A plan for sustainable MIR evaluation

Brian McFee*

Eric Humphrey

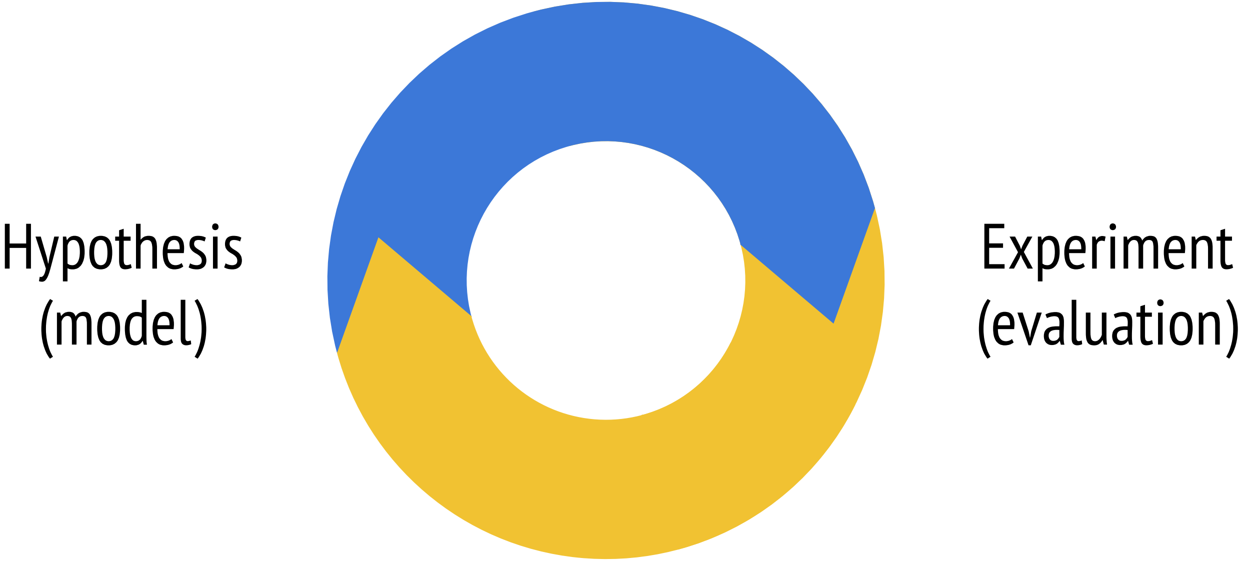
Julián Urbano

NYU | Center for Data Science

MARL

Spotify®

upf. Music Technology Group



Hypothesis
(model)

Experiment
(evaluation)

Progress depends on access to common data

mirex

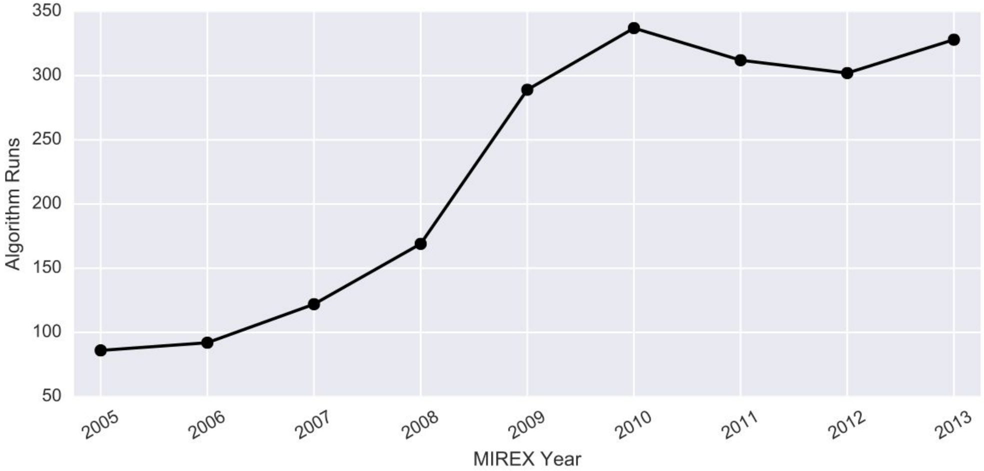
mirex

We’ve known this for a while

- Many years of MIREX!

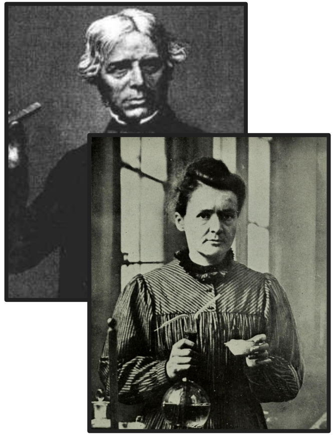
- Lots of participation

- It’s been great for the community

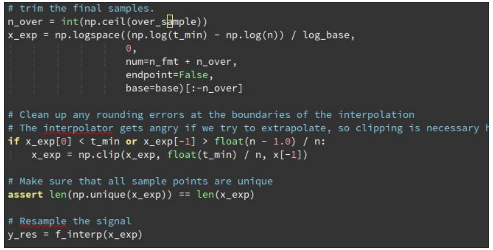


MIREX (cartoon form)

Scientists
(i.e., you folks)

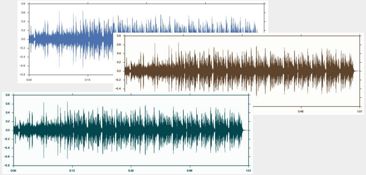


Code

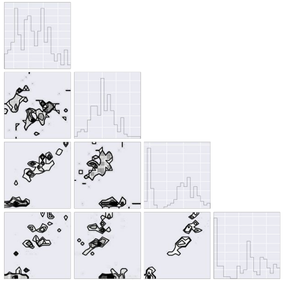


```
# trim the final samples.
n_over = int(np.ceil(over_sample))
x_exp = np.logspace((np.log(t_min) - np.log(n)) / log_base,
                    0,
                    num=n_fmt + n_over,
                    endpoint=False,
                    base=base)[:-n_over]

# Clean up any rounding errors at the boundaries of the interpolation
# The interpolator gets angry if we try to extrapolate, so clipping is necessary h
if x_exp[0] < t_min or x_exp[-1] > float(n - 1.0) / n:
    x_exp = np.clip(x_exp, float(t_min) / n, x[-1])

# Make sure that all sample points are unique
assert len(np.unique(x_exp)) == len(x_exp)

# Resample the signal
y_res = f_interp(x_exp)
```

MIREX machines
(and task captains)

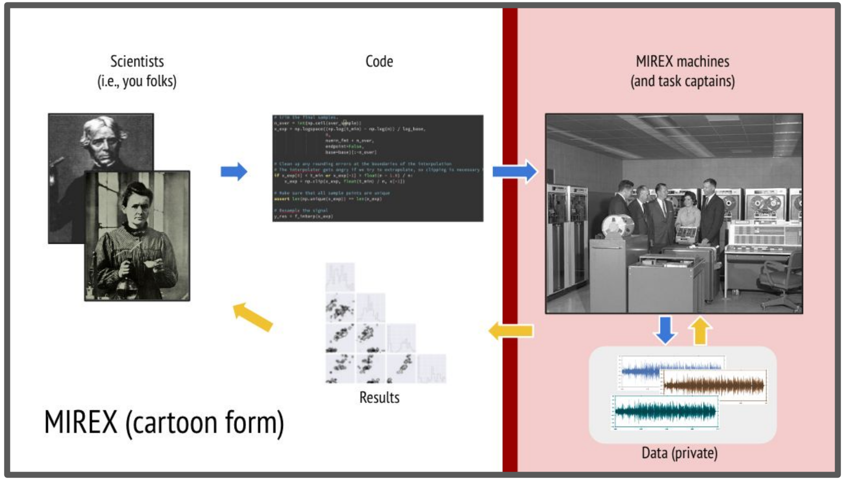


Data (private)



Results

Evaluating the evaluation model



Scientists
(i.e., you folks)

Code

```
# trim the final samples.
n_over = int(np.ceil(over_sample))
x_exp = np.logspace((np.log(t_min) - np.log(n)) / log_base,
                    0,
                    num=n_fft + n_over,
                    endpoint=False,
                    base=base)[:-n_over]

# Clean up any rounding errors at the boundaries of the interpolation
# The interpolator gets angry if we try to extrapolate, so clipping is necessary
if x_exp[0] < t_min or x_exp[-1] > float(n - 1.0) / n:
    x_exp = np.clip(x_exp, float(t_min) / n, x[-1])

# Make sure that all sample points are unique
assert len(np.unique(x_exp)) == len(x_exp)

# Resample the signal
y_res = f_interp(x_exp)
```

MIREX machines
(and task captains)

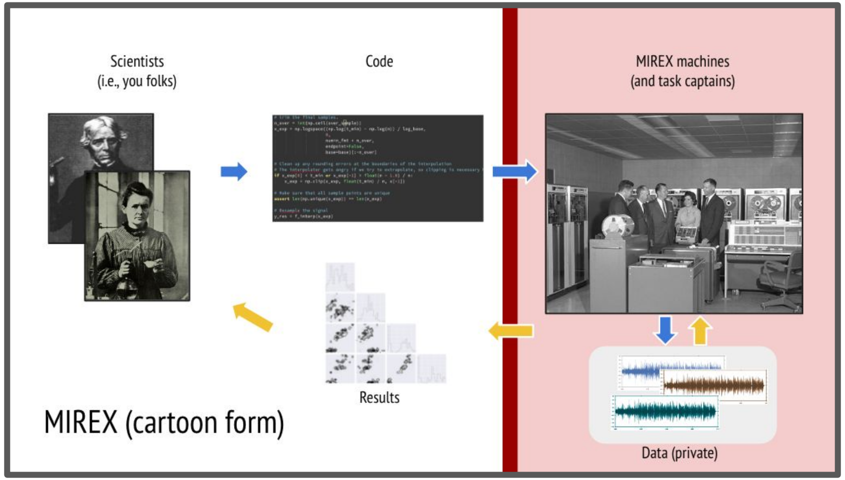
Results

Data (private)

MIREX (cartoon form)

We would not be where we are today without MIREX.

Evaluating the evaluation model



```
# trim the final samples.
n_over = int(np.ceil(over_sample))
x_exp = np.logspace((np.log(t_min) - np.log(n)) / log_base,
                    0,
                    num=n_fmt + n_over,
                    endpoint=False,
                    base=base)[:-n_over]

# Clean up any rounding errors at the boundaries of the interpolation
# The interpolator gets angry if we try to extrapolate, so clipping is necessary
if x_exp[0] < t_min or x_exp[-1] > float(n - 1.0) / n:
    x_exp = np.clip(x_exp, float(t_min) / n, x[-1])

# Make sure that all sample points are unique
assert len(np.unique(x_exp)) == len(x_exp)

# Resample the signal
y_res = f_interp(x_exp)
```

We would not be where we are today without MIREX.

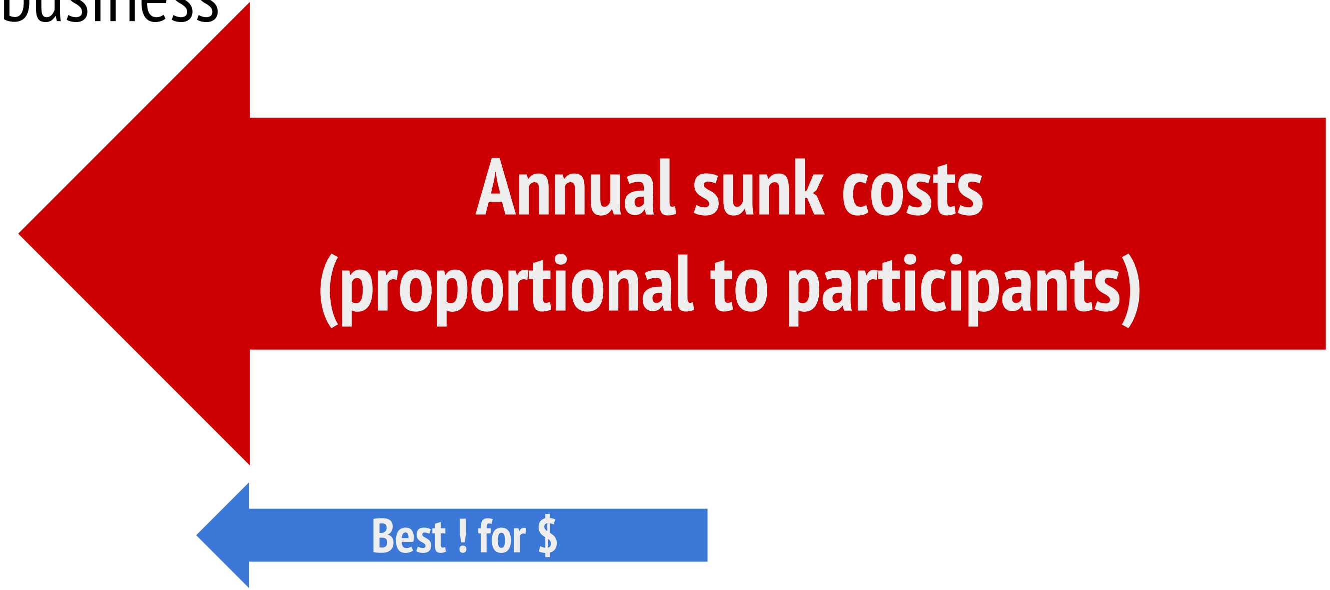
But this paradigm faces an uphill battle :'o(

Costs of doing business

- Computer time
- Human labor
- Data collection

Costs of doing business

- Computer time
- Human labor
- Data collection



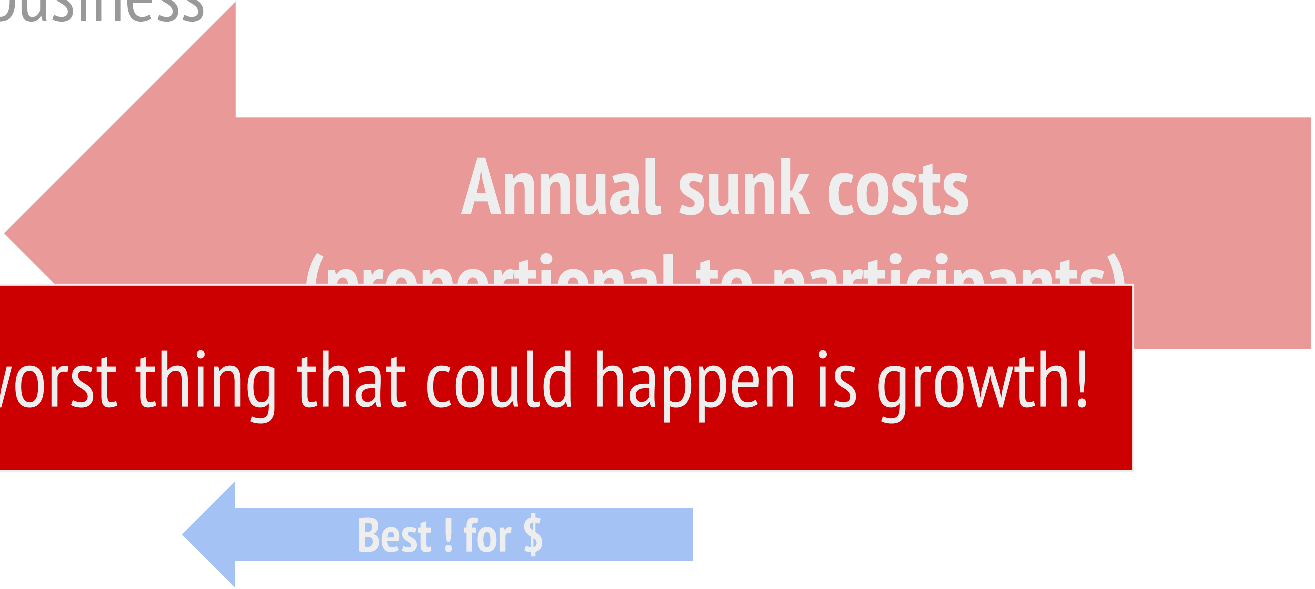
Annual sunk costs
(proportional to participants)

Best ! for $

*arrows are probably not to scale

Costs of doing business

- Computer time
- Human
- Data collection



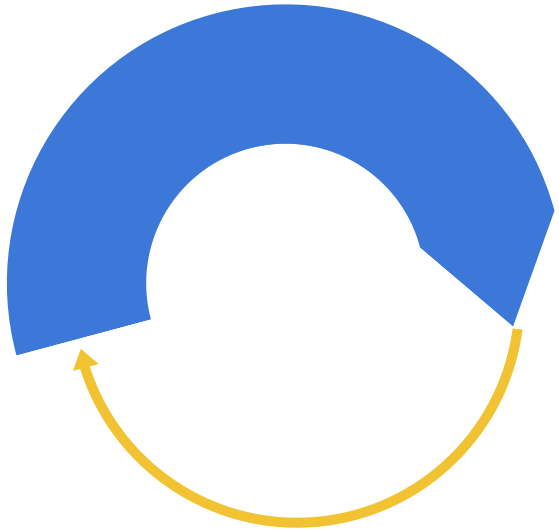
Annual sunk costs (proportional to participants)

The worst thing that could happen is growth!

Best ! for $

*arrows are probably not to scale

Limited feedback in the lifecycle



Hypothesis
(model)

Experiment
(evaluation)

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

Stale data implies bias

nine inch nails
Styx
OLDIES
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

EVERYONE KNOWS ROCK 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 (!!!)

kaggle™

What *is* a sustainable model?

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

- How it works:
  - Dc
  - Up
  - Ob

Distributed computation.

- 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

[Carterette & Allan, ACM-CIKM 2005]

- 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]

Incremental evaluation

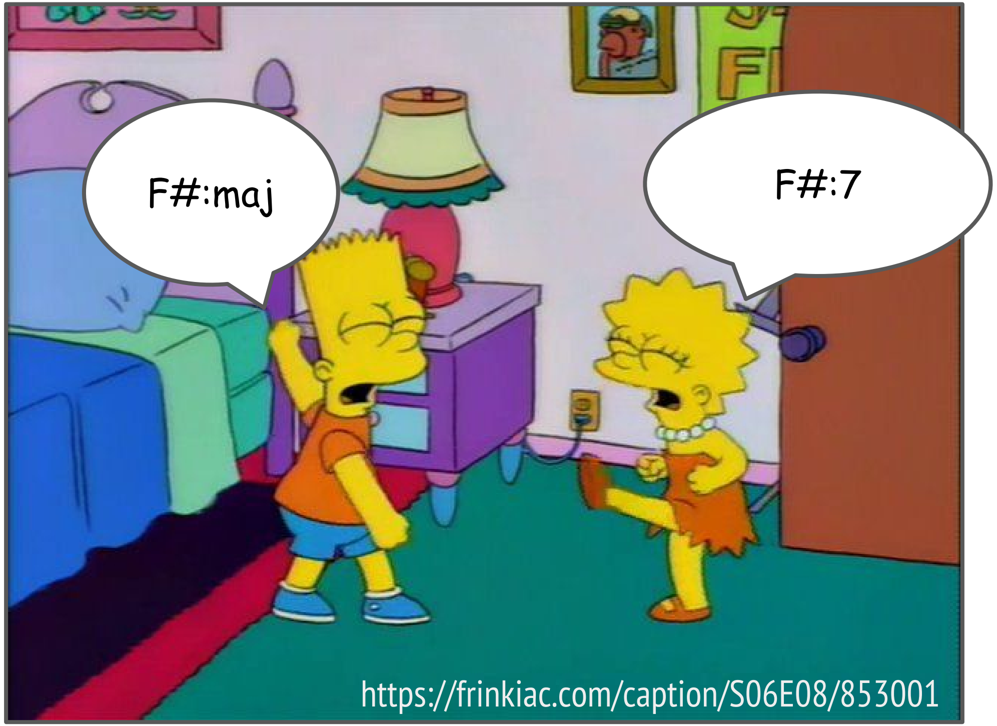
[Carterette & Allan, ACM-CIKM 2005]

- 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!

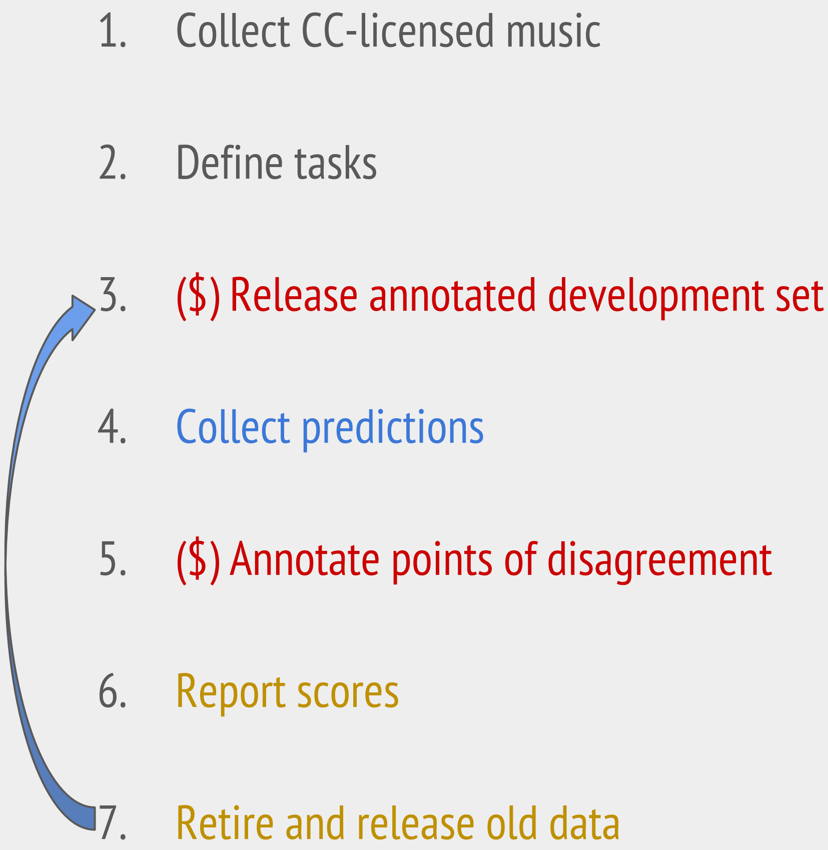
Disagreement can be informative



https://frinkiac.com/caption/S06E08/853001

The evaluation loop

Human costs ($) directly produce **data**



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

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
  - sed_eval https://github.com/TUT-ARG/sed_eval

- Submissions
  - CodaLab http://codalab.org/

- Annotation
  - Fork 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

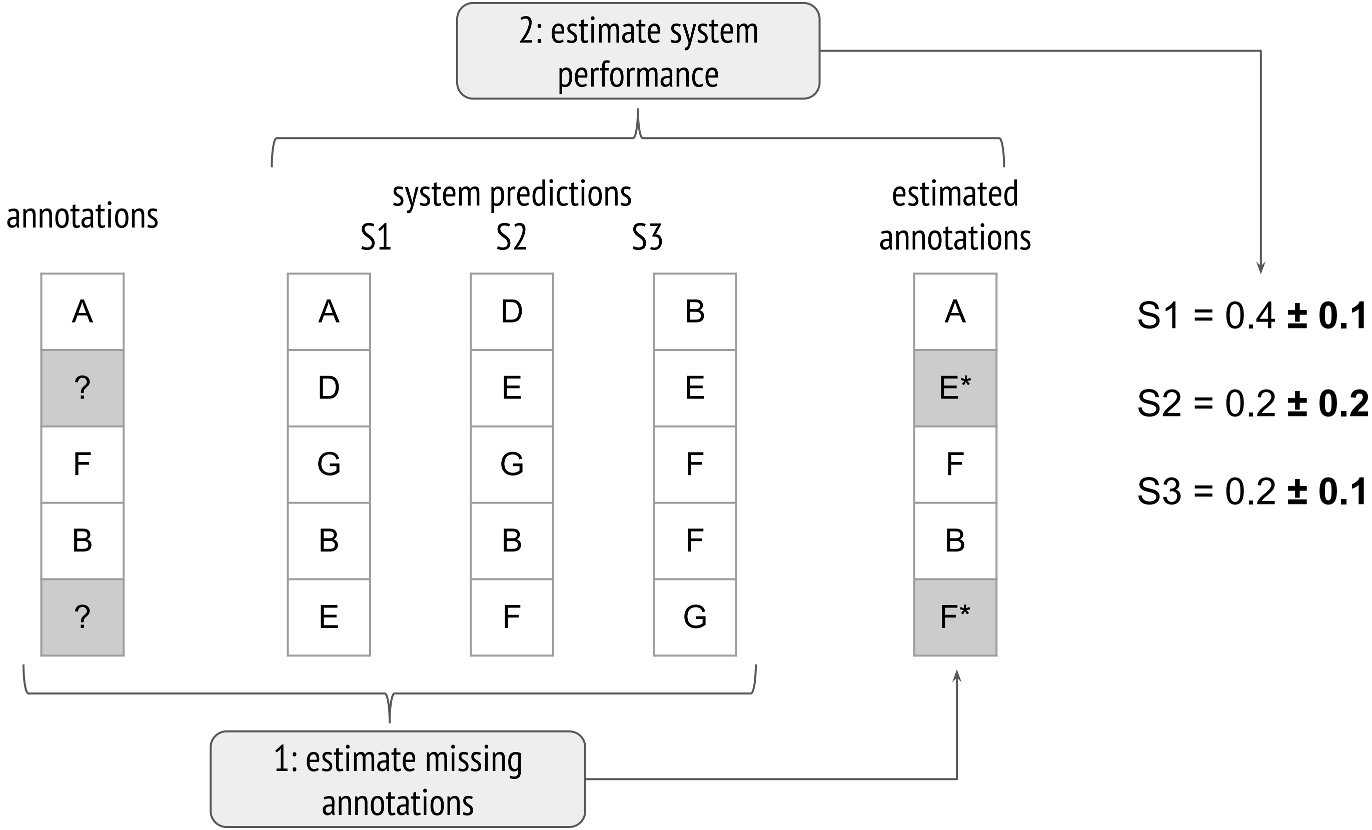
Where do annotations come from?

- 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



2: estimate system performance

annotations	system predictions			estimated annotations
	S1	S2	S3	
A	A	D	B	A
?	D	E	E	E*
F	G	G	F	F
B	B	B	F	B
?	E	F	G	F*

1: estimate missing annotations

S1 = 0.4 **± 0.1**

S2 = 0.2 **± 0.2**

S3 = 0.2 **± 0.1**