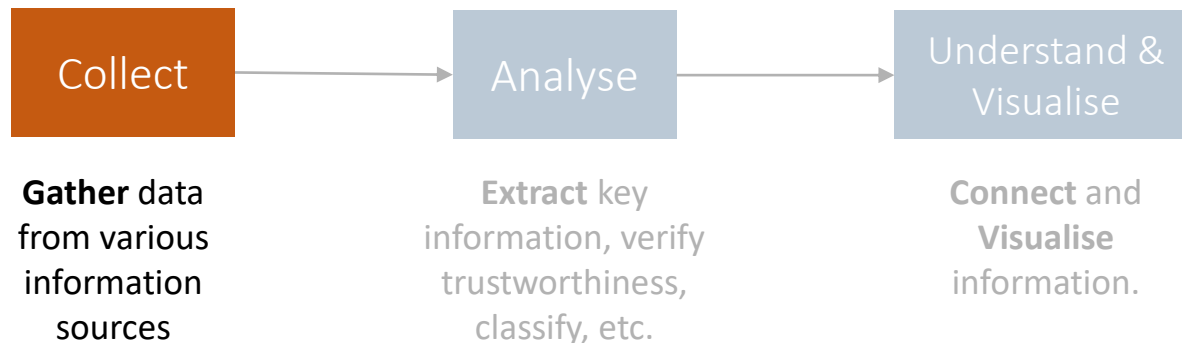


# SMASAC - Data Collection and Filtering

GRÉGOIRE BUREL, MAYANK KEJRIWAL AND  
PRASHANT KHARE



# Social Media Usage during Crises

- **Social media usage during crises varies** depending on:

1. Type of information shared (e.g., affected individuals, caution and advice, donation or volunteering, message of support, etc. ). (Olteanu et al, 2014)
2. Type of content shared (e.g., text, images, videos, links).
3. Content source (e.g., news organisation or journalist, eyewitness, government, NGO, company or for-profit organisation). (Olteanu et al, 2014)
4. Target audience (e.g., general public, other organisation, followers, friends/family).
5. Type of social media platform used (e.g., Facebook, Twitter, etc.)

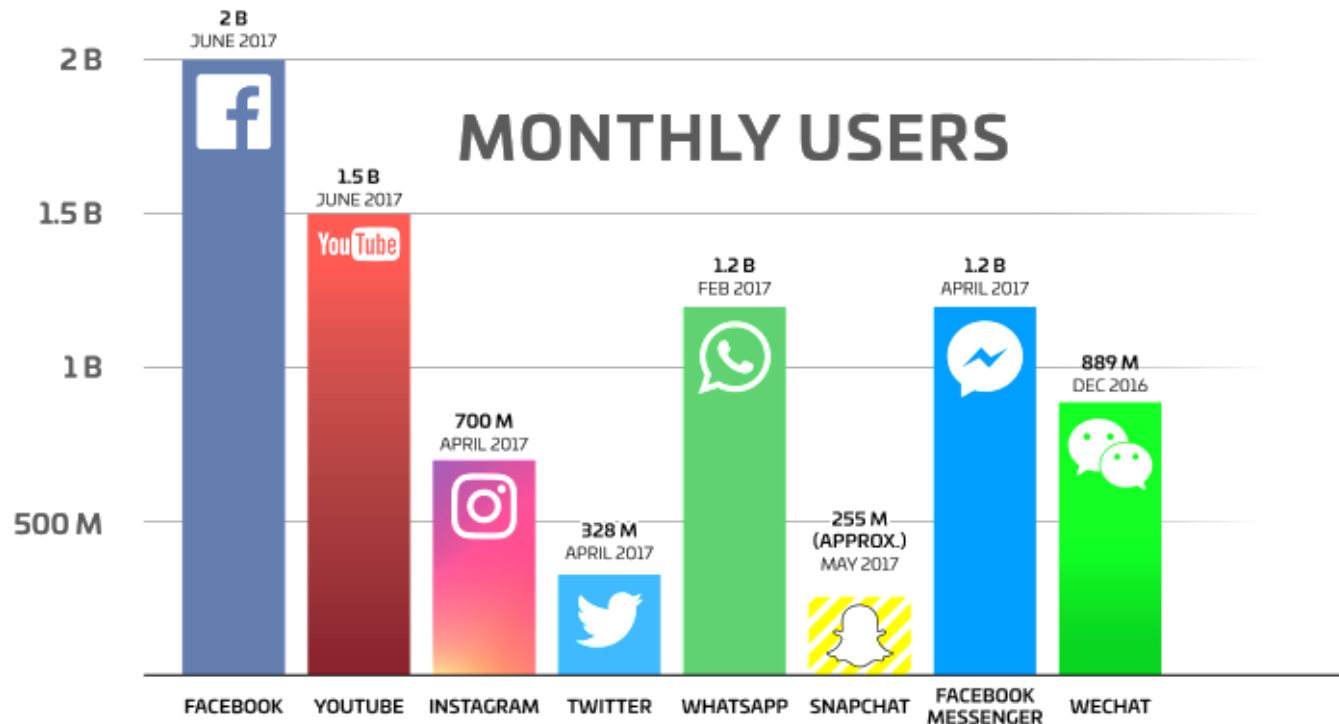


# Access to Social Media

- Automatic access to social media data **can be restricted** in different ways:
  1. Public / Non-public data: Most social media websites do not allow access to the information posted unless reading access is given explicitly by the information creator.
  2. Query restrictions: Data access can be limited by **API restrictions** (e.g., rate limiting, query allowance).
  3. Data Sampling: High velocity data is sometimes **sampled** by social media companies. As result, it is only possible to retrieve a portion of the relevant information.
  4. Query Filtering: Often data is retrieved using **query parameters** (e.g., keywords, geolocation, etc.). Access to relevant data can be negatively impacted.











# Most used Social Media Platforms







































Source: <https://techcrunch.com/2017/06/27/facebook-2-billion-users/>

# Most used Social Media Platforms

	Platform	API Access	Usage	Audience	
	Facebook	Low	High	Mostly Personal	
	YouTube	High	Medium	General	
	Instagram	High	-	Personal/General	
	Twitter	High	High	Mostly General	
	WhatsApp	NA	High	Personal	
	Snapchat	Low	-	Personal	
	Facebook Messenger	Low	-	Personal	
	WeChat	Low	-	Personal	

# Worldwide Social Media App Usage

Top App / Website	US	Canada	UK	France	Germany	Australia	Japan	China	India	Brazil	Mexico	South Africa	Saudi Arabia	Dubai	Jordan	Abu Dhabi
1																
2																
3																
4																
5																

# Data Collection/Processing Methods

- Manual Methods

- Go to social media website or solicit contributions.
- Check for relevant content.
- **Copy/Paste** content into spreadsheet.
- Annotate spreadsheet.
- Visualise the data.
- Take action.

- Automatic Methods

- **Use platform APIs**
- Perform query using the API (e.g., keyword search)
- Automatically populate a database or spreadsheet.
- Annotate manually or automatically the data.
- Visualise the data
- Take action

# Data Collection/Processing Methods

- Manual Methods

- Go to social media website or solicit contributions.
- Check for relevant content.
- **Copy/Paste** content into spreadsheet.
- Annotate spreadsheet.
- Visualise the data.
- Take action.

- Automatic Methods

- **Use platform APIs**
- Perform query using the API (e.g., keyword search)
- Automatically populate a database or spreadsheet.
- Annotate manually or automatically the data.
- Visualise the data
- Take action



# Automatic Data Collection APIs (Twitter)

- Automatic data collection generally relies on JSON APIs and OAuth credentials. For example, for Twitter, you need to:
  - Create a Twitter account (<https://twitter.com>).
  - Obtain an OAuth access credentials (i.e., access token, access secret, consumer key and consumer secret) (<https://apps.twitter.com/app/new>).
  - Use Search API for collecting tweets (<https://developer.twitter.com>).
  - Save Tweets in JSON or other format for later analysis.



```
Example
$ gem install twurl
$ twurl authorize --consumer-key key \
                  --consumer-secret secret
$ twurl authorize --consumer-key key \
                  --consumer-secret secret
$ twurl "/1.1/search/tweets.json?q=earthquake"
{
  "statuses": [
    {
      "created_at": "Mon Feb 12 10:58:42 +0000 2018",
      "id": 963004430915289100,
      "id_str": "963004430915289088",
      "text": "RT @Independent: Magnitude 4.4 earthquake strikes near Beijing https://t.co/usyGy0kyA",
      "truncated": false,
      "entities": {
        "hashtags": [],
        "symbols": [],
        "user_mentions": [
          {
            "screen_name": "Independent",
            "name": "The Independent",
            "id": 16973333,
            "id_str": "16973333",
            "indices": [
              3,
              15
            ]
          }
        ]
      },
      "urls": [
        {
          "url": "https://t.co/usyGy0kyA",
          "expanded_url": "http://www.independent.co.uk/news/world/asia/beijing-earthquake-latest-update-china-magnitude-hebei-tremors-capital-a8206336.html",
          "display_url": "independent.co.uk/news/world/asi...",
          "indices": [
            63,
            96
          ]
        }
      ]
    }
  ]
}
```

# Twitter Data and Crises



**Wildfire**



People of NSW, be careful because there's fires spreading! Stay safe everyone!

## CRISIS



Hundreds of volunteers in Mexico tried to unearth children they hoped were still alive beneath a school's ruins



**Earthquake**

**Floods**



Two trucks and one car in the water after a road collapse at Hwy 287 and Dillon. #cowx #boulderflood

Volume

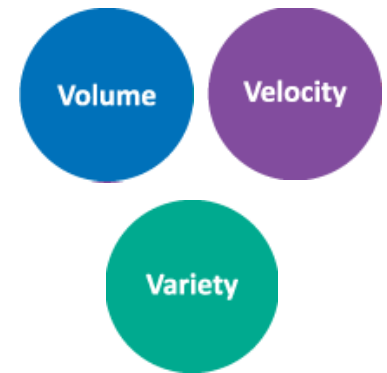
Variety

Velocity

Veracity

# Accessing Relevant Information

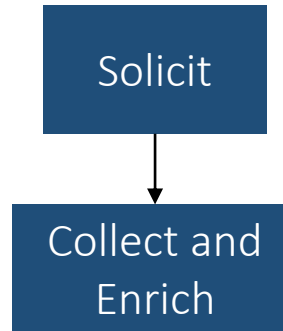
- Information overload:
  - During crises, a flood of data gets generated. For example:
    - Over a million tweets generated during Hurricane Harvey 2017.
    - 500% increase in the tweets bandwidth during 2011 Japan earthquake.
  - The characteristics of social media posts such as short length, colloquialism, syntactic issues pose additional challenges of processing the data.
  - Almost impossible to manually absorb and process the sheer volume.



How do we **separate crisis-related content** from irrelevant information?

How to **ensure** that a wide range of **crisis related topics across diverse crisis situations** are **filtered in**?

# Hurricane Harvey – Data Collection/Filtering



**Information  
solicited on  
social media.**

**Copy/Paste** and manually extract information in spreadsheet.



Timestamp	Name	Street Address	Apt #	City	Zip Code	Latitude	Longitude	Coordinate	Phone number	Twitter Handle	Location Comments
8/27/2017 2:00	John Doe	123 Main St		San Francisco	94102	37.7749	-122.4194		(415) 555-1234	@JohnDoe	
8/27/2017 2:05	Jane Smith	456 Oak Ave		San Francisco	94103	37.7850	-122.4050		(415) 555-5678	@JaneSmith	
8/27/2017 2:10	Bob Johnson	789 Pine St		San Francisco	94104	37.7950	-122.3950		(415) 555-9012	@BobJohnson	
8/27/2017 2:15	Alice Brown	101 Elm St		San Francisco	94105	37.8050	-122.3850		(415) 555-3456	@AliceBrown	
8/27/2017 2:20	Charlie Davis	202 Maple St		San Francisco	94106	37.8150	-122.3750		(415) 555-7890	@CharlieDavis	
8/27/2017 2:25	Diana Evans	303 Cedar St		San Francisco	94107	37.8250	-122.3650		(415) 555-2345	@DianaEvans	
8/27/2017 2:30	Frank Green	404 Birch St		San Francisco	94108	37.8350	-122.3550		(415) 555-6789	@FrankGreen	
8/27/2017 2:35	Grace Hill	505 Spruce St		San Francisco	94109	37.8450	-122.3450		(415) 555-0123	@GraceHill	
8/27/2017 2:40	Henry Lee	606 Ash St		San Francisco	94110	37.8550	-122.3350		(415) 555-4567	@HenryLee	
8/27/2017 2:45	Ivy King	707 Hickory St		San Francisco	94111	37.8650	-122.3250		(415) 555-8901	@IvyKing	
8/27/2017 2:50	Jack Lopez	808 Walnut St		San Francisco	94112	37.8750	-122.3150		(415) 555-2345	@JackLopez	
8/27/2017 2:55	Karen Miller	909 Chestnut St		San Francisco	94113	37.8850	-122.3050		(415) 555-6789	@KarenMiller	
8/27/2017 3:00	Leo Wilson	1010 Sycamore St		San Francisco	94114	37.8950	-122.2950		(415) 555-0123	@LeoWilson	
8/27/2017 3:05	Mia Young	1111 Magnolia St		San Francisco	94115	37.9050	-122.2850		(415) 555-4567	@MiaYoung	
8/27/2017 3:10	Noah Adams	1212 Dogwood St		San Francisco	94116	37.9150	-122.2750		(415) 555-8901	@NoahAdams	
8/27/2017 3:15	Olivia Baker	1313 Redwood St		San Francisco	94117	37.9250	-122.2650		(415) 555-2345	@OliviaBaker	
8/27/2017 3:20	Peter Clark	1414 Cypress St		San Francisco	94118	37.9350	-122.2550		(415) 555-6789	@PeterClark	
8/27/2017 3:25	Quinn Hall	1515 Juniper St		San Francisco	94119	37.9450	-122.2450		(415) 555-0123	@QuinnHall	
8/27/2017 3:30	Rachel King	1616 Fir St		San Francisco	94120	37.9550	-122.2350		(415) 555-4567	@RachelKing	
8/27/2017 3:35	Samuel Lee	1717 Willow St		San Francisco	94121	37.9650	-122.2250		(415) 555-8901	@SamuelLee	
8/27/2017 3:40	Tina Miller	1818 Cottonwood St		San Francisco	94122	37.9750	-122.2150		(415) 555-2345	@TinaMiller	
8/27/2017 3:45	Uma Wilson	1919 Alder St		San Francisco	94123	37.9850	-122.2050		(415) 555-6789	@UmaWilson	
8/27/2017 3:50	Victor Young	2020 Hawthorn St		San Francisco	94124	37.9950	-122.1950		(415) 555-0123	@VictorYoung	
8/27/2017 3:55	Wendy Adams	2121 Blackberry St		San Francisco	94125	38.0050	-122.1850		(415) 555-4567	@WendyAdams	
8/27/2017 4:00	Xavier Baker	2222 Raspberry St		San Francisco	94126	38.0150	-122.1750		(415) 555-8901	@XavierBaker	
8/27/2017 4:05	Yara Clark	2323 Strawberry St		San Francisco	94127	38.0250	-122.1650		(415) 555-2345	@YaraClark	
8/27/2017 4:10	Zoe Hall	2424 Tangerine St		San Francisco	94128	38.0350	-122.1550		(415) 555-6789	@ZoeHall	
8/27/2017 4:15	Adam King	2525 Lemon St		San Francisco	94129	38.0450	-122.1450		(415) 555-0123	@AdamKing	
8/27/2017 4:20	Bella Lee	2626 Lime St		San Francisco	94130	38.0550	-122.1350		(415) 555-4567	@BellaLee	
8/27/2017 4:25	Chris Miller	2727 Peach St		San Francisco	94131	38.0650	-122.1250		(415) 555-8901	@ChrisMiller	
8/27/2017 4:30	Diana Wilson	2828 Apple St		San Francisco	94132	38.0750	-122.1150		(415) 555-2345	@DianaWilson	
8/27/2017 4:35	Ethan Young	2929 Orange St		San Francisco	94133	38.0850	-122.1050		(415) 555-6789	@EthanYoung	
8/27/2017 4:40	Fiona Adams	3030 Grape St		San Francisco	94134	38.0950	-122.0950		(415) 555-0123	@FionaAdams	
8/27/2017 4:45	George Baker	3131 Lemon St		San Francisco	94135	38.1050	-122.0850		(415) 555-4567	@GeorgeBaker	
8/27/2017 4:50	Hannah Clark	3232 Lime St		San Francisco	94136	38.1150	-122.0750		(415) 555-8901	@HannahClark	
8/27/2017 4:55	Ian Hall	3333 Peach St		San Francisco	94137	38.1250	-122.0650		(415) 555-2345	@IanHall	
8/27/2017 5:00	Jessica King	3434 Apple St		San Francisco	94138	38.1350	-122.0550		(415) 555-6789	@JessicaKing	
8/27/2017 5:05	Kevin Lee	3535 Orange St		San Francisco	94139	38.1450	-122.0450		(415) 555-0123	@KevinLee	
8/27/2017 5:10	Laura Miller	3636 Grape St		San Francisco	94140	38.1550	-122.0350		(415) 555-4567	@LauraMiller	
8/27/2017 5:15	Michael Wilson	3737 Lemon St		San Francisco	94141	38.1650	-122.0250		(415) 555-8901	@MichaelWilson	
8/27/2017 5:20	Nancy Young	3838 Lime St		San Francisco	94142	38.1750	-122.0150		(415) 555-2345	@NancyYoung	
8/27/2017 5:25	Oscar Adams	3939 Peach St		San Francisco	94143	38.1850	-122.0050		(415) 555-6789	@OscarAdams	
8/27/2017 5:30	Pamela Baker	4040 Apple St		San Francisco	94144	38.1950	-121.9950		(415) 555-0123	@PamelaBaker	
8/27/2017 5:35	Quinn Clark	4141 Orange St		San Francisco	94145	38.2050	-121.9850		(415) 555-4567	@QuinnClark	
8/27/2017 5:40	Rachel Hall	4242 Grape St		San Francisco	94146	38.2150	-121.9750		(415) 555-8901	@RachelHall	
8/27/2017 5:45	Samuel King	4343 Lemon St		San Francisco	94147	38.2250	-121.9650		(415) 555-2345	@SamuelKing	
8/27/2017 5:50	Tina Lee	4444 Lime St		San Francisco	94148	38.2350	-121.9550		(415) 555-6789	@TinaLee	
8/27/2017 5:55	Uma Miller	4545 Peach St		San Francisco	94149	38.2450	-121.9450		(415) 555-0123	@UmaMiller	
8/27/2017 6:00	Victor Wilson	4646 Apple St		San Francisco	94150	38.2550	-121.9350		(415) 555-4567	@VictorWilson	
8/27/2017 6:05	Wendy Young	4747 Orange St		San Francisco	94151	38.2650	-121.9250		(415) 555-8901	@WendyYoung	
8/27/2017 6:10	Xavier Adams	4848 Grape St		San Francisco	94152	38.2750	-121.9150		(415) 555-2345	@XavierAdams	
8/27/2017 6:15	Yara Baker	4949 Lemon St		San Francisco	94153	38.2850	-121.9050		(415) 555-6789	@YaraBaker	
8/27/2017 6:20	Zoe Clark	5050 Lime St		San Francisco	94154	38.2950	-121.8950		(415) 555-0123	@ZoeClark	
8/27/2017 6:25	Adam Hall	5151 Peach St		San Francisco	94155	38.3050	-121.8850		(415) 555-4567	@AdamHall	
8/27/2017 6:30	Bella King	5252 Apple St		San Francisco	94156	38.3150	-121.8750		(415) 555-8901	@BellaKing	
8/27/2017 6:35	Chris Lee	5353 Orange St		San Francisco	94157	38.3250	-121.8650		(415) 555-2345	@ChrisLee	
8/27/2017 6:40	Diana Miller	5454 Grape St		San Francisco	94158	38.3350	-121.8550		(415) 555-6789	@DianaMiller	
8/27/2017 6:45	Ethan Wilson	5555 Lemon St		San Francisco	94159	38.3450	-121.8450		(415) 555-0123	@EthanWilson	
8/27/2017 6:50	Fiona Young	5656 Lime St		San Francisco	94160	38.3550	-121.8350		(415) 555-4567	@FionaYoung	
8/27/2017 6:55	George Adams	5757 Peach St		San Francisco	94161	38.3650	-121.8250		(415) 555-8901	@GeorgeAdams	
8/27/2017 7:00	Hannah Baker	5858 Apple St		San Francisco	94162	38.3750	-121.8150		(415) 555-2345	@HannahBaker	
8/27/2017 7:05	Ian Clark	5959 Orange St		San Francisco	94163	38.3850	-121.8050		(415) 555-6789	@IanClark	
8/27/2017 7:10	Jessica Hall	6060 Grape St		San Francisco	94164	38.3950	-121.7950		(415) 555-0123	@JessicaHall	
8/27/2017 7:15	Kevin King	6161 Lemon St		San Francisco	94165	38.4050	-121.7850		(415) 555-4567	@KevinKing	
8/27/2017 7:20	Laura Lee	6262 Lime St		San Francisco	94166	38.4150	-121.7750		(415) 555-8901	@LauraLee	
8/27/2017 7:25	Michael Miller	6363 Peach St		San Francisco	94167	38.4250	-121.7650		(415) 555-2345	@MichaelMiller	
8/27/2017 7:30	Nancy Wilson	6464 Apple St		San Francisco	94168	38.4350	-121.7550		(415) 555-6789	@NancyWilson	
8/27/2017 7:35	Oscar Young	6565 Orange St		San Francisco	94169	38.4450	-121.7450		(415) 555-0123	@OscarYoung	
8/27/2017 7:40	Pamela Adams	6666 Grape St		San Francisco	94170	38.4550	-121.7350		(415) 555-4567	@PamelaAdams	
8/27/2017 7:45	Quinn Baker	6767 Lemon St		San Francisco	94171	38.4650	-121.7250		(415) 555-8901	@QuinnBaker	
8/27/2017 7:50	Rachel Clark	6868 Lime St		San Francisco	94172	38.4750	-121.7150		(415) 555-2345	@RachelClark	
8/27/2017 7:55	Samuel Hall	6969 Peach St		San Francisco	94173	38.4850	-121.7050		(415) 555-6789	@SamuelHall	
8/27/2017 8:00	Tina King	7070 Apple St		San Francisco	94174	38.4950	-121.6950		(415) 555-0123	@TinaKing	
8/27/2017 8:05	Uma Lee	7171 Orange St		San Francisco	94175	38.5050	-121.6850		(415) 555-4567	@UmaLee	
8/27/2017 8:10	Victor Miller	7272 Grape St		San Francisco	94176	38.5150	-121.6750		(415) 555-8901	@VictorMiller	
8/27/2017 8:15	Wendy Wilson	7373 Lemon St		San Francisco	94177	38.5250	-121.6650		(415) 555-2345	@WendyWilson	
8/27/2017 8:20	Xavier Young	7474 Lime St		San Francisco	94178	38.5350	-121.6550		(415) 555-6789	@XavierYoung	
8/27/2017 8:25	Yara Adams	7575 Peach St		San Francisco	94179	38.5450	-121.6450		(415) 555-0123	@YaraAdams	
8/27/2017 8:30	Zoe Baker	7676 Apple St		San Francisco	94180	38.5550	-121.6350		(415) 555-4567	@ZoeBaker	
8/27/2017 8:35	Adam Clark	7777 Orange St		San Francisco	94181	38.5650	-121.6250		(415) 555-8901	@AdamClark	
8/27/2017 8:40	Bella Hall	7878 Grape St		San Francisco	94182	38.5750	-121.6150		(415) 555-2345	@BellaHall	
8/27/2017 8:45	Chris King	7979 Lemon St		San Francisco	94183	38.5850	-121.6050		(415) 555-6789	@ChrisKing	
8/27/2017 8:50	Diana Lee	8080 Lime St		San Francisco	94184	38.5950	-121.5950		(415) 555-0123	@DianaLee	
8/27/2017 8:55	Ethan Miller	8181 Peach St		San Francisco	94185	38.6050	-121.5850		(415) 555-4567	@EthanMiller	
8/27/2017 9:00	Fiona Wilson	8282 Apple St		San Francisco	94186	38.6150	-121.5750		(415) 555-8901	@FionaWilson	
8/27/2017 9:05	George Young	8383 Orange St		San Francisco	94187	38.6250	-121.5650		(415) 555-2345	@GeorgeYoung	
8/27/2017 9:10	Hannah Adams	8484 Grape St		San Francisco	94188	38.6350	-121.5550		(415) 555-6789	@HannahAdams	
8/27/2017 9:15	Ian Baker	8585 Lemon St		San Francisco	94189	38.6450	-121.5450		(415) 555-0123	@IanBaker	
8/27/2017 9:20	Jessica Clark	8686 Lime St		San Francisco	94190	38.6550	-121.5350		(415) 555-4567	@JessicaClark	
8/27/2017 9:25	Kevin Hall	8787 Peach St		San Francisco	94191	38.6650	-121.5250		(415) 555-8901	@KevinHall	
8/27/2017 9:30	Laura King	8888 Apple St		San Francisco	94192	38.6750	-121.5150		(415) 555-2345	@LauraKing	
8/27/2017 9:35	Michael Lee	8989 Orange St		San Francisco	94193	38.6850	-121.5050		(415) 555-6789	@MichaelLee	
8/27/2017 9:40	Nancy Miller	9090 Grape St		San Francisco	94194	38.6950	-121.4950		(415) 555-0123	@NancyMiller	
8/27/2017 9:45	Oscar Wilson	9191 Lemon St		San Francisco	94195	38.7050	-121.4850		(415) 555-4567	@OscarWilson	
8/27/2017 9:50	Pamela Young	9292 Lime St		San Francisco	94196	38.7150	-121.4750		(415) 555-8901	@PamelaYoung	
8/27/2017 9:55	Quinn Adams	9393 Peach St		San Francisco	94197	38.7250	-121.4650		(415) 555-2345	@QuinnAdams	
8/27/2017 10:00	Rachel Baker	9494 Apple St		San Francisco	94198	38.7350	-121.4550		(415) 555-6789	@RachelBaker	
8/27/2017 10:05	Samuel Clark	9595 Orange St		San Francisco	94199	38.7450	-121.4450		(415) 555-0123	@SamuelClark	
8/27/2017 10:10	Tina Hall	9696 Grape St		San Francisco	94200	38.7550	-121.4350		(415) 555-4567	@TinaHall	
8/27/2017 10:15	Uma King	9797 Lemon St		San Francisco	94201	38.7650	-121.4250		(415) 555-8901	@UmaKing	
8/27/2017 10:20	Victor Lee	9898 Lime St		San Francisco	94202	38.7750	-121.4150		(415) 555-2345	@VictorLee	
8/27/2017 10:25	Wendy Miller	9999 Peach St		San Francisco	9420						

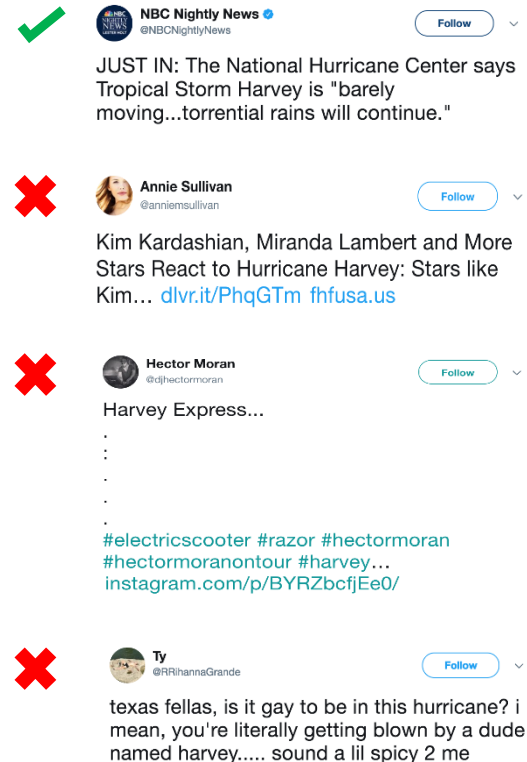
# Data Curation (Filtering)

- Content Curation:



- Methods and processes used for **gathering relevant information** relevant to a particular topic of interest.



- Information Filtering Systems:



- Approaches for removing irrelevant and unwanted information from an information stream using partially-automated or computerized methods **prior to presentation to a human user.**





The screenshot displays four tweets from a social media feed, each preceded by a green checkmark or a red X to indicate its status in a curation process. The first tweet, from NBC Nightly News, is marked with a green checkmark and contains a news report about Tropical Storm Harvey. The second tweet, from Annie Sullivan, is marked with a red X and contains a link to a video of celebrities reacting to Hurricane Harvey. The third tweet, from Hector Moran, is also marked with a red X and contains a link to an Instagram post. The fourth tweet, from Ty (@RihannaGrande), is marked with a red X and contains a humorous, inappropriate comment about Hurricane Harvey.

 **NBC Nightly News** @NBCNightlyNews [Follow](#)   
JUST IN: The National Hurricane Center says Tropical Storm Harvey is "barely moving...torrential rains will continue."

 **Annie Sullivan** @anniesullivan [Follow](#)   
Kim Kardashian, Miranda Lambert and More Stars React to Hurricane Harvey: Stars like Kim... [dlvr.it/PhqGTm](#) [fhfusa.us](#)

 **Hector Moran** @hjectormoran [Follow](#)   
Harvey Express...  
.  
.  
.  
.  
.  
[#electricscooter](#) [#razor](#) [#hjectormoran](#)  
[#hjectormoranontour](#) [#harvey...](#)  
[instagram.com/p/BYRZbcfjEe0/](#)

 **Ty** @RihannaGrande [Follow](#)   
texas fellas, is it gay to be in this hurricane? i mean, you're literally getting blown by a dude named harvey..... sound a lil spicy 2 me

# Filtering Methods



CrisisLex

- Query filtering using social media APIs:
  - Use hashtags, keywords, crisis specific phrases or lexicon (impacted location name, canonical form of disaster name - e.g. *Hurricane Harvey*).
- Post collection filtering (before or after storage):
  - Text search (similar to above).
  - Semantic search (requires entity extraction)\*.
  - Automatic categorisation / Tagging (clustering approaches, topic modelling, Machine Learning models)\*.

\* Entity extraction methods and automatic categorization is discussed later on in this tutorial

flood crisis, victims, flood  
victims, flood powerful,  
powerful storms, hoisted  
flood, storms amazing,  
explosion, amazing rescue,  
rescue women, flood cost,  
counts flood, toll rises,  
braces river, river peaks,  
crisis deepens, prayers,  
thoughts prayers, affected  
tornado, affected, death  
toll, tornado relief, photos  
flood, water rises, toll,  
flood waters, flood appeal,  
victims explosion, bombing  
suspect, massive explosion,  
affected areas, praying  
victims, injured, please  
join, join praying, prayers  
people, redcross, text  
redcross, visiting flood,  
lurches fire, video  
explosion, deepens death,  
aid, help flood,  
died explosions, marathon  
explosions, flood relief

# Filtering using Social Media APIs

- Twitter API:

- Use hashtags, keywords, crisis specific phrases or lexicon (e.g., CrisisLex lexicon).
- Use Geolocation constrains (but many documents do not have geolocation information).
- Use language restrictions.



# Filtering using APIs: Information Vs Noise

- Hashtag Hijacking:

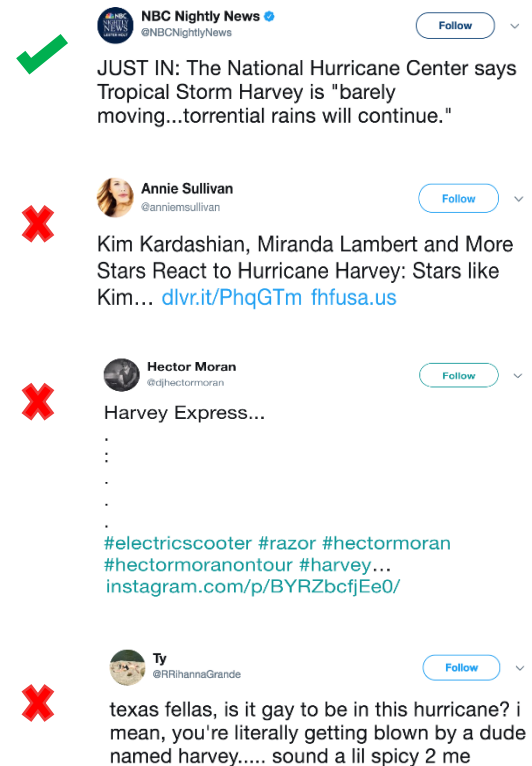
- Hashtags, Keywords can be a convenient way to create a broad set of data, but they **do not always reflect the crisis related information**.

- Overly specific query terms:

- The real crisis information is contained in a **very diverse** form.
- Document **geolocation** may be **inaccurate or missing**.

- Unknown situation / insufficient context:

- The subjects in the information range from health/well-being of affected individuals, infrastructure, donations etc.
- Obviously, various concepts together form a context which reflects relevance with the crisis situations.





# Post Collection Filtering

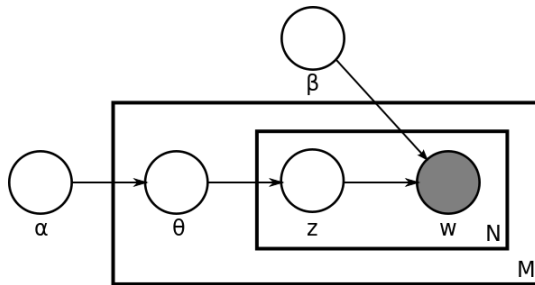
- Text search:
  - More complex queries can be used after indexing the data that may be not doable using the APIs (e.g. tokenisation, query expansion, etc.)
- Semantic search (requires entity extraction)\*:
  - Semantic search can be used after extracting entities of interest from documents (e.g., place names, actors, etc.)
- Automatic categorisation / Tagging (clustering approaches, topic modelling, Machine Learning models)\*:
  - Topic modelling and clustering can discover hidden topics in the collected documents.
  - Machine learning can be used for filtering data (e.g., relatedness, event type, information type).



\* Entity extraction methods and automatic categorization is discussed later on in this tutorial

# Text Clustering

- A simple approach for filtering textual data is to **group related documents automatically (clustering)**.
  - E.g., K-Nearest Neighbour (KNN), Latent Dirichlet Allocation (LDA) (Blei et al., 2003).



- Text clustering approaches are unsupervised: **manual annotations are unnecessary.**
- **Filtering** can then be done using the clusters generated by those models.
- Unsupervised approaches are limited and may not always produce relevant clusters.



# Semantic Search

Query based systems, where the **relevancy of a document is established based on the context of a query** (Mangold et al., 2007):

- ***Describe events as semantic queries and use knowledge graphs and ontologies to map the data and the query.***
- ***Use of Natural Language Processing techniques and external Knowledge Graphs.***
- ***Can be tailored to a certain type of events or broad category of events.***



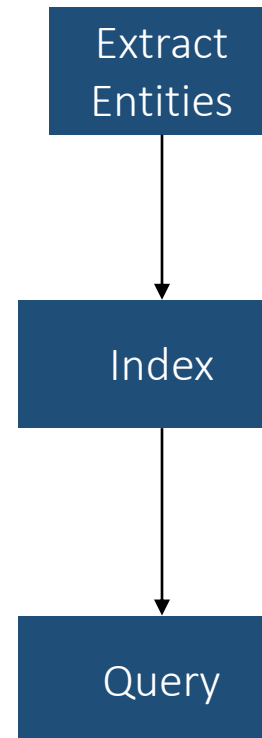
**WGS841**

# Semantic Search - Motivation

- Complex event/information types are **non-trivial to be described by single entity or keywords.**
- N-grams or bag of words do **not directly contain semantic information.**
- Search for content based on a theme/context instead of precise information.

# Semantic Search - Approach

- Entity extraction systems are used such as TextRazor, Dbpedia Spotlight, etc.
- An indexing and search system such as Lucene/Solr is used.
- A graph database store uniquely defined properties for each document

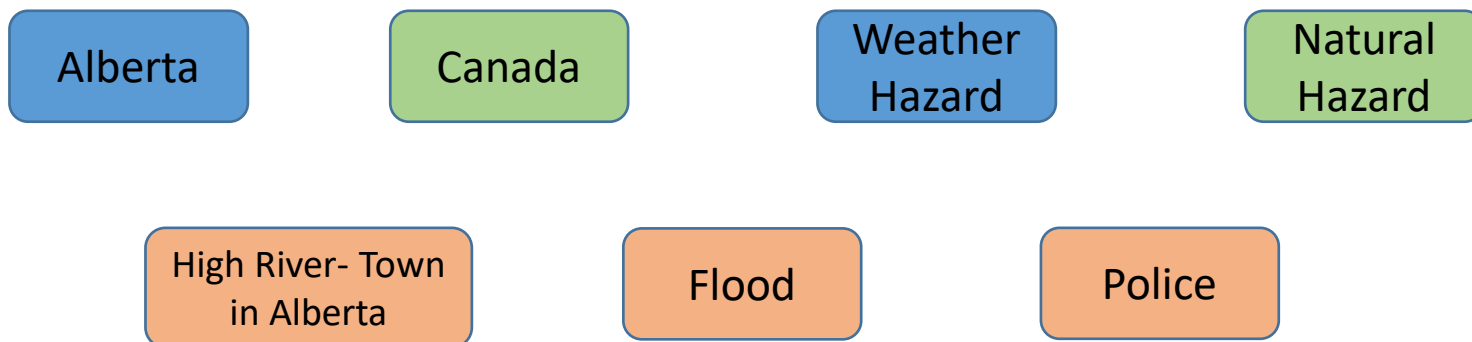


**Entity extraction** is used for obtaining fine grained information within documents.

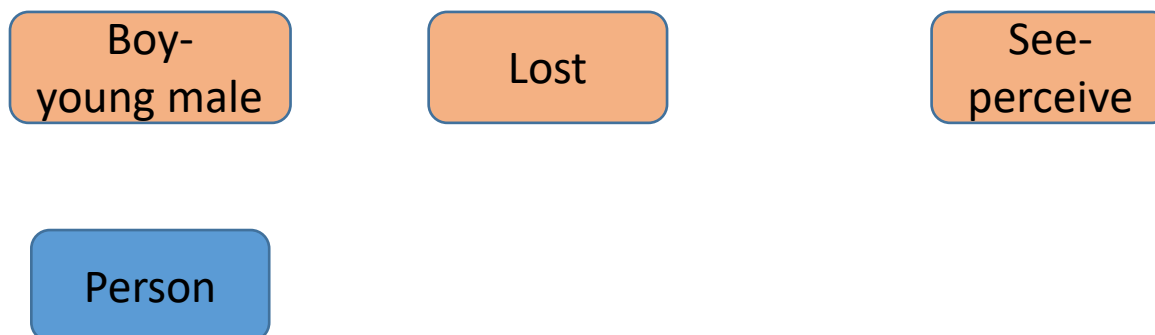
The data is **indexed** using regular information retrieval tools or stored in a graph database.

**Relevant information is retrieved** using entities and relations between related content.

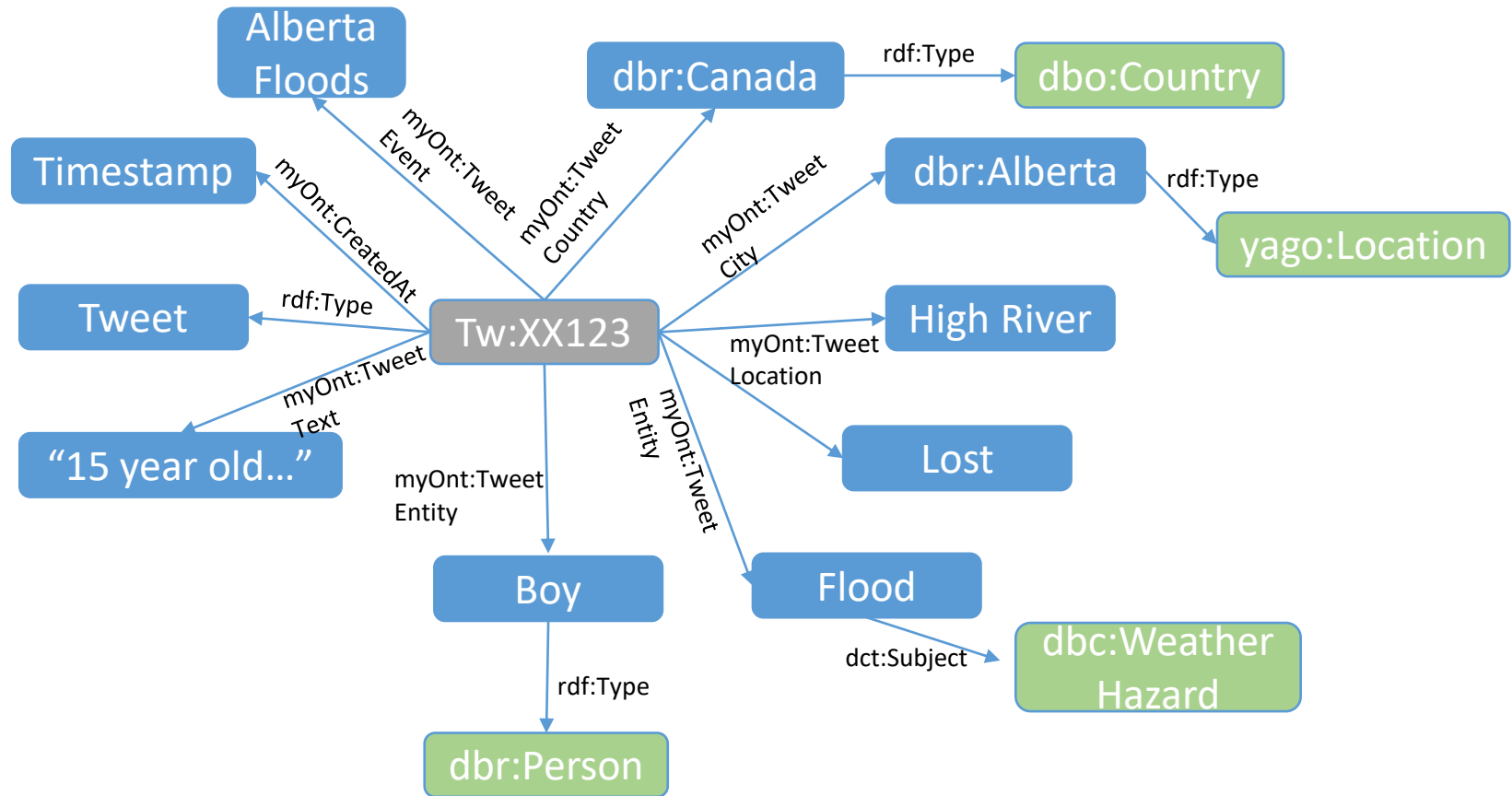
# Document Entities - Example



*"A 15 year old High River boy is missing due to flood. Call police if you see Eric St. Denis"*

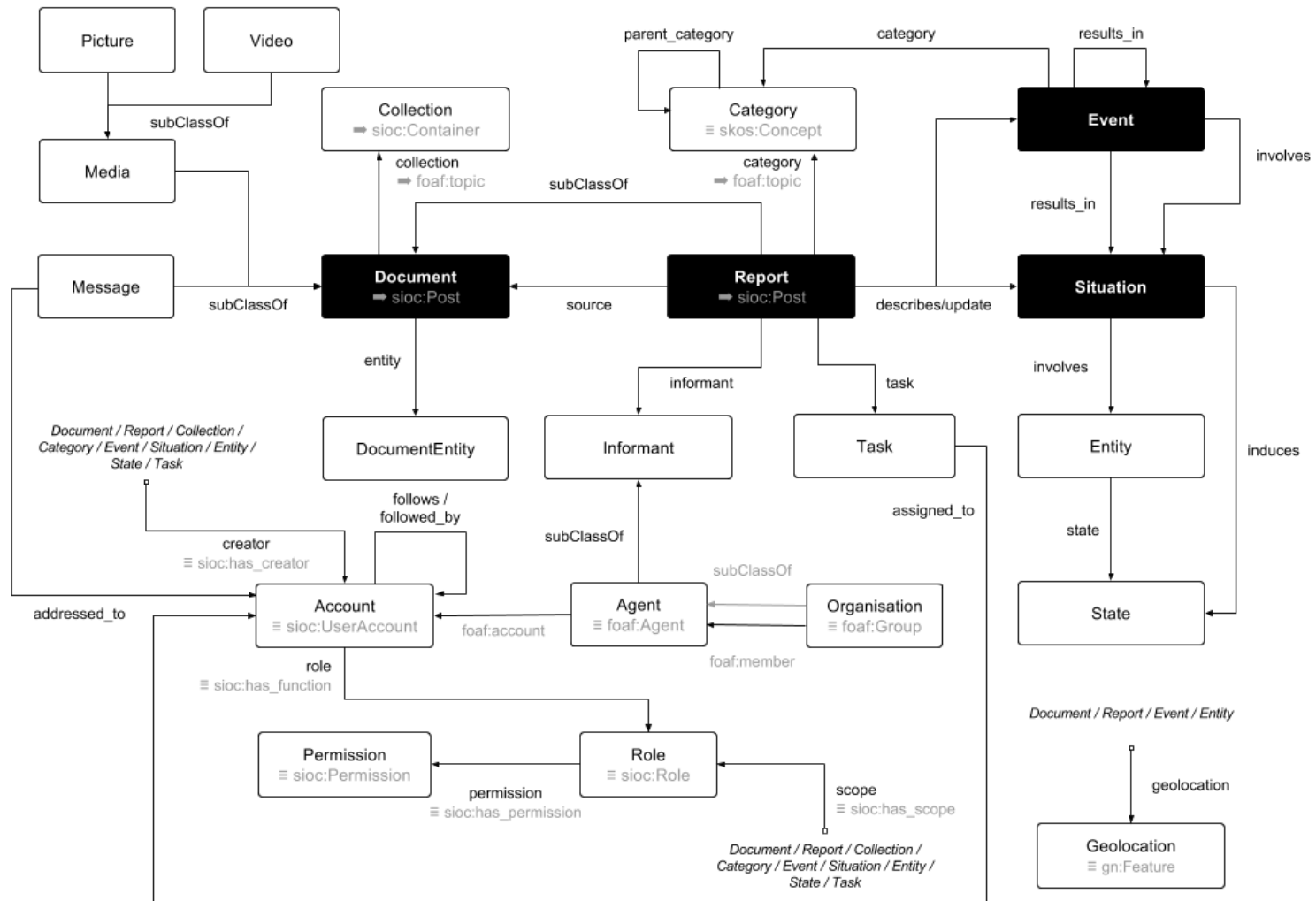


# Graph Representation of Tweet



Concept Source: Tonon, A., Cudré-Mauroux, P., Blarer, A., Lenders, V. and Motik, B., 2017, May. ArmaTweet: Detecting Events by Semantic Tweet Analysis. In *European Semantic Web Conference* (pp. 138-153). Springer, Cham.

# Ontological Representation





# Automatic Data Collection on Twitter

Hands-on

# Summary

- Data **collection can be done automatically** using the APIs of different social media platforms.
- **Social media platform usage varies** greatly across countries and demographics (e.g., target, information, etc.).
- Access to social media data **can be limited** using APIs restrictions.
- Multiple methods can be used for **filtering quickly irrelevant information** or focusing on specific content of interest.

