Lessons Learned in a Large-Scale Project to Digitize and Computationally Analyze Musical Scores

Cory McKay *(Marianopolis College, Canada)*
Julie Cumming *(McGill University, Canada)*
Ichiro Fujinaga *(McGill University, Canada)*

2019 Digital Humanities Conference
Utrecht, Netherlands

Topics

- Overview of the SIMSSA project
- General insights we have gained
  - Constructing datasets
  - Deep learning vs. feature-based approaches to machine learning
  - Sharing of data, software and results

Overview of the SIMSSA project

- SIMSSA (Single Interface for Music Score Searching and Analysis) is a large project involving:
  - Dozens of institutions in both Europe and North America
  - More than 125 researchers
  - Funding from 2014 to 2021
- Aims to unite, under a single framework, the ability to:
  - Automatically transform images of musical scores into digital symbolic representations using OMR (optical music recognition)
  - Automatically extract meaningful statistical information (features) from such symbolic music files
  - Use machine learning and statistical analysis to conduct musicological research using this data
  - Create a framework for searching symbolic scores based on both metadata and musical content
  - Make the resulting information and tools easily accessible to other researchers

Learning from our missteps (1/2)

- We have accomplished much since the SIMSSA concept was first presented at DH (Fujinaga and Hankinson 2013)
  - Also made some missteps
- Have noticed similar mistakes being made by others in fields our work has touched on:
  - Music information retrieval (MIR)
  - Computational musicology
  - Digital humanities
- We therefore wish to share our experiences, with the hope of helping other researchers avoid some of our mistakes

Learning from our missteps (2/2)

- Some of this advice may seem obvious, especially to domain specialists
  - Nonetheless, these issues continue to recur in work published in DH and related fields
- Such missteps are to be expected in such (wonderfully) multidisciplinary areas
  - Nobody can be a specialist in everything, so such problems are to be expected
  - However, we must as a community take steps to improve our digital methodologies

Dataset construction

- Humanities researchers sometimes simply combine digitized data as is, from whatever sources are readily available
  - Or digitize data themselves without first constructing a carefully considered workflow
- Can lead to erroneous conclusions:
  - False patterns may be observed due to inconsistent dataset construction practices
  - Meaningful patterns may be obscured in datasets that fail to capture essential information
- We encountered such problems when we carried out research on stylistic differences between Iberian and Franco-Flemish Renaissance music (McKay 2018)
  - Individual transcribers had encoded note durations differently
  - Rhythm was correlated more with the transcriber than with the underlying music

Data selection and balancing

- Selection and balancing of data are also essential
- Results can be compromised if a dataset:
  - Does not represent the full range of relevant instances
    - e.g. only an artist's early works
  - Contains uneven class distributions
    - e.g. many more works by one artist than another
- We observed in machine learning-based research on composer attribution (McKay et al. 2017b) that trained classification models would sometimes perform classifications based on genre rather than compositional style
  - The number of masses and motets were not evenly distributed across composers
  - Proper dataset balancing was necessary

Dataset encoding

- Unexpected problems can also be introduced during the encoding process
  - e.g. we observed that commercial score editing software sometimes confused the encoding of slurs and ties (Nápoles et al. 2018)
- We developed a set of best practices to help avoid bias when constructing datasets from historical documents (Cumming et al. 2018)

Deep learning vs. feature-based machine learning (1/3)

- Most current research involving machine learning employs deep learning (DL)
  - Models are typically trained on huge datasets
  - Data is processed in a relatively raw form
    - With, typically, some important pre-processing
- Contrasts with non-DL machine learning approaches:
  - Training often performed on hand-crafted statistical features that quantify specific qualities of domain interest
  - Sub-systems may sequentially process data in stages following a pre-defined workflow
- The current emphasis on deep learning is reasonable
  - Has been widely successful in many domains
  - e.g. our OMR performance improved substantially when we switched to a deep learning framework that processes pixel windows directly (Calvo-Zaragoza et al., 2018)

Deep learning vs. feature-based machine learning (2/3)

- However, deep learning's need for huge training sets can sometimes be a serious limitation when dealing with historical data with limited extant instances
  - e.g. early music
- Even clever data augmentation techniques can only help so much
  - Although they certainly can help

Deep learning vs. feature-based machine learning (3/3)

- Deep learning still also often results in black-box classifiers
  - Recent advances in model transparency are starting to help, but DL still tends to be opaque relative to feature-based approaches
- In contrast, feature-based systems (in conjunction with feature-selection algorithms) produce:
  - Data searchable by features in domain-meaningful ways
  - Directly accessible insights into how features differentiate classes
- In the humanities, these insights can be even more important than class label outputs themselves!
  - e.g. understanding what differentiates two composers stylistically can be more important than actually differentiating them
- Deep learning and feature-based learning each have different strengths and weaknesses
  - Must fully consider these before choosing which to utilize

Illustrative examples (1/2)

- Our jSymbolic software (McKay et al. 2018) extracts 1497 feature values from symbolic music
- Used these features to, with high accuracy:
  - Attribute the music of Renaissance composers (McKay et al. 2017b)
  - Identify the genre of Renaissance music (Cumming and McKay 2018)
  - Etc.
- More importantly, we analyzed the feature data to gain meaningful musicological insights into which musical characteristics statistically differentiate these classes
- We also used feature data to empirically test expert predications about musical style in these studies
  - 63% of these expert predictions were found to be inaccurate!
- There is a particular need for such testing in music (and in the humanities in general)
  - There are many generally accepted assertions that have never actually been properly validated empirically

Illustrative examples (2/2)

- We also used the jSymbolic features to provide content-based support (McKay et al. 2017a) for composer attribution confidence levels proposed previously by Rodin and Sapp (2015) based solely on historical evidence
  - A nice example of how computational and traditional humanities research can complement one another

Sharing data, software and results (1/3)

- It is essential to consider issues associated with making research data, software and results available, useable and attractive to other researchers in the humanities
  - Especially researchers not yet accustomed to computational approaches
- We must consult domain experts about what they need, as noted by Wiering (2017)
  - Rather than imposing decisions on them

Sharing data, software and results (2/3)

- Related priorities include:
  - Clean and consistent software and web interfaces
  - Extensive and accessible documentation
    - Including tutorials
  - Adoption of open accepted standards
  - Compatibility with diverse data formats
  - Facilitating extensibility for other researchers
  - Consider data and software in the context of international intellectual property laws

Sharing data, software and results (3/3)

- The better we become at facilitating the sharing of our work, the better we will be able to, across research groups:
  - Directly compare techniques and results
  - Engage in experimental repetition and validation
  - Make iterative refinements building on each other's work
- Such steps will in turn help us benefit from arguably the greatest advantages computational approaches bring to the humanities:
  - Subjecting long-standing assumptions to empirical validation
  - Exploring data in new and exciting ways

Thanks for your attention!

- **E-mail:** cory.mckay@mail.mcgill.ca
- **SIMSSA:** https://simssa.ca

McGill

Schulich School of Music
École de musique Schulich

CIR MMT Centre for Interdisciplinary Research in Music Media and Technology

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

MARIANOPOLIS COLLEGE

Fonds de recherche sur la société et la culture Québec

DDMAL DISTRIBUTED DIGITAL MUSIC ARCHIVES & LIBRARIES LAB

SIMSSA Single Interface for Music Score Searching and Analysis

jMIR