269",
              "indices": [0, 11]
            }
          ],
          "urls": []
        },
        "favorited": false,
        "retweeted": false,
        "lang": "en"
      },
      "metadata": {
        "iso_language_code": "in",
        "result_type": "recent"
      },
      "created_at": "Sat Jan 31 13:58:36 +0000 2015",
      "id": 561523946210996225,
      "id_str": "561523946210996225",
      "text": "RT @bemgipbsAINSTEK: Ini adl tengkorak hsl printer 3D yg tlh brhasil dbuat oleh Oxford Performance Materials. Wow keren yaah!! #technovasc \u0026",
      "source": "\u003ca href=\"https://mobile.twitter.com\" rel=\"nofollow\" \u003eMobile Web (M2)\u003c/a\u003e",
      "truncated": false,
      "in_reply_to_status_id": null,
      "in_reply_to_status_id_str": null,
      "in_reply_to_user_id": null,
      "in_reply_to_user_id_str": null,
      "in_reply_to_screen_name": null,
      "user": {
        "id": 363549024,
        "id_str": "363549024",
        "name": "Saepul Al-Malik",
        "screen_name": "saepulmalik27",
        "location": "SMANESPA'13/IPB(FIS)'50|Tasik",
        "profile_location": null,
        "description": "",
        "url": null,
        "entities": {
          "description": {
            "urls": []
          },
          "protected": false,
          "followers_count": 99,
          "friends_count": 109,
          "listed_count": 0,
          "created_at": "Sun Aug 28 08:11:41 +0000 2011",
          "favourites_count": 33,
          "utc_offset": null,
          "time_zone": null,
          "geo_enabled": false,
          "verified": false,
          "statuses_count": 429,
          "lang": "en",
          "contributors_enabled": false,
          "is_translator": false,
          "is_translation_enabled": false,
          "profile_background_color": "FAFAFA",
          "profile_background_image_url": "http://pbs.twimg.com/profile_background_images/37880000036376243/f06e0f73902a669a08339ca82af1f7c7.jpeg",
          "profile_background_image_url_https": "https://pbs.twimg.com/profile_background_images/37880000036376243/f06e0f73902a669a08339ca82af1f7c7.jpeg",
          "profile_background_tile": true,
          "profile_image_url": "http://pbs.twimg.com/profile_images/378800000235103788/0a5ccbfe20ba53c3b54fb803753017ed_normal.jpeg",
          "profile_image_url_https": "https://pbs.twimg.com/profile_images/378800000235103788/0a5ccbfe20ba53c3b54fb803753017ed_normal.jpeg",
          "profile_banner_url": "https://pbs.twimg.com/profile_banners/363549024/1375578137",
          "profile_link_color": "87EEF5",
          "profile_sidebar_border_color": "000000",
          "profile_sidebar_fill_color": "F6FFD1",
          "profile_text_color": "333333",
          "profile_use_background_image": true,
          "default_profile": false,
          "default_profile_image": false,
          "following": false,
          "follow_request_sent": false,
          "notifications": false
        },
        "geo": null,
        "coordinates": null,
        "place": null,
        "contributors": null,
        "retweet_count": 1,
        "favorite_count": 0,
        "entities": {
          "hashtags": [
            {
              "text": "technovasc",
              "indices": [106, 117]
            }
          ],
          "symbols": [],
          "user_mentions": [],
          "urls": [],
          "media": [
            {
              "id": 561523811372105728,
              "id_str": "561523811372105728",
              "indices": [118, 140],
              "media_url": "http://pbs.twimg.com/media/B8ru7m3CAAAsn6v.jpg",
              "media_url_https": "https://pbs.twimg.com/media/B8ru7m3CAAAsn6v.jpg",
              "url": "http://t.co/iVxZpe32nU",
              "display_url": "http://t.co/iVxZpe32nU"
            }
          ]
        }
      },
      "retweeted_status": {
        "metadata": {
          "iso_language_code": "in",
          "result_type": "recent"
        },
        "created_at": "Sat Jan 31 13:58:06 +0000 2015",
        "id": 561523821513957377,
        "id_str": "561523821513957377",
        "text": "Ini adl tengkorak hsl printer 3D yg tlh brhasil dbuat oleh Oxford Performance Materials. Wow keren yaah!! #technovasc http://t.co/iVxZpe32nU",
        "source": "\u003ca href=\"http://www.tweetcaster.com\" rel=\"nofollow\" \u003eTweetCaster for Android\u003c/a\u003e",
        "truncated": false,
        "in_reply_to_status_id": null,
        "in_reply_to_status_id_str": null,
        "in_reply_to_user_id": null,
        "in_reply_to_user_id_str": null,
        "in_reply_to_screen_name": null,
        "user": {
          "id": 2334787884,
          "id_str": "2334787884",
          "name": "SAINSTEK BEM G 2015",
          "screen_name": "bemgipbsAINSTEK",
          "location": "Bogor",
          "profile_location": null,
          "description": "Official account Departemen Sains dan Teknologi BEM FMIPA IPB 2015 | @exploscience @pestasains IPB #technovasc | Chief : @AkangTia (08562143993)",
          "url": null,
          "entities": {
            "description": {
              "urls": []
            },
            "protected": false,
            "followers_count": 110,
            "friends_count": 178,
            "listed_count": 0,
            "created_at": "Sun Feb 09 08:45:02 +0000 2014",
            "favourites_count": 0,
            "utc_offset": 25200,
            "time_zone": "Jakarta",
            "geo_enabled": false,
            "verified": false,
            "statuses_count": 143,
            "lang": "en",
            "contributors_enabled": false,
            "is_translator": false,
            "is_translation_enabled": false,
            "profile_background_color": "C0DEED",
            "profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png",
            "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme1/bg.png",
            "profile_background_tile": false,
            "profile_image_url": "http://pbs.twimg.com/profile_images/559762057575153664/BRvj-MHa_normal.jpeg",
            "profile_image_url_https": "https://pbs.twimg.com/profile_images/559762057575153664/BRvj-MHa_normal.jpeg",
            "profile_banner_url": "https://pbs.twimg.com/profile_banners/2334787884/1422290654",
            "profile_link_color": "0084B4",
            "profile_sidebar_border_color": "C0DEED",
            "profile_sidebar_fill_color": "DDEEF6",
            "profile_text_color": "333333",
            "profile_use_background_image": true,
            "default_profile": true,
            "default_profile_image": false,
            "following": false,
            "follow_request_sent": false,
            "notifications": false
          },
          "geo": null,
          "coordinates": null,
          "place": null,
          "contributors": null,
          "retweet_count": 1,
          "favorite_count": 0,
          "entities": {
            "hashtags": [
              {
                "text": "technovasc",
                "indices": [106, 117]
              }
            ],
            "symbols": [],
            "user_mentions": [],
            "urls": [],
            "media": [
              {
                "id": 561523811372105728,
                "id_str": "561523811372105728",
                "indices": [118, 140],
                "media_url": "http://pbs.twimg.com/media/B8ru7m3CAAAsn6v.jpg",
                "media_url_https": "https://pbs.twimg.com/media/B8ru7m3CAAAsn6v.jpg",
                "url": "http://t.co/iVxZpe32nU",
                "display_url": "http://t.co/iVxZpe32nU"
              }
            ]
          }
        }
      }
    }
  ]
}
```




```
{
  "search_metadata": {
    "count": 100,
    "completed_in": 0.085,
    "max_id_str": "561523971448123392",
    "since_id_str": "0",
    "next_results": "?max_id=561522257475170303&q=Oxford&count=100&include_entities=1&result_type=recent",
    "refresh_url": "?since_id=561523971448123392&q=Oxford&result_type=recent&include_entities=1",
    "since_id": 0,
    "query": "Oxford",
    "max_id": 561523971448123392
  },
  "statuses": [
    {
      "contributors": null,
      "truncated": false,
      "text": "@Gerrarrdus you really have to fear for the state of Oxford CMD after reading that article in the Church Times.",
      "in_reply_to_status_id": 561521412973035520,
      "id": 561523971448123392,
      "favorite_count": 0,
      "source": "<a href=\\\"http://twitter.com/#!/download/ipad\\\" rel=\\\"nofollow\\\">Twitter for iPad</a>",
      "retweeted": false,
      "coordinates": null,
      "entities": {
        "symbols": [],
        "user_mentions": [
          {
            "id": 206849269,
            "indices": [
              0,
              11
            ],
            "id_str": "206849269",
            "screen_name": "Gerrarrdus",
            "name": "Gerrarrdus"
          }
        ],
        "hashtags": [],
        "urls": []
      },
      "in_reply_to_screen_name": "Gerrarrdus",
      "in_reply_to_user_id": 206849269,
      "retweet count": 0,

```

FACEBOOK

- Graph API: developers.facebook.com/docs/graph-api
- Provides far less data than Twitter due to more restrictive privacy settings, but data about [public groups and pages](#) can be obtained.
- Need an [access token](#); Graph API explorer allows for short-term tokens: developers.facebook.com/tools/explorer
- Python tutorial: medium.com/towards-data-science/how-to-use-facebook-graph-api-and-extract-data-using-python-1839e19d6999

Access Token:  EAACEdEose0cBALx8ZCaJEMbRI0qyvadRNlxvu7nZBRAe2B5g20hqsanJ5pjuVADzGdqWfxWMJGSWhD5neRmu55KktqiKzTBxYpSUCdFX Get Token ▾

GET ▾ → /v2.10 ▾ /search?q=Brussels&type=group



▶ Submit

[Learn more about the Graph API syntax](#)

Node: search

```
{
  "data": [
    {
      "name": "Brussels SELL/SWAP/BUY/WANT OR GIVE IT AWAY",
      "privacy": "CLOSED",
      "id": "392213734210774"
    },
    {
      "name": "Brussels Griffon",
      "privacy": "CLOSED",
      "id": "199098950113055"
    },
    {
      "name": "Expats in Brussels",
      "privacy": "CLOSED",
      "id": "636490513125057"
    },
    {
      "name": "STUFF 4 SALE (Brussels)",
      "privacy": "OPEN",
      "id": "546038842095261"
    },
    {
      "name": "Greeks in Brussels",
      "privacy": "OPEN",
      "id": "242927652439242"
    },
    {
      "name": "Life in Brussels / Bruxelles ma belle",
      "privacy": "CLOSED",
      "id": "297313127142971"
    },
    {
      "name": "Colocation Brussels & Flatsharing Bruxelles",

```

OTHER APIs

- Instagram: [instagram.com/developer](https://www.instagram.com/developer)
- YouTube: developers.google.com/youtube/documentation
- Reddit: [reddit.com/dev/api](https://www.reddit.com/dev/api)
- Wikipedia: [wikidata.org/wiki/Wikidata:Data access](https://wikidata.org/wiki/Wikidata:Data_access)
- LINE: developers.line.me/en/docs/social-api/overview
- Weibo: open.weibo.com/wiki/API%E6%96%87%E6%A1%A3/en
- KakaoTalk: developers.kakao.com
- WeChat does not seem to have a public API.

WEB SCRAPING WITH PYTHON

- **Beautiful Soup** is a Python library for pulling data out of HTML and XML files:
crummy.com/software/BeautifulSoup/bs4/doc
- Incredibly **versatile**: can pull all URLs, only items in tables, only bolded headings, etc.
- Beautiful Soup for Python [gentle introduction]:
pythonforbeginners.com/beautifulsoup/beautifulsoup-4-python
- Scrape websites using Beautiful Soup [tutorial]:
medium.freecodecamp.org/how-to-scrape-websites-with-python-and-beautifulsoup-5946935d93fe

News Flows, Consciousness Streams: The Headwaters of a River of Words

By RANDY KENNEDY OCT. 25, 2007



COLLECTION & ANALYSIS TOOLS

- **Mozdeh:** mozdeh.wlv.ac.uk
 - Collection and analysis of YouTube and Twitter data (Windows)
- **Chorus:** chorusanalytics.co.uk
 - Collection and 'visual analytics suite' for Twitter data (Windows)
- **COSMOS:** socialdatalab.net/software
 - Collection, analysis, and visualisation of Twitter data (Windows, Mac, Linux)
- **YouTube Data Tools:** github.com/bernorieder/YouTube-Data-Tools/wiki
 - Collection of YouTube data (channels and videos).
- **Many other tools:** socialmediadata.wikidot.com

PREEXISTING DATASETS

- **Academic data repositories**

- The **Stanford Large Network Dataset Collection** has data from a wide range of platforms, in various formats: snap.stanford.edu/data
- The **Social Computing Data Repository** (Arizona State University) has datasets from a wide variety of social blog and news websites: socialcomputing.asu.edu/pages/datasets

- **Open data portals**

- kdnuggets.com/datasets/government-local-public.html

- **Industry-prepared datasets**

- YouTube-8M: research.google.com/youtube8m/index.html
- Yelp Challenge: yelp.com/dataset/challenge

- **CLARIN corpora**

ETHICS OF ONLINE RESEARCH

- Twitter's Terms of Service ('Your Rights' section):

By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). This license authorizes us to make your Content available to the rest of the world and to let others do the same. You agree that this license includes the right for Twitter to provide, promote, and improve the Services and to make Content submitted to or through the Services available to other companies, organizations or individuals for the syndication, broadcast, distribution, promotion or publication of such Content on other media and services, subject to our terms and conditions for such Content use. Such additional uses by Twitter, or other companies, organizations or individuals, may be made with no compensation paid to you with respect to the Content that you submit, post, transmit or otherwise make available through the Services.

ETHICS OF ONLINE RESEARCH

Key question: *In order to publish (anonymised) individual social media posts, is it first ethically necessary to contact the user and solicit their **informed consent**?* (Webb et al., 2017; Digital Wildfire project)

FACEBOOK CONTAGION STUDY



Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer^{a,1}, Jamie E. Guillory^{b,2}, and Jeffrey T. Hancock^{b,c}

^aCore Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of ^bCommunication and ^cInformation Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

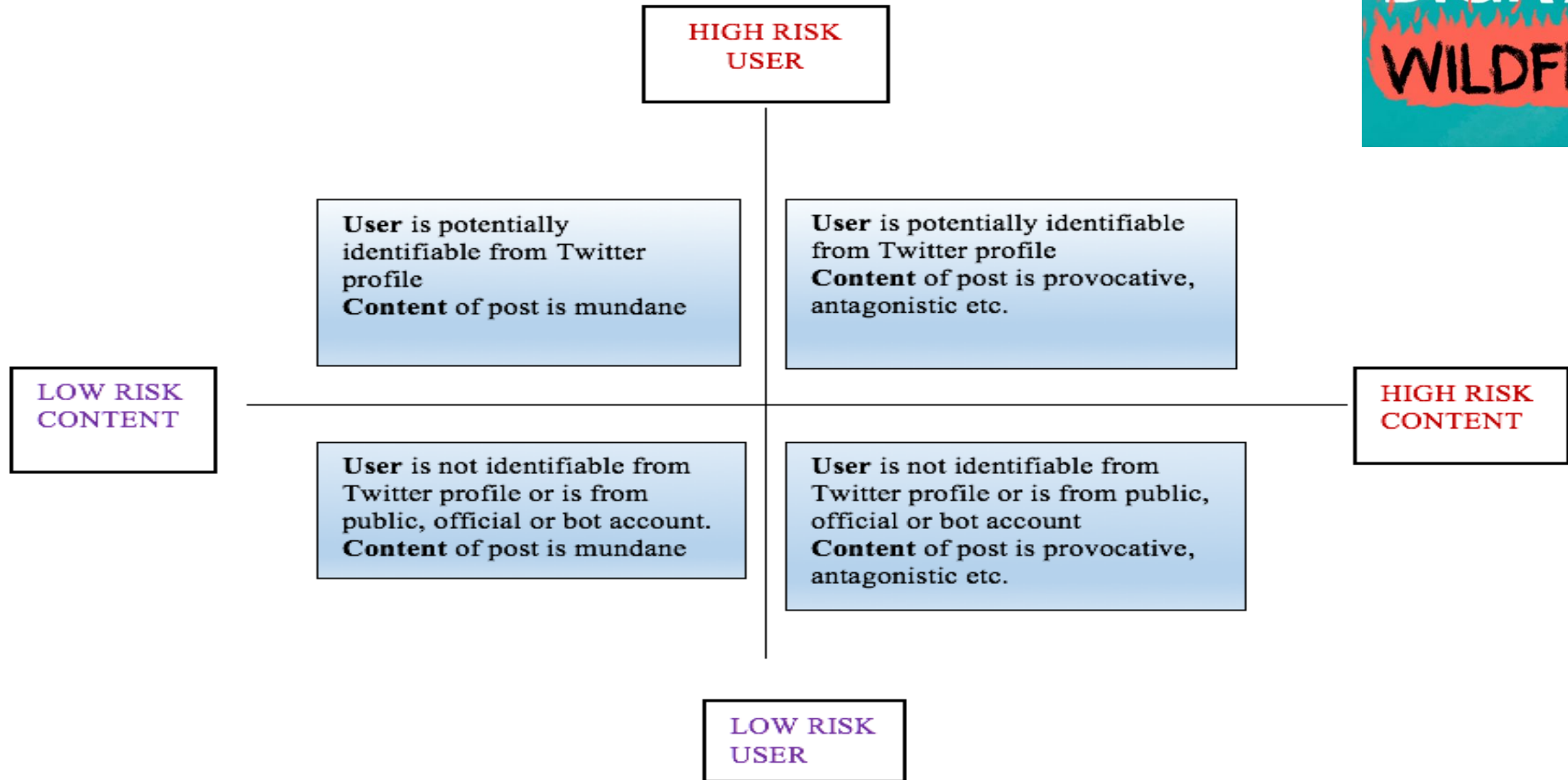
Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were re-

demonstrated that (i) emotional contagion occurs via text-based computer-mediated communication (7); (ii) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (iii) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experimenter and target.

On Facebook, people frequently express emotions, which are later seen by their friends via Facebook's "News Feed" product (8). Because people's friends frequently produce much more content than one person can view, the News Feed filters posts,

KEY CHALLENGES

- Potential **harms** from conducting research / publishing posts.
- Difficulty in achieving **meaningful anonymisation** (tweet can sometimes be found via Twitter advanced search).
- Difficulty in achieving **informed consent** (low level of replies, no contact information, user might not have capacity to consent).
- **Analytic integrity** (readers should be able to independently evaluate the analyses presented).
- **Regulatory vs commercial policies** (EU data protection policy vs. Twitter's broadcast guidelines).
- Absence of **academic consensus**.



From Webb et al. (2017), 'The ethical challenges of publishing Twitter data for research dissemination'

BEST PRACTICE

- Never quote identifiable **individual users** without informed consent.
 - Opt-in consent form for sensitive posts and/or vulnerable users; opt-out form otherwise.
- If informed consent cannot be obtained, represent content in **aggregate form**.
- Can quote from **public and commercial organisations** without obtaining informed consent.
- **Reproducibility**: Twitter's ToS only allows you to publish Tweet IDs, not the raw dataset. Other platforms have similar policies.

