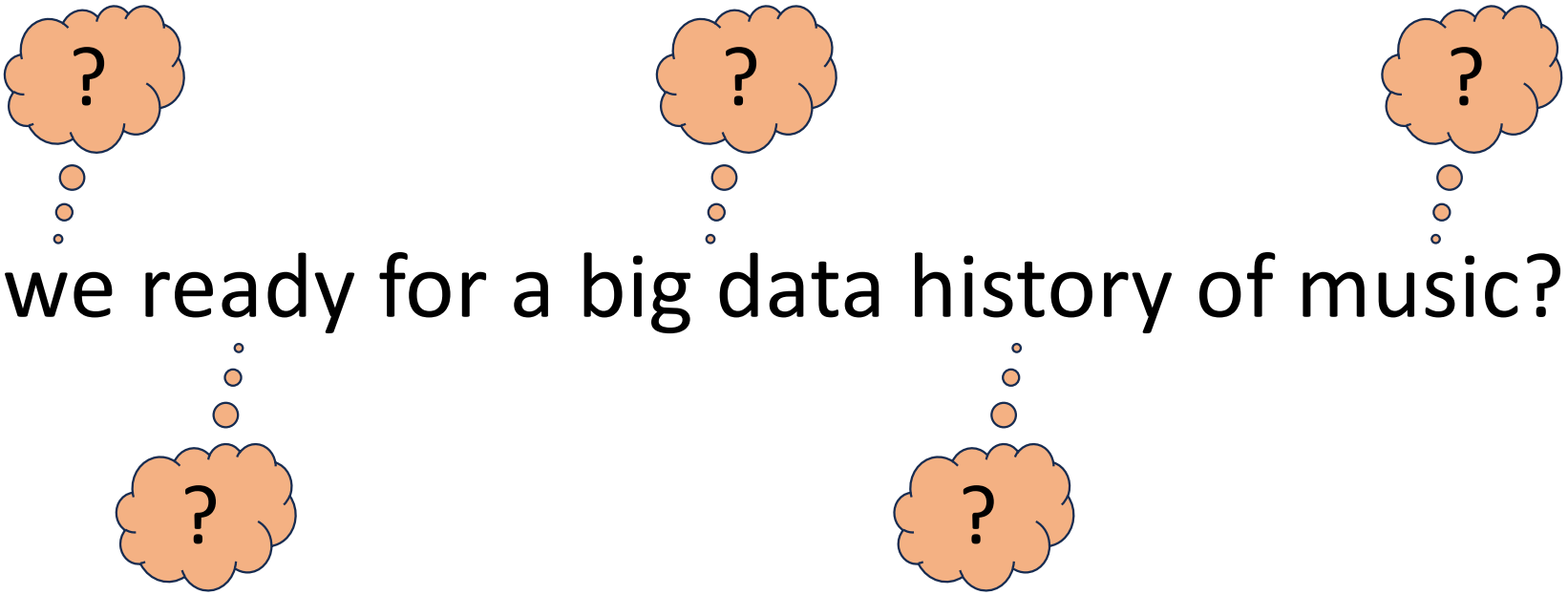
Are we ready for a big data history of music?

Athens, 22 June 2023

Frans Wiering, Utrecht University
f.wiering@uu.nl

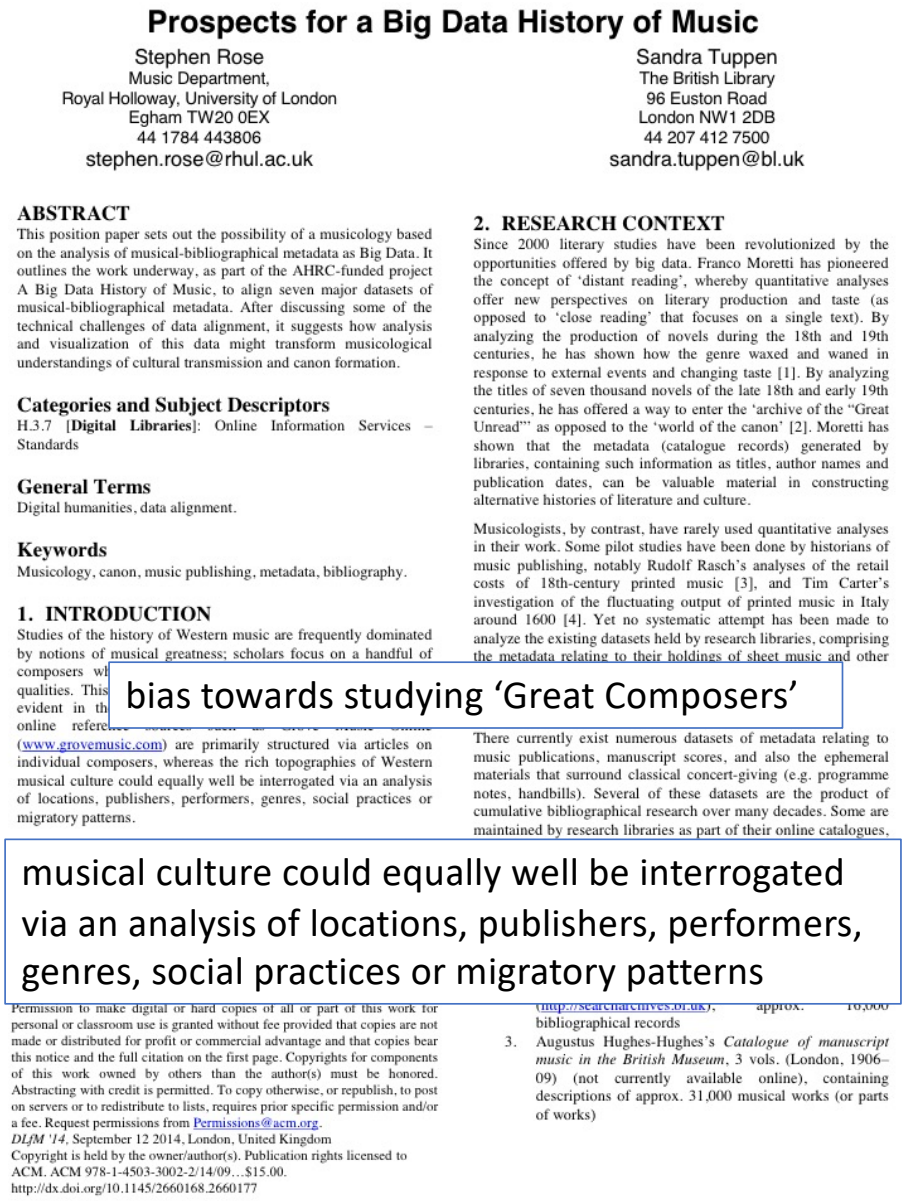
Utrecht University



Are we ready for a big data history of music?

Big data history of music

- expression coined by Stephen Rose and Sandra Tuppen (2014)
  - study of bibliographic data, mainly
- approach in key phrases
  - analyzing and vizualizing bibliographic data
  - quantitative analysis
  - examine developments and trends
  - use visualizations to test and develop hypotheses



Prospects for a Big Data History of Music

Stephen Rose
Music Department,
Royal Holloway, University of London
Egham TW20 0EX
44 1784 443806
stephen.rose@rhul.ac.uk

Sandra Tuppen
The British Library
96 Euston Road
London NW1 2DB
44 207 412 7500
sandra.tuppen@bl.uk

ABSTRACT
This position paper sets out the possibility of a musicology based on the analysis of musical-bibliographical metadata as Big Data. It outlines the work underway, as part of the AHRC-funded project A Big Data History of Music, to align seven major datasets of musical-bibliographical metadata. After discussing some of the technical challenges of data alignment, it suggests how analysis and visualization of this data might transform musicological understandings of cultural transmission and canon formation.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Online Information Services – Standards

General Terms
Digital humanities, data alignment.

Keywords
Musicology, canon, music publishing, metadata, bibliography.

1. INTRODUCTION
Studies of the history of Western music are frequently dominated by notions of musical greatness; scholars focus on a handful of composers wh
qualities. This
evident in th
online refere
(www.grovemusic.com) are primarily structured via articles on individual composers, whereas the rich topographies of Western musical culture could equally well be interrogated via an analysis of locations, publishers, performers, genres, social practices or migratory patterns.

bias towards studying 'Great Composers'

2. RESEARCH CONTEXT
Since 2000 literary studies have been revolutionized by the opportunities offered by big data. Franco Moretti has pioneered the concept of 'distant reading', whereby quantitative analyses offer new perspectives on literary production and taste (as opposed to 'close reading' that focuses on a single text). By analyzing the production of novels during the 18th and 19th centuries, he has shown how the genre waxed and waned in response to external events and changing taste [1]. By analyzing the titles of seven thousand novels of the late 18th and early 19th centuries, he has offered a way to enter the 'archive of the "Great Unread"' as opposed to the 'world of the canon' [2]. Moretti has shown that the metadata (catalogue records) generated by libraries, containing such information as titles, author names and publication dates, can be valuable material in constructing alternative histories of literature and culture.

Musicologists, by contrast, have rarely used quantitative analyses in their work. Some pilot studies have been done by historians of music publishing, notably Rudolf Rasch's analyses of the retail costs of 18th-century printed music [3], and Tim Carter's investigation of the fluctuating output of printed music in Italy around 1600 [4]. Yet no systematic attempt has been made to analyze the existing datasets held by research libraries, comprising the metadata relating to their holdings of sheet music and other

There currently exist numerous datasets of metadata relating to music publications, manuscript scores, and also the ephemeral materials that surround classical concert-giving (e.g. programme notes, handbills). Several of these datasets are the product of cumulative bibliographical research over many decades. Some are maintained by research libraries as part of their online catalogues,

musical culture could equally well be interrogated via an analysis of locations, publishers, performers, genres, social practices or migratory patterns

DLfM '14, September 12 2014, London, United Kingdom

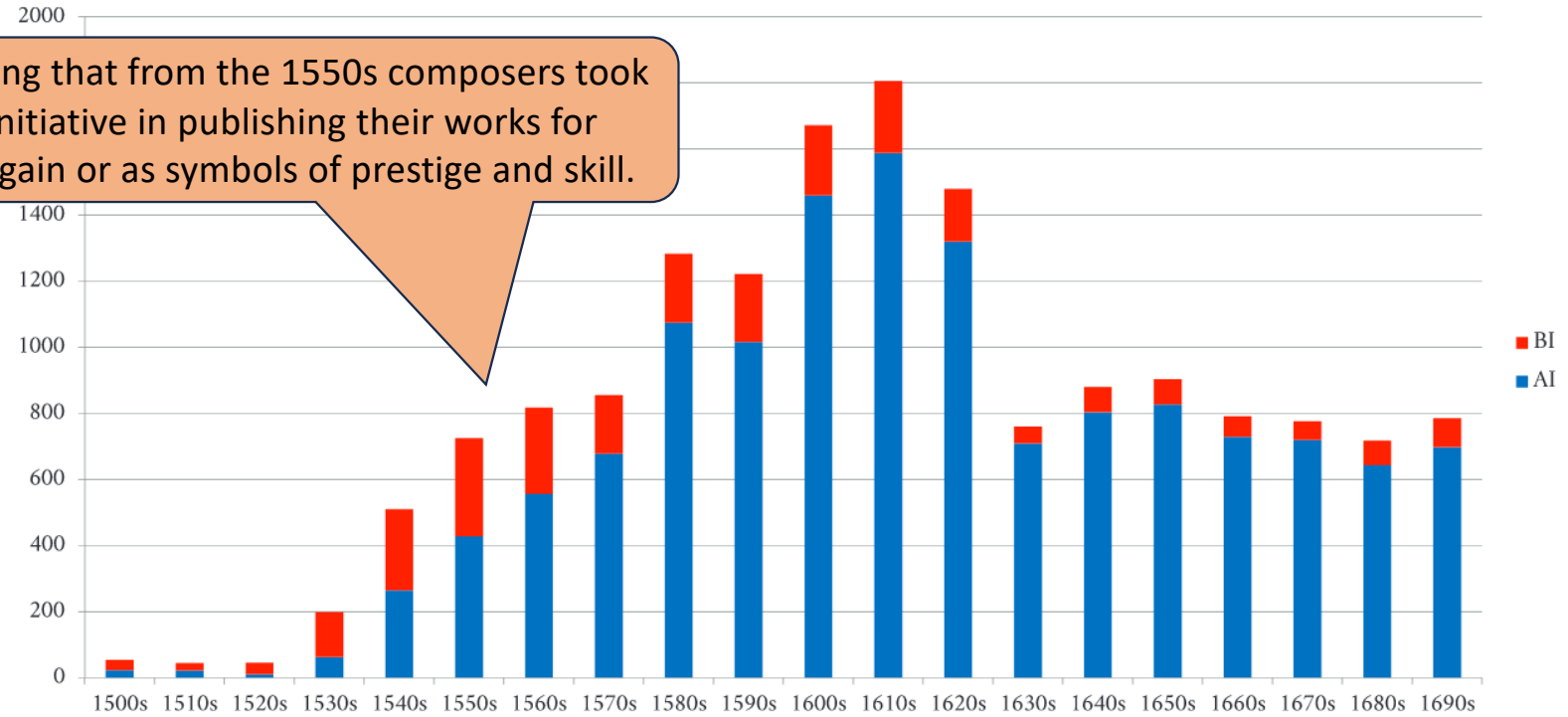
ACM 978-1-4503-3002-2/14/09…$15.00.
http://dx.doi.org/10.1145/2660168.2660177

(http://searcharchives.bl.uk), approx. 10,000 bibliographical records

3. Augustus Hughes-Hughes's *Catalogue of manuscript music in the British Museum*, 3 vols. (London, 1906–09) (not currently available online), containing descriptions of approx. 31,000 musical works (or parts of works)

Example result from Rose et al. (2015)

...suggesting that from the 1550s composers took more initiative in publishing their works for financial gain or as symbols of prestige and skill.



1 European output of printed music by decade, 1500–1699. Red shading denotes anthologies (data from RISM B/I) and blue shading indicates single-composer editions (data from RISM A/I)

My own big data history of music

CANTOSTREAM project (november 2021→)

- polyphony dating from 1500—1700
- musical objects as big data

- focus on ‘tonal structures in early music’ (Judd 1998), for example:
  - what was the role of the authentic-plagal distinction?
  - how did modality transition into tonality?
  - how did the Mixolydian modes disappear?

Mode Æolien and Sous-Æolien (on G) from *La rhétorique des Dieux* (1652), a 12-mode cycle consisting of 56 lute pieces

The question, once again

- are we ready for a big data history of music?
- what would, ideally, a ‘yes’ entail?

big data analysis is able to deliver new facts and insights that enrich the study of music history

- by confirming, questioning or providing alternatives to current views
- by shedding light on open issues in the discipline
- by suggesting new directions for research

Five questions about readiness

1. community: who are ‘we’?
2. data: what data are available?
3. processing: what methods and tools are available?
4. study: what research can be done?
5. persuasion: how convincing are the outcomes?

obviously, these are highly interconnected

Q1. community: who are ‘we’?

Computational musicologists?

1960-70s

1980s

today

- computational musicologists have been eager since at least 1967 (Logemann)
  - ‘heroic period’ of CM
  - focus on musical objects (as in musicology of the day)
- then musicology developed in a different direction
  - contextualisation rather than ‘the music itself’
- are computational musicologists ready to reconnect?

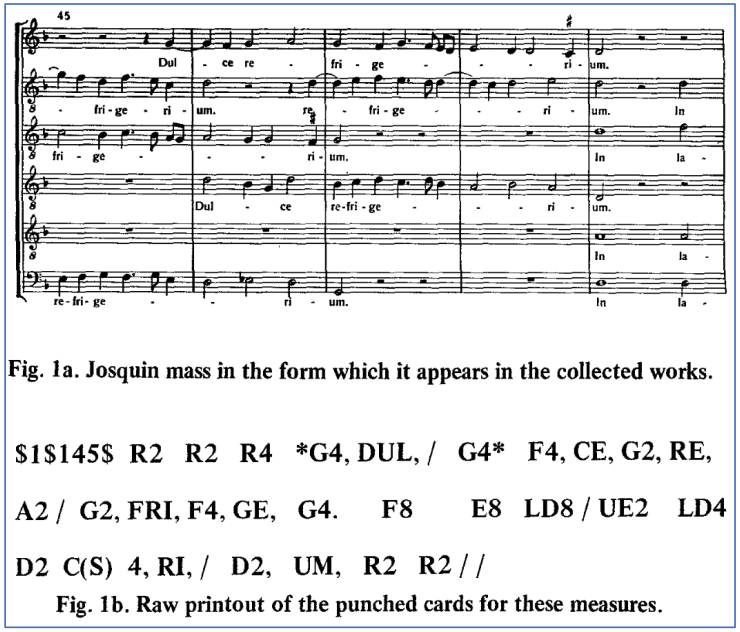


Fig. 1a. Josquin mass in the form which it appears in the collected works.

$1$145$ R2 R2 R4 *G4, DUL, / G4* F4, CE, G2, RE,

A2 / G2, FRI, F4, GE, G4. F8 E8 LD8 / UE2 LD4

D2 C(S) 4, RI, / D2, UM, R2 R2 / /

Fig. 1b. Raw printout of the punched cards for these measures.

The Princeton Josquin Project was first example of a big data approach in musicology (Mendel 1969)

Technology acceptance in musicology

- What Do Musicologists Do All Day 2015 outcomes (Inskip & Wiering 2015)
  - musicologists use technology widely and creatively
  - widely-shared feeling that this has changed musicology
  - access and discovery are very important
  - only weak interest in software for processing musical data (music notation programs excepted)

benefits of technology responses: terms related to processing are infrequent (n=625)

WDMDAD the pandemic edition (2022)

Q18: Which of the [technology-related] changes you have been forced to make are you going to keep after the pandemic is over?

amount of change	count
(almost) none	73
specific change(s)	210
most or all	19
dont't know	9
unclear or unusuable	6
total	**317**

area of change	count
human interaction: teaching, conferences, collaboration,...	130
digital and physical materials	76
remote work, travel	50
hardware, software, services	24
other changes: health, skills and attitudes, ...	21
total	**301**

strongest impacts on social structure of the discipline and on research materials

One possible storyline in the responses

perhaps none of them [my changes] huge or transformative on their own, but in sum the changes in habit and standard operating procedure have been huge. [176]

- tech skills were essential for survival
- many discovered potential of technology for research
- specifically, online access to materials (matched by concerns about study of physical resources)
- maybe a more collaborative attitude
- new research topics and methodological reflections
- promising for reconnection

Our tools shape the way we think - the increased speed of digitisation allows for more analysis of big data and more research questions based on questions related to quantifications in general

I am concerned about how reliance on digital resources can shape research questions and trends - a topic which to my knowledge no one has looked into.

Q2. data: what data are available?

Musical data come in many kinds

type	examples	number of musical items (wild estimate)
metadata	RISM, library catalogues	> 10.000.000
digitised sources	EMO, Gallica, DIAMM	100.000-500.000
digital editions (mostly PDF)	CPDL	50.000-200.000
MIDI	Classical Archives	100.000-500.000
encodings of notation	MEI, MusicXML, humdrum	20.000-100.000
audio recordings	CD, MP3, streaming	> 100.000.000
feature sets	Million Song Dataset, Essentia	2.000.000-5.000.000
born-digital materials	generative music; tags, tweets	> 10.000.000

Musical data come in many kinds

type	examples	number of musical items (wild estimate)
audio recordings	CD, MP3, streaming	> 100.000.000
born-digital materials	generative music; tags, tweets	> 10.000.000
metadata	RISM, library catalogues	> 10.000.000
feature sets	Million Song Dataset, Essentia	5.000.000-20.000.000
digitised sources	EMO, Gallica, DIAMM	100.000-500.000
MIDI	Classical Archives	100.000-500.000
digital editions (mostly PDF)	CPDL	50.000-200.000
encodings of notation	MEI, MusicXML, humdrum	20.000-100.000

what we need most is also the rarest

Situation of early music even more dramatic

- Harasim et al. (2021)
  - historical development of tonality
- MIDI representations
  - 12635 from Classical Archives
  - 777 from Du Chemin and ELVIS
- scarcity is bad news for CANTOSTREAM project

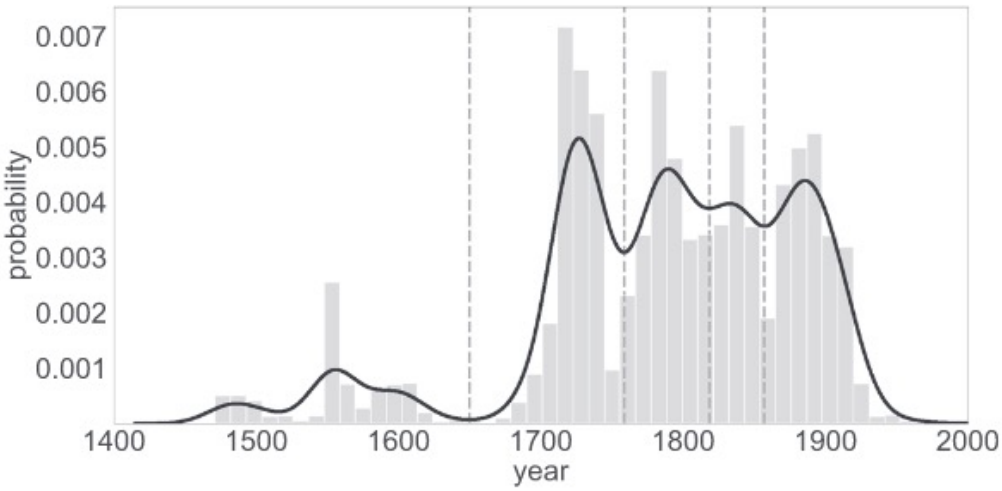


Fig. 1 Histogram of composition dates of 13,402 musical pieces in the dataset.

Early music datasets in recent research

publication	size	comments
Geelen et al. (2021). Clustering analysis of renaissance polyphony	1248 works	humdrum; Josquin Research Project
Harasim et al. (2021). Exploring the foundations of tonality	777+ works	midi; Lost Voices, CRIM, ELVIS
Long (2020). Hearing homophony	nearly 400 works	not a computational study; canzonettas by Gastoldi and others
Upham, Cumming (2020). Auditory streaming complexity	216 works, 1042 movements	midi; Josquin Research Project, RenComp7
Arthur (2021). Vicentino versus Palestrina	707 movements	humdrum; complete Palestrina Masses

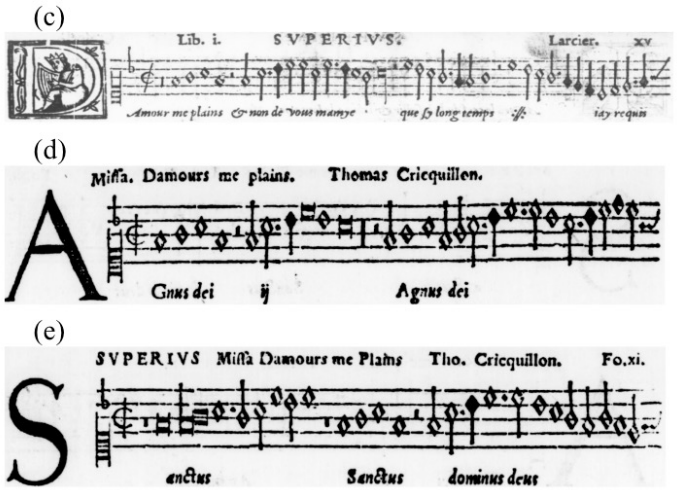
small datasets, some great composers

Treasure hunt (selection)

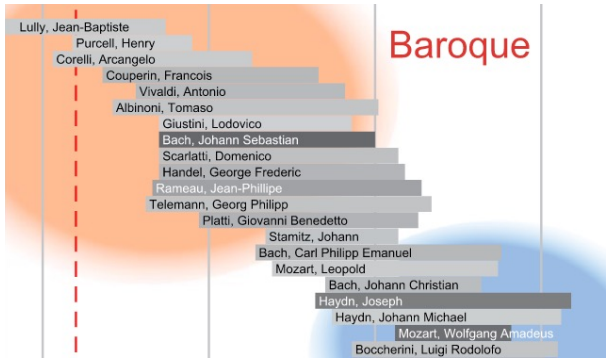
Dataset name	content	items before 1700	encoding system	url
Choral Public Domain Library	choral music of any age	unknown	various	https://www.cpdl.org/wiki/
ECOLM	16-18th century lute music	1619	tabcode	ecolm.org
ELVIS database	heterogeneous, aggregated	1000 (c)	various	https://database.elvisproject.ca/
Josquin Research project	works by Josquin and contemporaries	902	MusicXML, humdrum	https://josquin.stanford.edu/
Tasso in Music	musical settings of Tasso's poems	778	humdrum, MEI, MusicXML	https://www.tassomusic.org/
Neuma	multiple subcollections	600 (c)	MusicXML, MEI	http://neuma.huma-num.fr/
Lost Voices	16th c. French chansons	380 (c)	MEI	http://digitalduchemin.org/
CRIM	16th c. imitation masses and their models	250 (c)	MEI	https://crimproject.org/
Gesualdo online	complete works of Gesualdo	222	MEI	https://ricercar.gesualdo-online.cesr.univ-tours.fr/
miami Publication server of the University of Münster	3 digital editions of early 17th c. sacred music	230 (c)	Lilypond, MIDI	https://miami.uni-muenster.de/
Measuring polyphony	14th century motets	61	MEI	https://measuringpolyphony.org/
Computerized Mensural Music Editing	15-16th c. polyphony	59	CMME	http://www.cmme.org/
Furnace and fugue	digital edition of Atalanta fugiens (1618)	50	MEI	https://furnaceandfugue.org/

How to move forward?

- coordinated community initiative
- Optical Music Recognition
  - even the best OMR is far from perfect
  - imperfect data is usable in some scenarios (→ David Lewis)
- use audio recordings
  - much more data (including Great Composers)
  - much software development
  - intellectual property is a barrier



related items retrieved by searching OMR output (Crawford et al. 2018)

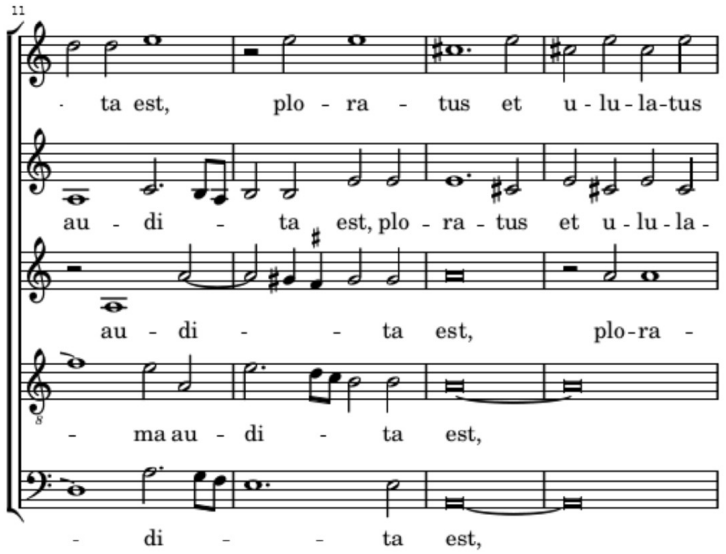


part of the audio corpus analysed by Weiß et al. (2019)

Q3. processing: what methods and tools are available?

Processing encodings

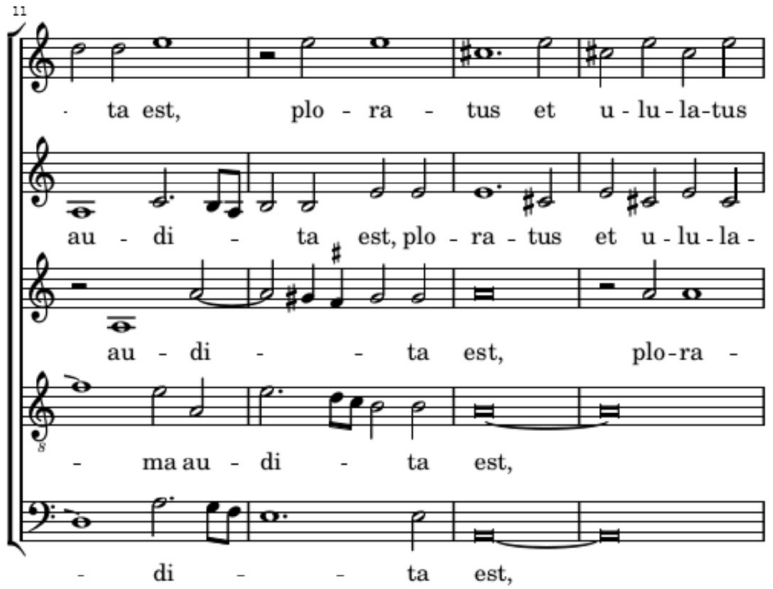
- quite a few toolboxes
  - humdrum, music21, jSymbolic, IDyOM, retrieval methods…
- weaker on detectors for high-level features
- example: cadence detection in renaissance polyphony
  - cadence is a marker for a boundary
  - and a manipulable musical object
  - studied in CRIM (Alex Morgan) and Polifonia (Christophe Guillotel-Nothmann)
- current status: cadence detection is an interactive process
  - another data requirement: expert annotations



Giaches de Wert, Vox in Rama, bars 11-14

Processing audio

- situation is more complex
- chroma features are marvellous
  - collect energy per pitch class
  - analysis of harmony
- polyphonic transcription is still an open problem
  - recent work on vocal polyphony (Cuesta et al. 2020, Rosenzweig 2022)
- most computational methods developed for popular music
  - e.g. boundary detectors perform rather weakly on polyphony
  - retraining needs yet another kind of data in quantity



Dm Am E A

chords extracted by chordify.net

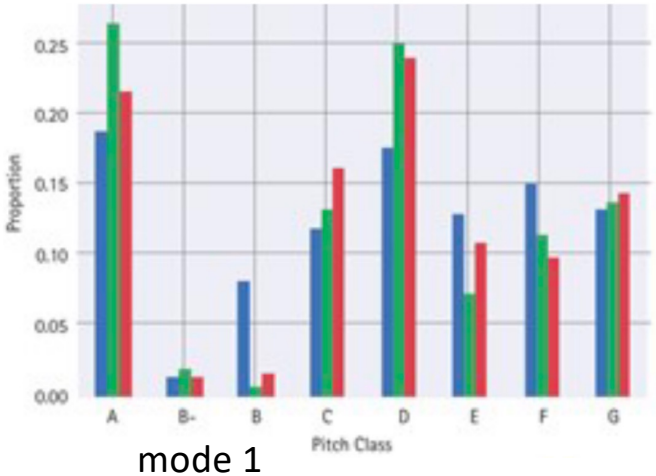
	Precision	Recall	F1
3 easy pieces	0.90	0.78	0.83
8 hard pieces	0.53	0.37	0.43

performance of OLDA segmentation algorithm (McFee & Ellis 2014) on renaissance vocal music

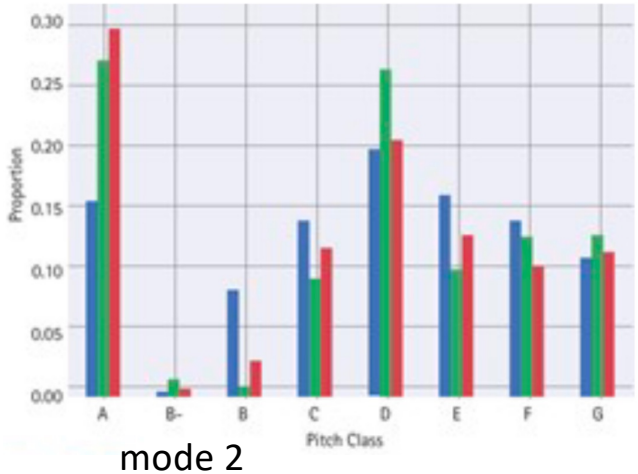
Q4. study: what research can be done?

Focussed studies

- attribution of anonymous works from *Leuven Chansonnier* (Geelen et al. 2021)
- Palestrina's counterpoint practice vs. Vicentino's theory (Arthur 2021)
- establishment of mode in renaissance duets (Arthur et al. 2022)
  - mode 'families' easy to distinguish
  - authentic-plagal distinction not so clear



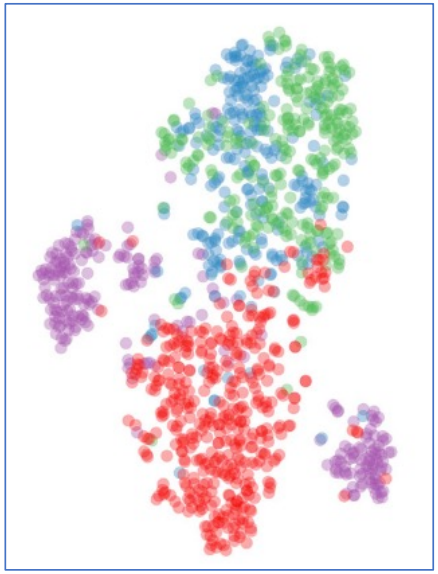
mode 1



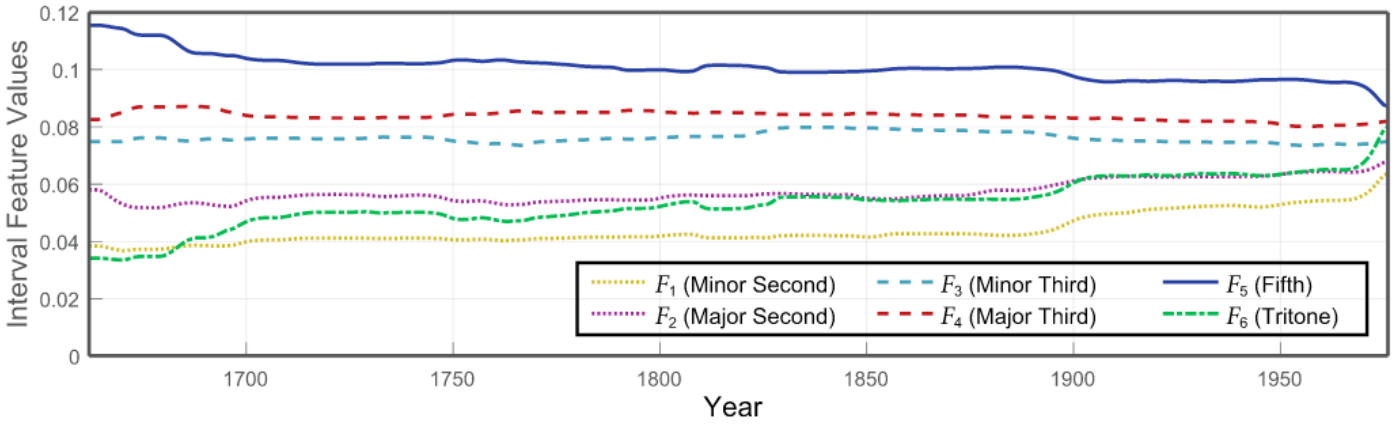
mode 2

Studying large-scale developments

- statistical cognitive modelling of mode (Harasim et al., 2021)
  - 500 years; c. 13.000 compositions
  - modal clarity at its highest in classical period
- stylistic information in pitch class distributions (Yust 2019)
  - period of nearly 400 years
- style evolution based on harmonic intervals extracted from audio (Weiß et al. 2019)
  - intervals, chord progressions,complexity
  - overall, outcomes align with Harasim's findings



weakly-separated modal clusters in renaissance music



interval category features distributed over the years

Observations

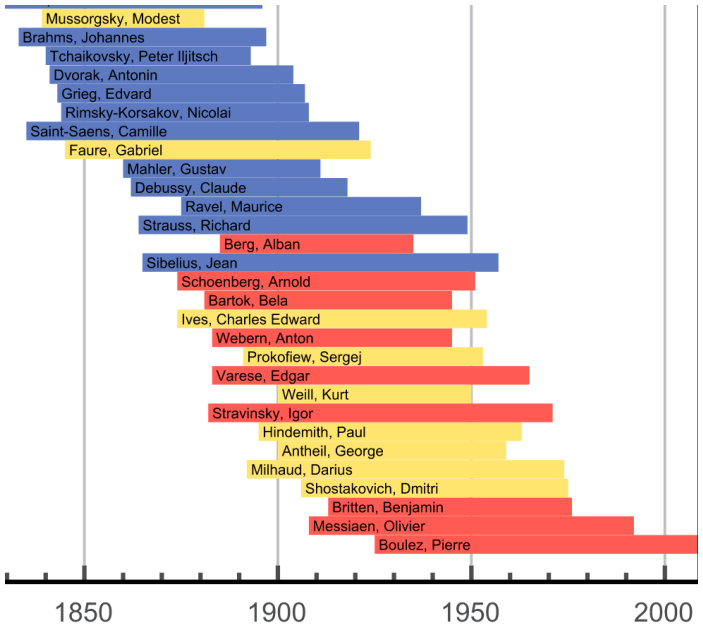
- focus on tonal aspects, less on rhythm, melody, genre
- mid-level studies seem to be rare
  - Long's (2020) study of the canzonetta is a 'manual' big data study
- these require much richer contextual data, such as
  - location
  - date
  - genre
  - performance information

  in short, interoperability with other (meta) data sets

Q5. persuasion: how convincing are the outcomes?

Some obstacles to persuasion

- questions that are too big
  - insights such as decreasing tonal clarity in the 19th century isn't really new musicological knowledge
- pushing things too far
  - perplexing classification outcomes are frequent
  - these are not musicologists' problems
- using whatever data is at hand
  - choice and quality of data matters hugely in musicological research
- reductionism
  - e.g. a mode is just another scale --> look at pitch class only
- conceptual shopping
  - not paying attention to the context from which a concept or idea originates



The assignment of Mussorgsky and Faure to this cluster is rather surprising since most of the late romantic composers (Mahler, Strauss) as well as the impressionists (Debussy, Ravel) are assigned to the Romantic cluster. This kind of unexpected observations could serve as an inspiration for musicological research (Weiß et al. 2019)

What makes computational results convincing

- choose problems at the right level of granularity
- make use of the richness of main concepts
  - mode, for example
- choose the right data
  - quality more important than size
  - appropriate contextual information
- maximise transparency and explainability
  - if necessary at the cost of performance
- apply data and tool criticism to investigate biases

Conclusion

Summary answers to the questions

- community: divided, but tech acceptation is growing
- data: problematic
- methods and tools: decent, but hoping for advances in audio analysis tools
- research: promising, creative
- persuasion and insight: rather weak, especially for studies of large-scale developments

Are we ready for a big data history of music?

We aren’t ready *yet*

Five easy steps towards maturity

1. take interdisciplinarity *very seriously*
   - connect to existing musicological expertise
2. sustainable integration of resources
   - projects like Polifonia are great but tend not to outlive their funding
   - lightweight: money will only become scarcer
3. practice data and tool criticism continuously
4. MEI-data button for all music notation programs
5. <...*insert your own favourite step here*...>

Thank you!

References

- Arthur, C. (2021). Vicentino versus Palestrina: A computational investigation of voice leading across changing vocal densities. *Journal of New Music Research*, *50*(1), 74–101.
- Arthur, C., Cumming, J. E., & Schubert, P. (2022). *The Role of Structural Tones in Establishing Mode in Renaissance Counterpoint*.
- Cuesta, H., McFee, B., & Gómez, E. (2020). Multiple F0 Estimation in Vocal Ensembles Usi Convolutional Neural Networks. *Proceedings ISMIR*, 302–309.
- Geelen, B., Burn, D., & De Moor, B. (2021). A Clustering Analysis of Renaissance Polyphony Using State-Space Models. *Journal of the Alamire Foundation*, *13*(1), 127–146.
- Harasim, D., Moss, F. C., Ramirez, M., & Rohrmeier, M. (2021). Exploring the foundations of tonality: Statistical cognitive modeling of modes in the history of Western classical music. *Humanities and Social Sciences Communications*, *8*(1), 1–11.
- Inskip, C., & Wiering, F. (2015). In their own words: Using text analysis to identify musicologists' attitudes towards technology. *Proceedings ISMIR*, 455–461.
- Judd, C. C. (1998). *Tonal structures in early music*. Garland.
- Logemann, G. W. (1967). The Canons in the Musical Offering of JS Bach: An Example of Computational Musicology. In *Elektronische Datenverarbeitung in der Musikwissenschaft* (pp. 63–87). Gustav Bosse Verlag Regensburg.

References, continued

- Long, M. K. (2020). *Hearing Homophony: Tonal Expectation at the Turn of the Seventeenth Century*. Oxford University Press, USA.
- McFee, B., & Ellis, D. P. W. (2014). Learning to segment songs with ordinal linear discriminant analysis. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5197–5201.
- Mendel, A. (1969). Some preliminary attempts at computer-assisted style analysis in music. *Computers and the Humanities*, *4*(1), 41–52.
- Rose, S., & Tuppen, S. (2014). Prospects for a big data history of music. *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, 1–3.
- Rosenzweig, S. (2022). *Interactive Signal Processing Tools for Analyzing Multitrack Singing Voice Recordings* [PhD Thesis]. Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).
- Upham, F., & Cumming, J. (2020). Auditory Streaming Complexity and Renaissance Mass Cycles. *Empirical Musicology Review*, *15*.
- Weiß, C., Mauch, M., Dixon, S., & Müller, M. (2019). Investigating style evolution of Western classical music: A computational approach. *Musicae Scientiae*, *23*(4), 486–507.
- Yust, J. (2019). Stylistic information in pitch-class distributions. *Journal of New Music Research*, *48*(3), 217–231.