Master in Sound and Music Computing

MIR Evaluation Practices

Marius Miron, Lorenzo Porcaro

SMC Master 20/21 (MTG-UPF)

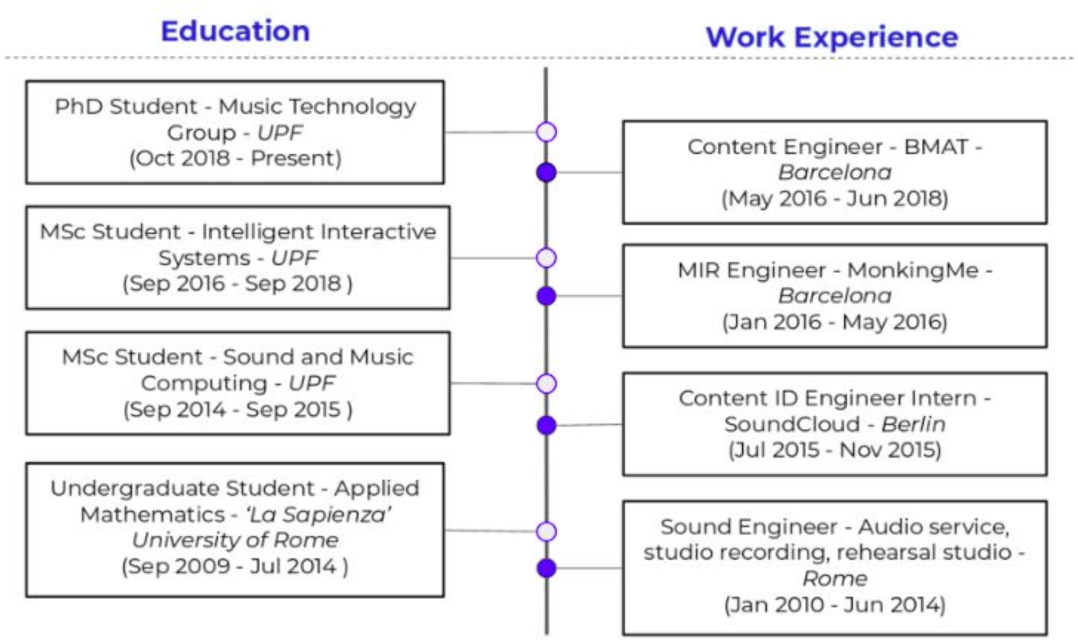
{marius.miron;lorenzo.porcaro}@upf.edu

Lorenzo Porcaro - 3rd year PhD student at the MIRLab (MTG/UPF), supervised by Emilia Gómez and Carlos Castillo (Web Science and Social Computing Group - UPF)

PhD Topic: Assessing the Impact of Music Recommendation Diversity

Main Interests:

- Music Recommender Systems
- MIR Evaluation
- Fairness, Accountability and Transparency in sociotechnical systems



Content

Day 1

1. **Human-centered MIR**
    a. Examples of problematic usages of MIR technologies
    b. Design principles
    c. Bias
    d. Fairness
    e. Interpretability
2. **Ethical Dimension in MIR - Grounding examples (introduction + discussion 1/2)**

Day 2

3. **MIR evaluation practices**
    a. Introduction to Evaluation in (M)IR
    b. Practical Lessons for MIR Evaluation
    c. MIREX, MediaEval
4. **Ethical Dimension in MIR - Grounding examples (discussion 2/2)**

Day 1

Problematic usages of MIR

Fingerprinting as a weapon

New Video Shows Beverly Hills Cops Playing Beatles to Trigger Instagram Copyright Filter

In at least three cases, Beverly Hills Cops have started playing music seemingly to prevent themselves from being filmed by an activist.

DT By Dexter Thomas

February 12, 2021, 3:34am Share Tweet Snap

MORE LIKE THIS

News

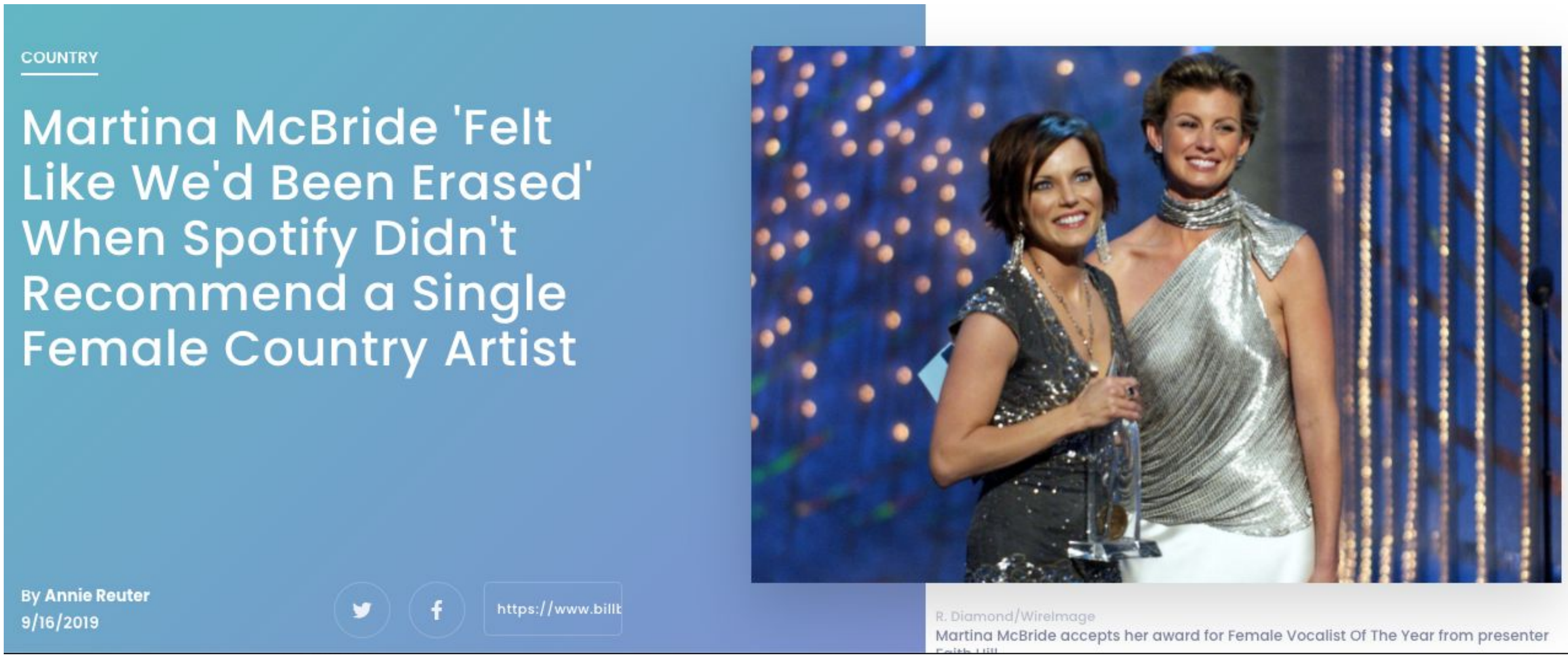
Cops in Minneapolis Can No Longer Turn Off Body Cameras Whenever They Want

VICE NEWS

02.03.21

Tech

Demographic disparity in music recommendation



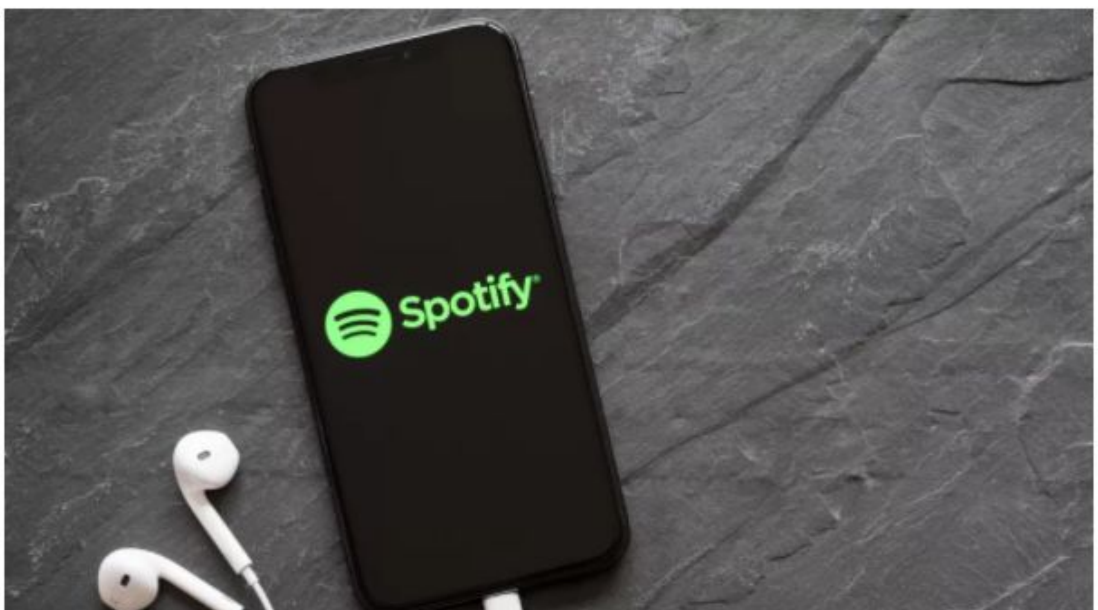
Affective computing and profiling

Spotify wants to know your 'emotional state' for music recommendations

By James Archer January 29, 2021

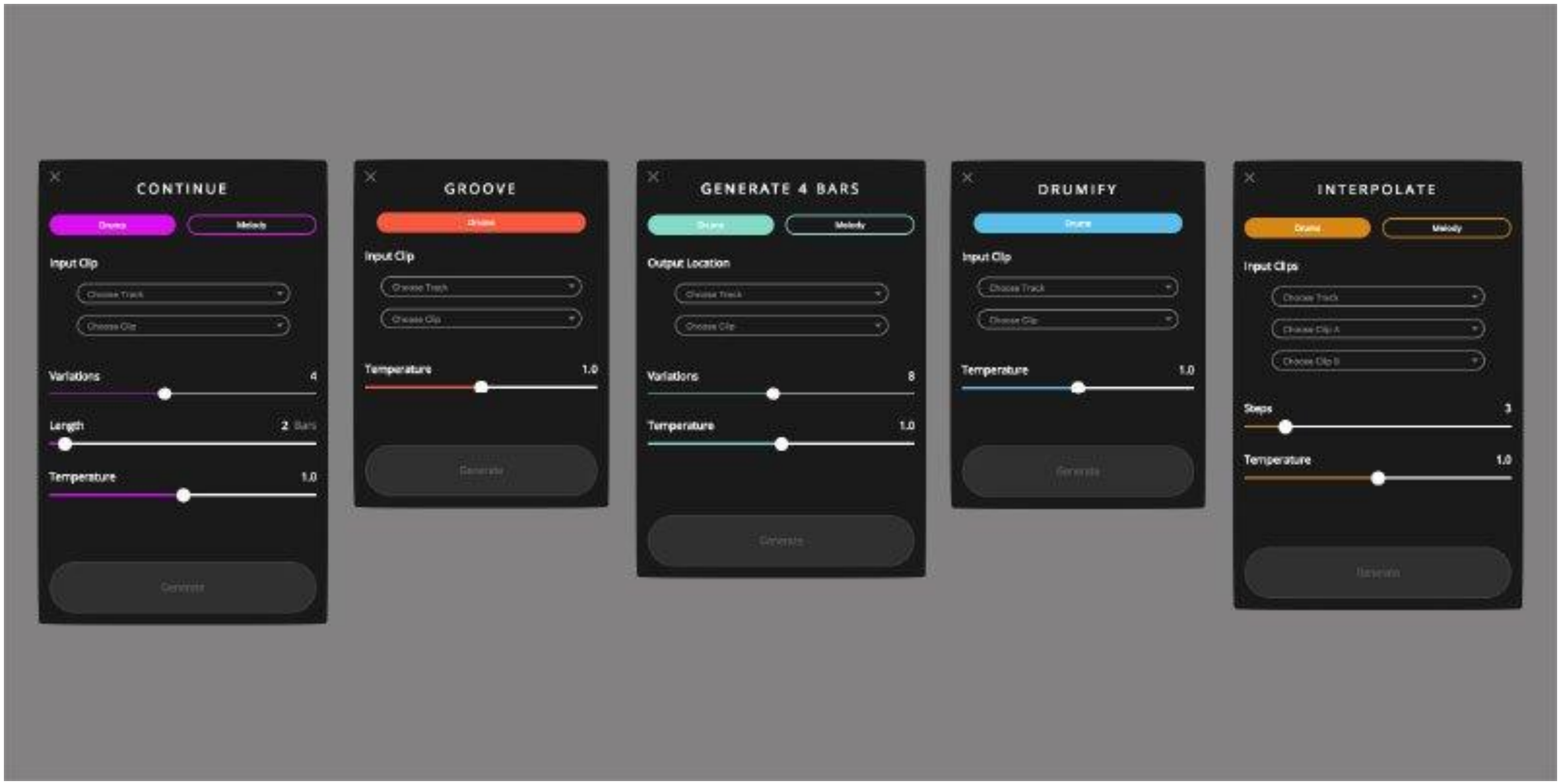
Spotify patent reveals speech recognition plans to make recommendations based on your mood

Comments (1)



(Image credit: Kaspars Grinvalds / Shutterstock)

Data-driven music generation



Master in Sound and Music Computing

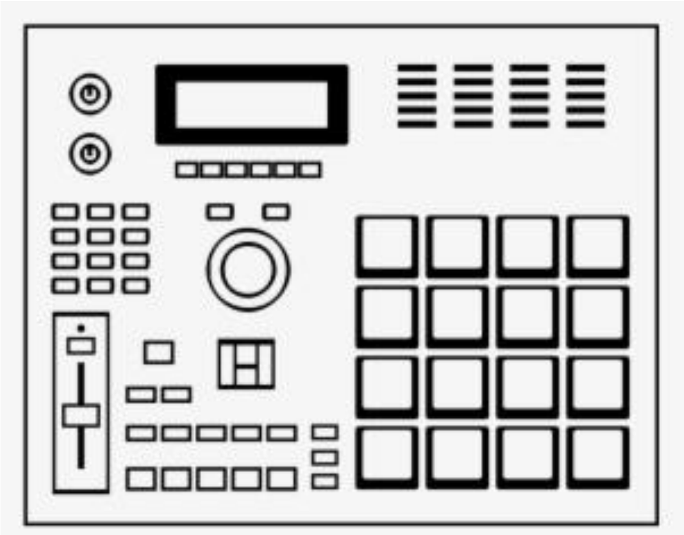
Day 1

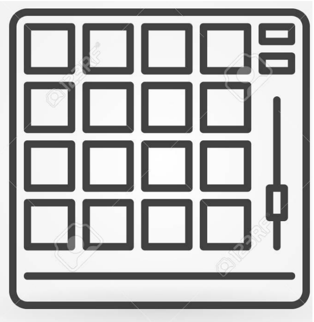
Design principles

Trustworthy AI

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

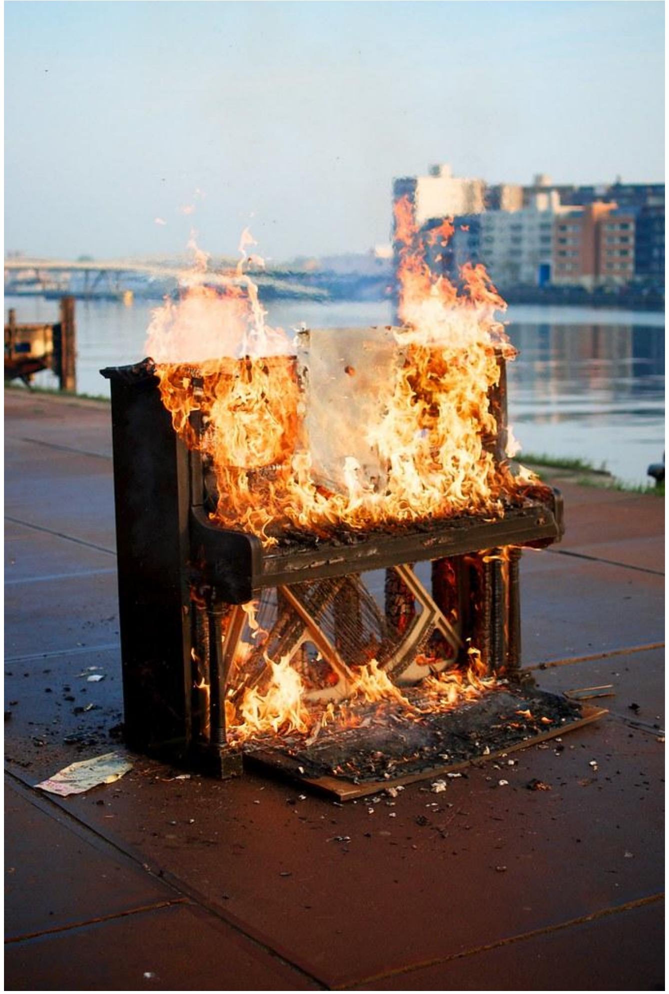
Human agency





https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Technical robustness

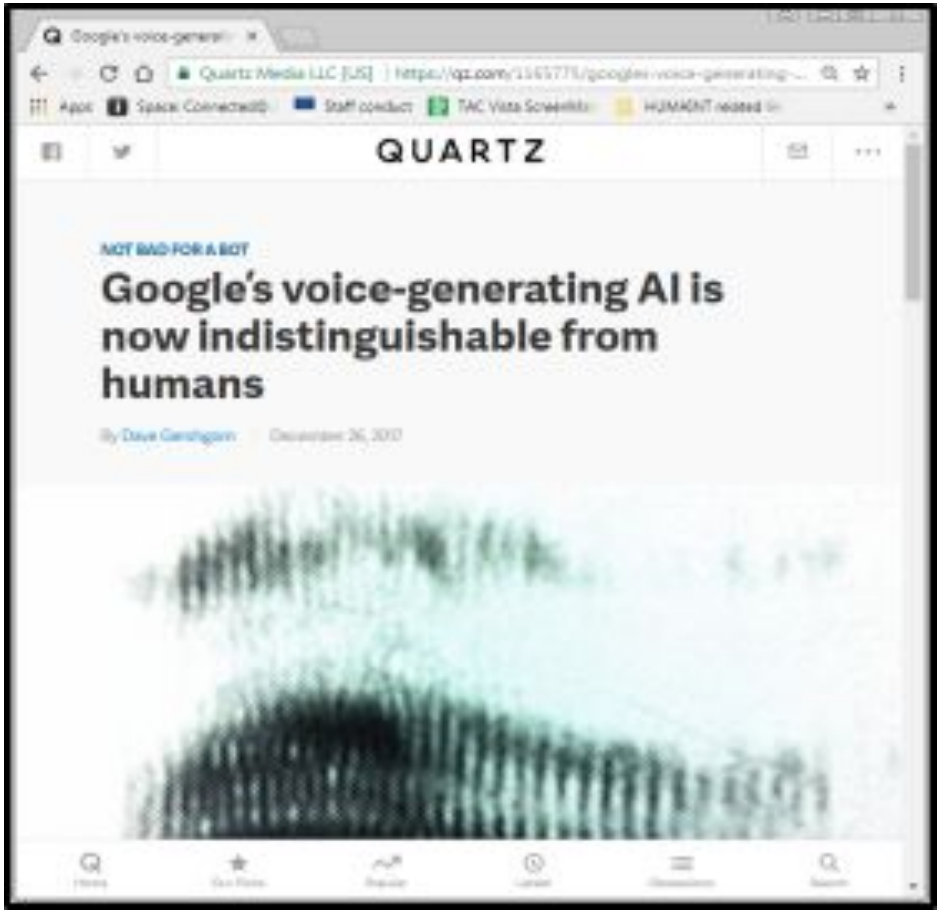


https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Privacy

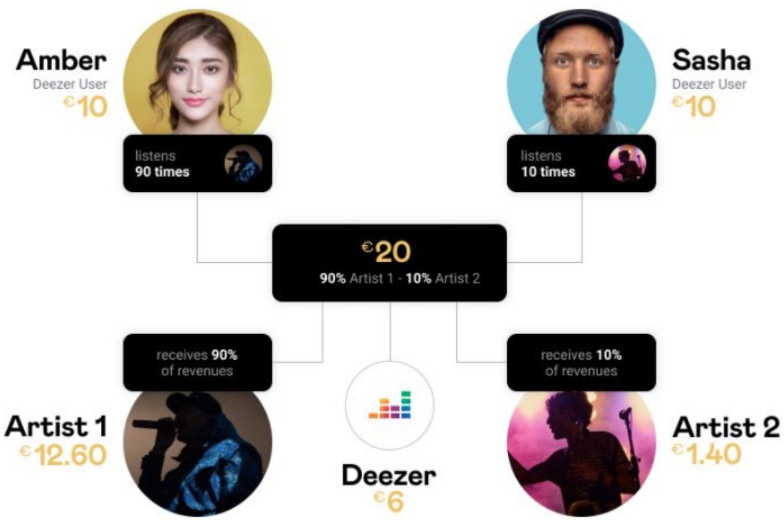
https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

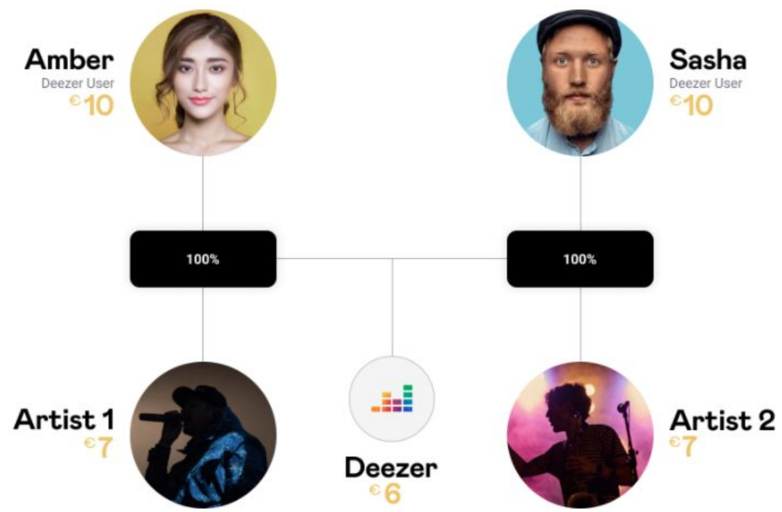
Transparency



https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Fairness





https://www.deezer.com/ucps

Well-being

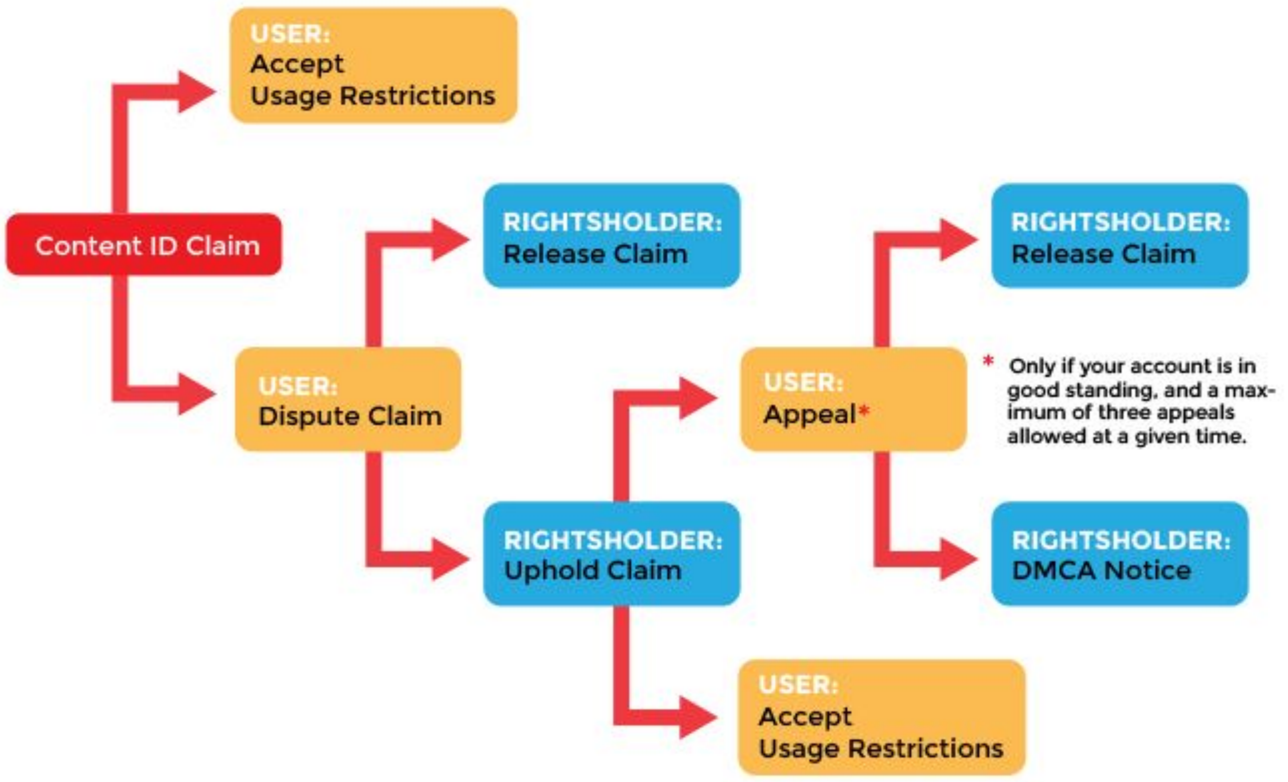
Emissions From Music Consumption Reach Unprecedented High, Study Shows

Overall plastic production has decreased in the streaming era while greenhouse gas emissions have reportedly increased

An Android smartphone with the Spotify music app onscreen, photo by Olly Curtis/Future Publishing via Getty Images

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Accountability

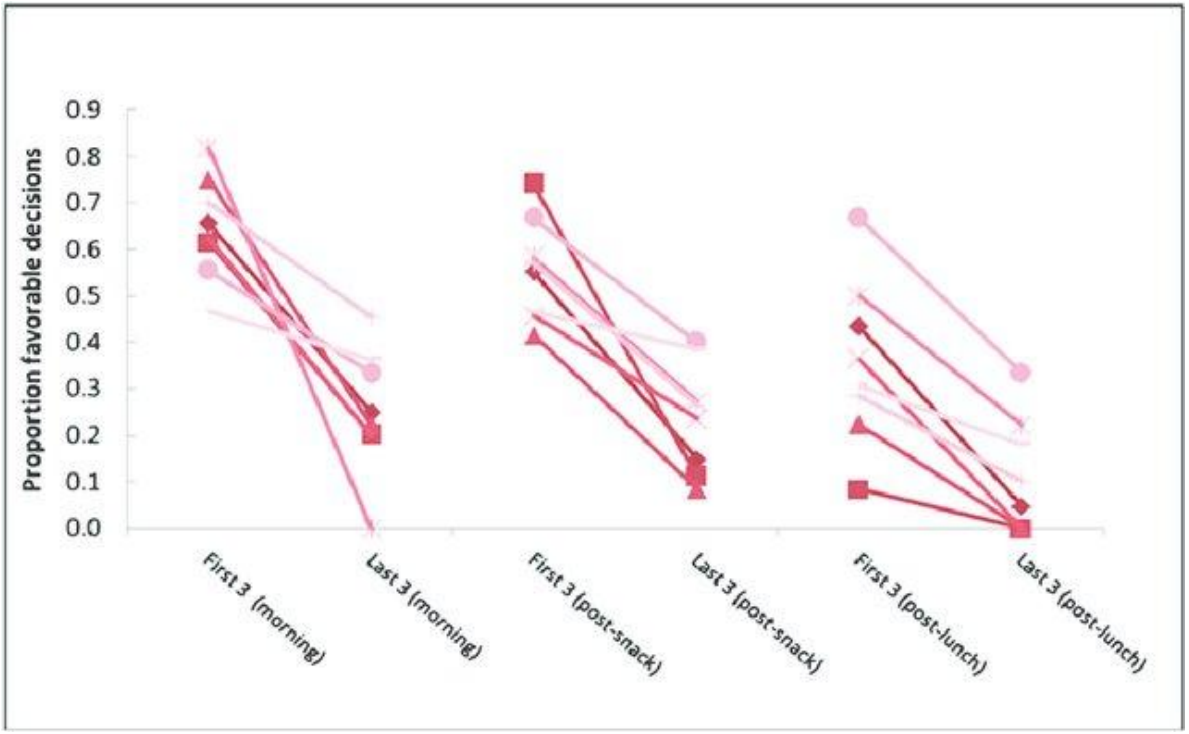


https://www.eff.org/issues/intellectual-property/guide-to-youtube-removals

Day 1

Bias, Fairness, Diversity

Human bias in decision making



Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, *108*(17), 6889–6892. https://doi.org/10.1073/pnas.1018033108

Algorithmic decision making

Algorithmic decision making

Automated underwriting **increased approval rates for minority** and low-income applicants by 30% while improving the overall accuracy of default predictions

Gates, S. W., Perry, V. G., & Zorn, P. M. (2002). Automated underwriting in mortgage lending: Good news for the underserved? Housing Policy Debate, 13(2), 369–391.

[..] results suggest **potentially large welfare gains**: one policy simulation shows crime reductions up to 24.7% with no change in jailing rates, or jailing rate reductions up to 41.9%

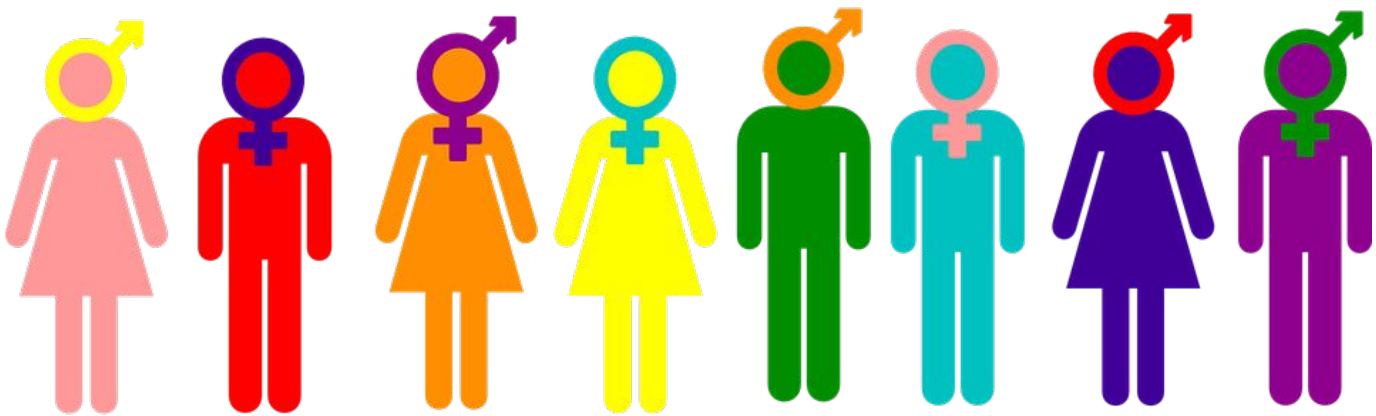
Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions*. The Quarterly Journal of Economics, 133(1), 237–293.

Bias may affect formal assessments and leave room for discrimination

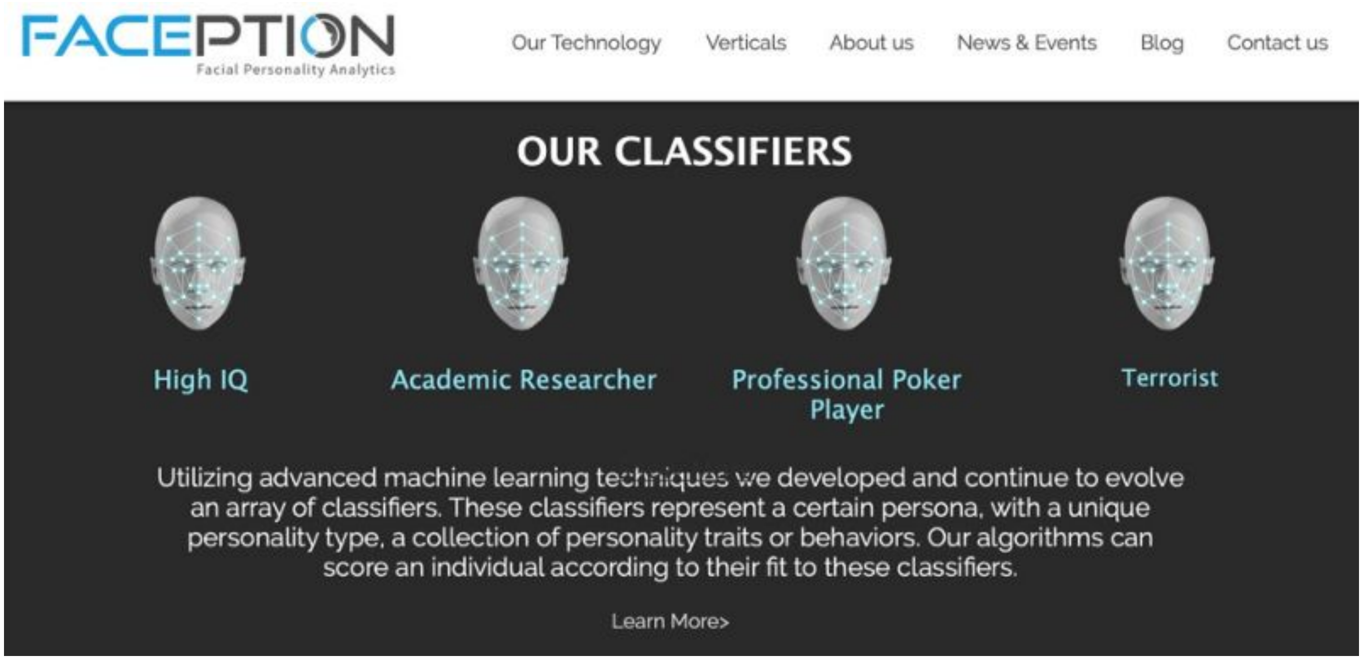
McKay, P. F., & McDaniel, M. A. (2006). A reexamination of black-white mean differences in work performance: More data, more moderators. Journal of Applied Psychology, 91(3), 538–554

Example 1

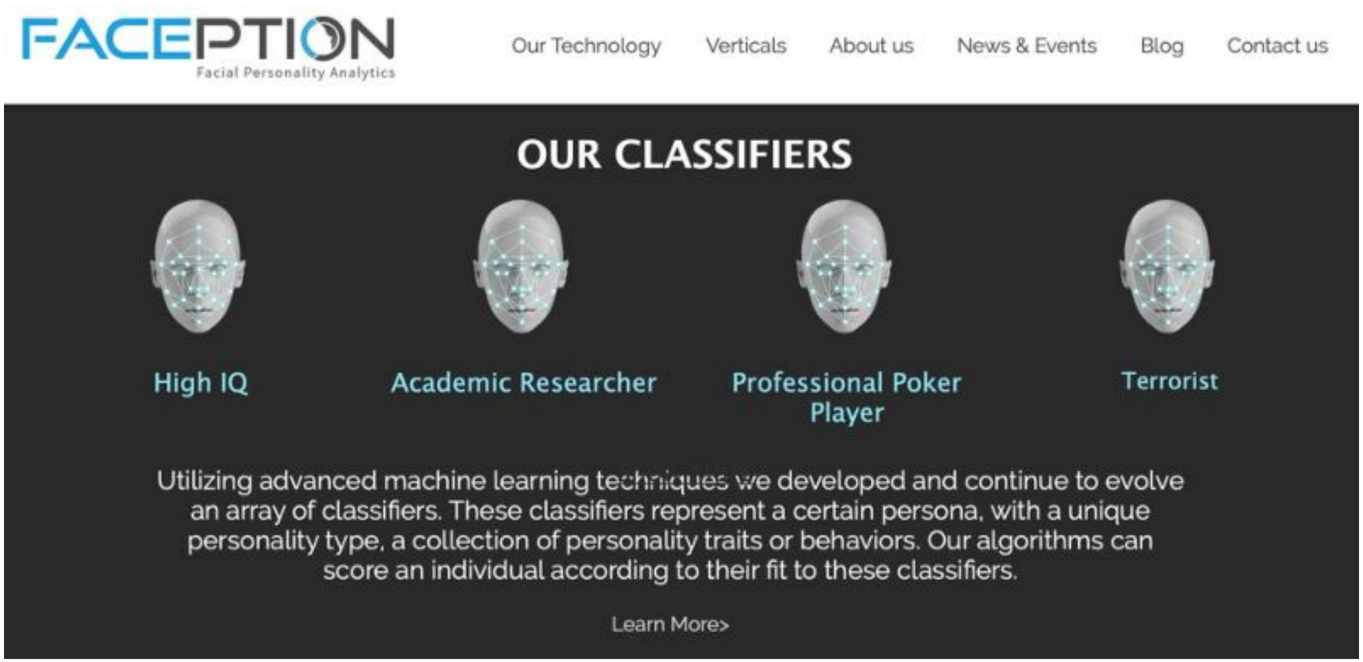
amazon.com®
and you're done.™



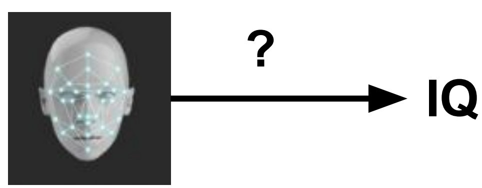
Example 2


FACEPTION
Facial Personality Analytics
Our Technology
Verticals
About us
News & Events
Blog
Contact us
OUR CLASSIFIERS
High IQ
Academic Researcher
Professional Poker Player
Terrorist
Utilizing advanced machine learning techniques we developed and continue to evolve an array of classifiers. These classifiers represent a certain persona, with a unique personality type, a collection of personality traits or behaviors. Our algorithms can score an individual according to their fit to these classifiers.
Learn More>

Example 2



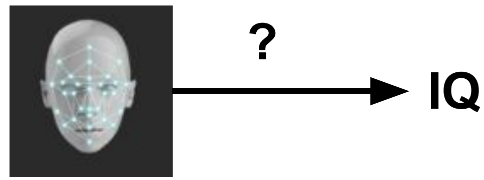
FACEPTION
Facial Personality Analytics
Our Technology
Verticals
About us
News & Events
Blog
Contact us
OUR CLASSIFIERS
High IQ
Academic Researcher
Professional Poker Player
Terrorist
Utilizing advanced machine learning techniques we developed and continue to evolve an array of classifiers. These classifiers represent a certain persona, with a unique personality type, a collection of personality traits or behaviors. Our algorithms can score an individual according to their fit to these classifiers.
Learn More>

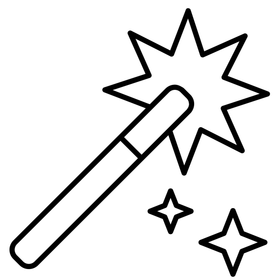


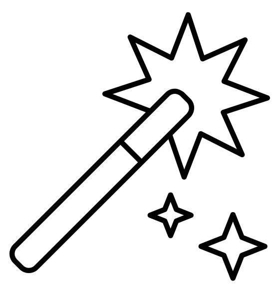
?
IQ

Example 2

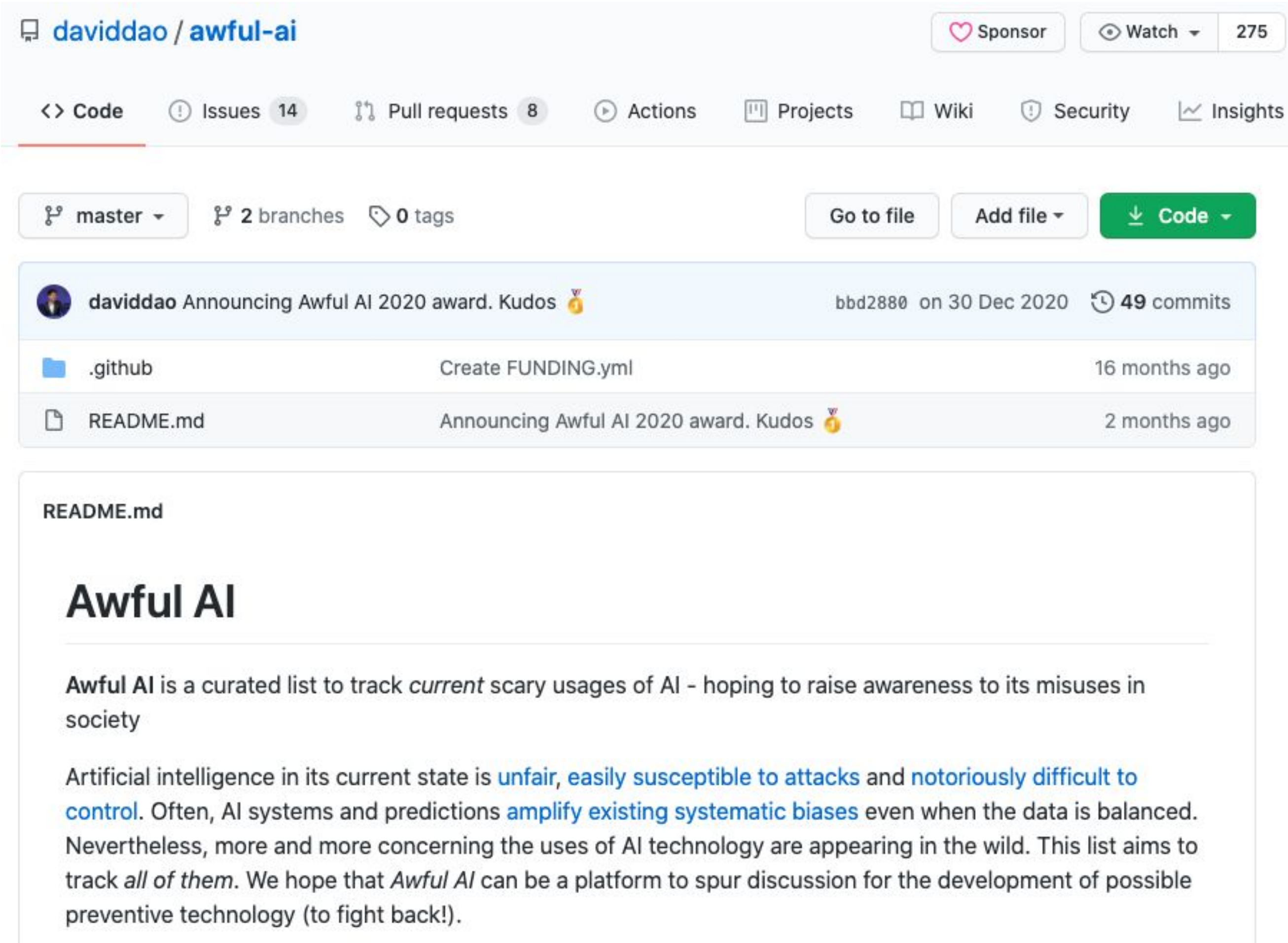
Machine learning magic



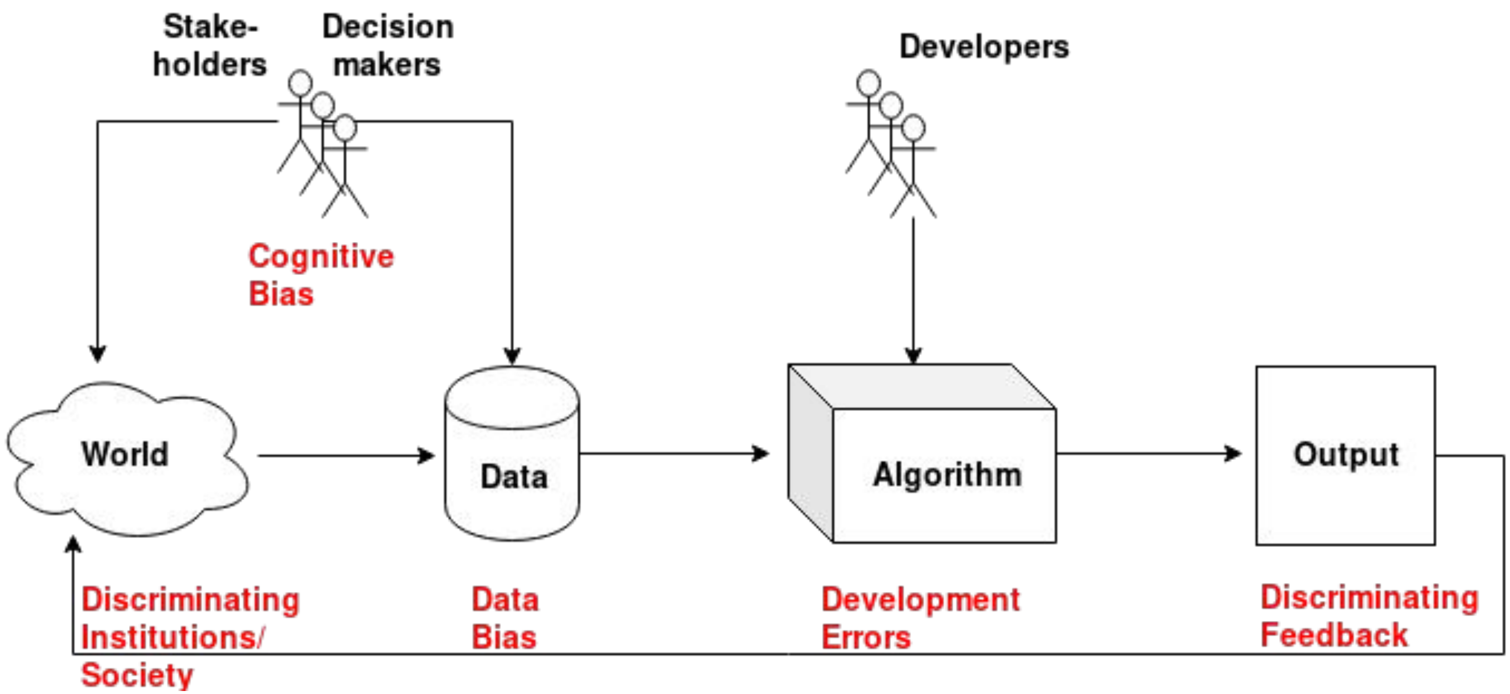
Data  Solved tasks

Data  Spurious correlations

More examples



Bias in socio-technical systems



Tolan S., Discrimination in Algorithmic Justice (2018)

What is bias?

Bias
A feature of statistical models. A systematic deviation from the truth.

Bias in data processing: selection bias, sampling bias, reporting bias

Bias in the machine learning model: bias of an estimator, inductive bias

What is bias?

Bias
A feature of statistical models. A systematic deviation from the truth.

Surprising view of computer scientists:

"The model summarizes the data correctly. If the data is biased it's not the algorithm's fault."

Data biases are inevitable. We must design algorithms that account for them.

Day 1

Fairness & Diversity

What is fairness?

Bias
A feature of statistical models. A systematic deviation from the truth.

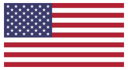
Fairness
A feature of value judgments. Discrimination: A legal concept based on group membership.

What is fairness?

Fairness
A feature of value judgments. Discrimination: A legal concept based on group membership*.

*sex, race, colour, ethnic or social origin, genetic features, language,religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation (Article 14, European Convention on Human Rights)

*sex, race, color, religion, national origin (Civil Rights Act of 1964), citizenship (Immigration Reform and Control Act), age (Age Discrimination in Employment Act of 1967), pregnancy (Pregnancy Discrimination Act), familial status (Civil Rights Act of 1968), disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990), veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act), genetic information (Genetic Information Nondiscrimination Act)



What is fairness?

Real challenge

Design systems that support human values.

Narayanan, 21 fairness definitions and their politics (2018) Tutorial at the ACM Conference on Fairness, Accountability, and Transparency (FAccT) 2018

Ethical dimension

“[..] machine learning should not be used for prediction, but rather to surface covariates that are fed into a causal model for understanding the social, structural and psychological drivers of crime.”

Barabas, C., Dinakar, K., Ito, J., Virza, M., & Zittrain, J. (2018). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. Journal of Machine Learning Research, July.

What is fairness?

Domain specific

How does this system/application affects people that use it/limits their opportunities?

Feature specific

The features have been used for “unjustified and systematically adverse treatment in the past”

Barocas and Hardt, Fairness in Machine Learning (2017). Tutorial at the Advances in Neural Information Processing Systems Conference (NeurIPS)

Disparate treatment

Formal or intentional discrimination

w.r.t a protected feature or proxy variable (e.g. zip code as a proxy for race)

Treatment depends on group membership

Barocas, S., & Selbst, A. D. (2014). Big Data's Disparate Impact. *California Law Review*, *671*, 671–732.

Disparate impact

Unjustified discrimination resulted from facially neutral practices

Outcome depends on group membership

The 80% rule (U.S. Equal Employment Opportunity Commission)

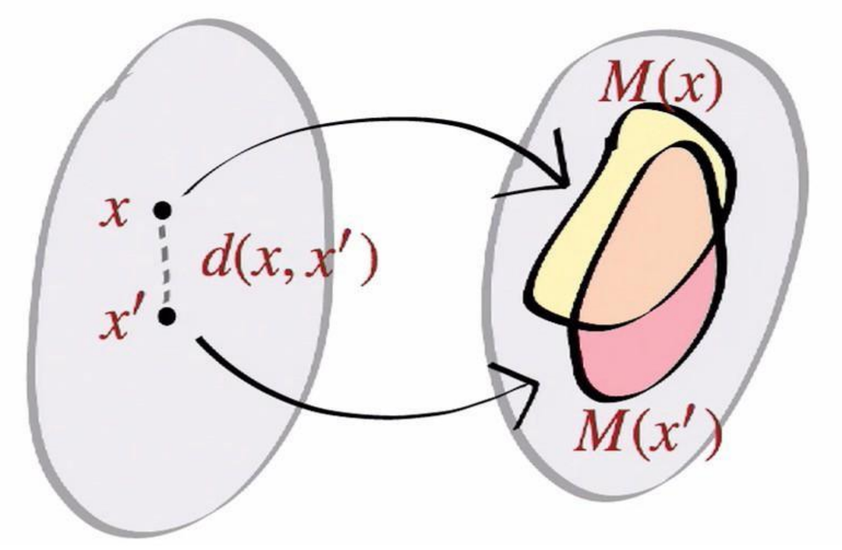
Must come with rigorous proof - account for confounders, exogenous effects

May come in conflict with disparate treatment (Ricci v. DeStefano)

Barocas, S., & Selbst, A. D. (2014). Big Data's Disparate Impact. *California Law Review*, *671*, 671–732.

Individual fairness

Similar individuals should be treated similarly
Assuming a dissimilarity measure d(x,x′), require similar individuals map to similar distributions over outcomes via map M:X→Δ(O)



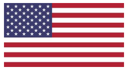
Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12). Association for Computing Machinery, New York, NY, USA, 214–226.

Group fairness

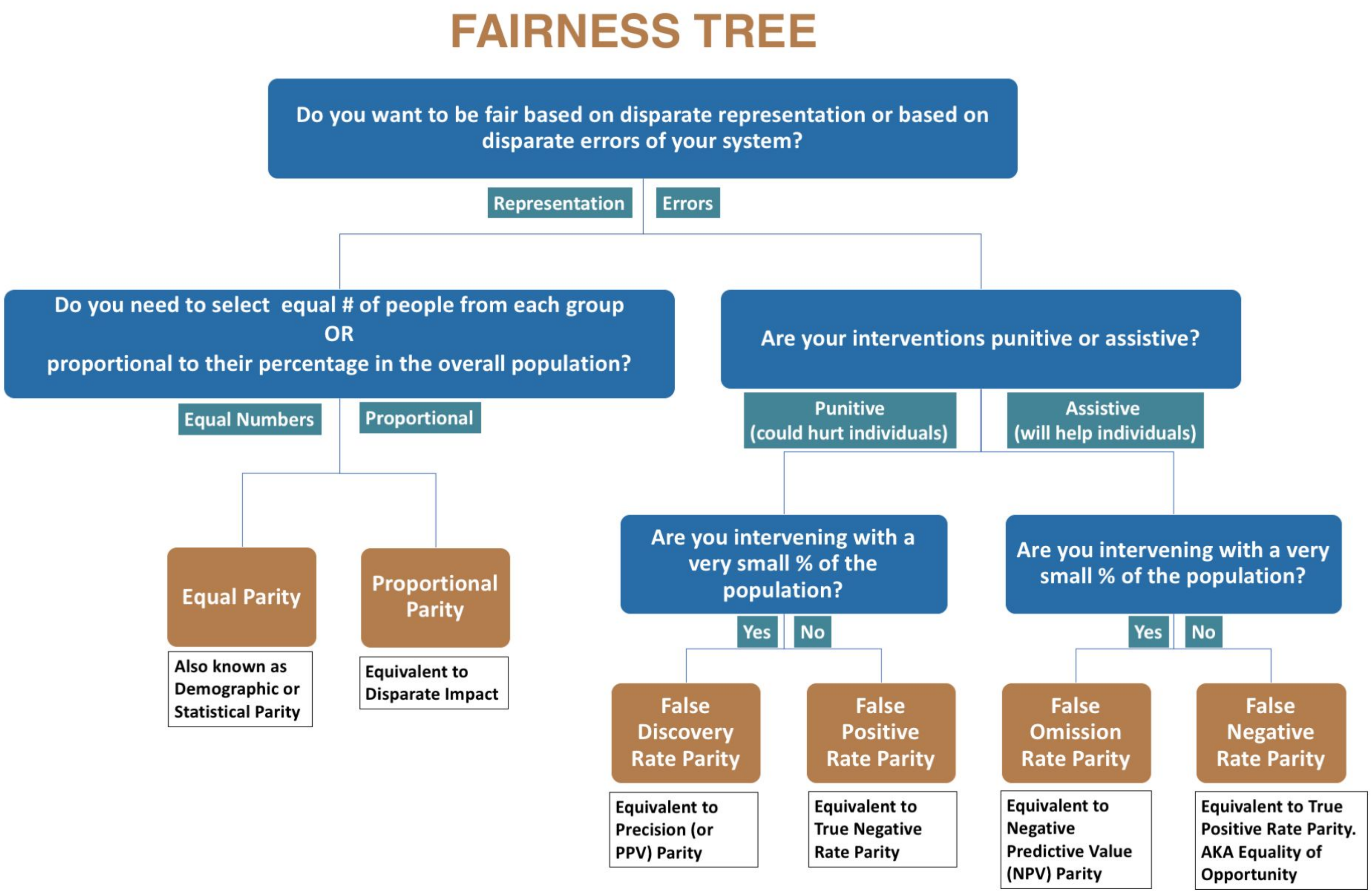
Fairness
A feature of value judgments. Discrimination: A legal concept based on group membership*.

*__sex, race, colour, ethnic or social origin__, genetic features, language,religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation (Article 14, European Convention on Human Rights)

*__sex, race, color, religion, national origin__ (Civil Rights Act of 1964), citizenship (Immigration Reform and Control Act), age (Age Discrimination in Employment Act of 1967), pregnancy (Pregnancy Discrimination Act), familial status (Civil Rights Act of 1968), disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990), veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act), genetic information (Genetic Information Nondiscrimination Act)



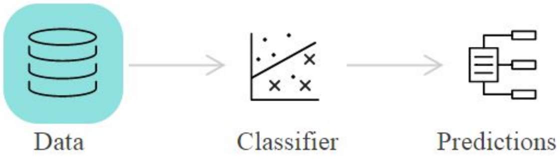
Domain specific



http://aequitas.dssg.io/

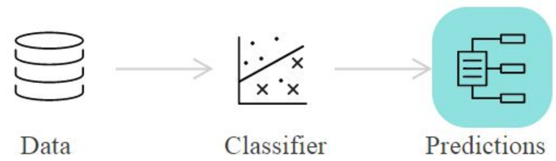
Mitigation

- Pre-processing



- In-processing

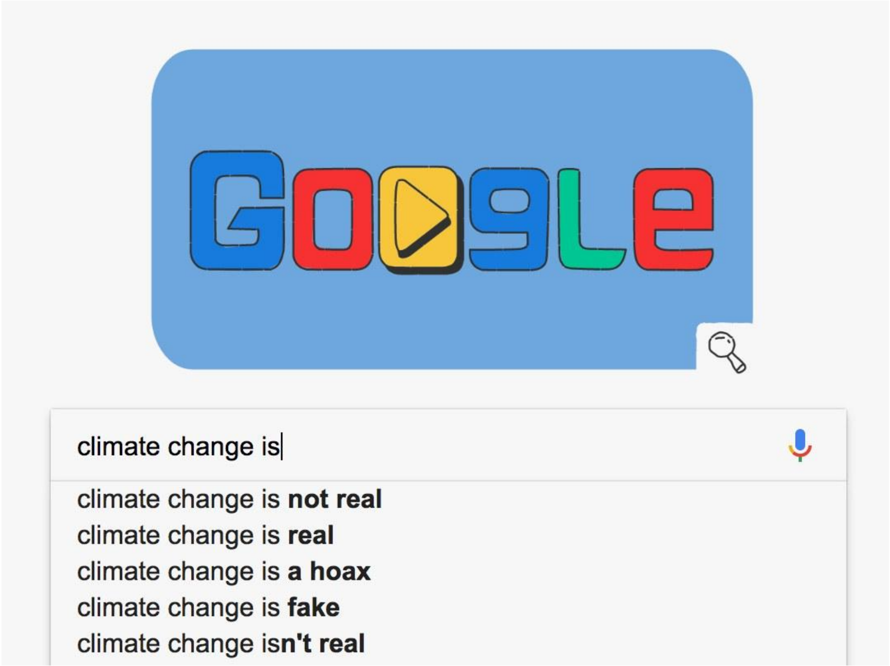
- Post-processing



http://aif360.mybluemix.net/data

Fairness in ranking

1. Demographic parity of protected groups in the top-k candidates (**Diversity**)
2. Some criterion of individual fairness
3. **Ensure no representational harm**



Carlos Castillo. 2019. Fairness and Transparency in Ranking. SIGIR Forum 52, 2 (December 2018), 64–71.

Fairness in recommendation

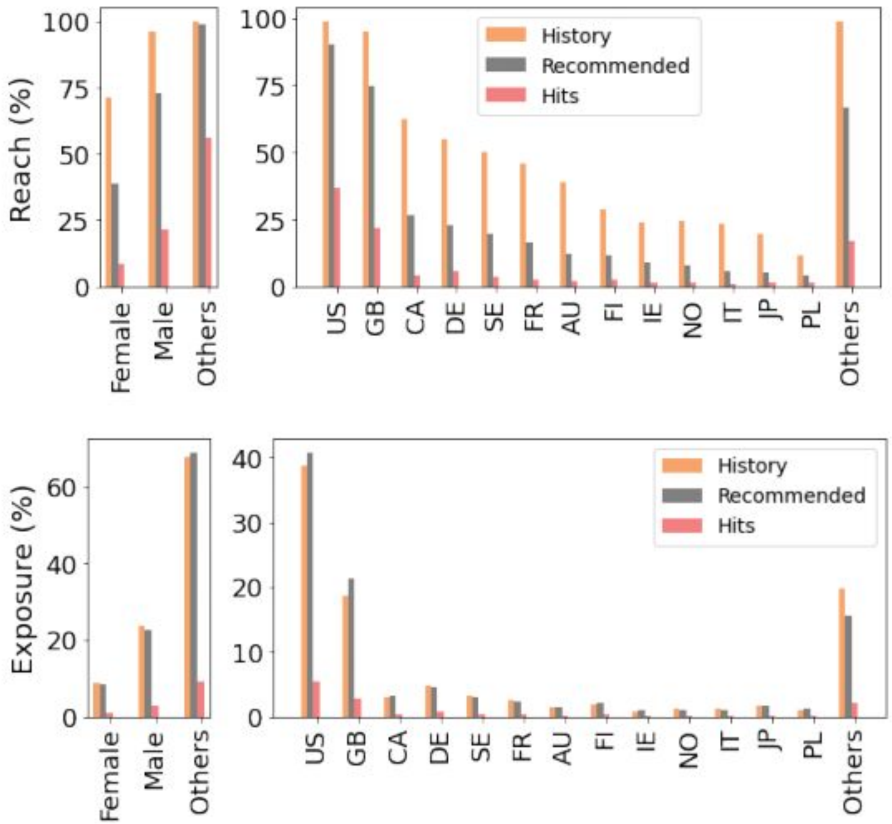
Multi-sided (Group) Fairness

Stakeholder 1	Stakeholder 2
Subject	Consumer
P-fairness	C-fairness
Diversity	

CP-fairness

Burke, Multisided Fairness for Recommendation, (2017) https://arxiv.org/abs/1707.00093

Fairness in music recommendation



Ferraro, Andrés, et al. "Artist biases in collaborative filtering for music recommendation." *Proceedings of the 37 th International Conference on Machine Learning; 2020 Jul 13-18; Vienna, Austria.[Vienna]: ICML; 2020.[3 p.]*. ICML, 2020.

Fairness in music recommendation

$$PR_S(G, C) = \frac{\sum_{u \in G} \sum_{i \in C} S(u, i)}{\sum_{u \in G} \sum_{i \in I} S(u, i)} \qquad BD(G, C) = \frac{PR_R(G, C) - PR_S(G, C)}{PR_S(G, C)}$$

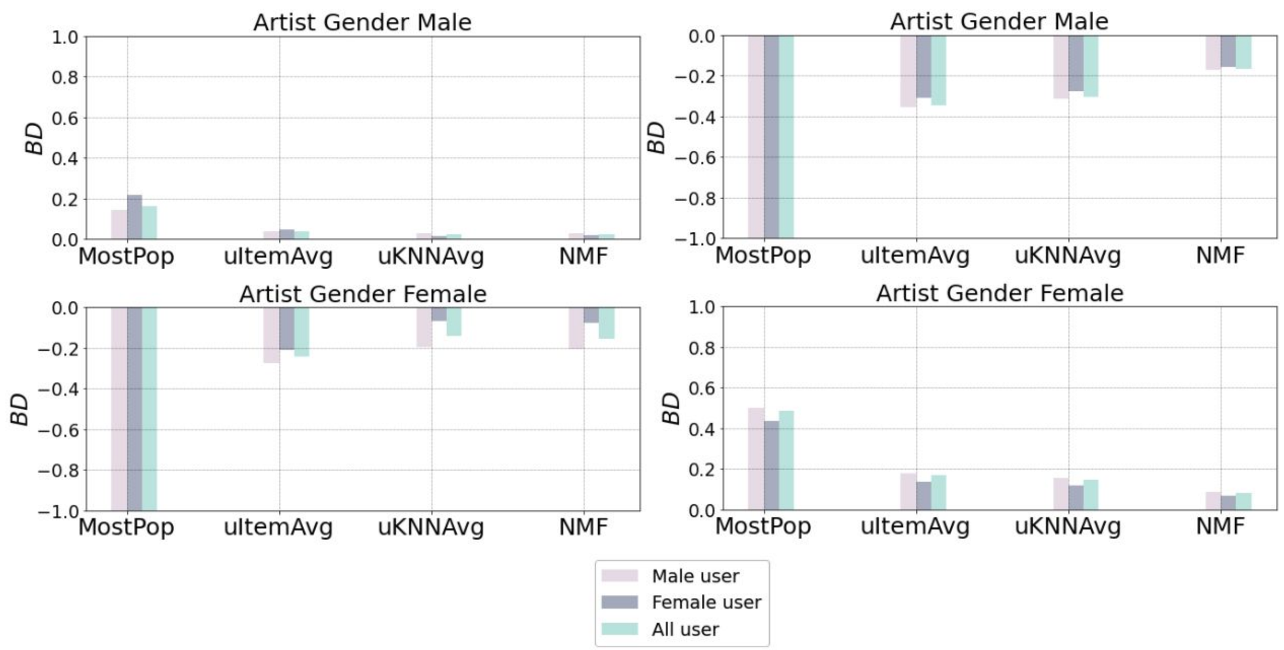


Figure 3: Bias Disparity (BD) results for LFM-1b dataset for experiment 1 (*left column*), and experiment 2 (*right column*).

Shakespeare, D., Porcaro, L., Gómez, E., & Castillo, C. (2020). Exploring Artist Gender Bias in Music Recommendation. 2nd Workshop on the Impact of Recommender Systems (ImpactRS20), Co-Located at RecSys2020.

Fairness in music recommendation

Cramer, H., Garcia-Gathright, J., Springer, A., & Reddy, S. (2018). Assessing and addressing algorithmic bias in practice. *Interactions*, *25*(6), 58-63.

Day 1

Interpretability

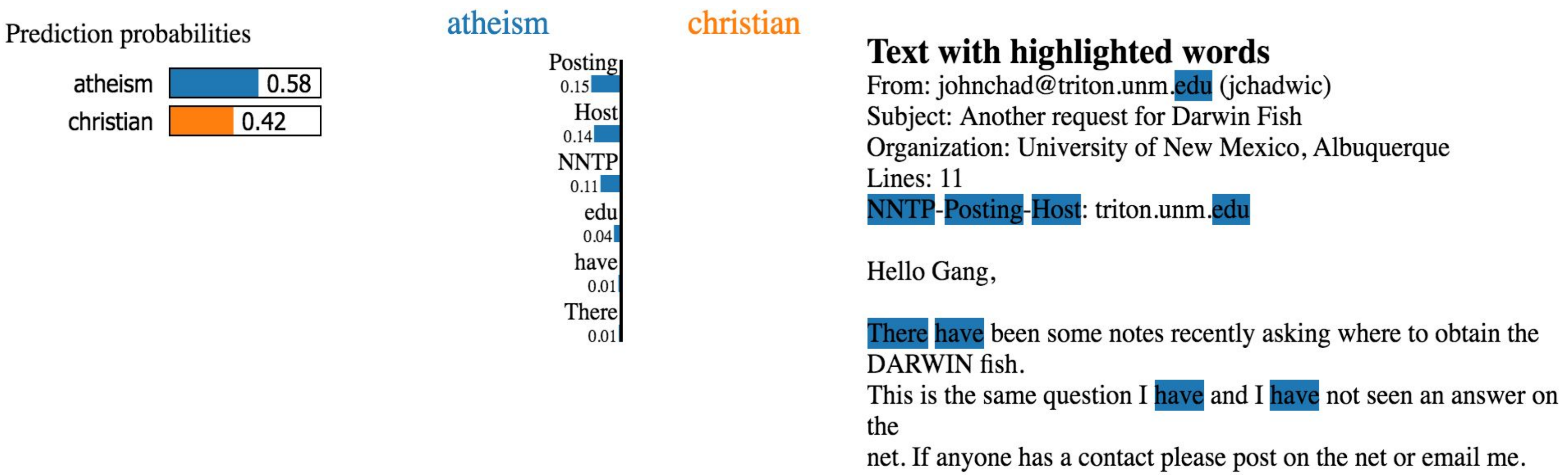
Post-hoc explanations LIME

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$)**, "Acoustic guitar"** ($p = 0.24$) **and "Labrador"** ($p = 0.21$)

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. " Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

Post-hoc explanations LIME



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. " Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

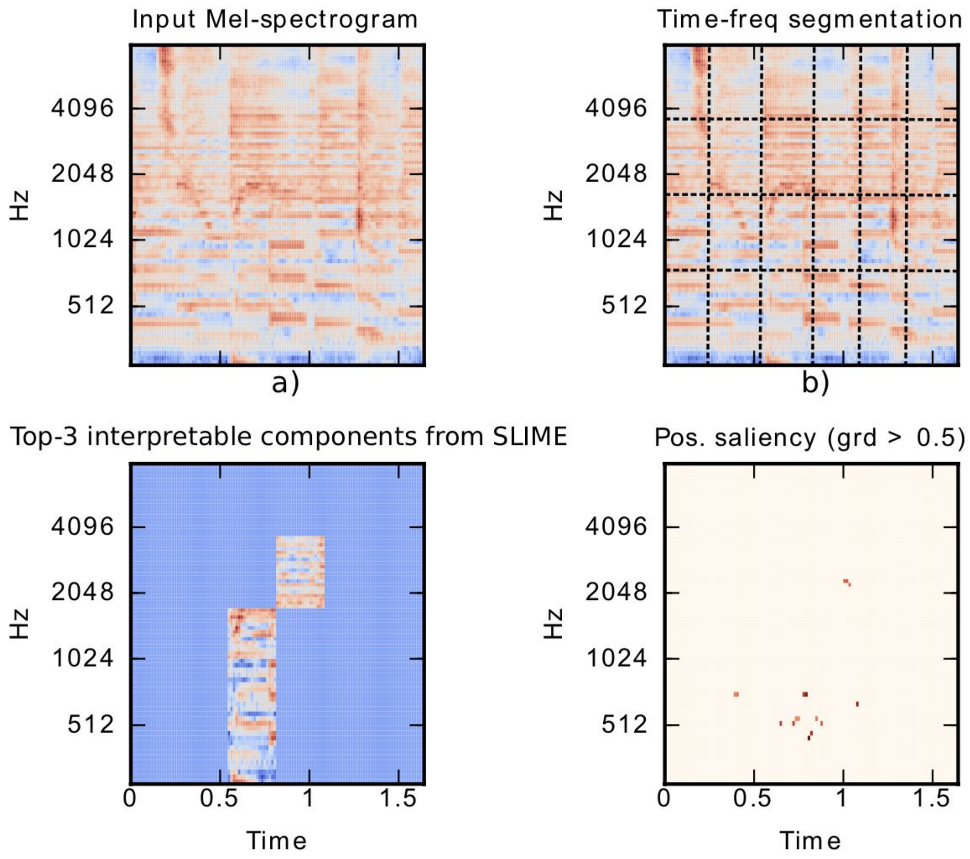
Intrinsic explanation SLIM

PREDICT MUSHROOM IS POISONOUS IF SCORE > 3

1.	$spore\_print\_color = green$	4 points		$\cdots\cdots$
2.	$stalk\_surface\_above\_ring = grooves$	2 points	$+$	$\cdots\cdots$
3.	$population = clustered$	2 points	$+$	$\cdots\cdots$
4.	$gill\_size = broad$	-2 points	$+$	$\cdots\cdots$
5.	$odor \in \{none, almond, anise\}$	-4 points	$+$	$\cdots\cdots$
	ADD POINTS FROM ROWS 1–5	**SCORE**	$=$	$\cdots\cdots$

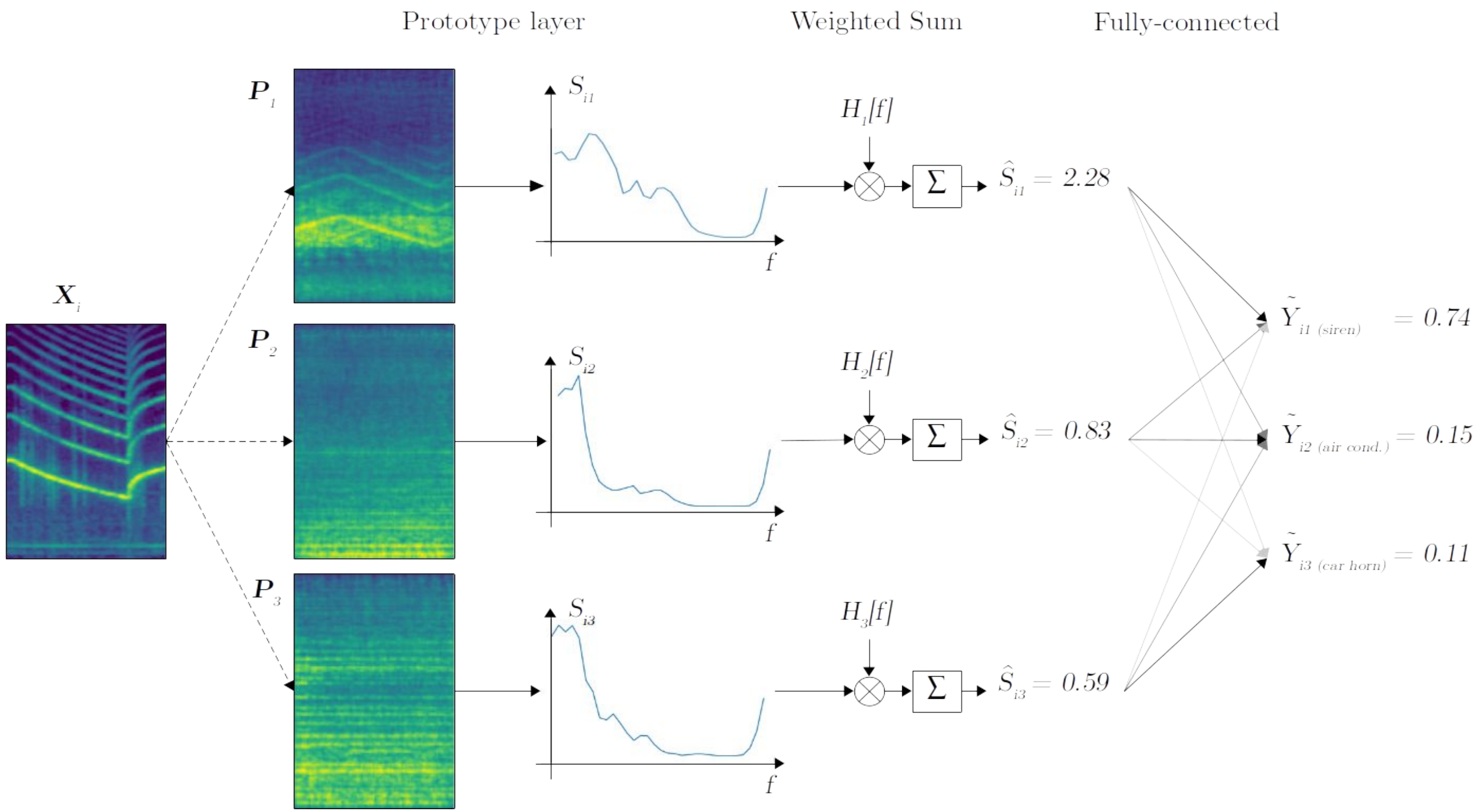
Ustun, Berk, and Cynthia Rudin. "Supersparse linear integer models for optimized medical scoring systems." *Machine Learning* 102.3 (2016): 349-391.

Sound LIME



Mishra, Saumitra, Bob L. Sturm, and Simon Dixon. "Local Interpretable Model-Agnostic Explanations for Music Content Analysis." *ISMIR*. 2017.

Sound classification with prototypes



Zinemanas P. et al. "An Interpretable Deep Learning Model for Automatic Sound Classification" *2021*

Day 1

Group exercise

Group Exercise: Ethical Considerations in MIR

Document: Ethical Considerations in MIR

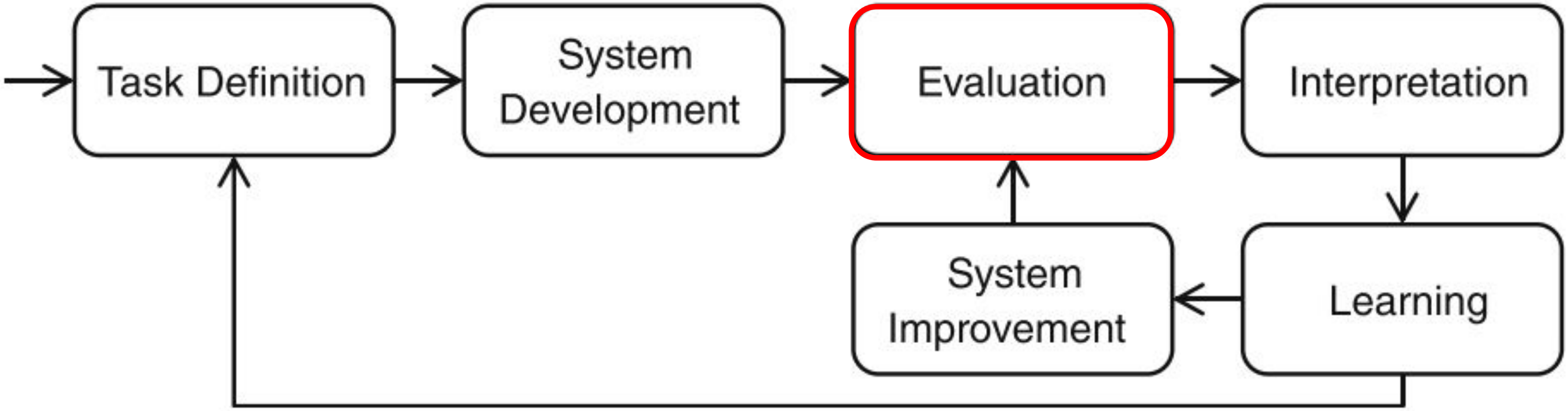
Instructions:

1. You will be assigned to a group and a topic/example.
2. Open the document and read the description of the exercise and the example assigned to your group
3. Focus on understanding what might be the use-case, applications, methodology or evaluation practices and the ethical considerations which may be linked. You can create your own example (related to your master thesis if you wish).
4. Present and discuss your thoughts.

Day 2

MIR Evaluation practices

(M)IR evaluation practices



The (M)IR research and development cycle

Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in music information retrieval. Journal of Intelligent Information Systems, 41(3), 345–369. https://doi.org/10.1007/s10844-013-0249-4

Why (proper) evaluation is important?

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Article | Authors | Metrics | Comments | Media Coverage

Abstract

Modeling the Framework for False Positive Findings

Bias

Testing by Several Independent Teams

Corollaries

Most Research Findings Are False for Most Research Designs and for Most Fields

Claimed Research Findings May Often Be Simply Accurate

Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLOS Medicine, 2(8). https://doi.org/10.1371/journal.pmed.0020124

(M)IR evaluation practices

System-centric

- Accuracy
- Precision-oriented
- Recall-oriented
- F-score
- RMSE
- ...

User-centric

- Satisfaction
- Usefulness
- Perceived Accuracy
- Transparency
- Redundancy
- ...

Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in music information retrieval. Journal of Intelligent Information Systems, 41(3), 345–369. https://doi.org/10.1007/s10844-013-0249-4

Schedl, M., Flexer, A., & Urbano, J. (2013). The neglected user in music information retrieval research. Journal of Intelligent Information Systems, 41(3), 523–539. https://doi.org/10.1007/s10844-013-0247-6

(M)IR evaluation practices

System-centric

- Accuracy
- Precision-oriented
- Recall-oriented
- F-score
- RMSE
- ...

User-centric

- Satisfaction
- Usefulness
- Perceived Accuracy
- Transparency
- Redundancy
- ...

Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in music information retrieval. Journal of Intelligent Information Systems, 41(3), 345–369. https://doi.org/10.1007/s10844-013-0249-4
Schedl, M., Flexer, A., & Urbano, J. (2013). The neglected user in music information retrieval research. Journal of Intelligent Information Systems, 41(3), 523–539. https://doi.org/10.1007/s10844-013-0247-6

Validity

Validity is the extent to which an experiment actually measures what the experimenter intended to measure.

Conclusion Validity: relationship found between our experimental treatments (systems) and our response variables (user-measures). *Can we conclude that the systems are different? How much different?*	**Internal Validity:** confounding factors that might cause the differences we attribute to the systems. *Are those differences caused by specific characteristics of the annotators or the queries?*
External Validity: generalization of that difference to other populations. *Would system differences remain for the wider realm of all genres and artists?*	**Construct Validity**: actual relationship between the system-measures and the user-measures. *Do differences in system-measures directly translate to the same differences in user-measures? How do those differences affect end users?*

Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in music information retrieval. Journal of Intelligent Information Systems, 41(3), 345–369. https://doi.org/10.1007/s10844-013-0249-4

Validity

Validity is the extent to which an experiment actually measures what the experimenter intended to measure.

Conclusion Validity: relationship found between our experimental treatments (systems) and our response variables (user-measures). *Can we conclude that the systems are different? How much different?*	**Internal Validity:** confounding factors that might cause the differences we attribute to the systems. *Are those differences caused by specific characteristics of the annotators or the queries?*
External Validity: generalization of that difference to other populations. *Would system differences remain for the wider realm of all genres and artists?*	**Construct Validity**: actual relationship between the system-measures and the user-measures. *Do differences in system-measures directly translate to the same differences in user-measures? How do those differences affect end users?*

Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in music information retrieval. Journal of Intelligent Information Systems, 41(3), 345–369. https://doi.org/10.1007/s10844-013-0249-4

Validity

Validity is the extent to which an experiment actually measures what the experimenter intended to measure.

Conclusion Validity: relationship found between our experimental treatments (systems) and our response variables (user-measures). *Can we conclude that the systems are different? How much different?*	**Internal Validity:** confounding factors that might cause the differences we attribute to the systems. *Are those differences caused by specific characteristics of the annotators or the queries?*
External Validity: generalization of that difference to other populations. *Would system differences remain for the wider realm of all genres and artists?*	**Construct Validity**: actual relationship between the system-measures and the user-measures. *Do differences in system-measures directly translate to the same differences in user-measures? How do those differences affect end users?*

Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in music information retrieval. Journal of Intelligent Information Systems, 41(3), 345–369. https://doi.org/10.1007/s10844-013-0249-4

Validity

Validity is the extent to which an experiment actually measures what the experimenter intended to measure.

Conclusion Validity: relationship found between our experimental treatments (systems) and our response variables (user-measures). *Can we conclude that the systems are different? How much different?*	**Internal Validity:** confounding factors that might cause the differences we attribute to the systems. *Are those differences caused by specific characteristics of the annotators or the queries?*
External Validity: generalization of that difference to other populations. *Would system differences remain for the wider realm of all genres and artists?*	**Construct Validity**: actual relationship between the system-measures and the user-measures. *Do differences in system-measures directly translate to the same differences in user-measures? How do those differences affect end users?*

Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in music information retrieval. Journal of Intelligent Information Systems, 41(3), 345–369. https://doi.org/10.1007/s10844-013-0249-4

Validity

Scenario

Can a machine learning model identify Blues, Country, and Reggae music?

- I create a labeled collection of 30-second music clips in each of these classes.
- I partition this collection into training and testing datasets.
- I train models A, B, C on the training dataset and compute their accuracies on the testing dataset.

Model	Accuracy
A	0.80
B	0.92
C	0.45

Flexer, A., Sturm, B. L.T. , & Urbano, J. (2020). Do We Care About the Validity of MIR research. ISMIR 2020, Special Session

Validity

Experiment

- To measure the effect of a treatment on a dependent variable

Treatment
manipulated by researcher
(independent variable)

Trained model (A, B or C)

Effect
measured by researcher
(dependent variable)

Inferred label
(summarized by accuracy)

Flexer, A., Sturm, B. L.T. , & Urbano, J. (2020). Do We Care About the Validity of MIR research. ISMIR 2020, Special Session

Validity

Can a machine learning model identify Blues, Country, and Reggae music?

Scenario

- I create a labeled collection of 30-second music clips in each of these classes.
- I partition this collection into training and testing datasets.
- I train models A, B, C on the training dataset and compute their accuracies on the testing dataset: A: 0.8, B: 0.92, C: 0.45
- I conclude:
    - *Model B is the best and can identify Blues, Country and Reggae music with 92% accuracy.*
    - *Features used by B are informative of Blues, Country and Reggae music.*
    - *The machine learning used by B is good for learning to identify Blues, Country and Reggae music.*

Flexer, A., Sturm, B. L.T. , & Urbano, J. (2020). Do We Care About the Validity of MIR research. ISMIR 2020, Special Session

Validity

Statistical conclusion validity

The validity of inferences about the correlation (covariation) between treatment and effect.

- Differences in accuracies between systems may not be statistically significant
- Differences in accuracies between systems may not be significant *with respect to users*

- What are some threats to this?
  - Low power of the experiment
  - Assumptions of test are violated
  - *p*-hacking

Flexer, A., Sturm, B. L.T. , & Urbano, J. (2020). Do We Care About the Validity of MIR research. ISMIR 2020, Special Session

Validity

Internal validity

Is the observed relationship causal or could confounding factors explain the relation?

Is a high accuracy *really* due to identifying these kinds of music? Can it be due to other things?

What are some threats to this?

- Data collection can introduce factors confounded with music label
  - *Infra-sonic information in GTZAN*
  - *Tempo information in BALLROOM*

Flexer, A., Sturm, B. L.T. , & Urbano, J. (2020). Do We Care About the Validity of MIR research. ISMIR 2020, Special Session

Validity

External validity

Do cause-effect relationships also hold for target populations beyond the sample used in the experiment?

- Bad generalization to out-of-sample data sets
- Bad generalization to marginally altered data (adversarial examples)
- What if different people create ground truth annotations?

- Major threats to this:
  - Sampling of population(s) not representative
  - Lack of internal validity
  - Lack of construct validity

Flexer, A., Sturm, B. L.T. , & Urbano, J. (2020). Do We Care About the Validity of MIR research. ISMIR 2020, Special Session

The Weirdest People in the World?

RatSWD Working Paper No. 139

69 Pages • Posted: 11 May 2010

Joe Henrich
University of British Columbia; Harvard University - Department of Human Evolutionary Biology

Steven J. Heine
University of British Columbia (UBC)

Ara Norenzayan
University of British Columbia (UBC)

Date Written: May 7, 2010

Keywords: external validity, population variability, experiments, cross-cultural research, culture, human universals, generalizability, evolutionary psychology, cultural psychology, behavioral economics

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The Weirdest People in the World? In RatSWD Working Paper (Issue 139). https://doi.org/10.2139/ssrn.1601785

Validity

Construct Validity

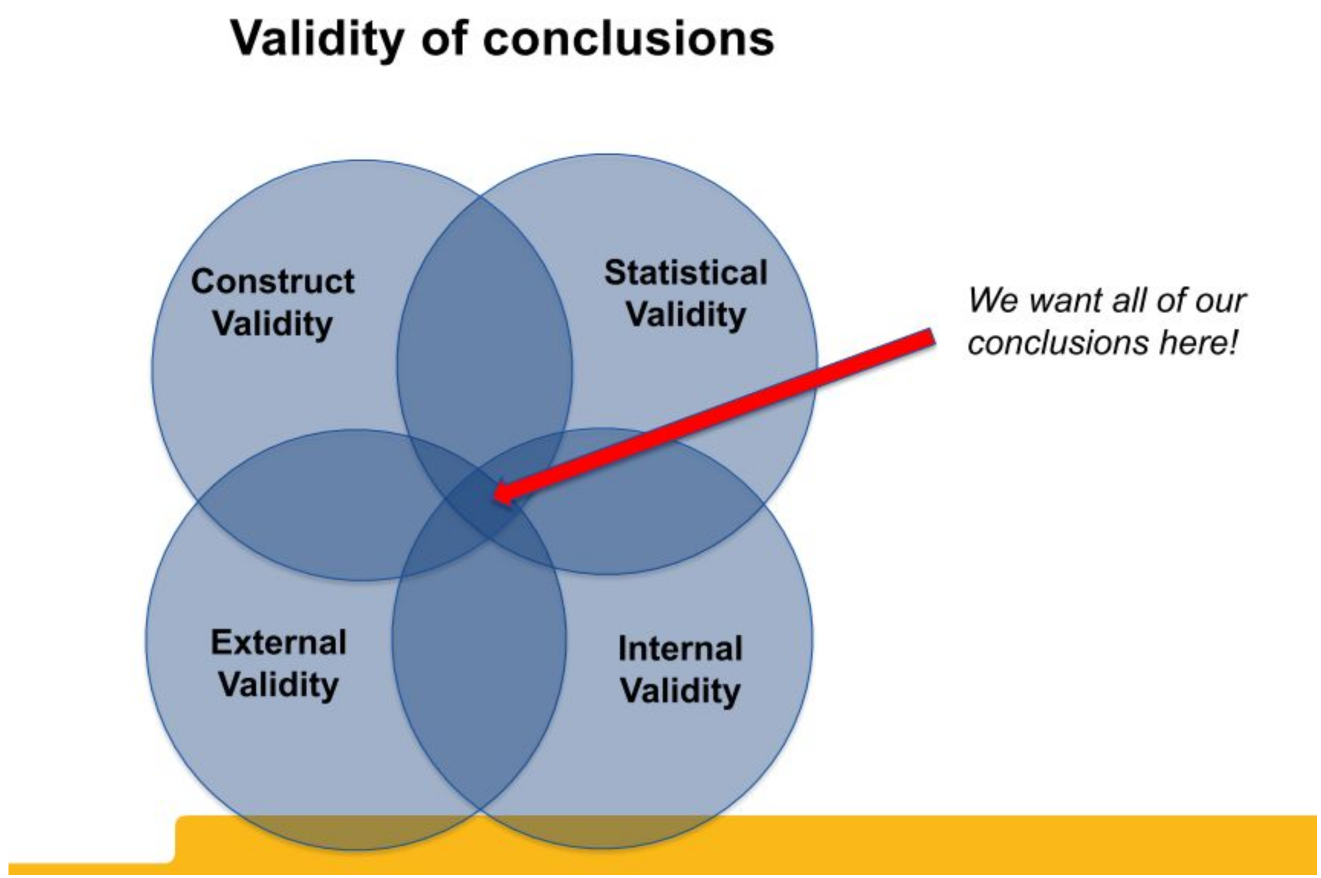
Are intentions and hypotheses of the experimenter represented in the actual experiment?

What are some threats to this?

- Are we aiming at genre classification in the small scenario data set or for all of Western Pop music?
- Should our system also be valid for slighly altered audio files?
- Do we want to model one specific annotator or a larger group of people?
- Is the accuracy measure relevant for our construct?
- Is the measure used confounded with another construct?

Flexer, A., Sturm, B. L.T. , & Urbano, J. (2020). Do We Care About the Validity of MIR research. ISMIR 2020, Special Session

Validity



Flexer, A., Sturm, B. L.T. , & Urbano, J. (2020). Do We Care About the Validity of MIR research. ISMIR 2020, Special Session

Reliability / Efficiency

Reliability is the extent to which the results of the experiment can be replicated.

Will we obtain similar results if we repeat the experiment with different sets of queries and annotators?

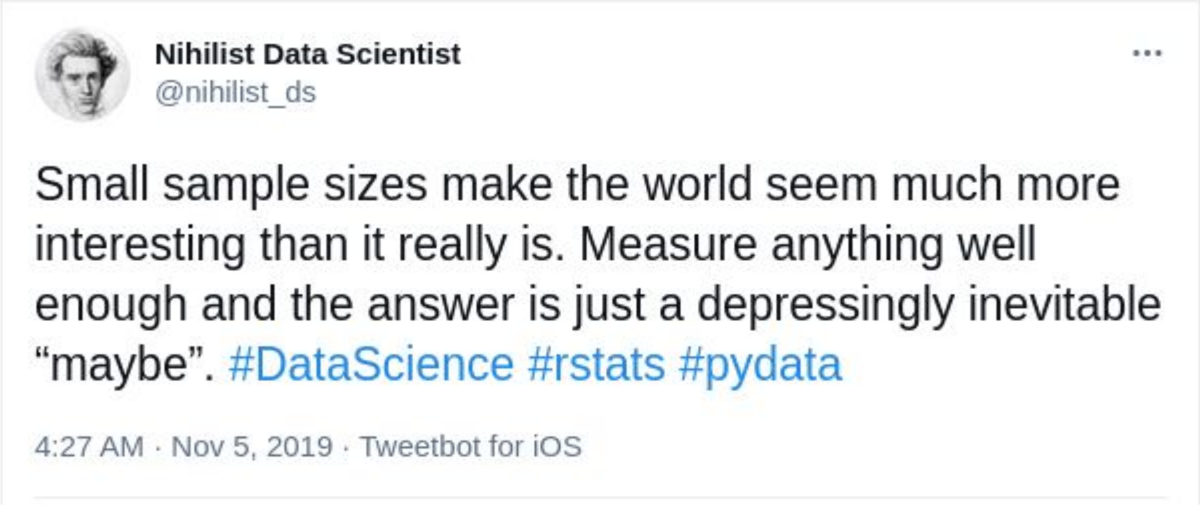
Efficiency is the extent to which the experimenter reaches a valid and reliable result at a low cost.

Are there other annotation procedures and alternative evaluation methods that result in a more cost-effective experiment?

Urbano, J., Schedl, M., & Serra, X. (2013). Evaluation in music information retrieval. Journal of Intelligent Information Systems, 41(3), 345–369. https://doi.org/10.1007/s10844-013-0249-4

In summary...

- Evaluation as fundamental step to advance scientific research
- System-centric VS User-centric
- Validity, Reliability, Efficiency (and much more…)



Day 2

Practical Lessons for MIR Evaluation

Practical lessons for MIR Evaluation

A Simple Method to Determine if a Music Information Retrieval System is a "Horse"

Bob L. Sturm, *Member, IEEE*

A "horse" is just a system that is not actually addressing the problem it appears to be solving

Sturm, B. L. (2014). A simple method to determine if a music information retrieval system is a "horse." IEEE Transactions on Multimedia, 16(6), 1636–1644. https://doi.org/10.1109/TMM.2014.2330697
https://en.wikipedia.org/wiki/Clever_Hans

Practical lessons for MIR Evaluation

General Idea: Testing the validity of experiments for the Music Genre Recognition (MGR) task.

Apply the Method of Irrelevant Transformations (MIT) (D: input space, S: MGR systems, T irrelevant transformation*)

1) Find the recordings in $\mathcal{D}$ that S maps "incorrectly"
2) Create irrelevant transformation T
3) Apply T to all recordings found in (1)
4) Have S map transformed recordings
5) Find the recordings that S maps "correctly"
6) For each recording in (1) that S now maps "correctly" in (5), replace it in $\mathcal{D}$ with its irrelevant transformation
7) Return to (1), repeat $20\times$, or until FoM of S is perfect.

*96-band near perfect reconstruction filterbank, 4 randomly choose several bands, and reduce their gains from 1 to 0.1

Sturm, B. L. (2014). A simple method to determine if a music information retrieval system is a "horse." IEEE Transactions on Multimedia, 16(6), 1636–1644. https://doi.org/10.1109/TMM.2014.2330697

Practical lessons for MIR Evaluation

Initial results

	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock	Pr
blues	80.00	2.00	4.00	0.00	2.00	0.00	0.00	0.00	6.00	6.00	80.00
classical	6.00	90.00	2.00	0.00	0.00	2.00	0.00	0.00	2.00	0.00	88.24
country	4.00	0.00	78.00	2.00	0.00	8.00	0.00	2.00	4.00	14.00	69.64
disco	2.00	0.00	2.00	66.00	4.00	0.00	0.00	8.00	6.00	10.00	67.35
hiphop	6.00	0.00	2.00	4.00	86.00	2.00	0.00	0.00	4.00	0.00	82.69
jazz	2.00	4.00	2.00	0.00	2.00	86.00	0.00	0.00	2.00	0.00	87.76
metal	0.00	0.00	0.00	0.00	4.00	0.00	100.00	2.00	4.00	20.00	76.92
pop	0.00	0.00	2.00	2.00	0.00	0.00	0.00	88.00	4.00	8.00	84.62
reggae	0.00	0.00	2.00	16.00	2.00	2.00	0.00	0.00	62.00	2.00	72.09
rock	0.00	4.00	6.00	10.00	0.00	0.00	0.00	0.00	6.00	40.00	60.61
F	80.00	89.11	73.58	66.67	84.31	86.87	86.96	86.27	66.67	48.19	77.60

(a)

Sturm, B. L. (2014). A simple method to determine if a music information retrieval system is a "horse." IEEE Transactions on Multimedia, 16(6), 1636–1644. https://doi.org/10.1109/TMM.2014.2330697

Practical lessons for MIR Evaluation

Initial results

	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock	Pr
blues	80.00	2.00	4.00	0.00	2.00	0.00	0.00	0.00	6.00	6.00	80.00
classical	6.00	90.00	2.00	0.00	0.00	2.00	0.00	0.00	2.00	0.00	88.24
country	4.00	0.00	78.00	2.00	0.00	8.00	0.00	2.00	4.00	14.00	69.64
disco	2.00	0.00	2.00	66.00	4.00	0.00	0.00	8.00	6.00	10.00	67.35
hiphop	6.00	0.00	2.00	4.00	86.00	2.00	0.00	0.00	4.00	0.00	82.69
jazz	2.00	4.00	2.00	0.00	2.00	86.00	0.00	0.00	2.00	0.00	87.76
metal	0.00	0.00	0.00	0.00	4.00	0.00	100.00	2.00	4.00	20.00	76.92
pop	0.00	0.00	2.00	2.00	0.00	0.00	0.00	88.00	4.00	8.00	84.62
reggae	0.00	0.00	2.00	16.00	2.00	2.00	0.00	0.00	62.00	2.00	72.09
rock	0.00	4.00	6.00	10.00	0.00	0.00	0.00	0.00	6.00	40.00	60.61
F	80.00	89.11	73.58	66.67	84.31	86.87	86.96	86.27	66.67	48.19	77.60

(a)

Results after applying MIT

	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock	Pr
blues	100.00	0.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	4.00	94.34
classical	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	98.04
country	0.00	0.00	98.00	0.00	4.00	0.00	0.00	0.00	2.00	8.00	87.50
disco	0.00	0.00	0.00	90.00	2.00	0.00	0.00	0.00	0.00	0.00	97.83
hiphop	0.00	0.00	0.00	0.00	94.00	0.00	0.00	0.00	0.00	0.00	100.00
jazz	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	100.00
metal	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	100.00
pop	0.00	0.00	0.00	4.00	0.00	0.00	0.00	100.00	0.00	2.00	94.34
reggae	0.00	0.00	0.00	6.00	0.00	0.00	0.00	0.00	96.00	0.00	94.12
rock	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	86.00	100.00
F	97.09	99.01	92.45	93.75	96.91	100.00	100.00	97.09	95.05	92.47	96.40

(a)

Sturm, B. L. (2014). A simple method to determine if a music information retrieval system is a "horse." IEEE Transactions on Multimedia, 16(6), 1636–1644. https://doi.org/10.1109/TMM.2014.2330697

Practical lessons for MIR Evaluation

Journal Of New Music Research, 2016
Vol. 45, No. 3, 239–251, http://dx.doi.org/10.1080/09298215.2016.1200631

Routledge
Taylor & Francis Group

The Problem of Limited Inter-rater Agreement in Modelling Music Similarity

Arthur Flexer and Thomas Grill

Austrian Research Institute for Artificial Intelligence (OFAI), Intelligent Music Processing and Machine Learning Group, Vienna, Austria

(Received 7 October 2015; accepted 25 May 2016)

Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. Journal of New Music Research, 45(3), 239–251. https://doi.org/10.1080/09298215.2016.1200631

Practical lessons for MIR Evaluation

General Idea: Testing if the annotators inter-agreement can define an upper-bound for the evaluation of the Audio Music Similarity (AMS) task.

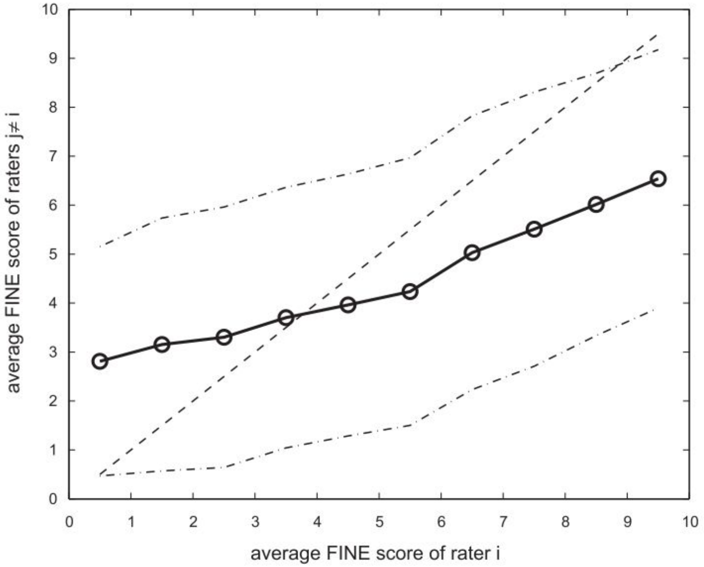
"if different human subjects are asked to rate the same song pairs according to their perceived similarity, only a certain amount of agreement can be expected due to a range of subjective factors."

Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. Journal of New Music Research, 45(3), 239–251. https://doi.org/10.1080/09298215.2016.1200631

Practical lessons for MIR Evaluation

General Idea: Testing if the annotators inter-agreement can define an upper-bound for the evaluation of the Audio Music Similarity (AMS) task.

"if different human subjects are asked to rate the same song pairs according to their perceived similarity, only a certain amount of agreement can be expected due to a range of subjective factors."



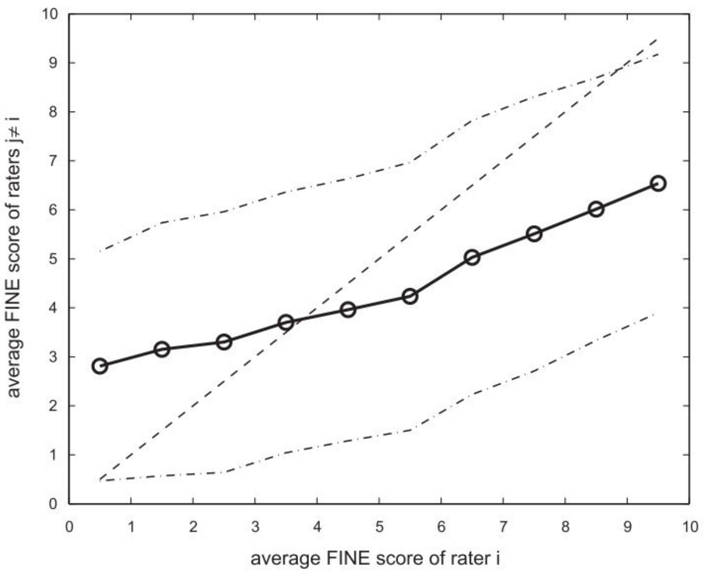
Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. Journal of New Music Research, 45(3), 239–251. https://doi.org/10.1080/09298215.2016.1200631

Practical lessons for MIR Evaluation

General Idea: Testing if the annotators inter-agreement can define an upper-bound for the evaluation of the Audio Music Similarity (AMS) task.

"if different human subjects are asked to rate the same song pairs according to their perceived similarity, only a certain amount of agreement can be expected due to a range of subjective factors."

"Inter-rater agreement present a natural upper bound for any algorithmic approach, since it is not meaningful to have computational models that go beyond the level of human agreement."



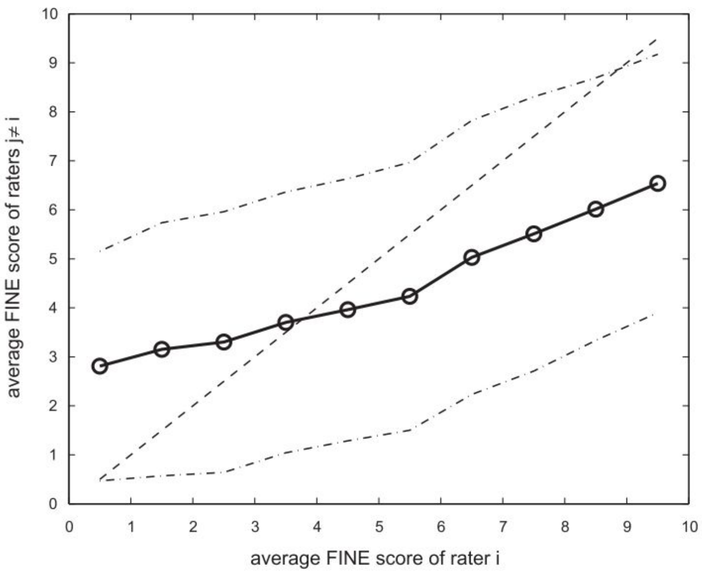
Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. Journal of New Music Research, 45(3), 239–251. https://doi.org/10.1080/09298215.2016.1200631

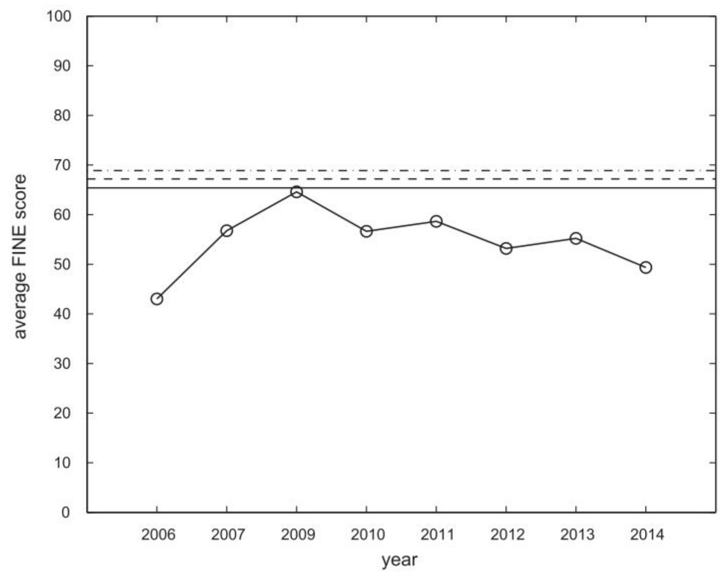
Practical lessons for MIR Evaluation

General Idea: Testing if the annotators inter-agreement can define an upper-bound for the evaluation of the Audio Music Similarity (AMS) task.

"if different human subjects are asked to rate the same song pairs according to their perceived similarity, only a certain amount of agreement can be expected due to a range of subjective factors."

"Inter-rater agreement present a natural upper bound for any algorithmic approach, since it is not meaningful to have computational models that go beyond the level of human agreement."





Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. Journal of New Music Research, 45(3), 239–251. https://doi.org/10.1080/09298215.2016.1200631

Practical lessons for MIR Evaluation

General Idea: Testing if the annotators inter-agreement can define an upper-bound for the evaluation of the Audio Music Similarity (AMS) task.

Some Issues

Ask more specific questions → It is probably necessary to research what the concept of music similarity actually means to human listeners.

Care about confounding variables → Examples for confounding variables are the level of expertise of the human graders or their familiarity with the music pieces that are part of the evaluation.

Flexer, A., & Grill, T. (2016). The Problem of Limited Inter-rater Agreement in Modelling Music Similarity. Journal of New Music Research, 45(3), 239–251. https://doi.org/10.1080/09298215.2016.1200631

Practical lessons for MIR Evaluation

TISMIR

Schreiber, H., et al. (2020). Music Tempo Estimation: Are We Done Yet? *Transactions of the International Society for Music Information Retrieval*, 3(1), pp. 111–125. DOI: https://doi.org/10.5334/tismir.43

OVERVIEW ARTICLE

Music Tempo Estimation: Are We Done Yet?

Hendrik Schreiber*, Julián Urbano† and Meinard Müller*

With the advent of deep learning, global tempo estimation accuracy has reached a new peak, which presents a great opportunity to evaluate our evaluation practices. In this article, we discuss presumed and actual applications, the pros and cons of commonly used metrics, and the suitability of popular datasets. To guide future research, we present results of a survey among domain experts that investigates today's applications, their requirements, and the usefulness of currently employed metrics. To aid future evaluations, we present a public repository containing evaluation code as well as estimates by many different systems and different ground truths for popular datasets.

Schreiber, H., Urbano, J., & Müller, M. (2020). Music Tempo Estimation: Are We Done Yet? Transactions of the International Society for Music Information Retrieval, 3(1), 111. https://doi.org/10.5334/tismir.43

Practical lessons for MIR Evaluation

General Idea: in the context of tempo estimation, understand how applications, use-case and metrics/dataset are linked.

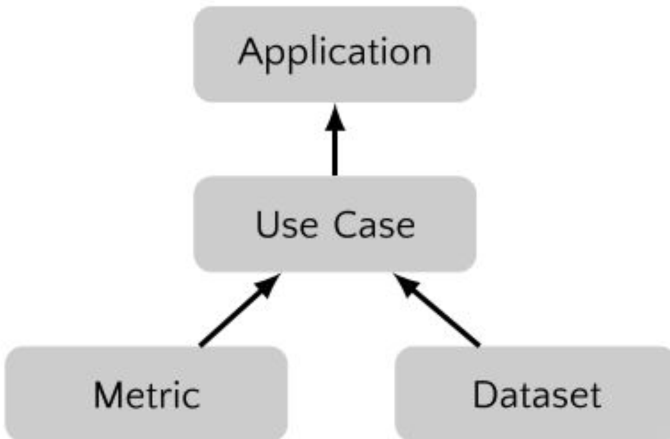


Figure 6: Dependencies between application, use case, metric, and dataset (an arrow from *A* to *B* denotes that *A* depends on *B*).

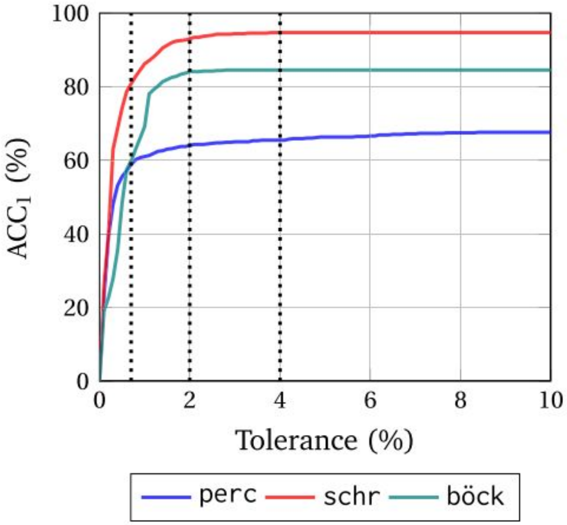
Schreiber, H., Urbano, J., & Müller, M. (2020). Music Tempo Estimation: Are We Done Yet? Transactions of the International Society for Music Information Retrieval, 3(1), 111. https://doi.org/10.5334/tismir.43

Practical lessons for MIR Evaluation

General Idea: in the context of tempo estimation, understand how applications, use-case and metrics/dataset are linked.

→ ACC_1 computes a 0 or 1 score per track, which indicates the correctness of an estimate, allowing a 4% tolerance.

→ The Just-Noticeable Difference (JND) for music tempi is approximately 4% and therefore '4% is probably the highest precision level that should be considered.'



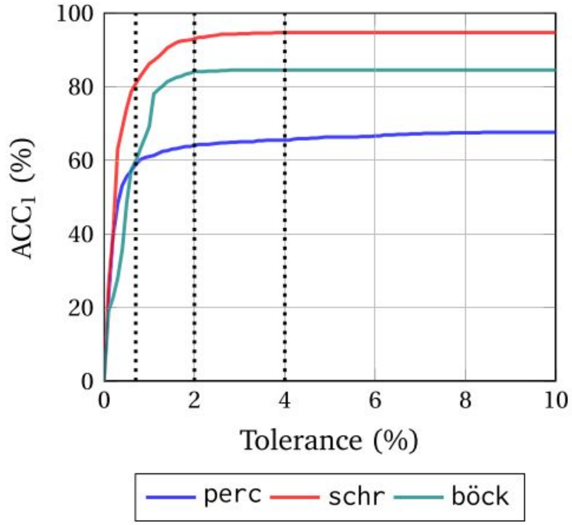
Schreiber, H., Urbano, J., & Müller, M. (2020). Music Tempo Estimation: Are We Done Yet? Transactions of the International Society for Music Information Retrieval, 3(1), 111. https://doi.org/10.5334/tismir.43

Practical lessons for MIR Evaluation

General Idea: in the context of tempo estimation, understand how applications, use-case and metrics/dataset are linked.

→ ACC_1 computes a 0 or 1 score per track, which indicates the correctness of an estimate, allowing a 4% tolerance.

→ The Just-Noticeable Difference (JND) for music tempi is approximately 4% and therefore '4% is probably the highest precision level that should be considered.'



Issues

1. The threshold is usually arbitrary.
2. It does not tell us how wrong an estimate is, nor in which direction.
3. It is blind to small systematic errors below the threshold.
4. It may overemphasize differences between systems.

Schreiber, H., Urbano, J., & Müller, M. (2020). Music Tempo Estimation: Are We Done Yet? Transactions of the International Society for Music Information Retrieval, 3(1), 111. https://doi.org/10.5334/tismir.43

Practical lessons for MIR Evaluation

General Idea: in the context of tempo estimation, understand how applications, use-case and metrics/dataset are linked.

→ ACC_2 additionally allows estimates to be wrong by the factors 2, 3, ½ or ⅓ (so-called octave errors).

→ Justified because used annotations may not match the perception of human listeners.

Schreiber, H., Urbano, J., & Müller, M. (2020). Music Tempo Estimation: Are We Done Yet? Transactions of the International Society for Music Information Retrieval, 3(1), 111. https://doi.org/10.5334/tismir.43

Practical lessons for MIR Evaluation

General Idea: in the context of tempo estimation, understand how applications, use-case and metrics/dataset are linked.

→ ACC_2 additionally allows estimates to be wrong by the factors 2, 3, ½ or ⅓ (so-called octave errors).

→ Justified because used annotations may not match the perception of human listeners.

Issues

1. It says nothing about a system's ability to help a user to distinguish between slow and fast tracks (useless for applications like playlist generation based on tempo continuity or when searching for slow music).

Schreiber, H., Urbano, J., & Müller, M. (2020). Music Tempo Estimation: Are We Done Yet? Transactions of the International Society for Music Information Retrieval, 3(1), 111. https://doi.org/10.5334/tismir.43

Practical lessons for MIR Evaluation

General Idea: in the context of tempo estimation, understand how applications, use-case and metrics/dataset are linked.

The mismatch between metric and usefulness illustrates that:

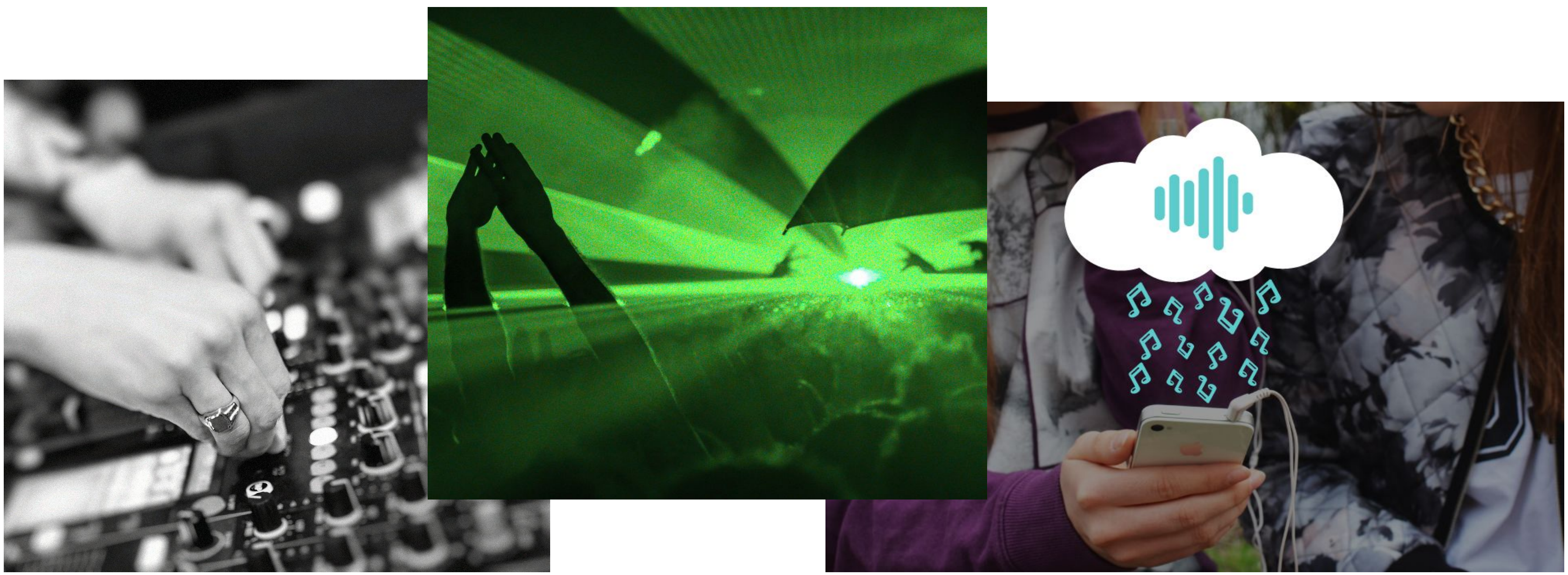
The correlation between use case, success criteria, and the employed metric is far from perfect for the mentioned use cases.

(construct validity)

Schreiber, H., Urbano, J., & Müller, M. (2020). Music Tempo Estimation: Are We Done Yet? Transactions of the International Society for Music Information Retrieval, 3(1), 111. https://doi.org/10.5334/tismir.43

In summary...

- Evaluation goes beyond accuracy (accuracy can be uninformative).
- Evaluation considers all the aspects of a technology.
- Music as social construct implies human-centered evaluation in MIR.



MIR Evaluation Campaigns

https://www.music-ir.org/mirex/wiki/MIREX_HOME

https://multimediaeval.github.io
https://multimediaeval.github.io/2018-AcousticBrainz-Genre-Task

Q&A