A plan for sustainable MIR evaluation

Brian McFee*

Eric Humphrey

Julián Urbano

NYU Center for Data Science

Spotify

upf. Music Technology Group

MARR
Hypothesis (model)

Experiment (evaluation)

Progress depends on access to common data
We’ve known this for a while

- Many years of MIREX!
- Lots of participation
- It’s been great for the community
MIREX (cartoon form)

Scientists (i.e., you folks)

Code

MIREX machines (and task captains)

Results

Data (private)
Evaluating the evaluation model

We would not be where we are today without MIREX.
Evaluating the evaluation model

We would not be where we are today without MIREX.

But this paradigm faces an uphill battle :’(o
Costs of doing business

- Computer time
- Human labor
- Data collection
Costs of doing business

- Computer time
- Human labor
- Data collection

Annual sunk costs (proportional to participants)

*arrows are probably not to scale*
Costs of doing business

- Computer time
- Human
- Data collection

Annual sunk costs (proportional to participants)

The worst thing that could happen is growth!

Best! for $

*arrows are probably not to scale*
Limited feedback in the lifecycle

Hypothesis
(model)

Performance metrics (always)
Estimated annotations (sometimes)
Input data (almost never)

Experiment
(evaluation)
Stale data implies bias

BUT YOU'VE GOT ALL THE TOP BANDS IN HERE. STYX?
I JUST HEARD THEM ON "THE KING BISCUIT FLOUR HOUR."

https://frinkiac.com/caption/S07E24/252468
Stale data implies bias

EVE R Y O N E K N O W S R O C K
ATTAINED PERFECTION IN
1974. IT'S A SCIENTIFIC
FACT.

https://frinkiac.com/caption/S07E24/288671
The current model is unsustainable

- Inefficient distribution of labor
- Limited feedback
- Inherent and unchecked bias
What is a sustainable model?

- Kaggle is a data science evaluation community (sound familiar?)

- How it works:
  - Download data
  - Upload predictions
  - Observe results

- The user-base is huge
  - 536,000 registered users
  - 4,000 forum posts per month
  - 3,500 competition submissions per day (!!!)
What *is* a sustainable model?

- Kaggle is a data science evaluation community (sound familiar?)

- How it works:
  - Download data
  - Upload predictions
  - Observe results

- The user-base is huge
  - 536,000 registered users
  - 4,000 forum posts per month
  - 3,500 competition submissions per day (!!!)
Open content

- Participants need unfettered access to audio content
- Without input data, error analysis is impossible
- Creative commons-licensed music is plentiful on the internet!
  - FMA: 90K tracks
  - Jamendo: 500K tracks
The Kaggle model is sustainable

- Distributed computation
- Open data means clear feedback
- Efficient allocation of human effort
But what about annotation?
Incremental evaluation

- Which tracks do we annotate for evaluation?
  - None, at first!

- Annotate the most informative examples first
  - Beats: [Holzapfel et al., TASLP 2012]
  - Similarity: [Urbano and Schedl, IJMI 2013]
  - Chords: [Humphrey & Bello, ISMIR 2015]
  - Structure: [Nieto, PhD thesis 2015]
Incremental evaluation

- Which tracks do we annotate for evaluation?
  - None, at first!

- Annotate the most informative examples first
  - Beats: [Holzapfel et al., TASLP 2012]
  - Similarity: [Urbano and Schedl, IJMIR 2013]
  - Chords: [Humphrey & Bello, ISMIR 2015]
  - Structure: [Nieto, PhD thesis 2015]

This is already common practice in MIR.
Let's standardize it!

[Carterette & Allan, ACM-CIKM 2005]
Disagreement can be informative

F#:maj

F#:7

https://frinkiac.com/caption/S06E08/853001
The evaluation loop

1. Collect CC-licensed music
2. Define tasks
3. ($) Release annotated development set
4. Collect predictions
5. ($) Annotate points of disagreement
6. Report scores
7. Retire and release old data

Human costs ($) directly produce data
What are the drawbacks here?

- Loss of algorithmic transparency
- Potential for cheating?
- CC/PD music isn’t “real” enough
What are the drawbacks here?

- Loss of algorithmic transparency
- Potential for cheating?
- CC/PD music isn’t “real” enough

- Linking to source makes results verifiable and replicable!
- What’s the incentive for cheating?
- Even if people do cheat, we still get the annotations.
- For which tasks?
Proposed implementation details (please debate!)

- **Data exchange**
  - OGG + JAMS

- **Evaluation**
  - mir_eval: [https://github.com/craffel/mir_eval](https://github.com/craffel/mir_eval)
  - sed_eval: [https://github.com/TUT-ARG/sed_eval](https://github.com/TUT-ARG/sed_eval)

- **Submissions**
  - CodaLab: [http://codalab.org/](http://codalab.org/)

- **Annotation**
  - Fork NYPL transcript editor?: [https://github.com/NYPL/transcript-editor](https://github.com/NYPL/transcript-editor)
A trial run in 2017: mixed instrument detection

- Complements what is currently covered in MIREX
- Conceptually simple task for annotators
- A large, well-annotated data set would be valuable for the community

To-do:
  a. Collect audio
  b. Define label taxonomy
  c. Build annotation infrastructure
  d. Stretch goal: secure funding for annotators (here’s looking at you, industry folks ;o)
Get involved!

- This only works with community backing

- Help shape this project!

- Lots of great research problems here:
  - Develop web-based annotation tools
  - How to minimize the amount of annotations
  - How to integrate disagreements over many tasks/metrics
  - Evaluate crowd-source accuracy for different tasks
  - Incremental evaluation with ambiguous/subjective data
Thanks!

Let’s discuss at the evaluation town hall and unconference!

http://slido.com

#ismir2016eval
Where do annotations come from?

- Crowd-sourcing can work for some tasks
  - ... but we’ll probably have to train and pay annotators for the difficult ones

- This use of funding is efficient, and a good investment for the community
  - Grants or industrial partnerships can help here
  - Idea: increase/divert ISMIR membership fees toward data creation?

- Point of reference: annotating MedleyDB cost $12/track ($1240 total)
  - $5 per attendee = a new MedleyDB each year
Incremental evaluation

1: estimate missing annotations

2: estimate system performance

Annotations

S1 S2 S3

A D B
? E E
F G F
B B F
? E G

System predictions

Estimated annotations

S1 = 0.4 ± 0.1
S2 = 0.2 ± 0.2
S3 = 0.2 ± 0.1