JCDL 2018 @ Fort Worth, Texas, USA

Digital History meets Microblogging: Analyzing Collective Memories in Twitter

○Yasunobu Sumikawa
Tokyo Metropolitan Univ.

Adam Jatowt
Kyoto Univ.

Marten During
Luxembourg Centre for Contemporary and Digital History

Motivation

- History is important.
  - It helps us understand the processes which shape the present
  - Children have history lessons from elementary school.

- Importance of supporting history learning
  - To support
    1. obtaining knowledge and
    2. using it in real situation.
  - History education researchers argue that applying historical knowledge for proposing creative solutions to present issues (historical analogy) is important.

P. Lee. 2005. Historical Literacy: Theory and Research. International Journal of Historical Learning, Teaching and Research 5, 1 (2005), 25–40.

Related Works

- How does our society remember the past?

  C.-m. Au Yeung and A. Jatowt. *Studying How the Past is Remembered: Towards Computational History Through Large Scale Text Mining*. CIKM '11, 1231–1240.

- Analyzing edit history of Wikipedia's event articles.

  M. Ferron and P. Massa. *Collective Memory Building in Wikipedia: The Case of North African Uprisings.* WikiSym '11, 114–123.

Focusing on memory triggers that

1. cause forgotten
2. vaguely remembered events

to be brought back into *social attention*.

Research Questions Guiding this Study

- What past events do Twitter users focus on?
  1. How do people refer to history in microblogs?
  2. What is the time horizon of history-related references?
  3. How are collective memories expressed in Twitter?
  4. What are the key remembered events and entities?




Benefits of Analyzing History-related Tweets

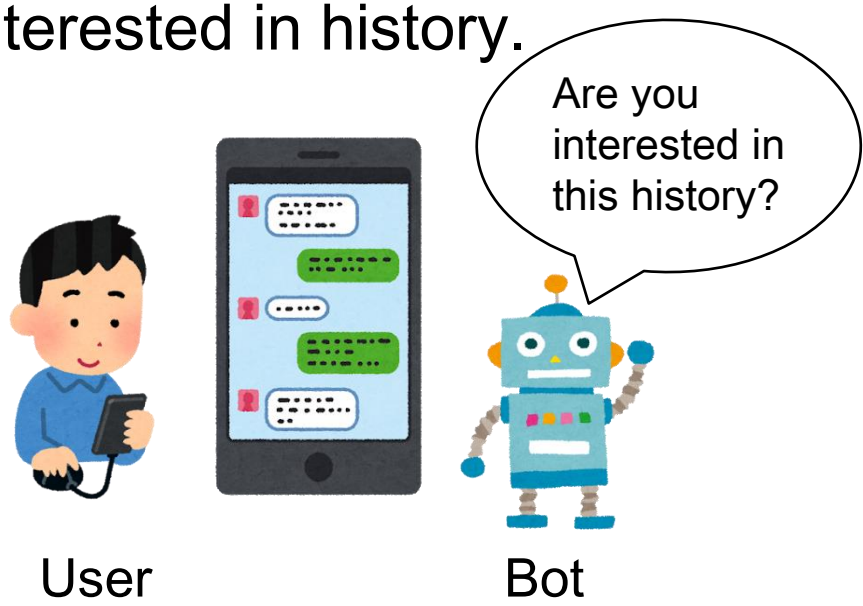
1. Understanding collective memories in Wikipedia, news and social media.

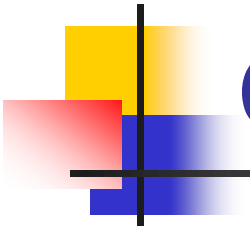
   Understanding what types of entities are commonly shared.

2. Analyzing what each user is interested in history.

Can be useful for

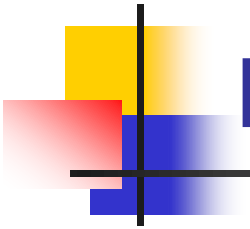
1. Recommendation system
2. History-focused chatbots





Contributions

1. Large scale history-related microblog analysis.
2. Novel findings how collective memories are maintained and formed.
3. Proposal of new categorization of historical references.
4. Outline of novel research directions and potential applications utilizing history-related content.



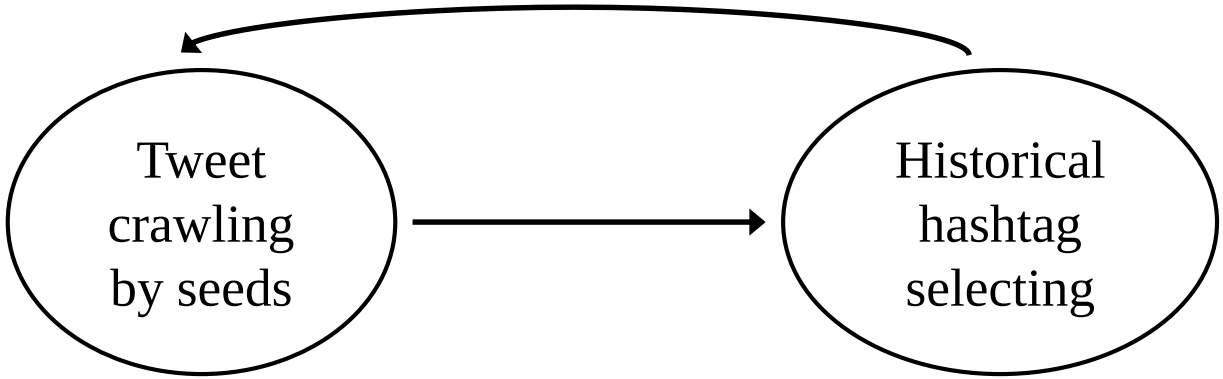
Presentation Schedule

- Data Collection
- General Analysis
  - Temporal Analysis
  - Entity Analysis
- Category Based Analysis
  - Definitions
  - Temporal Category Analysis
- Discussions
- Conclusions

How to Collect History-related Tweets?

- History-related hashtags based crawling.
  - Ex. `#history`, `#WmnHist`, `#HistoryTeacher`, and so on.
  - We use them as *seed* hashtags.

- Bootstrapping applied to increase the coverage
  - Initial seeds: historians' definitions

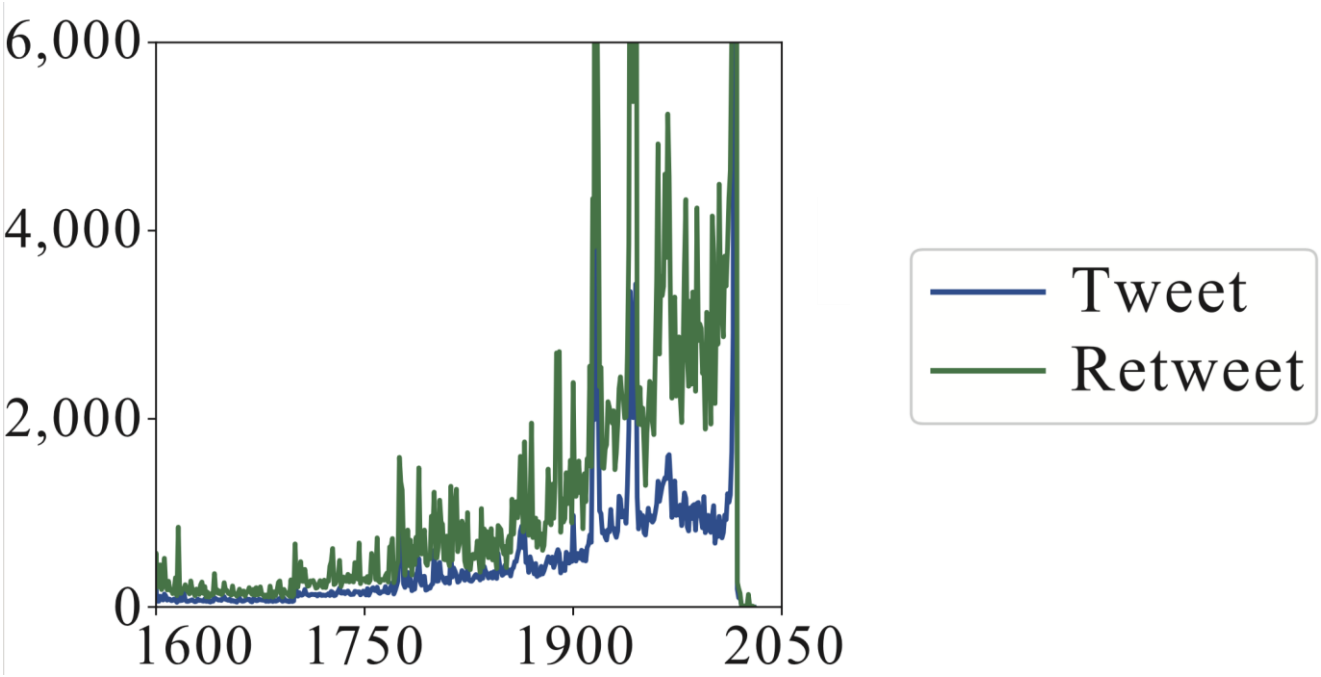


Statistics of Data Collection

Num. of historical hashtags	147
Num. of tweets	888,251
Period of timestamps	8 Mar. 2016 ~ 24 Feb. 2017
Period of time references*	8156 BC ~ 2019

* Years mentioned in tweets

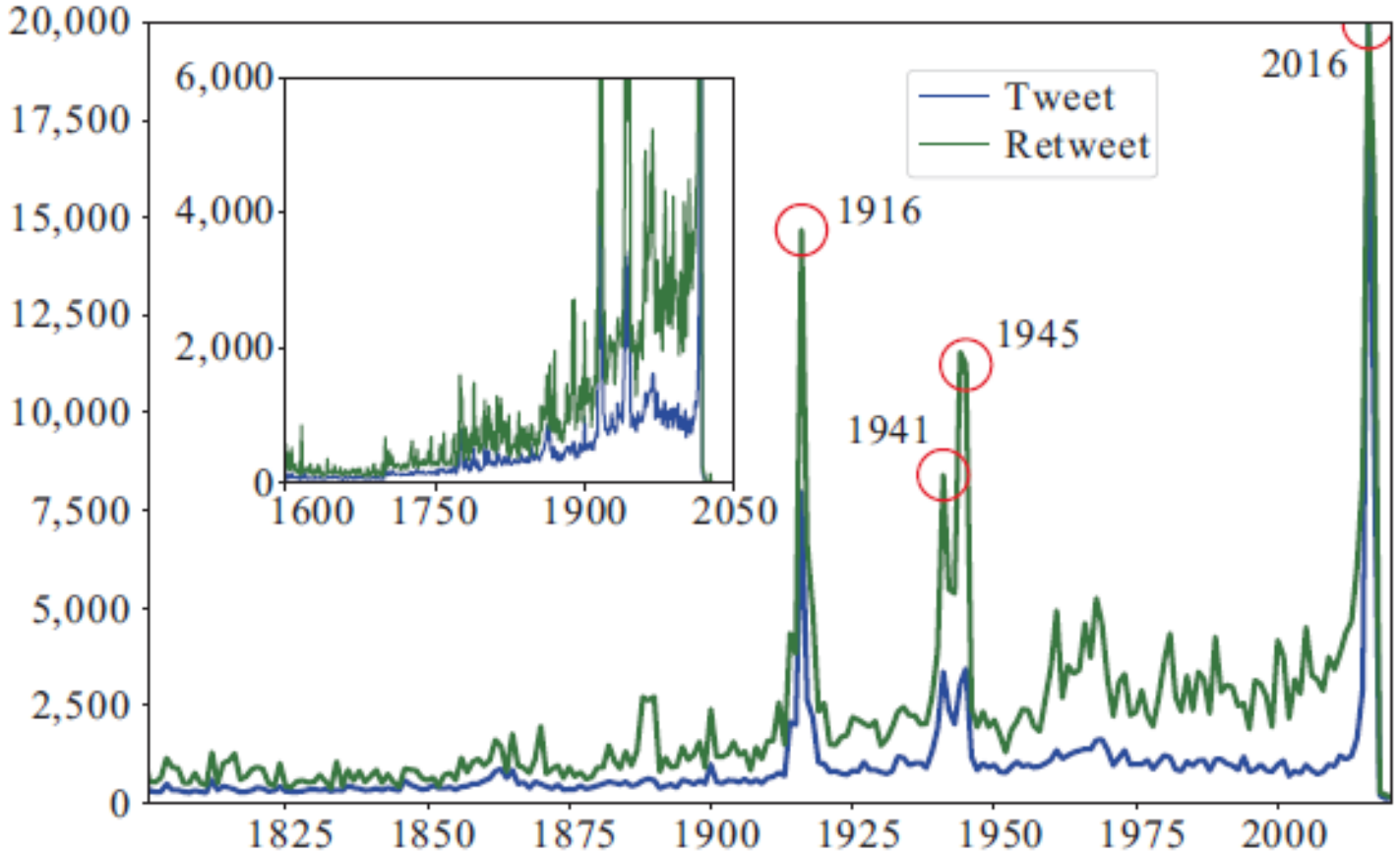
Distribution of Time References



Remembering curve:

1. the recent past is referred to more than the distant past,
2. the memory decay is fastest in the recent years.

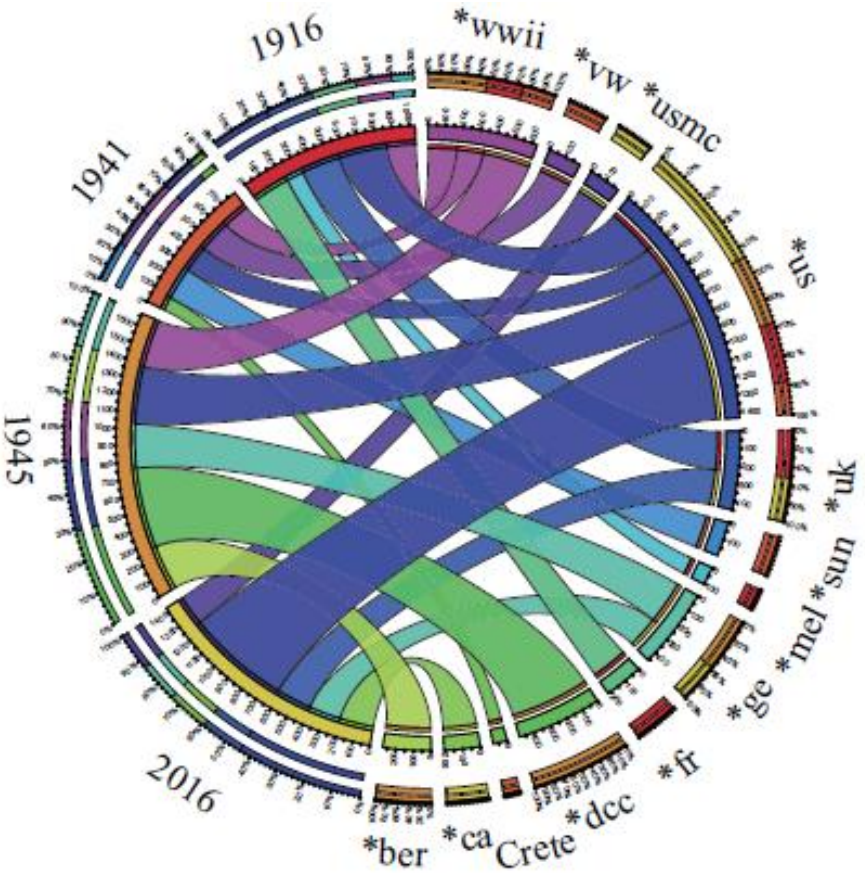
Distribution of Time References



4 large peaks.

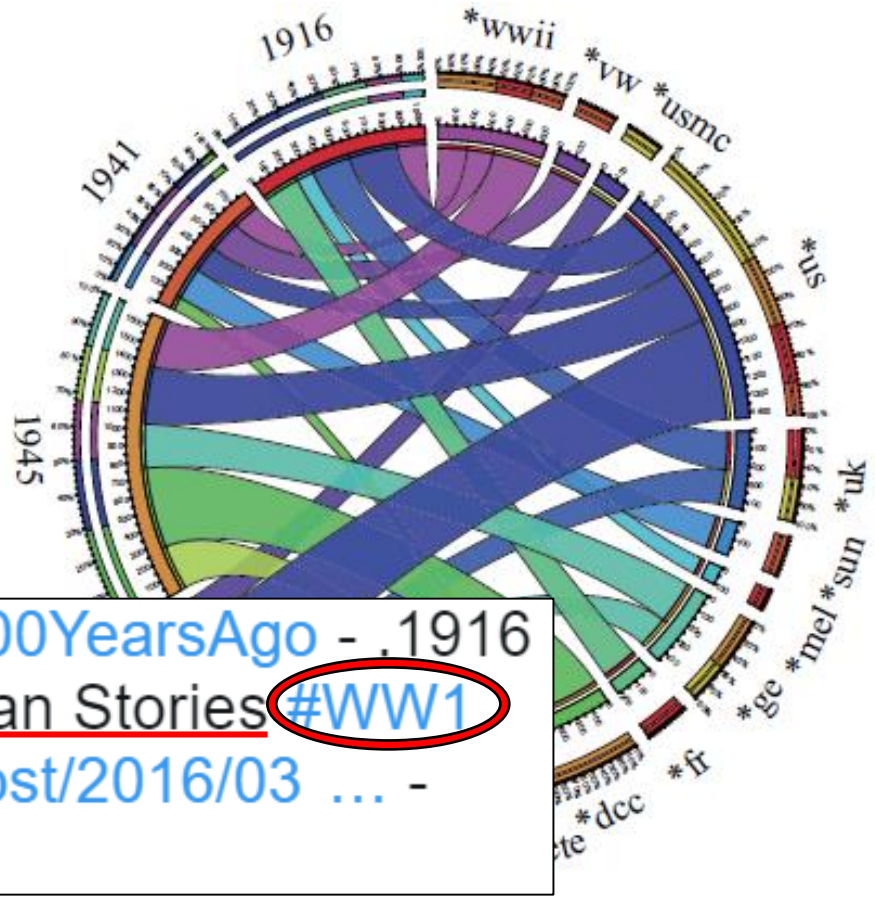
What Do the Peaks Represent?
Top 5 Entities Mentioned identified by AIDA

1. USA is often mentioned.
2. 1916: WWII
3. 1941 and 1945: WWII
4. 2016: 4 countries
   - USA, UK, Germany, Canada



What Do the Peaks Represent?

1. USA is often mentioned.
2. 1916: WWII
   - Talk together WWI and WWII
3. 1941 and 1945: WWII
4. 2
   - #BattleOfTheSomme - #100YearsAgo - .1916 - David Doughty - Australian Stories #WW1 #WW2 ddoughty.com/1/post/2016/03 … - ddoughty.com



* An example of mentioning both of two world wars

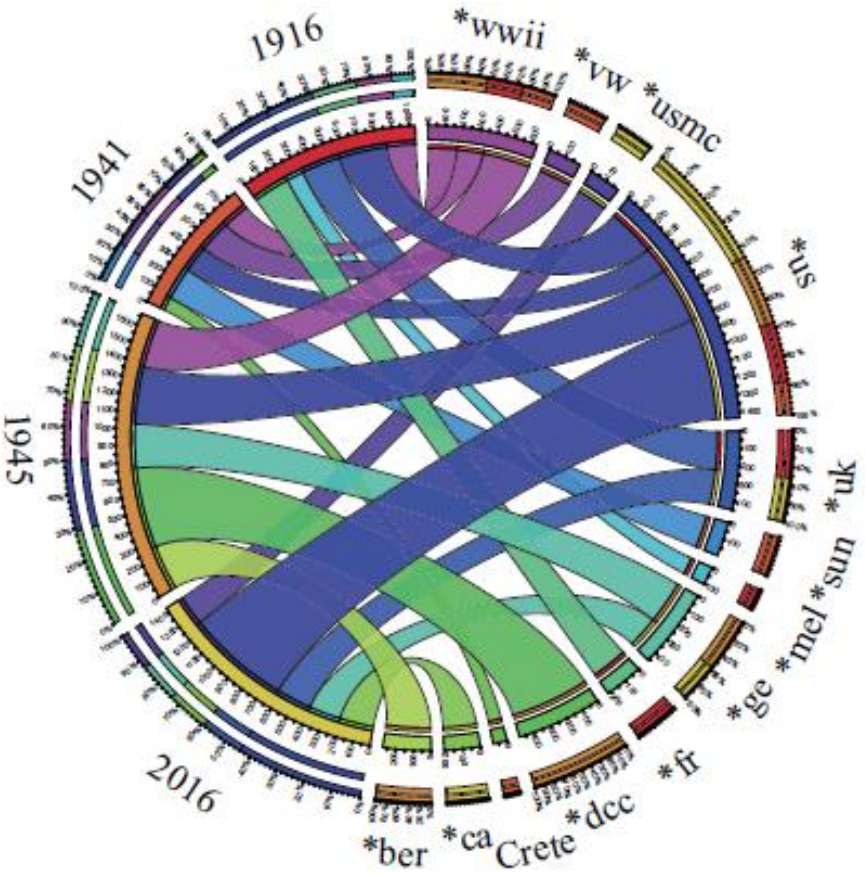
What Do the Peaks Represent?

1. USA is often mentioned

2. 1916: 4 places
   - USA, UK, Melbourne, France

3. 1941 and 1945:
   - 3 and 4 places

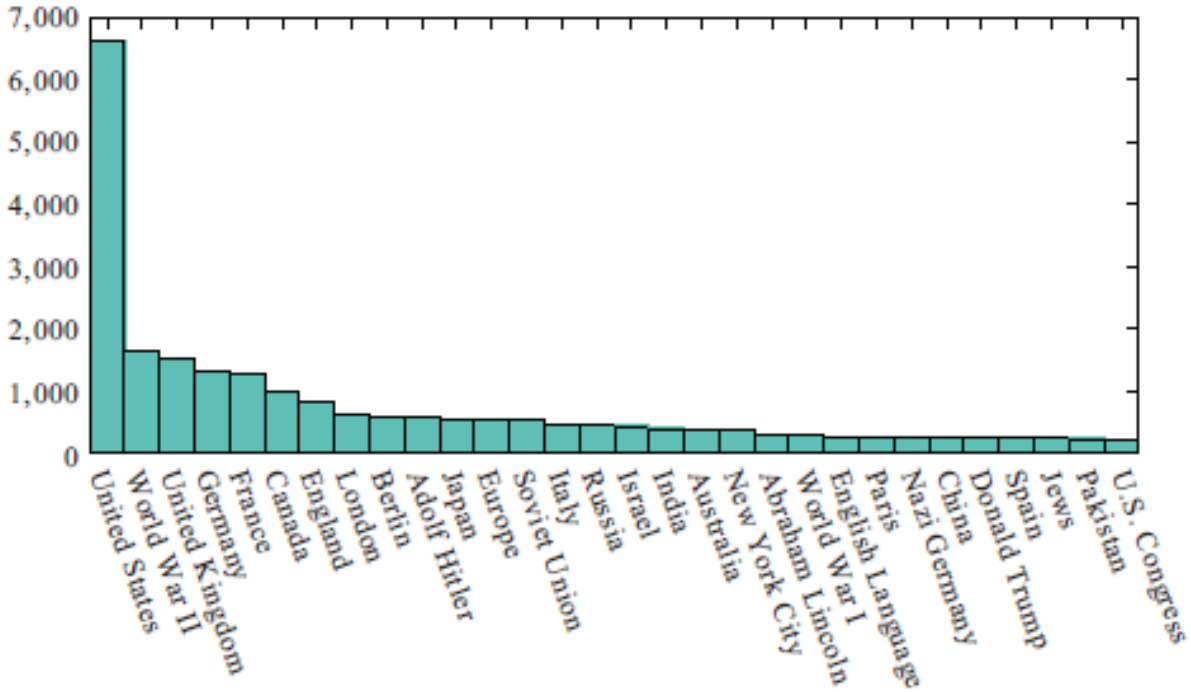
4. 2016: 4 countries
   - USA, UK, Germany, Canada



→ Users often mention places of events occurrence

Top 30 Entities Mentioned in Dataset

- 22 countries
  - USA is the most mentioned entity

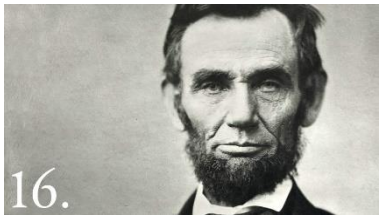


Place as a "bridge" between past and present entities

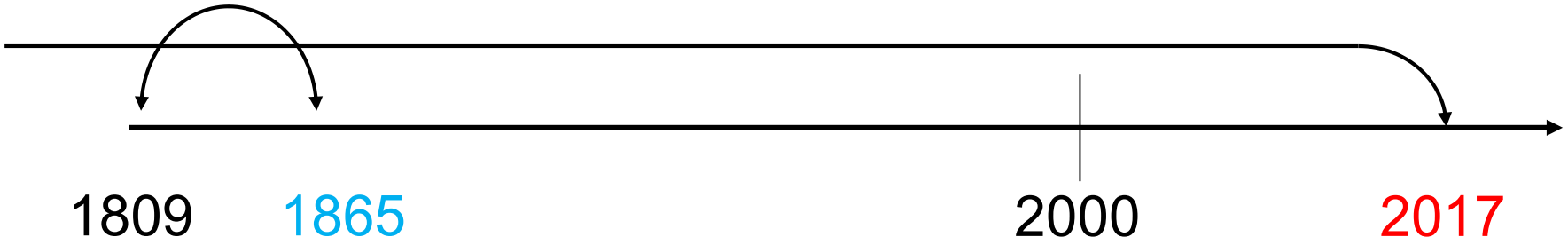
Definitions of Past and Present Entities

- Past: anything with its "end time" before 2000.
  - If end time is not recorded, we set it to 2017.
- Present: entities with their "end time" after 2000

Past person



Present place



Top 3 Entities Co-occurring with Top 5 Mentioned Places

Place	1	2	3
USA	WWII	Battles of Saratoga	Soviet Union
UK	WWII	James II of England	Soviet Union
Germany	Adolf Hitler	John Cudahy	Soviet Union
France	WWII	Napoleon	Louis XIV of France
Canada	WWII	WWI	Dom. Of Newfoundland

Top 3 Entities Co-occurring with Top 5 Mentioned Places

Place	1	2	3
USA	WWII	Battles of Saratoga	Soviet Union
UK	WWII	James II of England	Soviet Union
Germany	Adolf Hitler	John Cudahy	Soviet Union
France	WWII	Napoleon	Louis XIV of France
Canada	WWII	WWI	Dom. Of Newfoundland

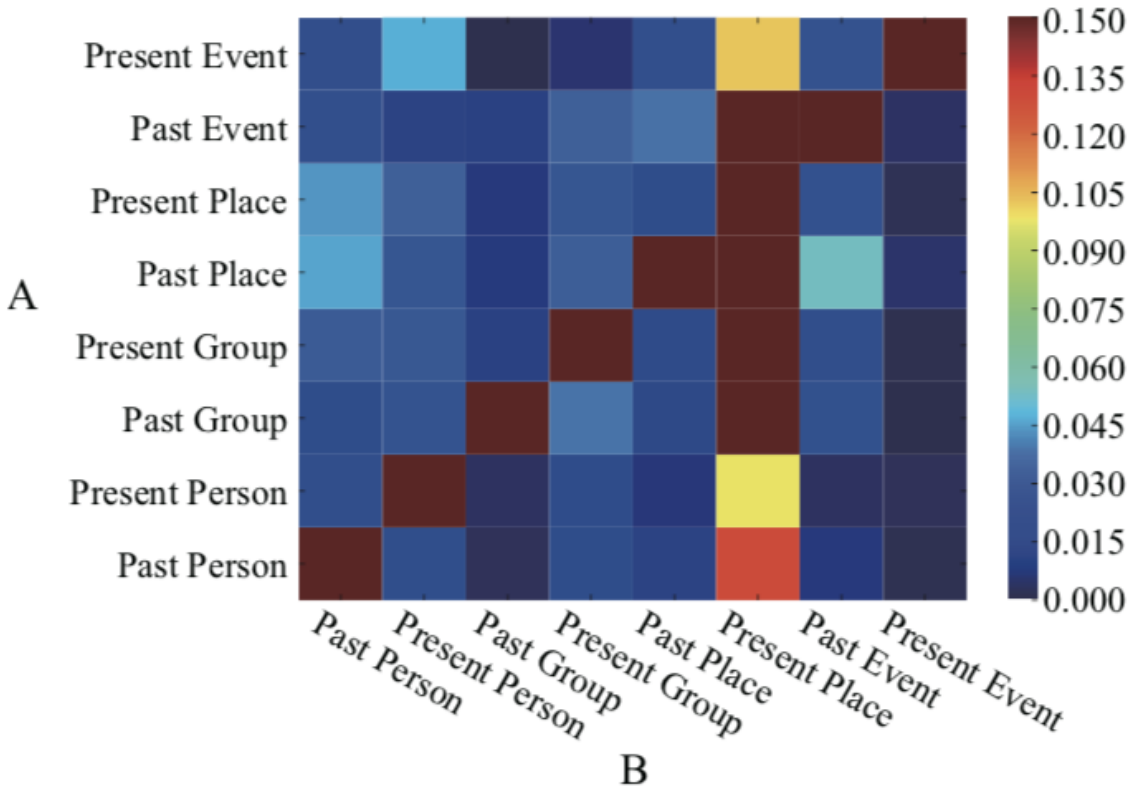
How Often Different Kinds of Entities are Mentioned Together?

- Entity type resource: DBpedia
  - Entity types: place, person, event, group and other.
- Conditional probabilities $P(B|A)$
  - Each cell represents a conditional probability for x axis.

$$P(A) = \frac{|T_A|}{|T|}, \qquad P(B|A) = \frac{P(T_A \wedge T_B)}{P(T_A)}$$

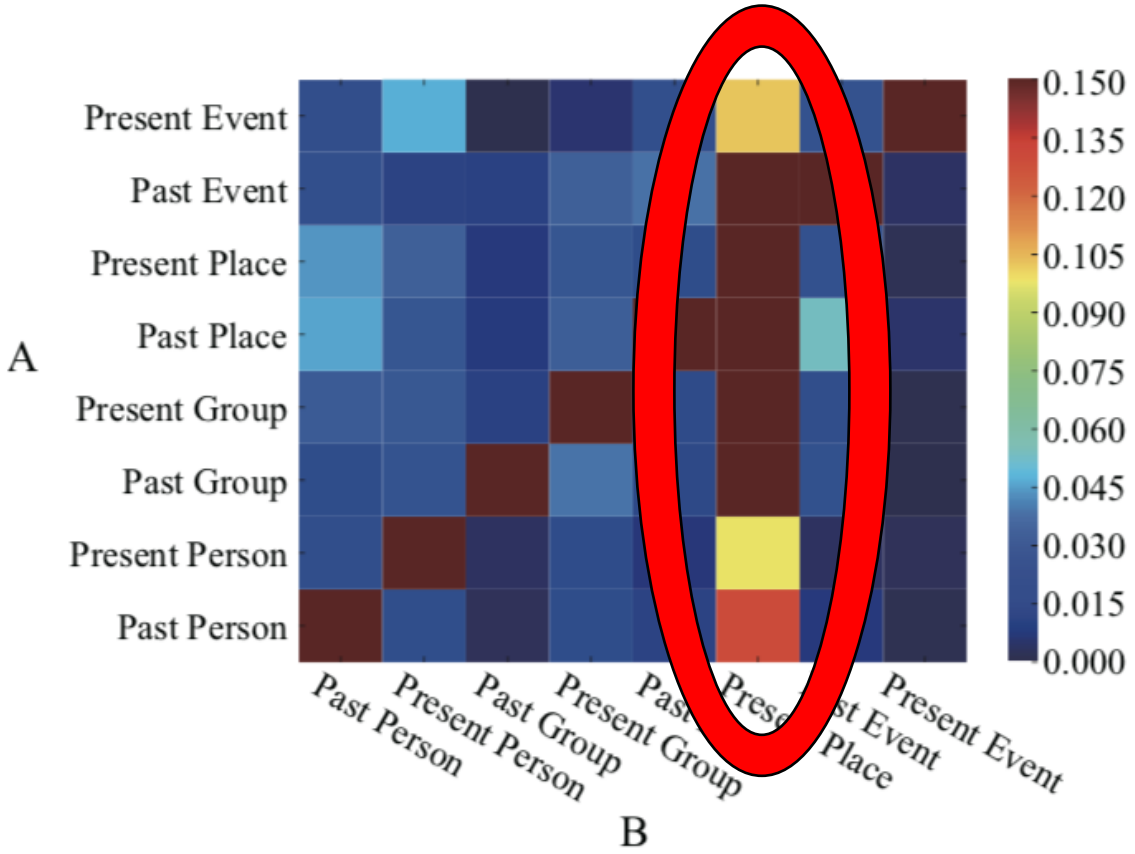
Scores of Conditional Probabilities

Red: higher
Blue: lower



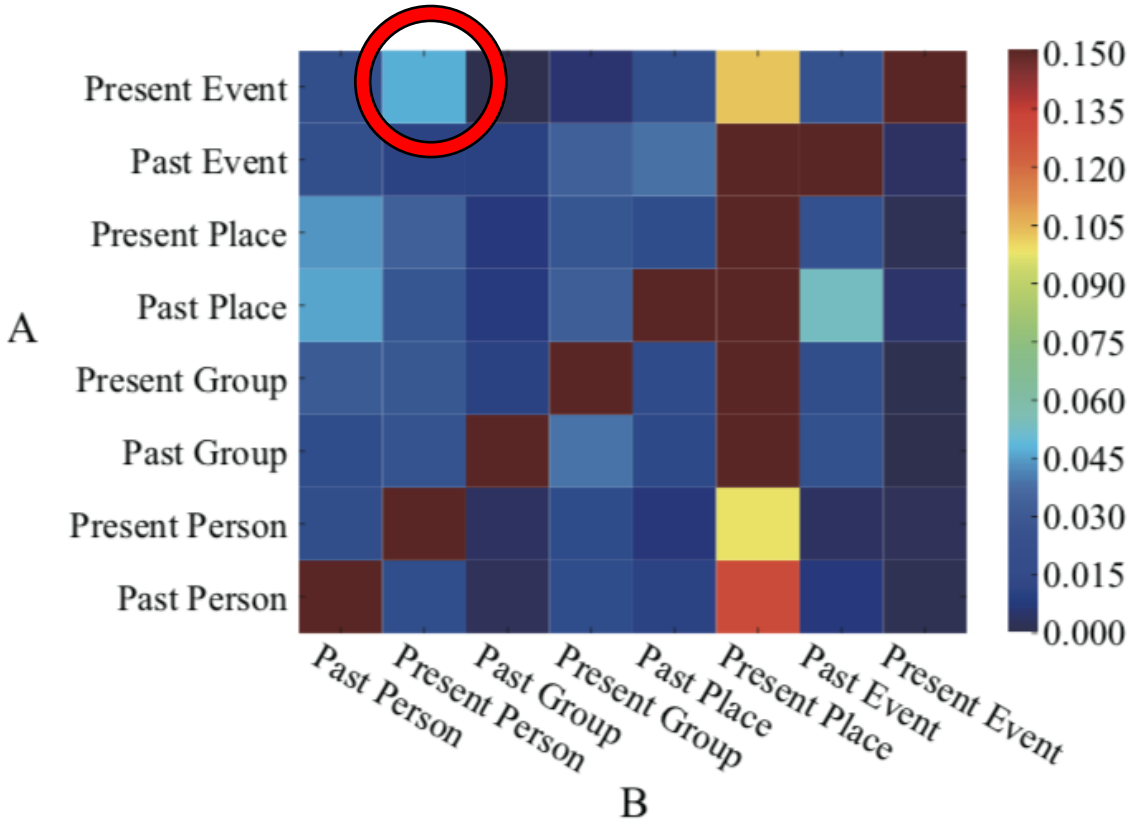
Conditional Probabilities of Present Place

- All are relatively high



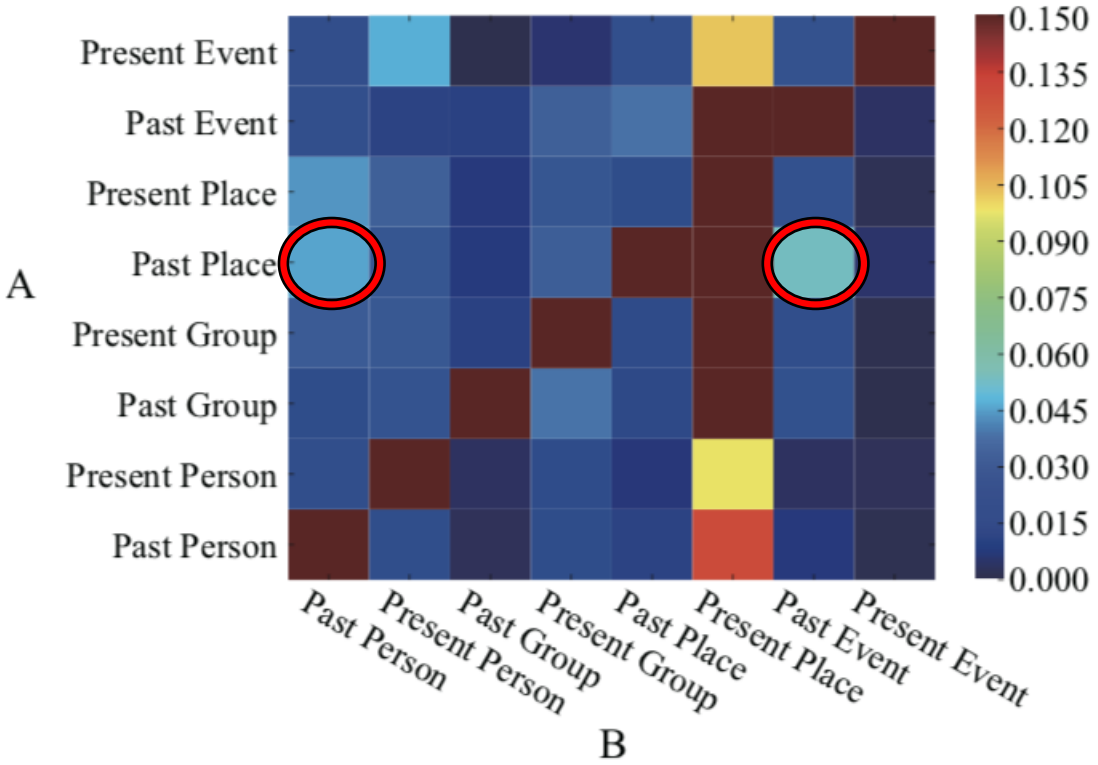
Conditional Probabilities of Present Person

- Present persons tend to be mentioned with present events.



Conditional Probabilities of Past Person and Event

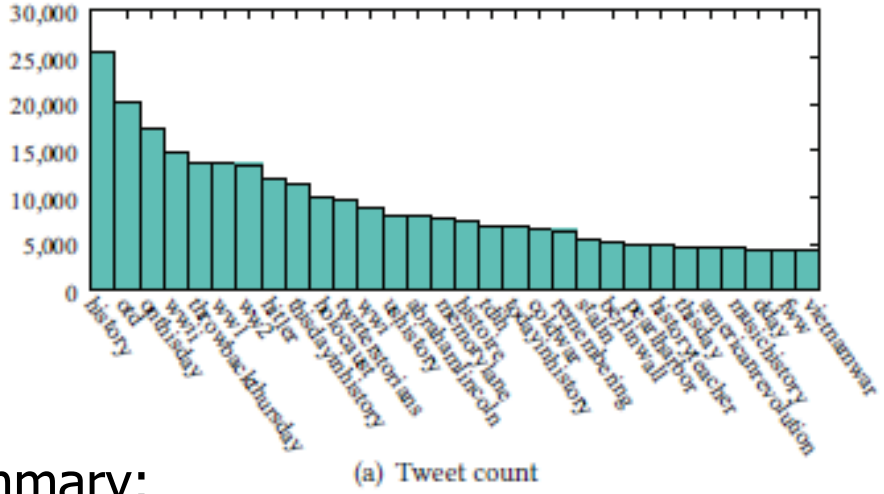
- Past events and persons are mentioned with past places.
  - Examples of past places: Ottoman Empire, Holy Roman Empire



Why Places and Persons are Mentioned?

- Counting number of tweets including hashtags.

1. #history
2. #otd
3. #onthisday
4. #WWII
5. #ThrowBackThursday
6. #WW1
7. #WW2
8. #hitler
9. #ThisDayInHistory
10. #Holocaust



(a) Tweet count

Summary:

4 commemorative days

4 events

1. Locations of event occurrences or historical buildings
2. Areas where people lived

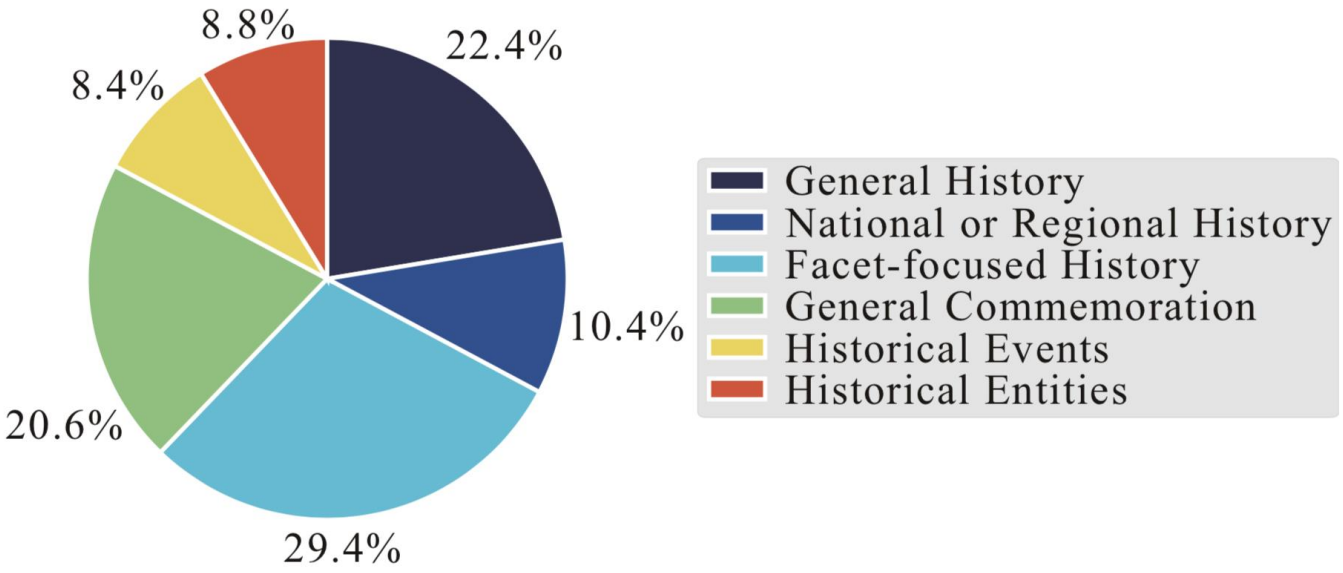
Categorizing Historical References

Hashtag Category	Examples of hashtag
General history (General)	#history, #worldhistory, #historicalcontext
National or regional history (National)	#ushistory, #ottomanempire, #jewishhistory
Facet-focused history (Facet)	#ArtHistory, #DigitalHistory
General commemoration (Comm.)	#OnThisDay, #ThrowbackThursday
Historical events (Events)	#ww1, #ww2, #ColdWarHist
Historical entities (Entities)	#Stalin, #Napoleon, #AbrahamLincoln

Rate of Each Category based on Tweets

Ranking

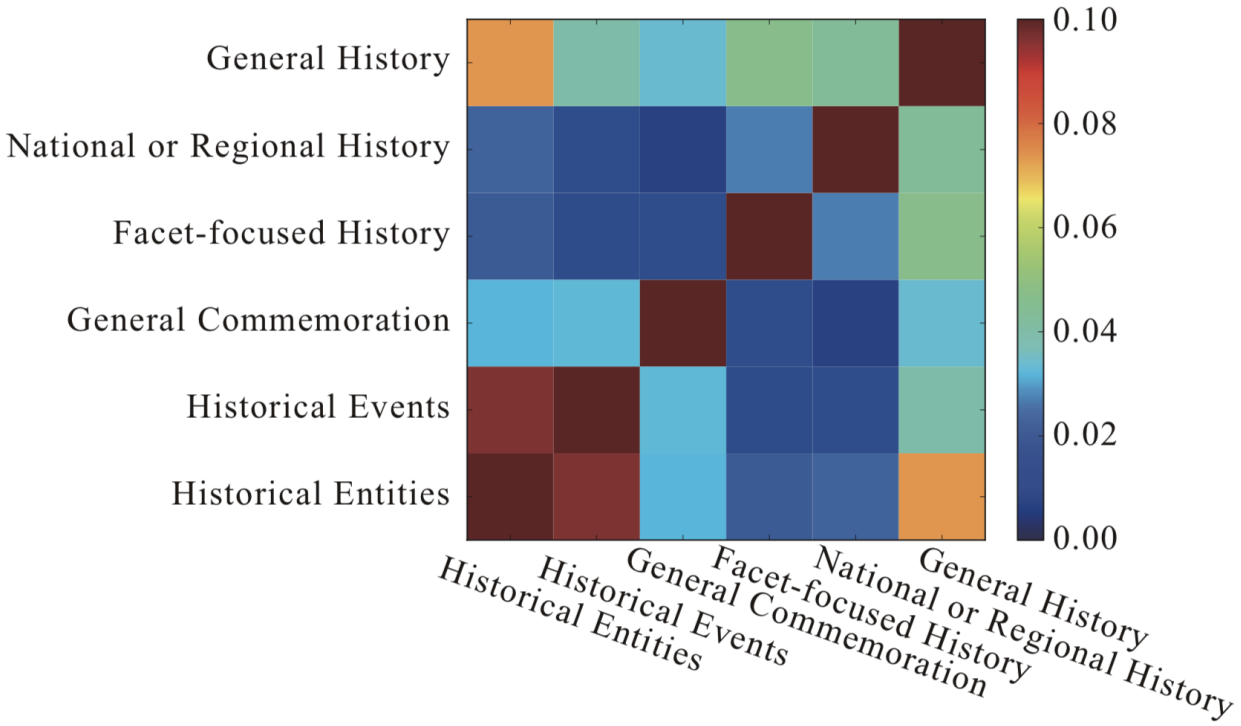
1. Facet
2. General
3. Comm.



Relatively significant amount of specialized history-related content besides broad and general history related content.

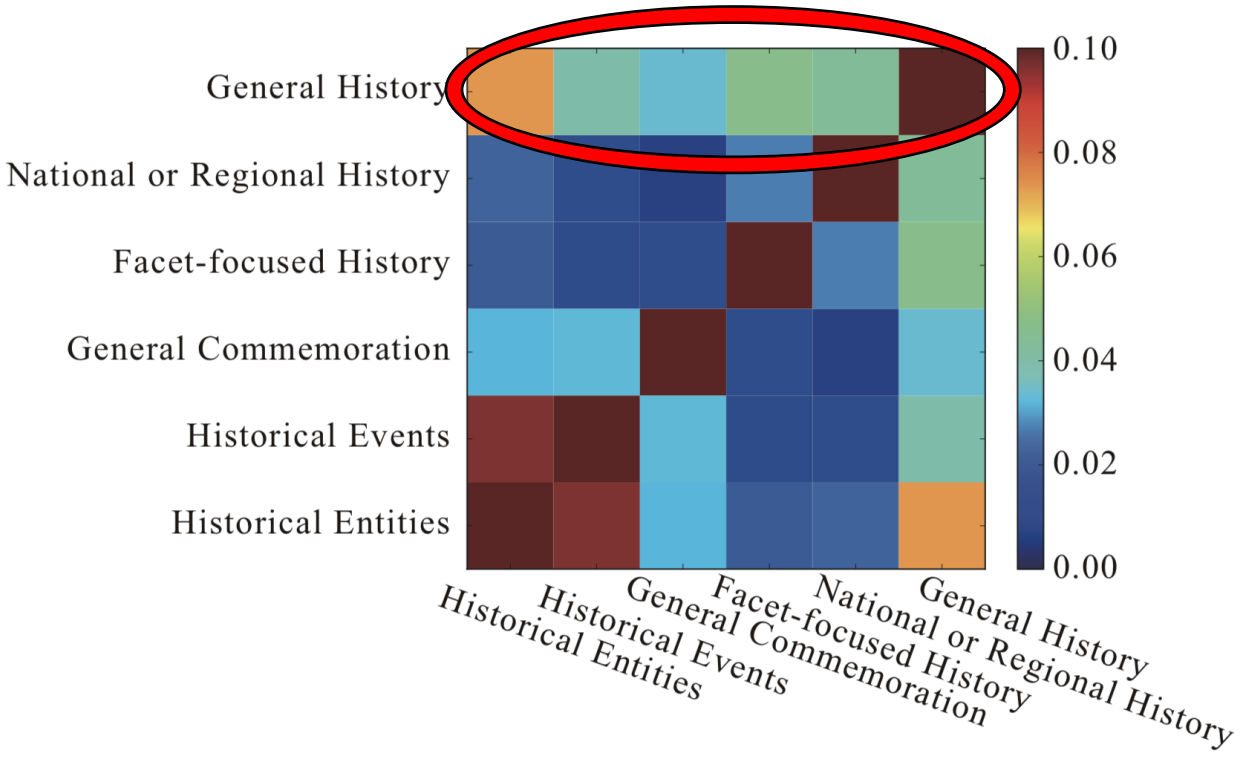
Similarities of the Categories
Co-occurrence of Different Categories

- Measurement: Jaccard coefficient
  - Number of tweets including hashtags in different classes.
- Scores:
  - Red: high
  - Blue: low



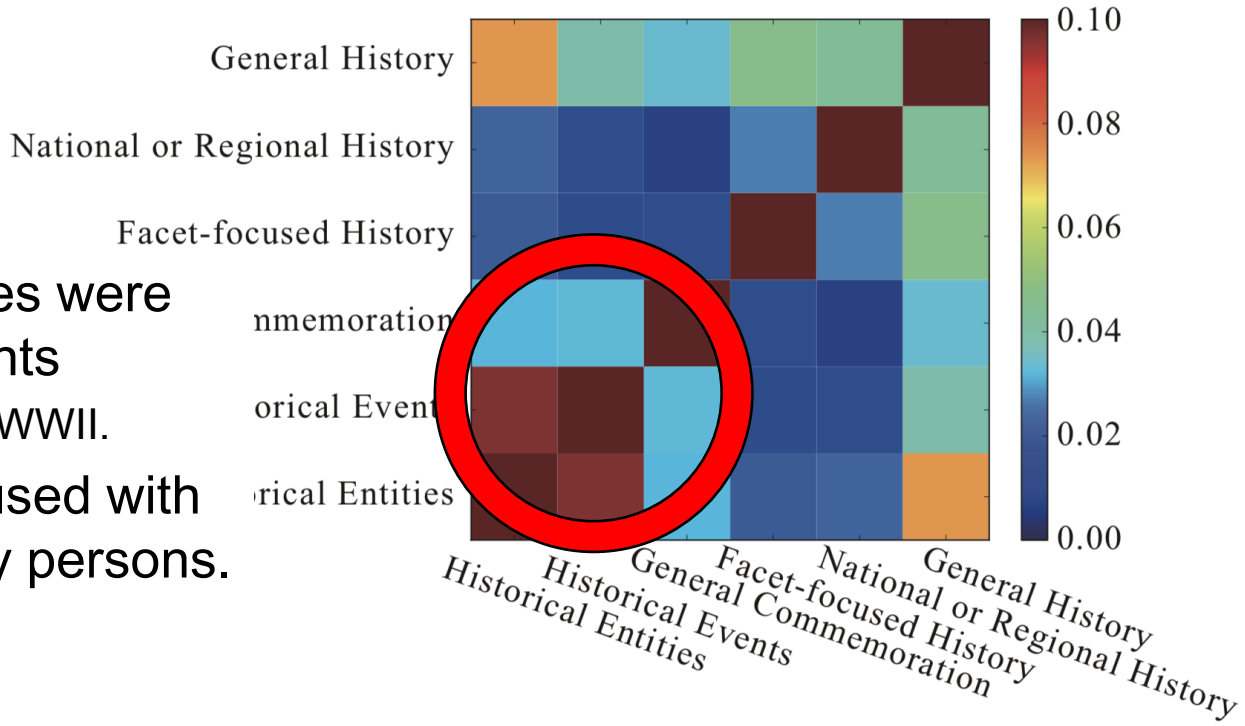
Co-occurrence of Different Categories

- General: truly general as all scores are high



Co-occurrence of Different Categories

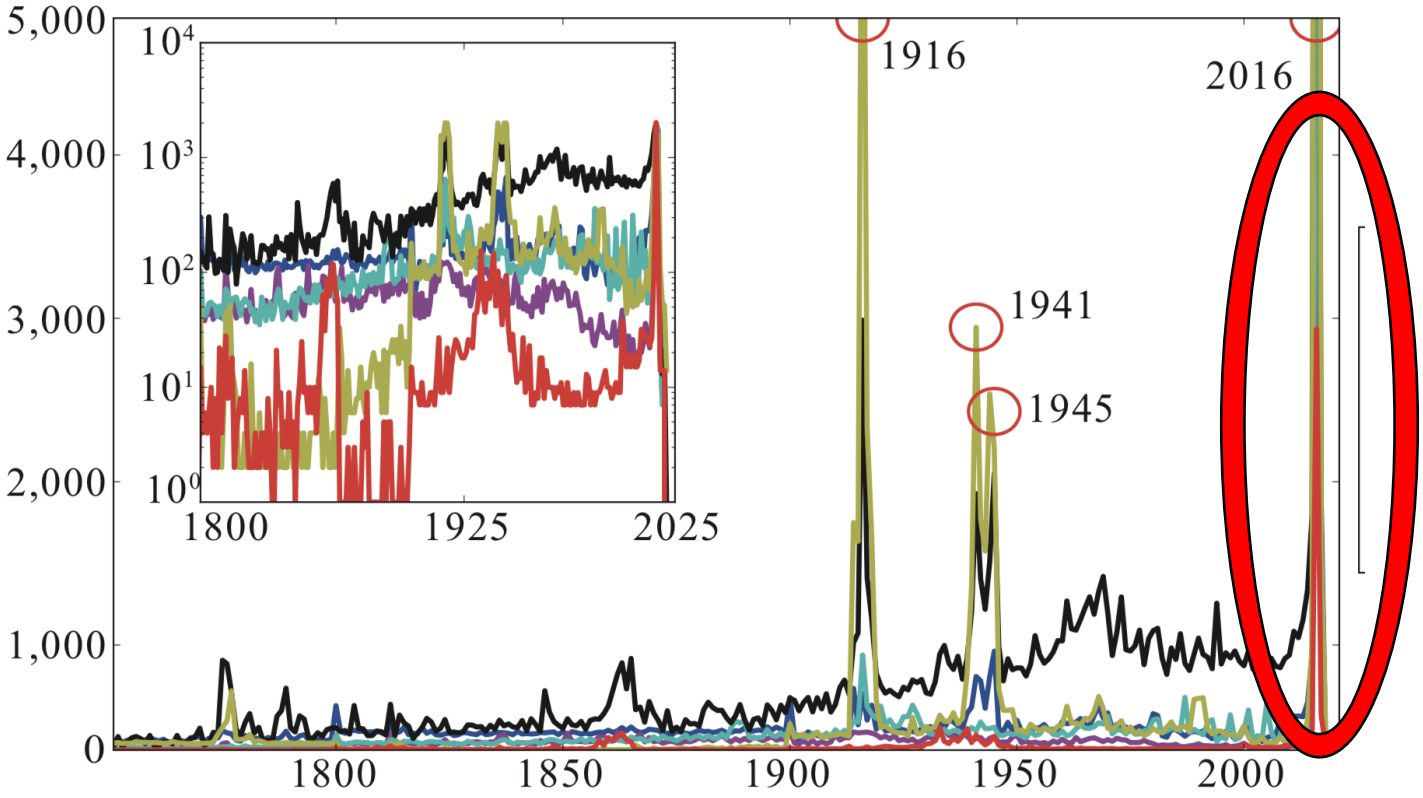
- Entities, Events and Comm.: are sometimes used together



- many famous entities were involved in key events
  - Ex., Stalin, Hitler in WWII.
- Comm. tend to be used with past events and key persons.

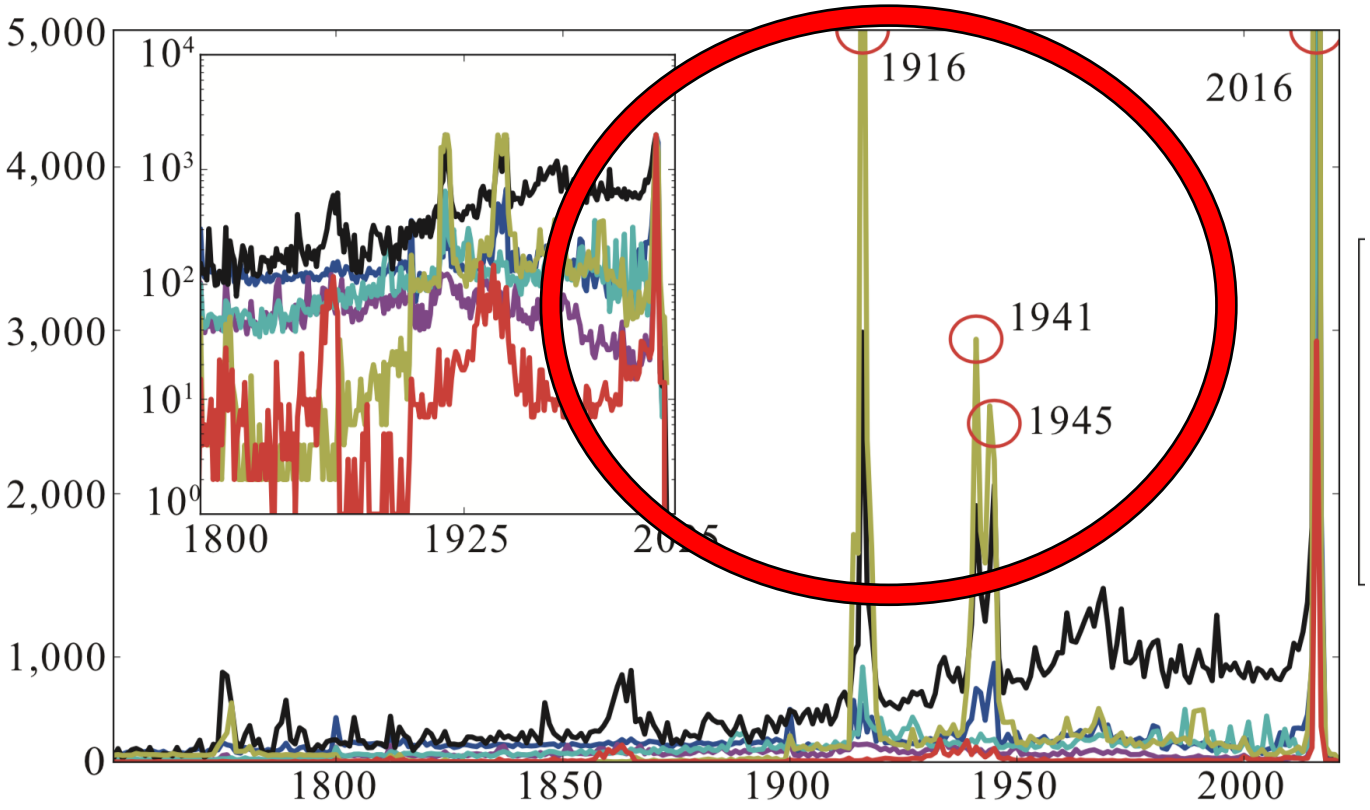
Temporal Category Analysis

- All categories have strong relation to the present



Temporal Category Analysis

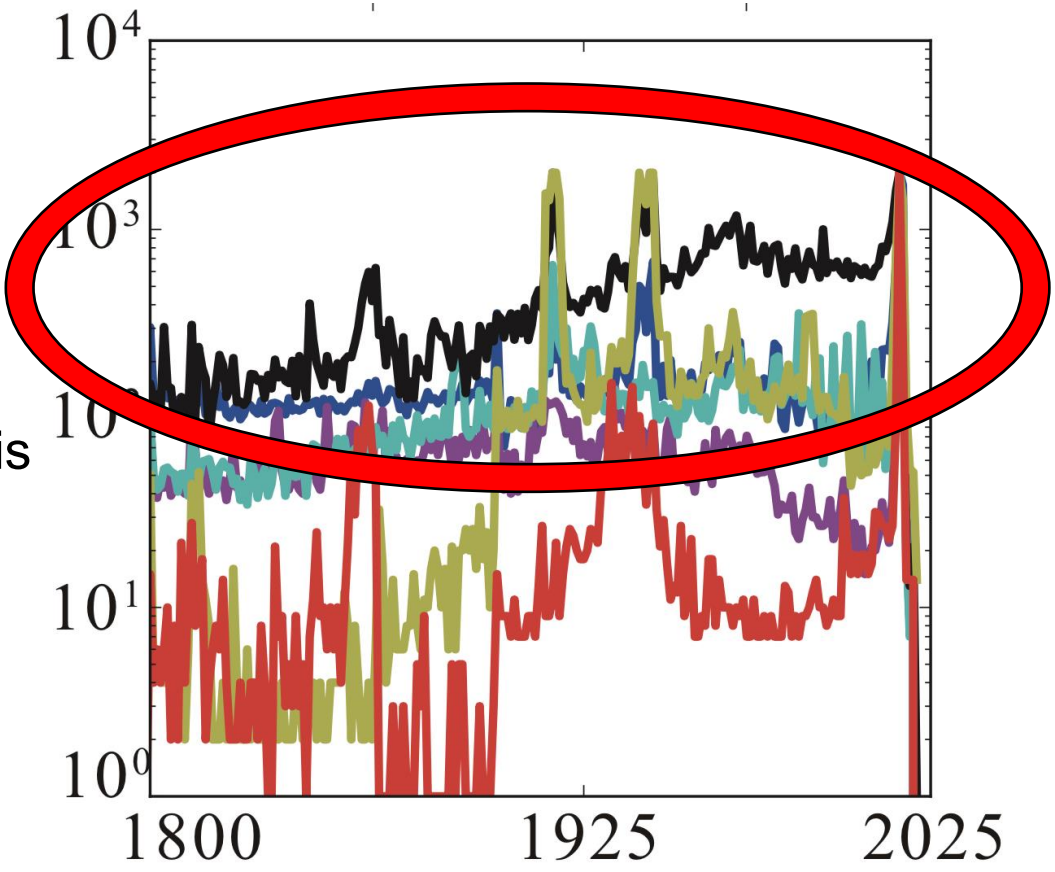
- Event (yellow line): strongly focusing on WWI and WWII



Temporal Category Analysis

- Comm. (black line): many diverse years in the past

Log scale y axis



Discussions | Limitations

1. Data collection
   - Difficult to collect history-related tweets if there is no hashtags.
   - The coverage of hashtags is also limited.

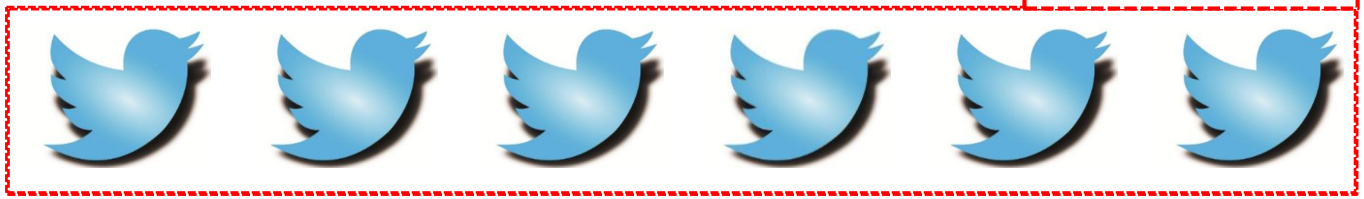
   → Classification (Historical vs. Not tweets) should be done.

2. User-focused analysis
   - Who posts historical tweets, bots, experts or learners?
   - Important for recommendation system.

3. Fine-grained analysis
   - Ex.: analyzing popularizes of entities to identify tendencies.

Discussions | Future Works

This study



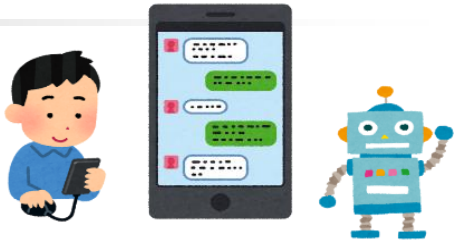
Future Work ①:
More content analysis

Future Work ③: Bridging two data
1. Qualitative approach
2. Educational approach

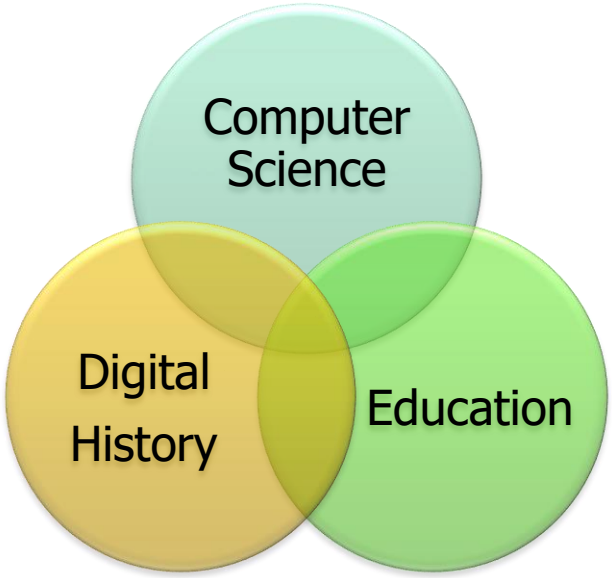
Future Work ②: User analysis

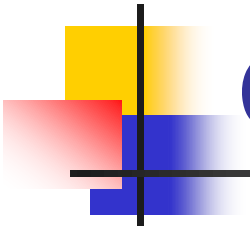
Discussions | Potential Applications

1. Historical contents recommendation



2. Finding, summarizing and explaining past entities
   - This should be useful for analogy determination





Conclusions

- We collected and analyzed history-related tweets.
- Our tweet crawling method was based on historical hashtags.
  - Bootstrapping procedures

- Contributions
  - Large scale history-related microblogging.
  - Findings *how collective memories are maintained and formed*.
  - Proposal *new categorization of historical references*.
  - Outline *novel research directions* and *potential applications utilizing history-related content*.

Thank you!