Web Studies Involving User Data

Dr. Daniel Wolff, TIDO Music UK Ltd.

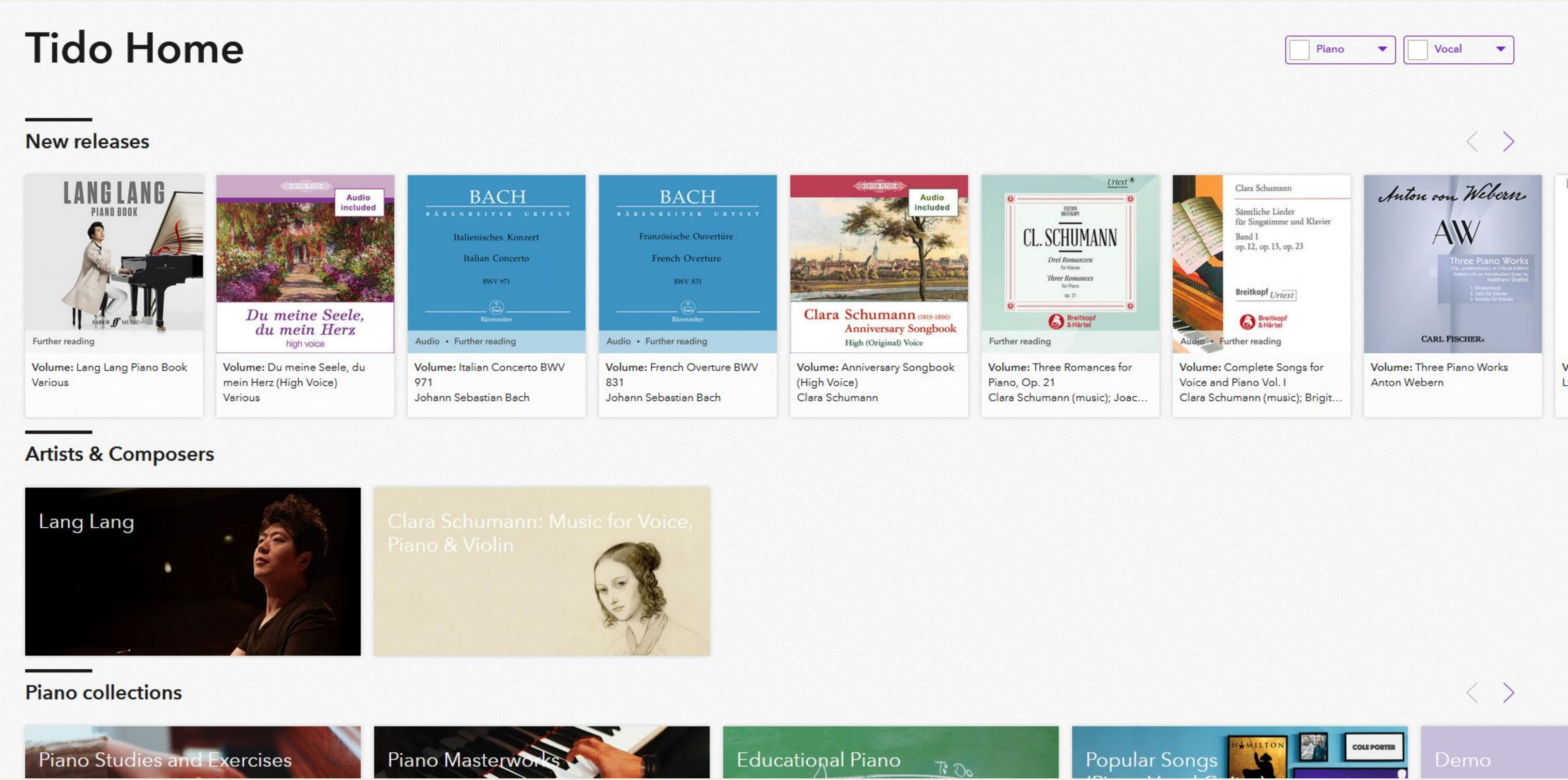
daniel.wolff@tido-music.com

Introduction

- User data is valuable in many applications
  - Validation for user-targeted application functionality
  - (Automatic) adaptation to preference and behaviour
  - Data-centered approach to research and design
- User data is critical
  - Data protection
  - Users can revoke data access
- User data is expensive
  - Participant acquisition
  - Data management

www.tido-music.com - a data-driven music platform

- Music platform with highly interactive apps on iOs and the web
- Combining different media such as
    - Notation
    - Audio
    - Multi-perspective video
    - Teacher and performer commentary
    - Rich information about the composer and pieces
- Editors produce content with help of MIR, e.g. audio-score alignment, image processing ...
- User feedback essential for app development and design



Tido Home

New releases

Further reading

Volume: Lang Lang Piano Book
Various

Volume: Du meine Seele, du mein Herz (High Voice)
Various

Audio • Further reading

Volume: Italian Concerto BWV 971
Johann Sebastian Bach

Audio • Further reading

Volume: French Overture BWV 831
Johann Sebastian Bach

Volume: Anniversary Songbook (High Voice)
Clara Schumann

Further reading

Volume: Three Romances for Piano, Op. 21
Clara Schumann (music); Joac...

Audio • Further reading

Volume: Complete Songs for Voice and Piano Vol. I
Clara Schumann (music); Brigit...

Volume: Three Piano Works
Anton Webern

Artists & Composers

Lang Lang

Clara Schumann: Music for Voice, Piano & Violin

Piano collections

Piano Studies and Exercises

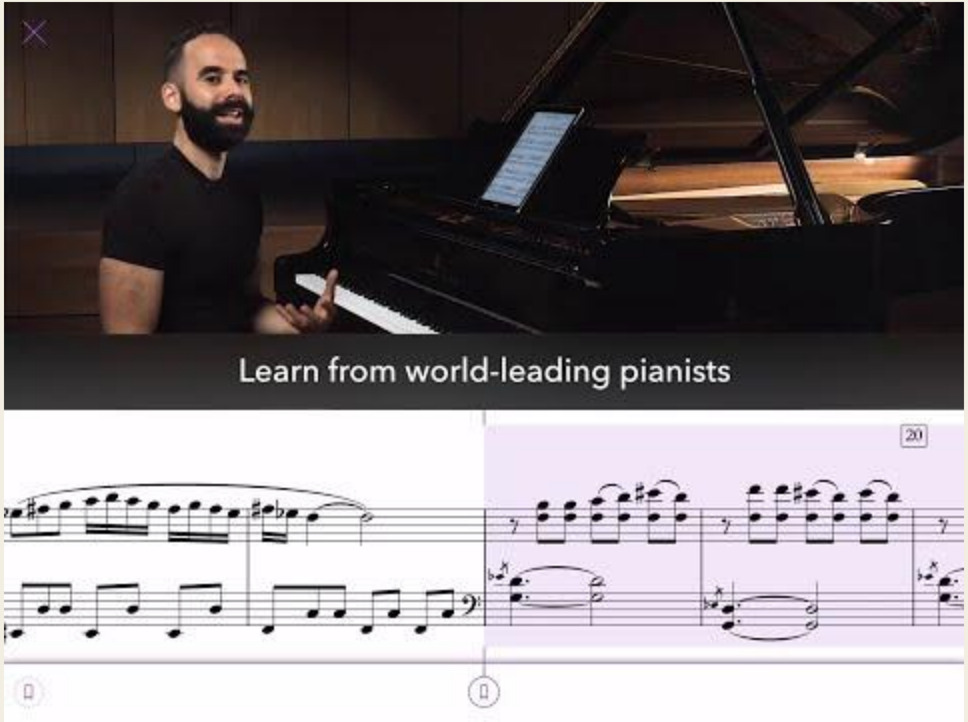
Piano Masterworks

Educational Piano

Popular Songs

Demo

TIDO Music



Learn from world-leading pianists

Common Types of MIR Studies on the Web

- Listening tests (rate loudness, similarity, add tags, …)
- Playlist collection (sequence, grouping)
- Interactive tests (tapping etc...)
- Exploration (2d/3d music maps & worlds)

Participants: Sample size & distribution

- Need to monitor / control demographics
- Distribution / requirements on statistical representation
    - What group to analyse / predict
    - What distribution of attributes within that group
- Requirements for machine learning
    - Minimum sample number per class / target …

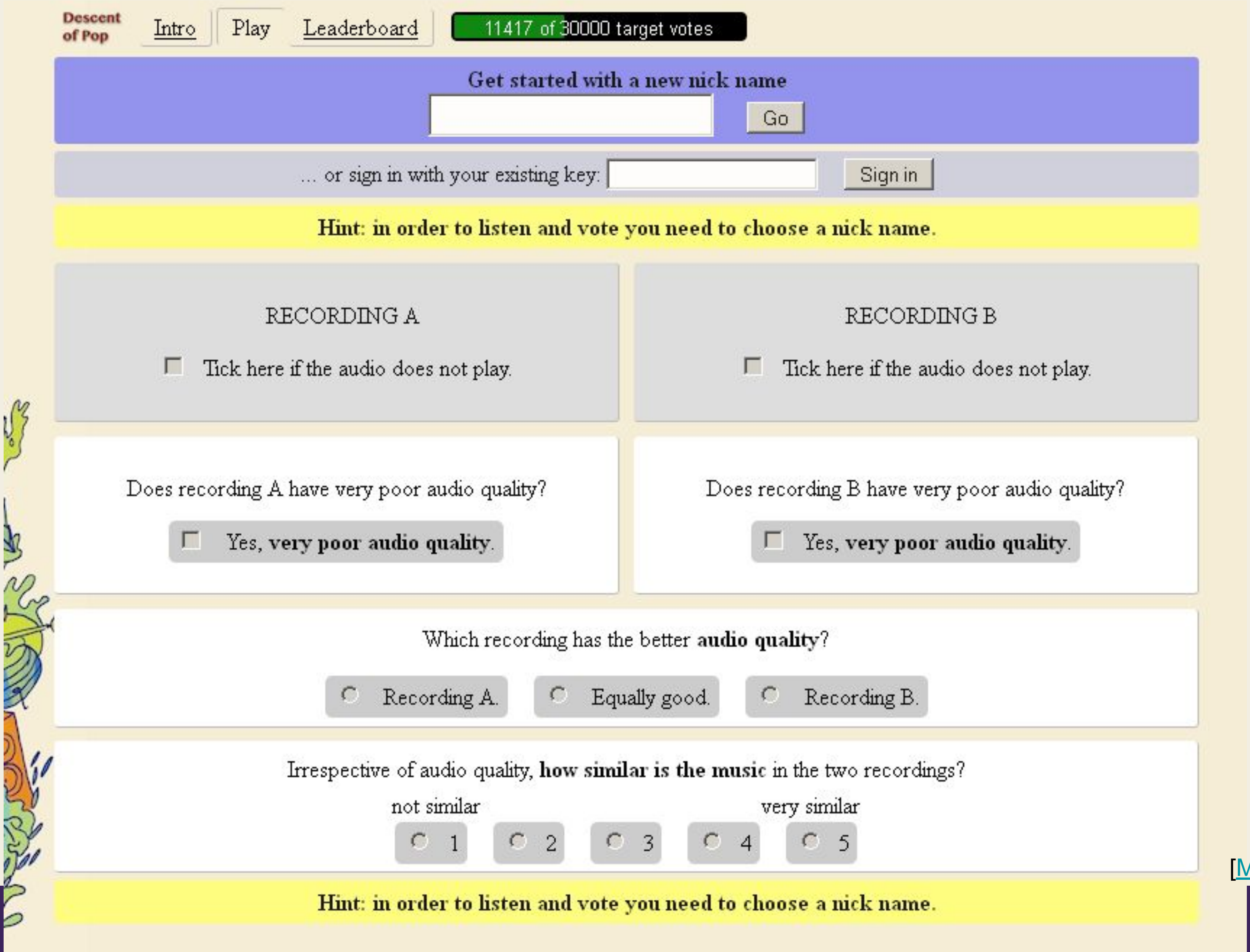
Participant acquisition and motivation

- Students
- Special interest groups (ISMIR-Community, MIREX, Kaggle …)
- Integration into existing application
- Integration into social network
- Game with a purpose (fun)
- Payment / vouchers ($$$)
- Paid platform (e.g. Prolific Academic, Amazon Mechanical Turk)

Interactive audio tests: Some examples

- MIREX, e.g. audio similarity task. [Downie et al. 2014]
- Survey on pleasurable moments in music [de Fleurian 2018]
- Subjective comparison of music production practices using the Web Audio Evaluation Tool [De Man et al. 2016]

- BBC: How Musical Are You [BBC Labs 2011]
- Magnatagatune [Law et al. 2009], HerdIt [Barrington et al. 2009]
- Spot the Odd Song Out [Wolff et al. 2013]
- KKBOX Tag Game



Descent of Pop | Intro | Play | Leaderboard | 11417 of 30000 target votes

Get started with a new nick name

[ ] Go

... or sign in with your existing key: [ ] Sign in

Hint: in order to listen and vote you need to choose a nick name.

RECORDING A

□ Tick here if the audio does not play.

RECORDING B

□ Tick here if the audio does not play.

Does recording A have very poor audio quality?

□ Yes, **very poor audio quality**.

Does recording B have very poor audio quality?

□ Yes, **very poor audio quality**.

Which recording has the better **audio quality**?

○ Recording A. ○ Equally good. ○ Recording B.

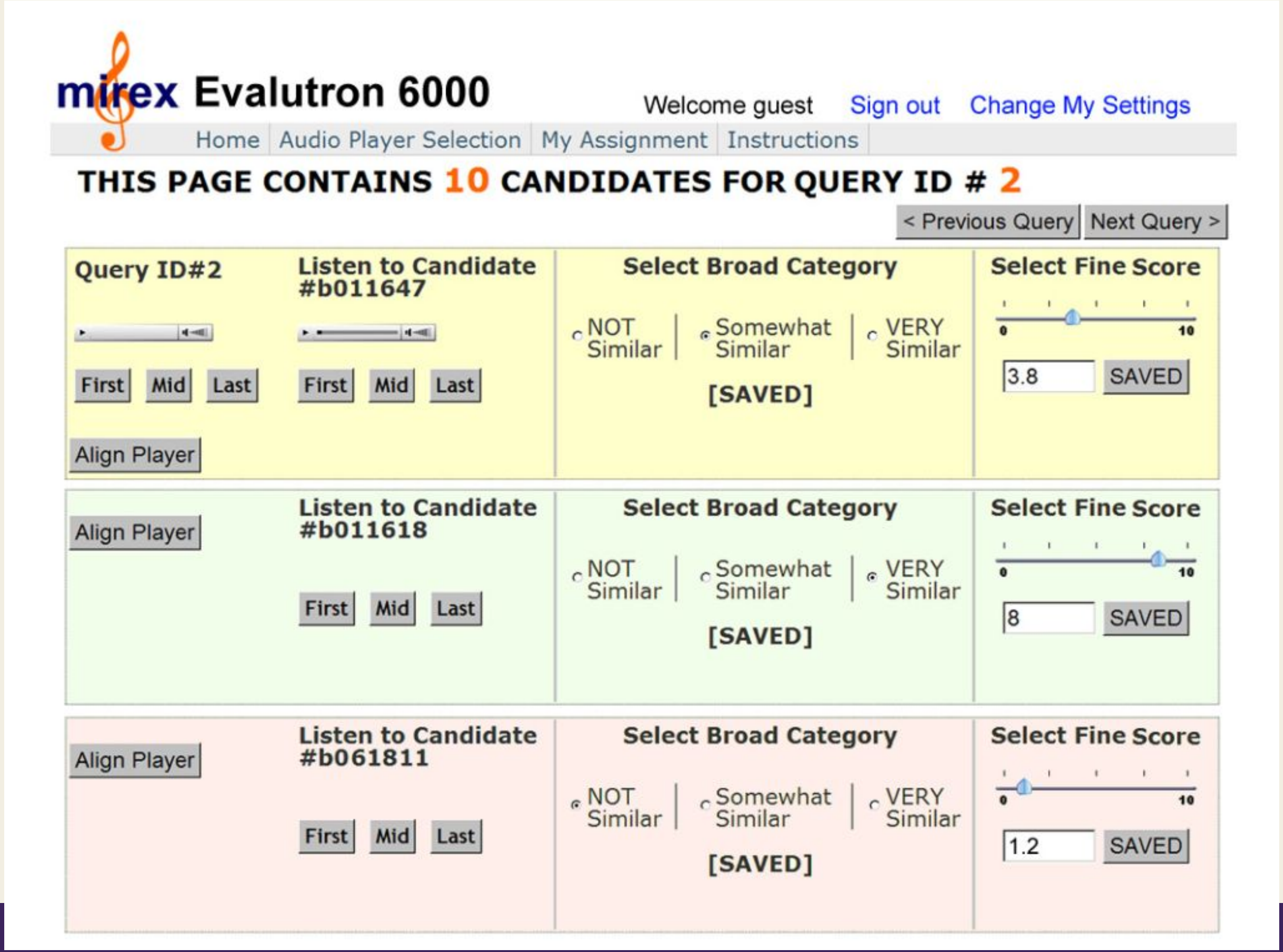
Irrespective of audio quality, **how similar is the music** in the two recordings?

not similar ... very similar

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Hint: in order to listen and vote you need to choose a nick name.

[Mauch et al. 2013]



THIS PAGE CONTAINS 10 CANDIDATES FOR QUERY ID # 2

Query ID#2

Listen to Candidate #b011647

Select Broad Category: ○ NOT Similar | ● Somewhat Similar | ○ VERY Similar

[SAVED]

Select Fine Score: 0 – 10: 3.8 SAVED

Listen to Candidate #b011618

Select Broad Category: ○ NOT Similar | ○ Somewhat Similar | ● VERY Similar

[SAVED]

Select Fine Score: 0 – 10: 8 SAVED

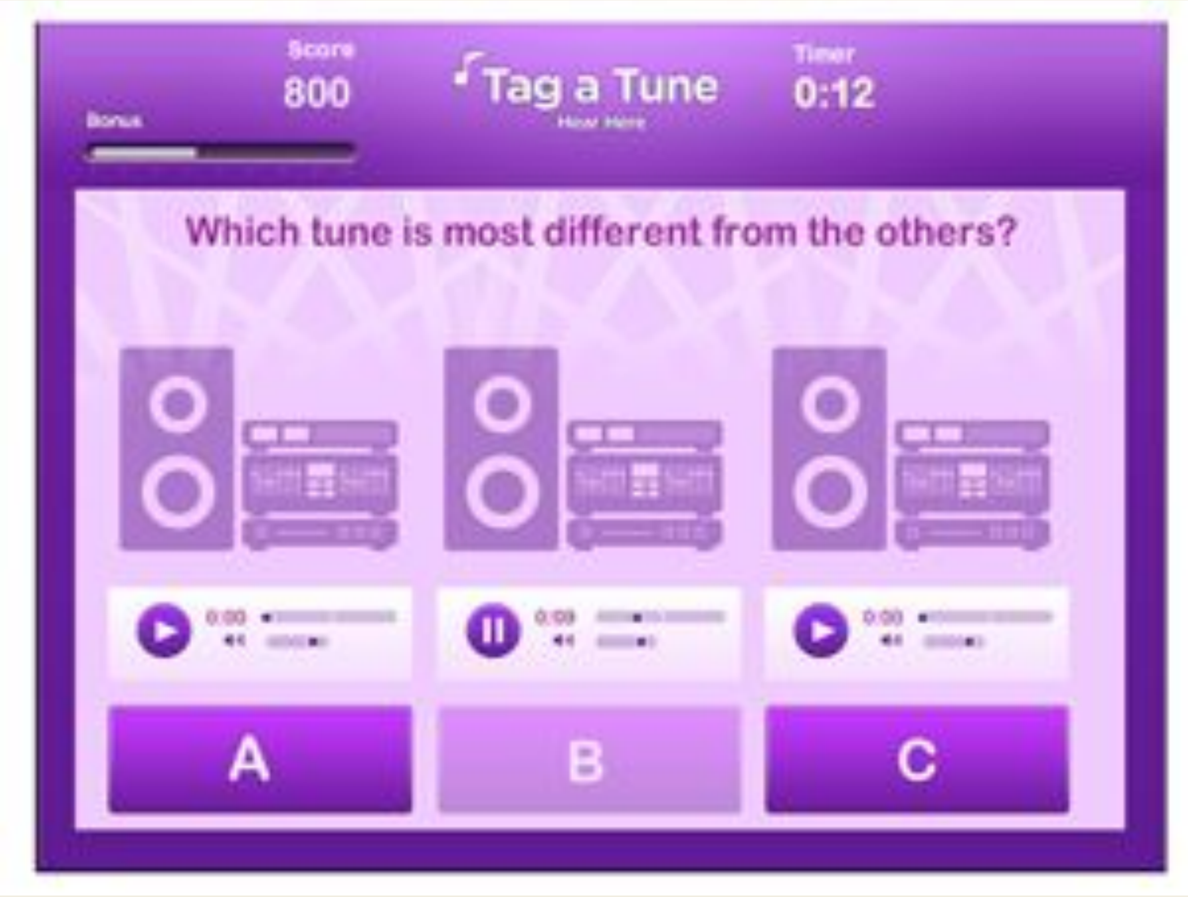
Listen to Candidate #b061811

Select Broad Category: ● NOT Similar | ○ Somewhat Similar | ○ VERY Similar

[SAVED]

Select Fine Score: 0 – 10: 1.2 SAVED

[Gruzd et al. 2007]



Which tune is most different from the others?

[Law et al. 2009]

MIR-Specific Requirements

- Strongly depend on task at hand

- Audio calibration & adjustment with data report
- Playback jitter and quality assurance
- Synchronisation between audio & video playback
- Restricted and / or monitored playback controls
- Anonymisation of recorded data

Quick and Simple: Existing Platforms

- Well-developed tools for form-based surveys
  - Google forms,
  - Qualtrix,
  - Survey Monkey ...
- MIR-specific web/survey frameworks exist
  - Web Audio Evaluation Tool [Jillings et al. 2015]
  - JS-XTRACT: A realtime audio feature extraction library for the web
  - CASimIR [Wolff et al. 2013.]

Brew your own?

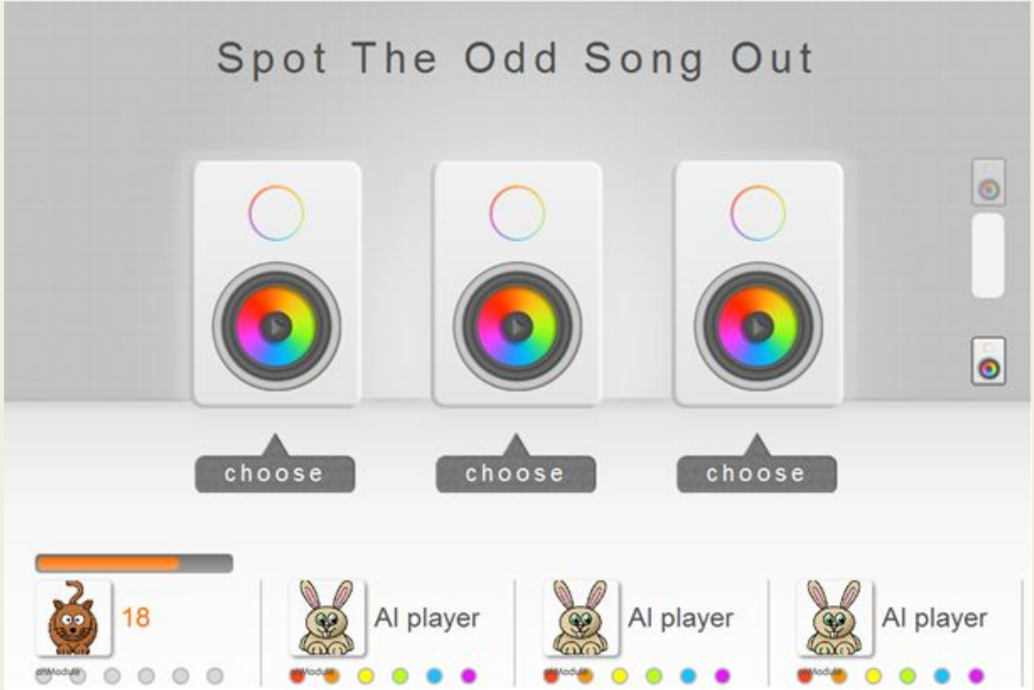
- MIR has many very specific use-cases with requirements on
    - Data collected (e.g. response timing, audio loudness …)
    - Music dataset format and access
- Tempting to (re) implement large parts of the collection system
    - Benefits: custom everything, control
    - Drawbacks: maintenance, portability, shareability, testing
- Suggestion:
    - Re-use existing and maintained projects
    - Keep custom part (UI) implementation simple with few dependencies

BYO: Front-end Frameworks

- Use html5 media containers & web-audio where possible

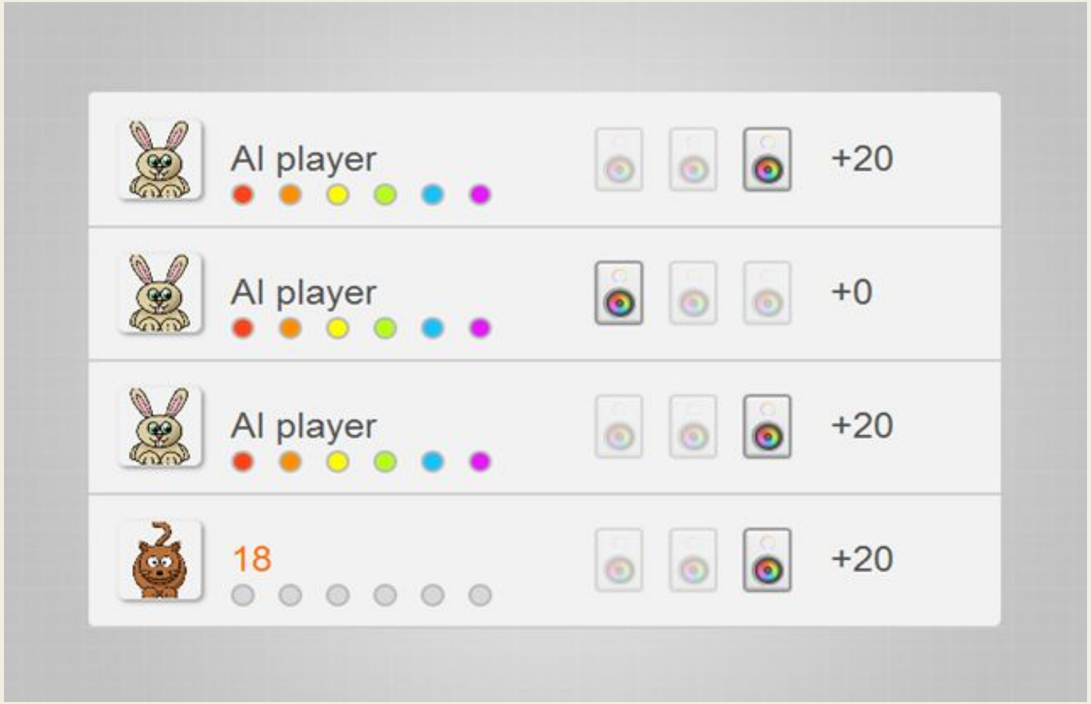
- Consider security features
  - Https
  - Avoid cross-site scripting
- Consider limitations on mobile
  - Screen sizes
  - Interaction necessary for automatic playback, download,
  - limited control on when / whether playback starts

Case analysis: Spot The Odd Song Out



[Wolff et al. 2013]

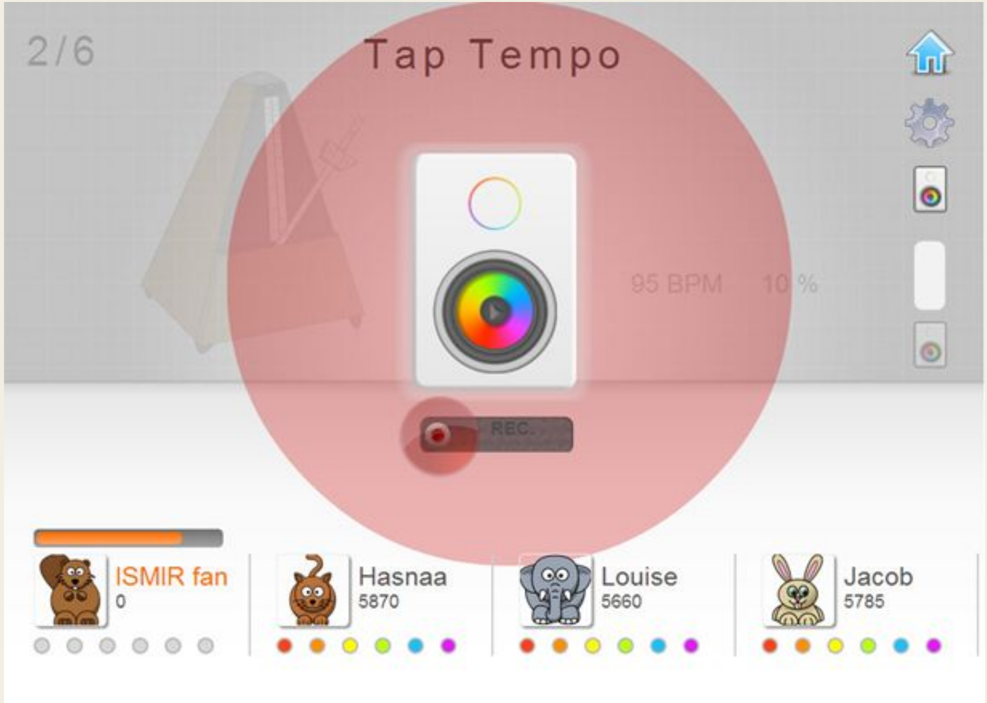
Case analysis: Spot The Odd Song Out



AI player	+20
AI player	+0
AI player	+20
18	+20

[Wolff et al. 2013]

Case analysis: Spot The Odd Song Out



[Wolff et al. 2013]

BYO: Hosting

- Need to assure reliability, security, and development access
- Cloud infrastructure exists in AWS/Google cloud
  - ++ : encryption enabled, user authentification, security certification, reliable back-up
  - -- : data is “on the web”, 3rd party has (some) access
- Alternative: University infrastructure
  - ++ : cheaper (hopefully), “closer” access to data and admin
  - -- : less streamlined method, depends on local resources

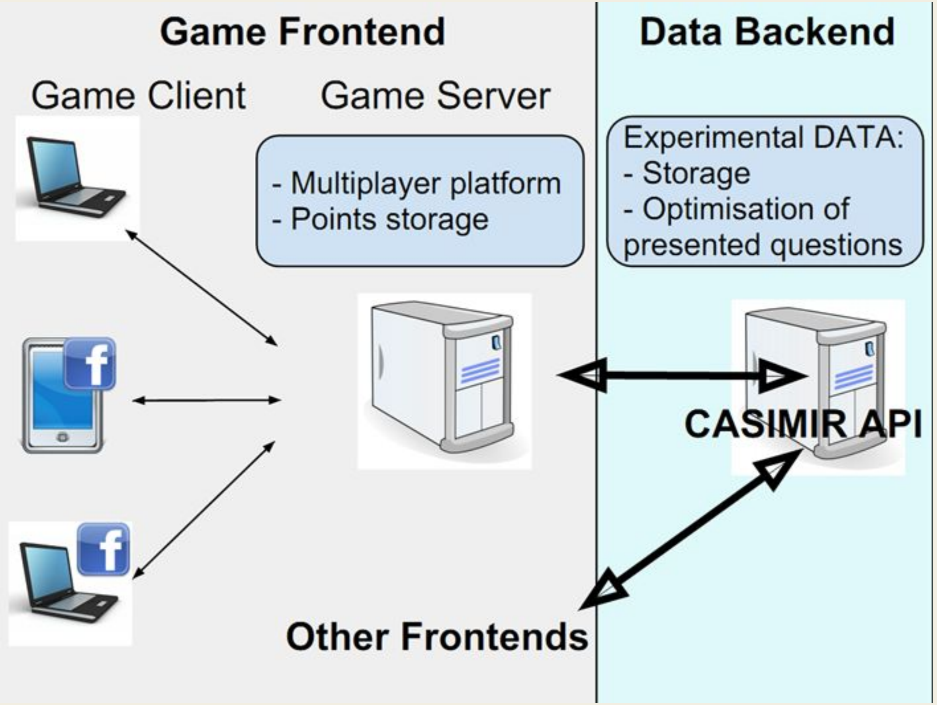
BYO: Back-end

- Keep data storage back-end independent of front-end / UI

- Consider scaling to many users (1000s or more ?)
- Consider portability to other servers
- Popular **python (flask/django)** or **node** frameworks

- Consider data storage and export options : MYSQL; NOSQL; MongoDB

Case analysis: Spot The Odd Song Out



[Wolff et al. 2013]

Back-end Data Security

- Back-up (clone/snapshot, automation)
- Integrity (real-time/across snapshots)
- Access restrictions
- Encryption
- Anonymisation

Handling Participant Consent

- Check University Ethical Guidelines
- Participant needs to know:
  - What will they be doing
  - Are there any risks or specific requirements
  - How long will it take
  - What are the benefits to them or society
  - Contact details for later questions

Participant Data & Consent

- Check University Ethical Guidelines
- **Informed** consent necessary to collect personal data
    - Anonymity / possible ways of (re)identification
    - **Type** of data collected
    - Data storage place and duration of retainment
    - People having access to data (if to be made public make explicit)
    - Any data handlers (e.g. Amazon AWS if stored there)
    - Mechanism to request deletion of data (even after de-identification)
    - Note that deletion of published or anonymised data becomes impossible

Participant Data

- Typical categorisation of data in terms of protection:

- **Personal Identifiable Information**: Participant is identifiable
- **De-identified data**: Extra information is kept to re-identify the participant
- **Anonymised data**: This part of data cannot be re-identified easily
- **Anonymous data**: Data has never been identifiable
- **Data aggregation**: Data combined from different sources
- **Re-identification**: Participant again linked to a data sample through combination of data sources

Participant Data: Potential PID

- University Guidelines: Categorisation for sampled data often still under development

- Identifiability often depends on context
  - Linkable data in other datasets
  - Amount of data collected per user
  - Uniqueness of data with user

Privacy-Preserving Machine Learning

- Modern deep networks need large amounts of data
- Large models can copy large amounts of data
  - => transform data such that it cannot be identified prior to training
  - => reduce probability of private data being stored in the model
- In multi-server computation, data is shared
  - between servers
  - across networks
  - => shape computation such that privated data is not shared

Summary

- User data is helpful in adapting applications to the real world
- User data can be collected easily through the web
- Personal data needs to be protected and requires consent
- Platforms for studies exist, but complex tasks need development
- Code on the web can reach many, but it ages fast
- Re-use the wheel
- Web studies give your work great exposure