HPC in the Humanities

How high-performance computing supports and enables humanities (and social science) research

Mark J. Hill, 2 July 2025
mark.j.hill@kcl.ac.uk
https://github.com/markjhill/

Comhis
Helsinki Computational History Group

KING'S College LONDON

Why do I use HPC?

1. Big, messy, data
   - Historical archives going digital (or linked)
   - Social media as cultural record
   - Cross-platform, multi-format sources
   - Resource-Intensive Methods
2. Resource-intensive methods
   - GPU-dependent language models
   - Large-scale network analysis
   - Real-time processing pipelines

But fundamentally: No just 'faster' - the ability to ask fundamentally different questions

My HPC use cases

1. HPC as **Convenience**
    - Could do on desktop, but HPC is more practical
    - Example: Running RStudio remotely while traveling
2. HPC as **Enabler**
    - Technically possible locally, but HPC makes it feasible
    - Example: Parallel analyses on millions of documents
    - Example: The ability to re-run computationally intensive scripts iteratively
3. HPC as **Necessity**
    - Simply cannot be done without HPC resources
    - Example: Processing data that won't fit into memory
    - Example: Using GPUs for LLM tasks (sentiment analysis, toxicity detection)

These categories are NOT mutually exclusive

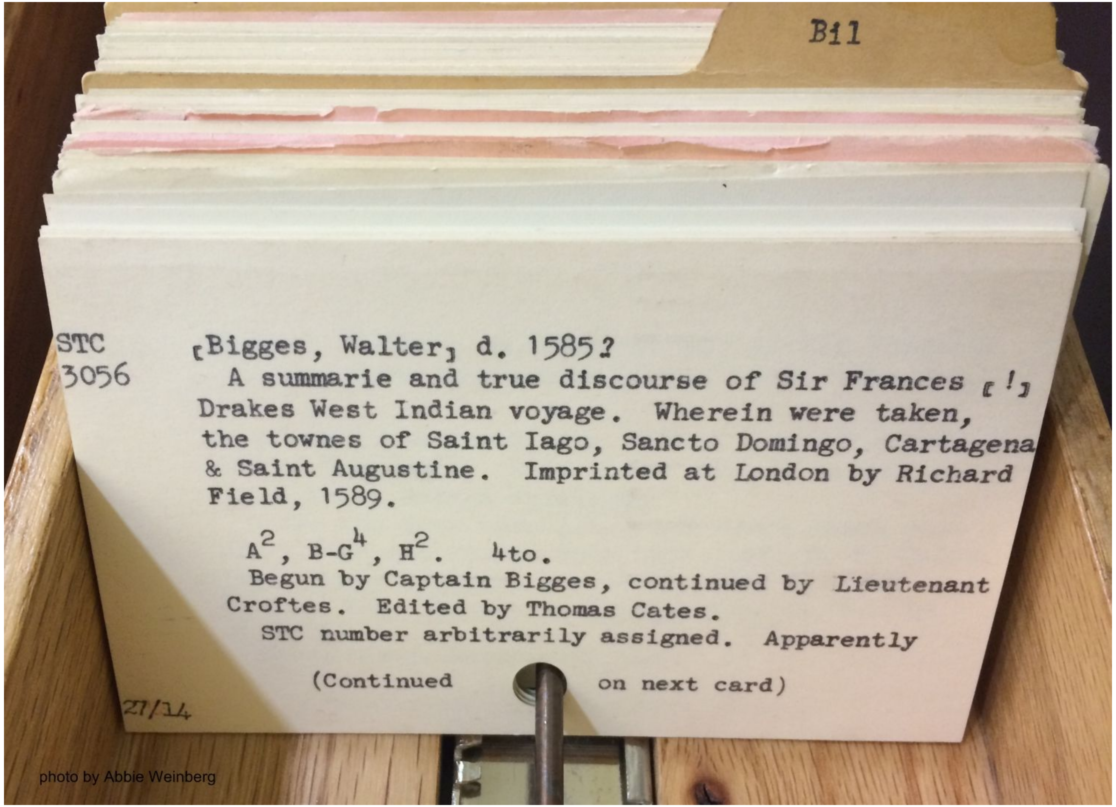
Example 1: ESTC

ENGLISH
SHORT
TITLE
CATALOGUE

BRITISH
LIBRARY

English Short Title Catalogue (ESTC)

- Bibliographic database
- Chronologically, its scope extends from the earliest printed work in British Isles (ca. 1473) through the last item printed in 1800
- Geographically:
  - British Isles
  - North America
  - British governed territories
  - Items printed in English, any part of the world
- Held by over 2000 institutions in North America, the United Kingdom, Europe, Australia and New Zealand
- 483,331 documents



STC
3056

[Bigges, Walter] d. 1585?
A summarie and true discourse of Sir Frances [!] Drakes West Indian voyage. Wherein were taken, the townes of Saint Iago, Sancto Domingo, Cartagena & Saint Augustine. Imprinted at London by Richard Field, 1589.

A^2, $B\text{-}G^4$, H^2. 4to.

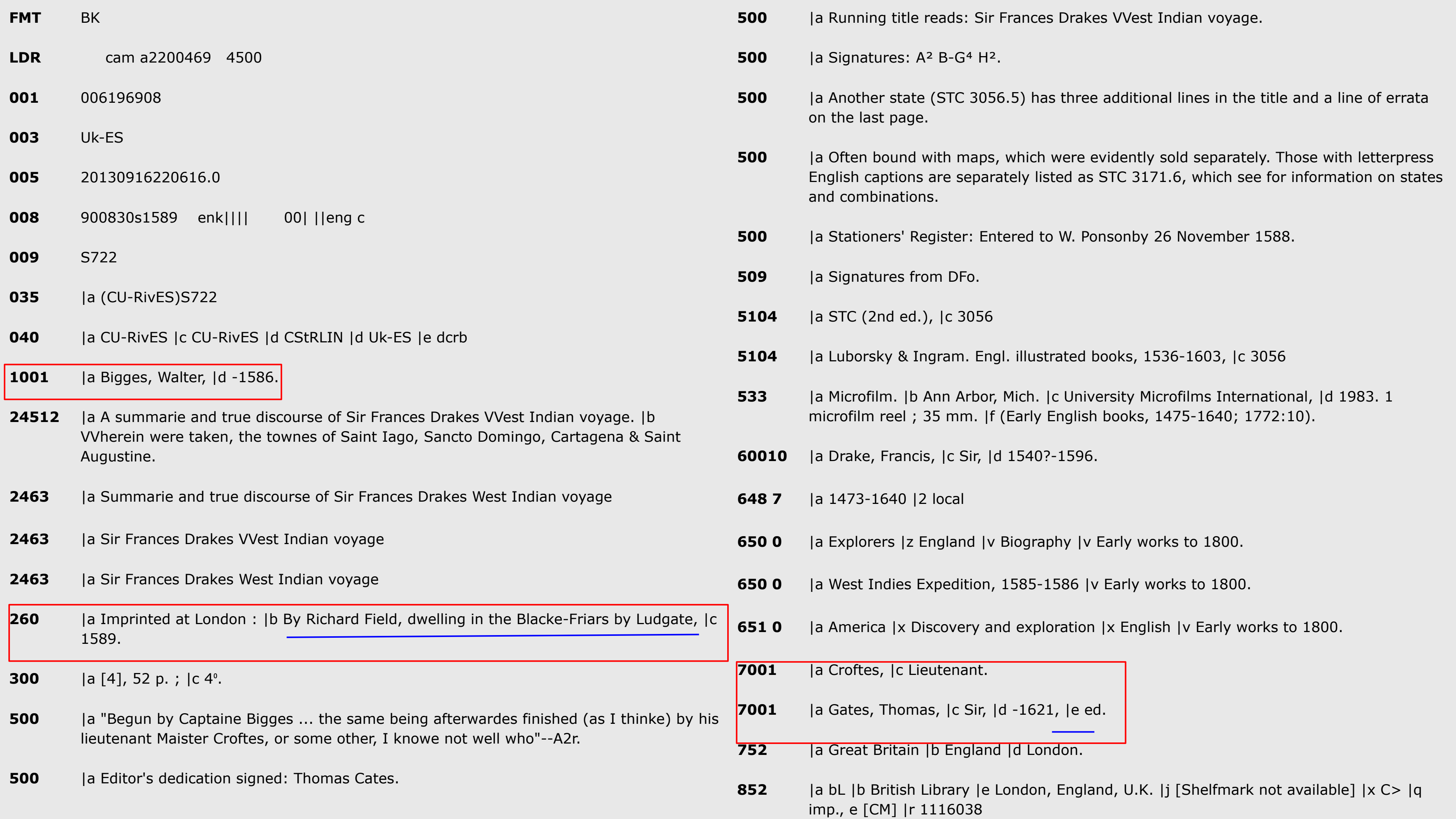
Begun by Captain Bigges, continued by Lieutenant Croftes. Edited by Thomas Cates.

STC number arbitrarily assigned. Apparently

(Continued on next card)

27/14

photo by Abbie Weinberg



```
FMT     BK

LDR        cam a2200469   4500

001     006196908

003     Uk-ES

005     20130916220616.0

008     900830s1589    enk||||       00| ||eng c

009     S722

035     |a (CU-RivES)S722

040     |a CU-RivES |c CU-RivES |d CStRLIN |d Uk-ES |e dcrb

1001    |a Bigges, Walter, |d -1586.

24512   |a A summarie and true discourse of Sir Frances Drakes VVest Indian voyage. |b
        VVherein were taken, the townes of Saint Iago, Sancto Domingo, Cartagena & Saint
        Augustine.

2463    |a Summarie and true discourse of Sir Frances Drakes West Indian voyage

2463    |a Sir Frances Drakes VVest Indian voyage

2463    |a Sir Frances Drakes West Indian voyage

260     |a Imprinted at London : |b By Richard Field, dwelling in the Blacke-Friars by Ludgate, |c
        1589.

300     |a [4], 52 p. ; |c 4º.

500     |a "Begun by Captaine Bigges ... the same being afterwardes finished (as I thinke) by his
        lieutenant Maister Croftes, or some other, I knowe not well who"--A2r.

500     |a Editor's dedication signed: Thomas Cates.

500     |a Running title reads: Sir Frances Drakes VVest Indian voyage.

500     |a Signatures: A² B-G⁴ H².

500     |a Another state (STC 3056.5) has three additional lines in the title and a line of errata
        on the last page.

500     |a Often bound with maps, which were evidently sold separately. Those with letterpress
        English captions are separately listed as STC 3171.6, which see for information on states
        and combinations.

500     |a Stationers' Register: Entered to W. Ponsonby 26 November 1588.

509     |a Signatures from DFo.

5104    |a STC (2nd ed.), |c 3056

5104    |a Luborsky & Ingram. Engl. illustrated books, 1536-1603, |c 3056

533     |a Microfilm. |b Ann Arbor, Mich. |c University Microfilms International, |d 1983. 1
        microfilm reel ; 35 mm. |f (Early English books, 1475-1640; 1772:10).

60010   |a Drake, Francis, |c Sir, |d 1540?-1596.

648 7   |a 1473-1640 |2 local

650 0   |a Explorers |z England |v Biography |v Early works to 1800.

650 0   |a West Indies Expedition, 1585-1586 |v Early works to 1800.

651 0   |a America |x Discovery and exploration |x English |v Early works to 1800.

7001    |a Croftes, |c Lieutenant.

7001    |a Gates, Thomas, |c Sir, |d -1621, |e ed.

752     |a Great Britain |b England |d London.

852     |a bL |b British Library |e London, England, U.K. |j [Shelfmark not available] |x C> |q
        imp., e [CM] |r 1116038
```

Actor Fields (100, 110, 700, 710)

- Extracted 557,847 actors from 397,061 documents (for which there were named actors)
- Cleaned up and standardized unicode.
- Created individual actor records per document.
- Assigned roles when known.
- Harmonized by string matching (when appropriate) and with **Virtual International Authority File (VIAF)**
  - Problems: VIAF often has duplicate records; single records are clearly for multiple individuals, IDs change.
- **558,243 actor records harmonized into 92.044 unique actors.**

Imprint Field (260)

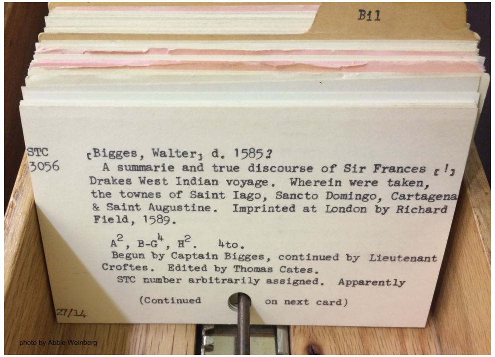
- Extracted named entities from ESTC field 260 (imprint/publisher statement)
  - **printed for Bernard Lintott at the Cross-Keys, between the two Temple-Gates, in Fleet-Street. The Double Gallant: Or, the Sick Lady's Cure. A Comedy. Written by Mr. Cibber**
  - **printed by E: Coates. 1655. Sould by Thomas Heath in Covent garden, and Henry Herringman at the Ancker on the lowest side of the New-Exchange.**
- On above used **Stanford NLP Parser** to:
  - Assign roles (publisher, printer, bookseller) and addresses
  - Corrected and enriched names
  - Using town, address, matching initials and name, name combinations, years of activity, etc, harmonized and expand on existing named entities.
  - Verified against BBTI (British Book Trade Index), VIAF, and our own data.
- **332,410 named actors in tag 260 unified as 35,252 unique actors..**

ESTC as historic, quantitative, and network data source

Extracted roughly 1,000,000 actors from the 400k+ documents for which there are named actors.

Got those actors through time consuming harmonisation process down to 142,407 unique actors.

Create a graph (social network) of those actors where the document they worked on (authored, published, edited) links them to other historical actors.



B11

STC
3056

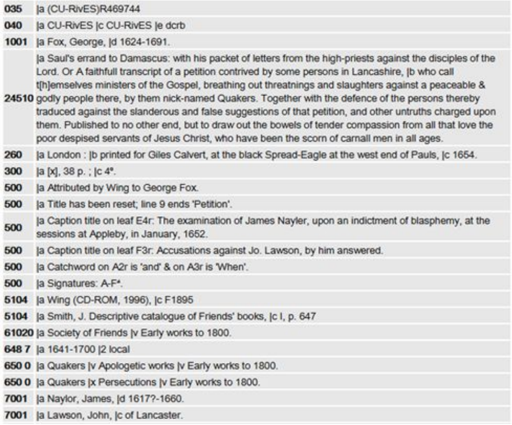
[Bigges, Walter] d. 1585?
A summarie and true discourse of Sir Frances [!]
Drakes West Indian voyage. Wherein were taken,
the townes of Saint Iago, Sancto Domingo, Cartagena
& Saint Augustine. Imprinted at London by Richard
Field, 1589.

A^2, B-G^4, H^2. 4to.
Begun by Captain Bigges, continued by Lieutenant
Croftes. Edited by Thomas Cates.
STC number arbitrarily assigned. Apparently

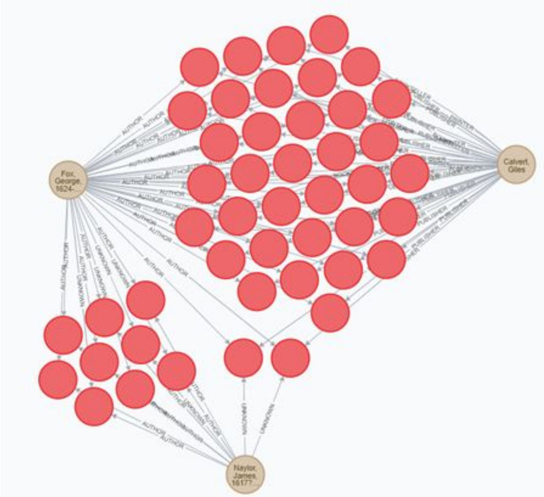
(Continued on next card)

27/14

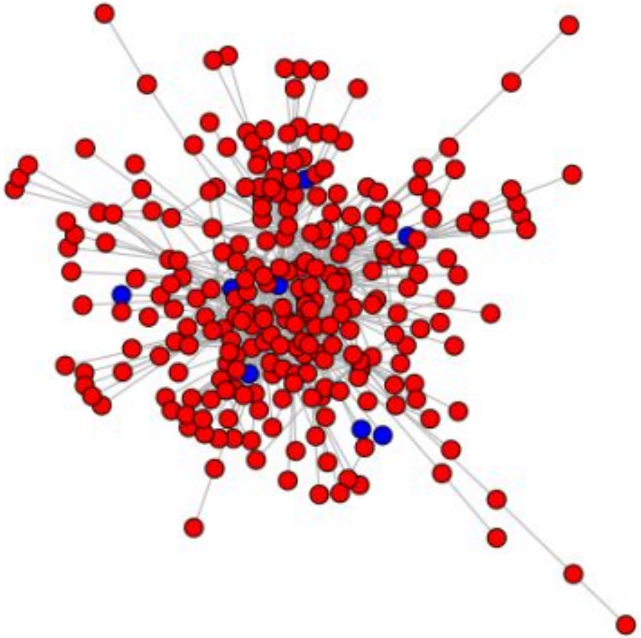
photo by Abbie Weinberg



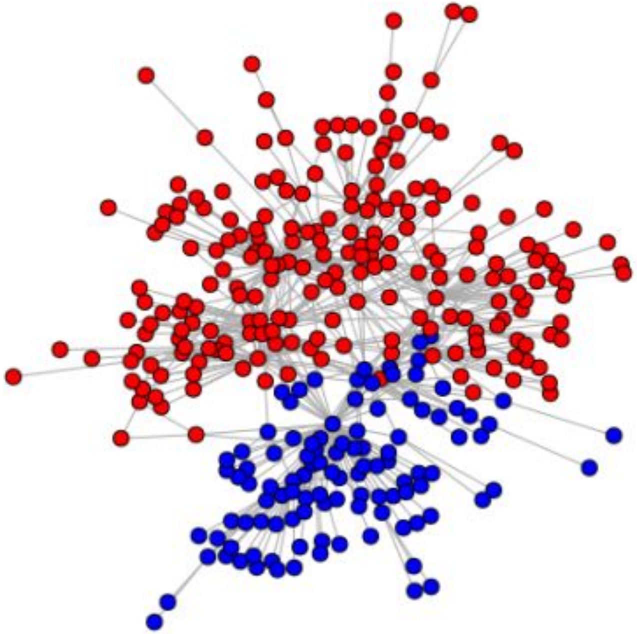
<table>
<tr><td>035</td><td>|a (CU-RivES)R469744</td></tr>
<tr><td>040</td><td>|a CU-RivES |c CU-RivES |e dcrb</td></tr>
<tr><td>1001</td><td>|a Fox, George, |d 1624-1691.</td></tr>
<tr><td>24510</td><td>|a Saul's errand to Damascus: with his packet of letters from the high-priests against the disciples of the Lord. Or A faithfull transcript of a petition contrived by some persons in Lancashire, |b who call t[h]emselves ministers of the Gospel, breathing out threatnings and slaughters against a peaceable &amp; godly people there, by them nick-named Quakers. Together with the defence of the persons thereby traduced against the slanderous and false suggestions of that petition, and other untruths charged upon them. Published to no other end, but to draw out the bowels of tender compassion from all that love the poor despised servants of Jesus Christ, who have been the scorn of carnall men in all ages.</td></tr>
<tr><td>260</td><td>|a London : |b printed for Giles Calvert, at the black Spread-Eagle at the west end of Pauls, |c 1654.</td></tr>
<tr><td>300</td><td>|a [x], 38 p. ; |c 4°.</td></tr>
<tr><td>500</td><td>|a Attributed by Wing to George Fox.</td></tr>
<tr><td>500</td><td>|a Title has been reset; line 9 ends 'Petition'.</td></tr>
<tr><td>500</td><td>|a Caption title on leaf E4r: The examination of James Nayler, upon an indictment of blasphemy, at the sessions at Appleby, in January, 1652.</td></tr>
<tr><td>500</td><td>|a Caption title on leaf F3r: Accusations against Jo. Lawson, by him answered.</td></tr>
<tr><td>500</td><td>|a Catchword on A2r is 'and' &amp; on A3r is 'When'.</td></tr>
<tr><td>500</td><td>|a Signatures: A-F⁴.</td></tr>
<tr><td>5104</td><td>|a Wing (CD-ROM, 1996), |c F1895</td></tr>
<tr><td>5104</td><td>|a Smith, J. Descriptive catalogue of Friends' books, |c I, p. 647</td></tr>
<tr><td>61020</td><td>|a Society of Friends |v Early works to 1800.</td></tr>
<tr><td>648 7</td><td>|a 1641-1700 |2 local</td></tr>
<tr><td>650 0</td><td>|a Quakers |v Apologetic works |v Early works to 1800.</td></tr>
<tr><td>650 0</td><td>|a Quakers |x Persecutions |v Early works to 1800.</td></tr>
<tr><td>7001</td><td>|a Naylor, James, |d 1617?-1660.</td></tr>
<tr><td>7001</td><td>|a Lawson, John, |c of Lancaster.</td></tr>
</table>



Community split (1645-1659)



1645-1654

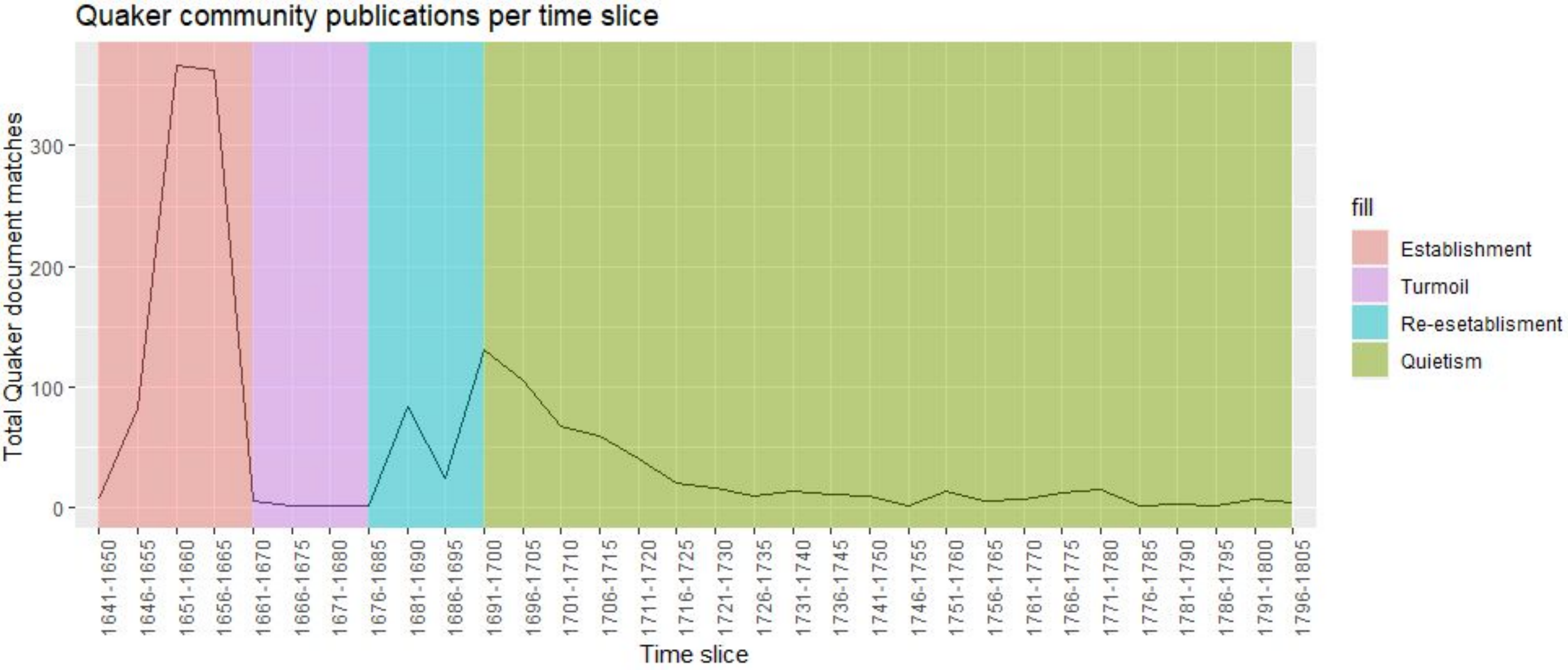


1650-1659

Left: One community of general religious actors. Blue nodes: Quaker founders George Fox, Edward Burrough, and Francis Howgill.

Right: Two communities. One made up of Quakers (blue) - nearly half of the “Valiant Sixty” - and a second general religious community (red).

Early Modern Quakers as a case study



Ryan, Y., & Tolonen, M. (2024). Networks of Influence in Scottish Enlightenment Publishing. *Connections*, *44*(1), 33-46. https://doi.org/10.21307/connections-2019.034

Rosson, D. E., Mäkelä, E., Vaara, V., Mahadevan, A., Ryan, Y. C., & Tolonen, M. (2023). Reception Reader: Exploring Text Reuse in Early Modern British Publications. *Journal of open humanities data*, *9*. https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.101

Marjanen, J. (2023). Quantitative Conceptual History: On Agency, Reception, and Interpretation. *Contributions to the History of Concepts*, *18*(1), 46-67. https://www.berghahnjournals.com/view/journals/contributions/18/1/choc180103.xml

Zhang, J., Ryan, Y. C., Rastas, I., Ginter, F., Tolonen, M., & Babbar, R. (2022). Detecting Sequential Genre Change in Eighteenth-Century Texts. In F. Karsdorp, A. Lassche, & K. Nielbo (Eds.), *Proceedings of the Computational Humanities Research Conference 2022* (pp. 243-255). (CEUR Workshop Proceedings; Vol. 3290). CEUR-WS.org. https://ceur-ws.org/Vol-3290/short_paper2630.pdf

Umerle, T., Colavizza, G., Herden, E., Jagersma, R., Kiraly, P., Koper, B., Lahti, L., Lindemann, D., Łubocki, J. M., Malínek, V., Milanova, A., Péter, R., Rißler-Pipka, N., Romanello, M., Roszkowski, M., Siwecka, D., Tolonen, M., & Vimr, O. (2023). *AN ANALYSIS OF THE CURRENT BIBLIOGRAPHICAL DATA LANDSCAPE IN THE HUMANITIES: A Case for the Joint Bibliodata Agendas of Public Stakeholders*. Czech Academy of Sciences. https://doi.org/10.5281/zenodo.6559857

Tiihonen, I. L. I., Ryan, Y. C., Pivovarova, L., Liimatta, A., Säily, T., & Tolonen, M. (2022). Distinguishing discourses: A data-driven analysis of works and publishing networks of the Scottish Enlightenment. In K. Berglund, M. La Mela, & I. Zwart (Eds.), *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)* (pp. 120-134). (CEUR Workshop Proceedings; Vol. 3232). CEUR-WS.org. http://ceur-ws.org/Vol-3232/paper09.pdf

Rastas, I., Ryan, Y. C., Tiihonen, I. L. I., Qaraei, M., Repo, L., Babbar, R., Mäkelä, E., Tolonen, M., & Ginter, F. (2022). Explainable Publication Year Prediction of Eighteenth Century Texts with the BERT Model. In N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky, & L. Borin (Eds.), *PROCEEDINGS OF THE 3RD INTERNATIONAL WORKSHOP ON COMPUTATIONAL APPROACHES TO HISTORICAL LANGUAGE CHANGE 2022 (LCHANGE 2022)* (pp. 68–77). The Association for Computational Linguistics. https://aclanthology.org/2022.lchange-1.7.pdf

Oberbichler, S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H., & Tolonen, M. (2022). Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology*, *73*(2), 225-239. https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24565

Tolonen, M., Mäkelä, E., & Lahti, L. (2022). The Anatomy of Eighteenth Century Collections Online (ECCO). *Eighteenth-century studies*, *56*(1), 95-123. https://muse.jhu.edu/article/867734

Sandberg, K., Andrushchenko, M., Turunen, R., Marjanen, J., Kurunmäki, J., Peltonen, J., Nummenmaa, T., & Nummenmaa, J. (2022). Analyzing Temporalities in Parliamentary Speech about Ideologies Using Dependency Parsed Data. teoksessa K. Berglund, M. La Mela, & I. Zwart (Toimittajat), *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)* (Sivut 406-414). (CEUR Workshop Proceedings; Nro 3232). CEUR-WS.org. https://ceur-ws.org/Vol-3232/paper40.pdf

Hill, M. J., & Tolonen, M. (2021). A Computational Investigation into the Authorship of Sister Peg. *Eighteenth-century studies*, *54*(4), 861-885. https://muse.jhu.edu/article/802445

Hengchen, S., Ros, R., Marjanen, J., & Tolonen, M. (2021). A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital scholarship in the humanities* , *36*(Suppl. 2), ii109-ii126. https://academic.oup.com/dsh/article/36/Supplement_2/ii109/6421793?login=false

Turunen, R. J. (2021). *Shades of Red: Evolution of the Political Language of Finnish Socialism from the 19th Century until the Civil War of 1918*. (Papers on Labour History).

Tolonen, M., Hill, M. J., Ijaz, A., Vaara, V., & Lahti, L. (2021). Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production. In I. Baird (Ed.), *Data Visualization in Enlightenment Literature and Culture* (pp. 63-119). Palgrave Macmillan. https://link.springer.com/chapter/10.1007/978-3-030-54913-8_3

Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M., Kaislaniemi, S., & Nevalainen, T. (2020). Wrangling with non-standard data. In S. Reinsone, I. Skadiņa, A. Baklāne, & J. Daugavietis (Eds.), *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference: Riga, Latvia, October 21-23, 2020* (pp. 81-96). (CEUR Workshop Proceedings; No. 2612). CEUR-WS.org. http://ceur-ws.org/Vol-2612/paper6.pdf

Hill, M. J. & Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*. https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqz024/5476122

Hill, M. J., Vaara, V., Säily, T., Lahti, L., & Tolonen, M. (2019). Reconstructing Intellectual Networks: From the ESTC's bibliographic metadata to historical material. In: Navarretta, C., Agirrezabal, M. and Maegaard, B. (eds.). *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5–8, 2019*. Aachen: CEUR Workshop Proceedings vol. 2364: 201-219. http://ceur-ws.org/Vol-2364/19_paper.pdf [Best paper award]

Kurunmäki, J. & Marjanen, J. (2018). A Rhetorical View of Isms: An Introduction. *Journal of Political Ideologies*, 23(3): 241-255. DOI:10.1080/13569317.2018.1502939

Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019). Bibliographic Data Science and the History of the Book (c. 1500–1800). *Cataloging & Classification Quarterly*, 57(1): 5–23. DOI: 10.1080/01639374.2018.1543747

Mäkelä, E., Tolonen, M., Marjanen, J., Kanner, A., Vaara, V., & Lahti, L. (2019). Interdisciplinary collaboration in studying newspaper materiality. In: Krauwer, S. and Fišer, D. (eds.). *Proceedings of the Twin Talks Workshop, co-located with Digital Humanities in the Nordic Countries (DHN 2019)*. Aachen: CEUR Workshop Proceedings vol. 2365: 55–66. http://ceur-ws.org/Vol-2365/07-TwinTalks-DHN2019_paper_7.pdf

Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., & Tolonen, M. (2019). A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917. *Journal of European Periodical Studies,* 4(1): 54–77. DOI: 10.21825/jeps.v4i1.10483

Tolonen, M., Lahti, L., Roivainen, H., & Marjanen, J. (2019). A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828. *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 52(1): 57-78. DOI: 10.1080/01615440.2018.1526657

Example 2: Social Media Data

Catching Stray Balls

Research Question: Do real-world events trigger toxic behaviour that spreads across unrelated online communities?

The Scale:

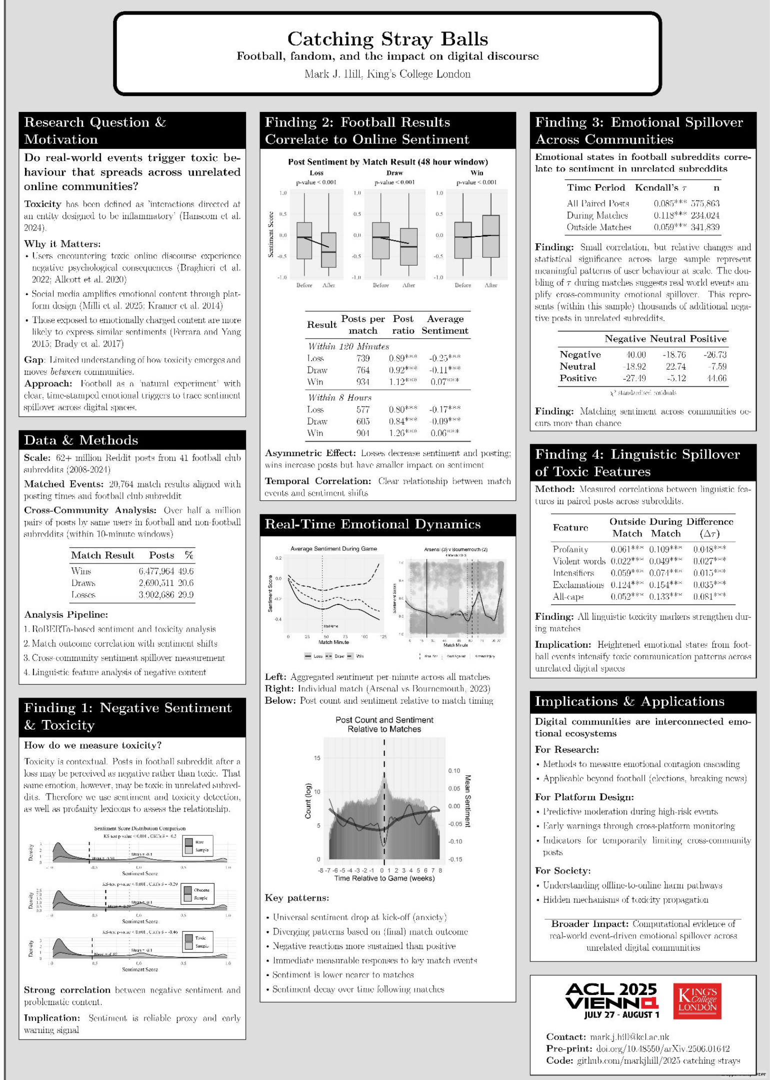
62+ million Reddit posts from 41 football club subreddits (2008-2024)

20,764 match results aligned with posting times

575,863 paired posts by same users in football and non-football subreddits

10-minute time windows for cross-community analysis

Why Football? **A "natural experiment" with clear, time-stamped emotional triggers**



Sentiment Score Distribution Comparison
KS-test p-value < 0.001 , Cliff's δ = -0.5
Hate
Sample
Mean = -0.58
Mean = -0.1
Density
Sentiment Score
KS-test p-value < 0.001 , Cliff's δ = -0.29
Obscene
Sample
Mean = -0.37
KS-test p-value < 0.001 , Cliff's δ = -0.46
Toxic
Mean = -0.52
Post Sentiment by Match Result (48 hour window)
Loss
Draw
Win
p-value < 0.001
Sentiment Score
Before
After
Average Sentiment During Game
Half-time
Match Minute
Loss
Draw
Win
Arsenal (3) v Bournemouth (2)
Goal For
Goal Against
Arsenal Injury
Post Count and Sentiment Relative to Matches
Count (log)
Mean Sentiment
Time Relative to Game (weeks)

Catching Stray Balls

Football, fandom, and the impact on digital discourse

Mark J. Hill, King's College London

Research Question & Motivation

Do real-world events trigger toxic behaviour that spreads across unrelated online communities?

Toxicity has been defined as 'interactions directed at an entity designed to be inflammatory' (Hanscom et al. 2024).

Why it Matters:

- Users encountering toxic online discourse experience negative psychological consequences (Braghieri et al. 2022; Allcott et al. 2020)
- Social media amplifies emotional content through platform design (Milli et al. 2025; Kramer et al. 2014)
- Those exposed to emotionally charged content are more likely to express similar sentiments (Ferrara and Yang 2015; Brady et al. 2017)

Gap: Limited understanding of how toxicity emerges and moves *between* communities.

Approach: Football as a 'natural experiment' with clear, time-stamped emotional triggers to trace sentiment spillover across digital spaces.

Data & Methods

Scale: 62+ million Reddit posts from 41 football club subreddits (2008-2024)

Matched Events: 20,764 match results aligned with posting times and football club subreddit

Cross-Community Analysis: Over half a million pairs of posts by same users in football and non-football subreddits (within 10-minute windows)

Match Result	Posts	%
Wins	6,477,964	49.6
Draws	2,690,511	20.6
Losses	3,902,686	29.9

Analysis Pipeline:

1. RoBERTa-based sentiment and toxicity analysis
2. Match outcome correlation with sentiment shifts
3. Cross-community sentiment spillover measurement
4. Linguistic feature analysis of negative content

Finding 1: Negative Sentiment & Toxicity

How do we measure toxicity?

Toxicity is contextual. Posts in football subreddit after a loss may be perceived as negative rather than toxic. That same emotion, however, may be toxic in unrelated subreddits. Therefore we use sentiment and toxicity detection, as well as profanity lexicons to assess the relationship.

Strong correlation between negative sentiment and problematic content.

Implication: Sentiment is reliable proxy and early warning signal

Finding 2: Football Results Correlate to Online Sentiment

Result	Posts per match	Post ratio	Average Sentiment
Within 120 Minutes			
Loss	739	0.89***	-0.25***
Draw	764	0.92***	-0.11***
Win	934	1.12***	0.07***
Within 8 Hours			
Loss	577	0.80***	-0.17***
Draw	605	0.84***	-0.09***
Win	904	1.26***	0.06***

Asymmetric Effect: Losses decrease sentiment and posting; wins increase posts but have smaller impact on sentiment

Temporal Correlation: Clear relationship between match events and sentiment shifts

Real-Time Emotional Dynamics

Left: Aggregated sentiment per-minute across all matches
Right: Individual match (Arsenal vs Bournemouth, 2023)
Below: Post count and sentiment relative to match timing

Key patterns:

- Universal sentiment drop at kick-off (anxiety)
- Diverging patterns based on (final) match outcome
- Negative reactions more sustained than positive
- Immediate measurable responses to key match events
- Sentiment is lower nearer to matches
- Sentiment decay over time following matches

Finding 3: Emotional Spillover Across Communities

Emotional states in football subreddits correlate to sentiment in unrelated subreddits

Time Period	Kendall's τ	n
All Paired Posts	0.085***	575,863
During Matches	0.118***	234,024
Outside Matches	0.059***	341,839

Finding: Small correlation, but relative changes and statistical significance across large sample represent meaningful patterns of user behaviour at scale. The doubling of τ during matches suggests real world events amplify cross-community emotional spillover. This represents (within this sample) thousands of additional negative posts in unrelated subreddits.

	Negative	Neutral	Positive
Negative	40.00	-18.76	-26.73
Neutral	-18.92	22.74	-7.59
Positive	-27.49	-5.12	44.66

χ^2 standardised residuals

Finding: Matching sentiment across communities occurs more than chance

Finding 4: Linguistic Spillover of Toxic Features

Method: Measured correlations between linguistic features in paired posts across subreddits.

Feature	Outside Match	During Match	Difference ($\Delta\tau$)
Profanity	0.061***	0.109***	0.048***
Violent words	0.022***	0.049***	0.027***
Intensifiers	0.059***	0.074***	0.015***
Exclamations	0.124***	0.154***	0.035***
All-caps	0.052***	0.133***	0.081***

Finding: All linguistic toxicity markers strengthen during matches

Implication: Heightened emotional states from football events intensify toxic communication patterns across unrelated digital spaces

Implications & Applications

Digital communities are interconnected emotional ecosystems

For Research:

- Methods to measure emotional contagion cascading
- Applicable beyond football (elections, breaking news)

For Platform Design:

- Predictive moderation during high-risk events
- Early warnings through cross-platform monitoring
- Indicators for temporarily limiting cross-community posts

For Society:

- Understanding offline-to-online harm pathways
- Hidden mechanisms of toxicity propagation

Broader Impact: Computational evidence of real-world event-driven emotional spillover across unrelated digital communities

ACL 2025 VIENNA JULY 27 - AUGUST 1

Contact: mark.j.hill@kcl.ac.uk

Pre-print: doi.org/10.48550/arXiv.2506.01642
Code: gitlab.com/markjhill/2025-catching-strays

How HPC made this research possible

1. **Initial Data Parsing**
   - 4TB+ compressed Reddit data (billions of posts)
   - Streaming identification of football-related posts
   - Result: 62M posts from 41 club subreddits
2. **GPU-Intensive Text Analysis**
   - RoBERTa-based sentiment models on tens of millions of posts
   - Transformer architecture requires GPU acceleration
   - Parallel processing across multiple nodes
3. **Cross-Community Matching**
   - Re-parsing billions of posts to find same users in non-football subreddits
   - 575,863 paired posts within 10-minute windows
   - Complex temporal alignment across communities

Digital communities as interconnected emotional ecosystems

Key Findings:

1) Asymmetric emotional response: Losses decrease sentiment more than wins increase it
2) Cross-community spillover
3) Linguistic toxicity markers strengthen during live matches
4) Real-time emotional contagion measurable minute-by-minute

The "Different Questions" Enabled:

Traditional: "How do fans react to their team losing?"

HPC-Enabled: "How do emotional states cascade across unrelated digital communities in real-time?"

Ownership as discourse

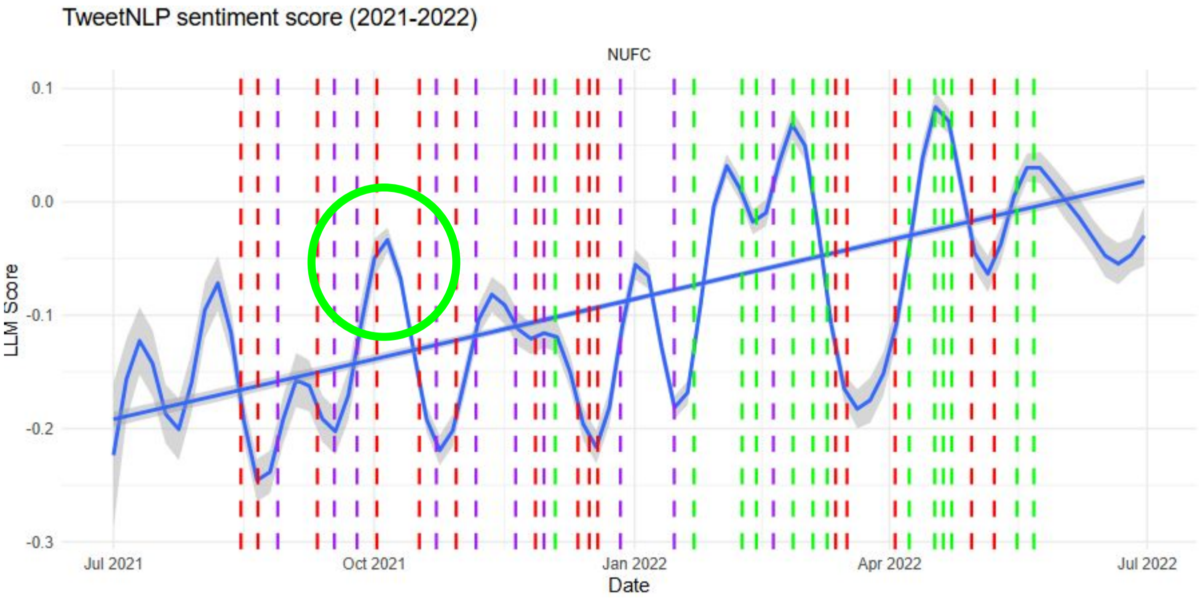


Image from *Newcastle Fans TV*, showing fans celebrating outside St James' Park on 7 October 2021, following change of ownership from Mike Ashley to Saudi backed PID (Source: wikipedia)

What HPC Enables for Humanities (or me)

Traditional Humanities Questions:

"How do people talk about X?"

"What does this text mean?"

"How does this community behave?"

HPC-Enabled Questions:

"How do conceptualizations vary across contexts and communities?"

"How do meanings shift over time and space?"

"How do emotional states cascade across digital ecosystems?"

1) Scale reveals patterns invisible to traditional methods
2) Enables computational grounded theory approaches
3) From case studies to systematic cultural analysis

Thanks!