20th International Society for Music Information Retrieval Conference
Delft, The Netherlands, November 4-8, 2019

Fairness, Accountability and Transparency in Music Information Research (FAT-MIR)

Tutorial

Emilia Gomez, Andre Holzapfel, Marius Miron, Bob L. T. Sturm

upf. MTG Music Technology Group
European Commission Joint Research Centre
KTH VETENSKAP OCH KONST

FAT-MIR in a tutorial

- Introduction
- Ethical principles in practical MIR scenarios

BREAK

- Fairness in machine learning
- Transparency/Explicability in MIR
- Discussion

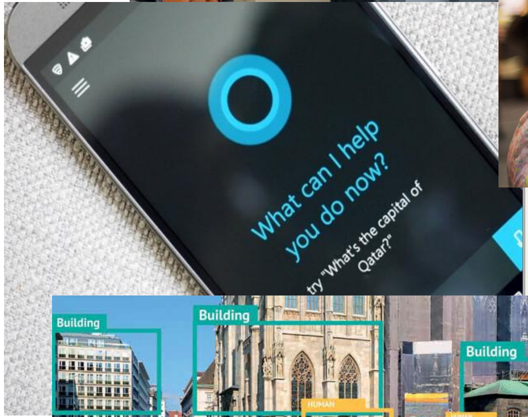
FAT-MIR in a tutorial

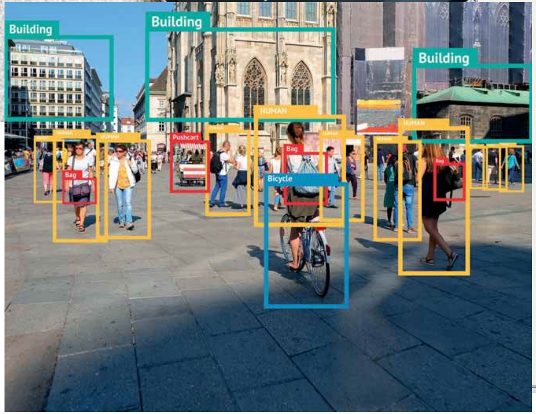
- Introduction (Emilia)
- Ethical principles in practical MIR scenarios

BREAK

- Fairness in machine learning
- Transparency/Explicability in MIR
- Discussion

Artificial Intelligence: *machines or agents capable of observing its environment and taking decisions towards a certain goal.*





NOVEMBER 15, 2017

Stanford algorithm can diagnose pneumonia better than radiologists

Stanford researchers have developed a deep learning algorithm that evaluates chest X-rays for signs of disease. In j over a month of development, their algorithm outperformed expert radiologists at diagnosing pneumonia.

BY TAYLOR KUBOTA

Stanford researchers have developed an algorithm that offers diagnoses based off chest X-ray images. It can diagnose up to 14 types of medical conditions and is able to diagnose pneumonia better than expert radiologists working alone. A paper about the algorithm, called CheXNet, was published Nov. 14 on the open-access, scientific preprint website arXiv.

"Interpreting X-ray images to diagnose pathologies like pneumonia is very challenging, and we know that there's a lot of variability in the diagnoses radiologists arrive at," said Pranav Rajpurkar, a graduate student in the Stanford Machine Learning Group and co-lead author of the paper. "We became interested in developing machine learning algorithms that could learn from hundreds of thousands of chest X-ray diagnoses and make accurate

TRUSTED INSIGHTS FOR COMPUTING'S LEADING PROFESSIONALS | ACM.org | Join ACM | About Communications | ACM Resources | Alerts & Feeds

SIGN IN

COMMUNICATIONS OF THE ACM

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEARCH | PRACTICE | CAREERS | ARCHIVE | VIDEOS

Home / Magazine Archive / December 2018 (Vol. 61, No. 12) / AI Judges and Juries / Full Text

NEWS

AI Judges and Juries

By Logan Kugler
Communications of the ACM, December 2018, Vol. 61 No. 12, Pages 19-21
10.1145/3283222
Comments

VIEW AS: SHARE:

Credit: Andrey Popov

When the head of the U.S. Supreme Court says artificial intelligence (AI) is having a significant impact on how the legal system in this country works, you pay attention. That's exactly what happened when Chief Justice John Roberts was asked the following question:

"Can you foresee a day when smart machines, driven with artificial intelligences, will assist with courtroom fact-finding or, more controversially even, judicial decision-making?"

His answer startled the audience.

"It's a day that's here and it's putting a significant strain on how the judiciary goes about doing things," he said, as reported by *The New York Times*.

In the last decade, the field of AI has experienced a renaissance. The field was long in the grip of an "AI winter," in which progress and funding dried up for decades, but technological breakthroughs in AI's power and accuracy changed all that. Today, giants like Google, Microsoft, and Amazon rely on AI to power their current and future profit centers.

Yet AI isn't just affecting tech giants and cutting-edge startups; it is transforming one of the oldest disciplines on the planet: the application of the law.

AI is already used to analyze documents and data during the legal discovery process, thanks to its ability to

SIGN IN for Full Access
User Name
Password
» Forgot Password?
» Create an ACM Web Account
SIGN IN

ARTICLE CONTENTS:
Introduction
The Predictable, Reliable Choice?
"Unbiased" Machines Created by Biased Humans
Author

MORE NEWS & OPINIONS
Apple CEO Tim Cook on Screen Time Controls, Working with China
NPR
How NASA Was Born 60 Years Ago from Panic Over a 'Second Moon'
CNET

From lab to market

Music listening

Well-being and therapy

Digital libraries

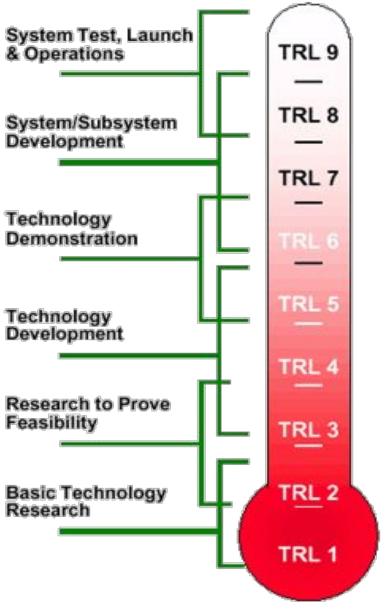
Music composition, performance and production

Education

Business, marketing

Gaming

Industry - Real world



Academia - Lab

User-centred MIR

Domain knowledge

Computational tasks

Music theory

Music similarity

NASA Technology Readiness Levels (Wikipedia)

Technology impact assessment

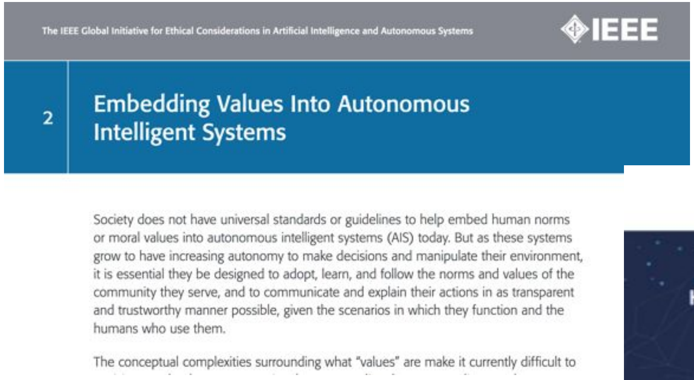
1. Who are the people affected?
2. Who are the ‘winners’ (benefit), who the ‘losers’ (cost)?
3. How many lives can be saved?
4. How much money/jobs can be saved?
5. What are the short-term and long-term costs/benefits?

Technology is not neutral

(Dusek 2006)

- *Human-centred*
- *Trustworthy*
- *For good*

***Ethics**, also called **moral philosophy**, the <u>discipline</u> concerned with what is morally good and bad and morally right and wrong.*
https://www.britannica.com/topic/ethics-philosophy


The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

IEEE

2 Embedding Values Into Autonomous Intelligent Systems

Society does not have universal standards or guidelines to help embed human norms or moral values into autonomous intelligent systems (AIS) today. But as these systems grow to have increasing autonomy to make decisions and manipulate their environment, it is essential they be designed to adopt, learn, and follow the norms and values of the community they serve, and to communicate and explain their actions in as transparent and trustworthy manner possible, given the scenarios in which they function and the humans who use them.

The conceptual complexities surrounding what "values" are make it currently difficult to

Different proposals for ethical frameworks: public, private, civil organizations

(Hand 2018, IEEE SA 2017, Bryson and Winfield 2017)

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

Trustworthy AI

- Lawful
- Robust
- Ethical

7 principles

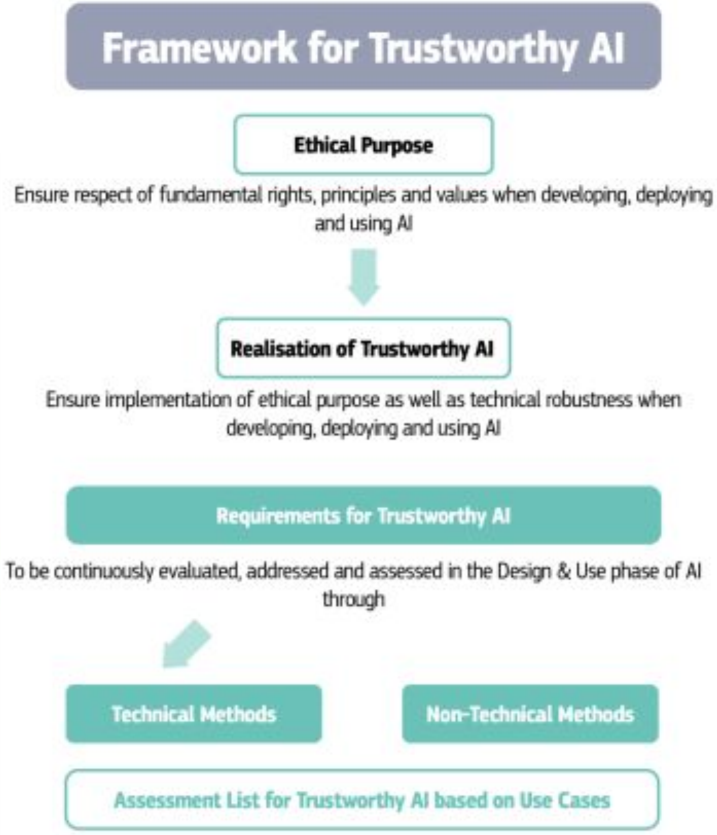


Figure 1: The Guidelines as a framework for Trustworthy AI

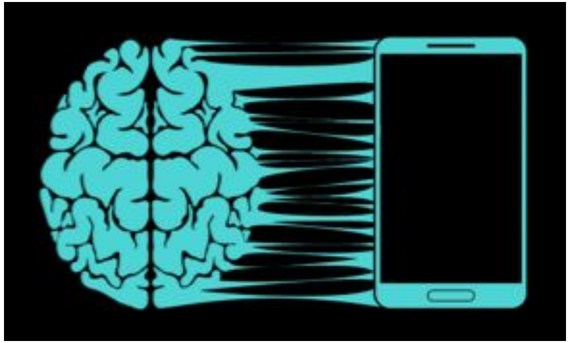
1: Human agency and oversight

AI systems should <u>empower</u> human beings.

Proper oversight mechanisms need to be ensured: human-in-the-loop, human-on-the-loop, and human-in-command approaches.

Extended mind

(Vold 2018)



https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

2: Technical robustness and safety

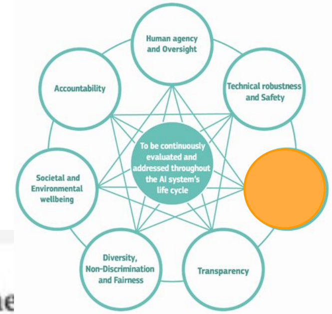
AI systems need to be resilient, safe, accurate, reliable and reproducible.

Technical and social robustness

Even with good intentions, AI systems can cause unintentional harm.

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

3: Privacy and data governance



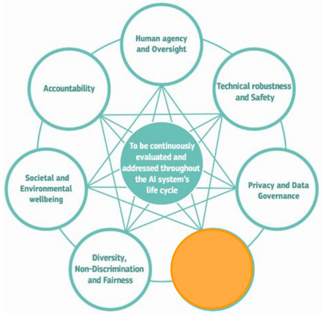
Ensure full respect for privacy and data protection, and adequate data governance, e.g. quality and integrity, legitimised access.

New Data Privacy Laws Could Slow The Music Business—But Might Help The Next Beatles

Bill Hochberg Contributor
Hollywood & Entertainment
I write about the business and law of music.

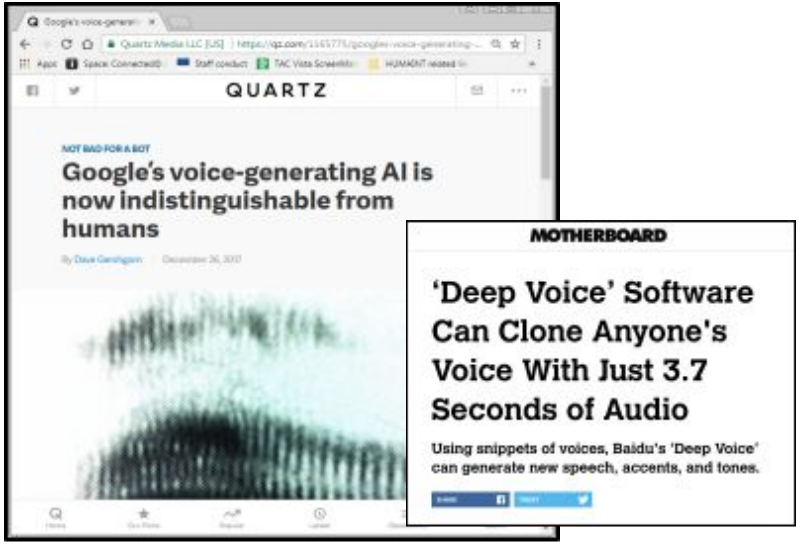
https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

4: Transparency



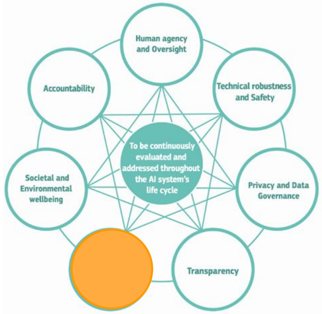
- Purpose of a system.
- Capabilities, limitations.
- Processes of operation.
- Explainable to those directly and indirectly affected.

Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.



Gómez, Blaauw, Bonada, Chandna, Cuesta. Deep Learning for Singing Processing: Achievements, Challenges and Impact on Singers and Listeners
arxiv.org/abs/1807.03046

5: Diversity, non discrimination, fairness



- Equal and just distribution of benefits/costs, equal opportunities.
- Free from bias.
- Balancing of competing interests and objectives.
- Accessible to all, involved different views.

6: Societal and environmental well-being

AI systems should benefit human beings, future generations, be sustainable and environmentally friendly.

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

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

> Review

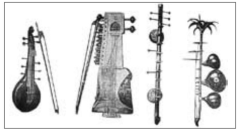
Music and Manipulation

Brown, Steven and Ulrik Volgsten eds. 2006. *Music and Manipulation. On the Social Uses and Social Control of Music.* New York, London: Berghahn Books. 376 pages. ISBN 1 57181 489 2

Bob van der Linden

Among human beings (and animals), music has always been a key mode of communication, being able to influence individual and group behaviour and to create social cohesion as well as conflict. Rhythm, harmony and melody manipulate and can be manipulated. The interdisciplinary anthology under review contains theoretical analyses by sociologists, humanists and psychologists about the use and control of music in society. It is the first volume 'to address the social ramifications of music's behaviourally manipulative effects, its morally questionable uses and control mechanisms, and its economic and artistic management through commercialisation, thus highlighting not only music's diverse uses at the social level, but also the ever-fragile relationship between aesthetics and

shrines of Sufi saints (which often also serve as music schools): This music, made famous in the West by the Pakistani singer Nusrat Fateh Ali Khan, is based not on the text of the Quran but on Sufi poetry. As in Independent India, it seemed that initially the Pakistani government was going to support the broadcast of classical music. Yet the clergy's opposition to music prevented this, mainly because the texts of many of the classical songs were connected either with Hindu deities or with the separation of lovers. Accordingly, the Pakistani government adopted a more easy-going attitude, with the result that the market for classical music gradually diminished and popular (mainly film) music became utterly dominant. In 1974, however, the government did establish the Institute of Folk Heritage in Islamabad, which among other things did much for the conservation of



Traditional Indian instruments. www.hinduwisdom.info

munication system and take the social production of music rather than music itself as a starting-point for the understanding of the relationship between music and society. Music can create both consensus and conflict as it is 'a major tool for propagating group ideologies identities, and as such serves as an important device for reinforcing collective actions and for delineating the lines of inclusion for social groups'

Hargreaves unsurprisingly make clear that 'it is extremely difficult to predict how customers or staff will react to a particular piece of music because any response to music is determined by three interacting factors, namely, the music itself, the listener, and the listening situation' (p. 117). Steven Brown and Töres Theorell question the validity of the dogma that "good music is good for you". In their opinion, twentieth cen-

Average: 198,5 Count: 5 Sum:

Strubell, E., Ganesh, A., and McCaullm, A. Energy and Policy Considerations for Deep Learning in NLP, 2019 https://arxiv.org/abs/1906.02243

7: Accountability

Ensure responsibility

Provide ability to contest and seek effective redress/remedy against decisions made by AI systems → accountable entity.

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Music Information Retrieval

ISMIR

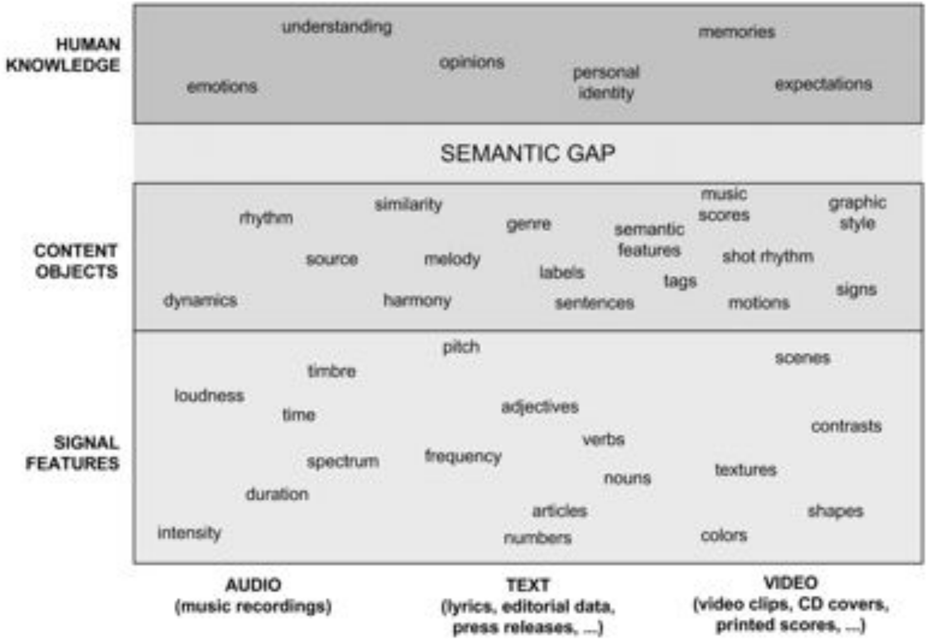
As a field, music information retrieval focuses on the research and development of computational systems to ***help humans better make sense of music data,*** *drawing from a* ***diverse*** *set of disciplines, including, but my no means limited to, music theory, computer science, psychology, neuroscience, library science, electrical engineering, and artificial intelligence and machine learning*

diversity in culture, gender, …

The technology-centred motivation

(Celma et al., 2006)

- Facilitate access to large music collections.
- Provide data-driven understanding of music.
- Bridge the semantic gap.



M. Schedl, E. Gómez, and J. Urbano, "Music Information Retrieval: Recent Developments and Applications," Foundations and Trends® in Information Retrieval, vol. 8, no. 2–3, pp. 127-261, Sep. 2014. doi: 10.1561/1500000042

The human-centred motivation

HUMAINT
HUman behaviour and MAchine INTelligence

European Commission
Joint Research Centre

#HUMAINT

- Facilitate access to large music collections?
- Provide data-driven understanding of music?
- Be aware of IMPACT and establish adequate means that ensure that our systems are developed WITH people and FOR people’s welfare

https://trompamusic.eu/

TROMPA

Some questions to start

How does MIR impacts music and the various participants contributing to and benefiting from music: composers, musicians, educators, listeners, and organisations?

1. In many areas technology leads to more efficient production lines and increased profit but human redundancy and deskilling. Can the same happen in music?
2. Who (and how) is accountable for the MIR systems?
3. Should listeners be informed about the involvement of AI in the music and playlists they listen to, much the same way ingredients of food products are communicated? How should this information be presented in a transparent way, and to what level of detail?
4. Are music recommendation algorithms fair?
5. Who owns the rights to the music generated by AI models? What is their artistic value?

Bob L.T. Sturm, Maria Iglesias, Oded Ben-Tal, Marius Miron and Emilia Gómez. Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis. Arts 2019, 8(3)

Some practical questions: data biases

Engineers share the responsibility for the resulting outcomes, positive and negative, intended and unintended

- Are we aware that our data encodes existing biases and our methods can unintentionally perpetuate these biases or introduce new ones? (Barocas and Selbst 2016)

Bob L.T. Sturm, Maria Iglesias, Oded Ben-Tal, Marius Miron and Emilia Gómez. Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis. Arts 2019, 8(3)

Some practical questions: data transparency

Engineers share the responsibility for the resulting outcomes, positive and negative, intended and unintended

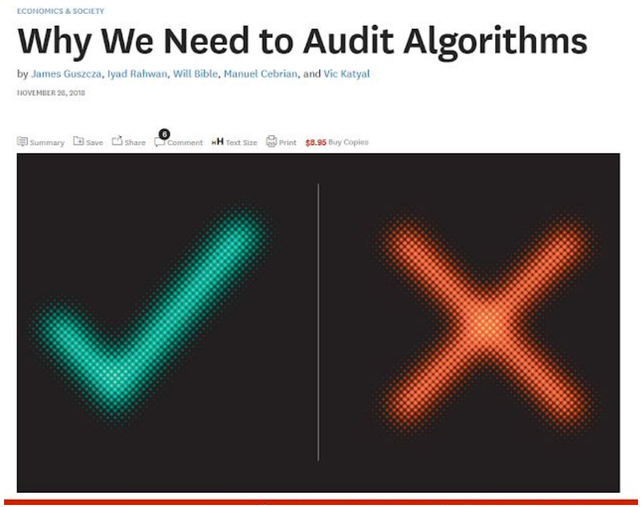
- Do we follow proper mechanisms for transparent data collection?
- Datasheet for datasets `(Gebru et al. 2017)`
    - `The motivation for dataset creation`
    - `The composition of the dataset`
    - `The data collection process`
    - `The preprocessing of the data`
    - `The distribution of the data`
    - `The maintenance of the data`
    - `The legal and ethical considerations`

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, Kate Crawford. *Datasheets for Datasets, 2018* https://arxiv.org/abs/1803.09010?context=cs

Some practical questions: algorithm auditing

Do we know that transparency ≠ open source?

- Study, evaluate and document algorithm working principles and limitations (Schedl et al. 2014; Sturm 2016)
- Select metrics (societal values under them). Performance (accuracy, precision, reliability) → metrics reflecting impact (e.g. diversity)
- Auditing tools (Crawford, 2017)



Check our paper tomorrow! **20 years of playlists: A statistical analysis on popularity and diversity** (L Porcaro, E Gomez)

FAT-MIR in a tutorial

- Introduction
- Ethical principles in practical MIR scenarios (André)

BREAK

- Fairness in machine learning
- Transparency/Explicability in MIR
- Discussion

FAT-MIR in a tutorial

- Introduction
- Ethical principles in practical MIR scenarios

BREAK

- Fairness in machine learning (Marius)
- Transparency/Explicability in MIR
- Discussion

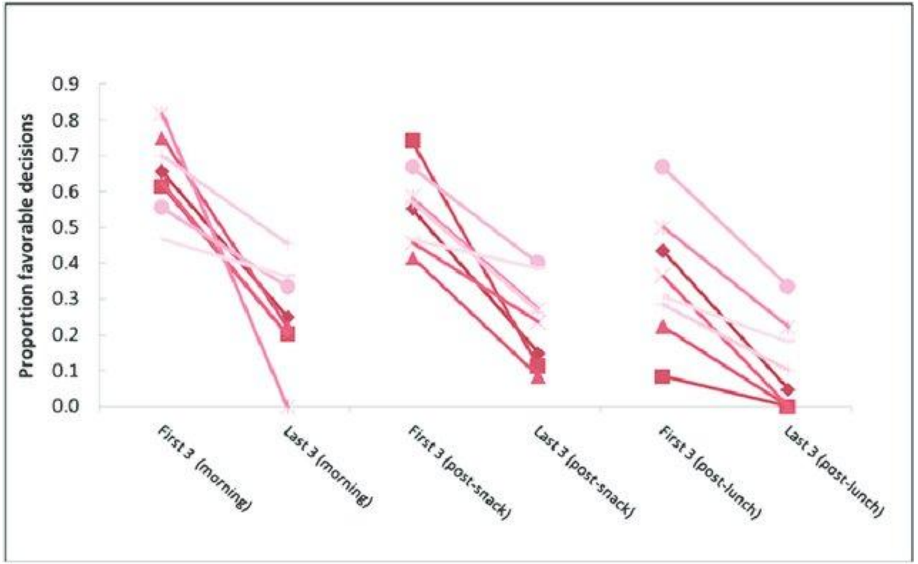
FAT-MIR in a tutorial

- Introduction
- Ethical principles in practical MIR scenarios

BREAK

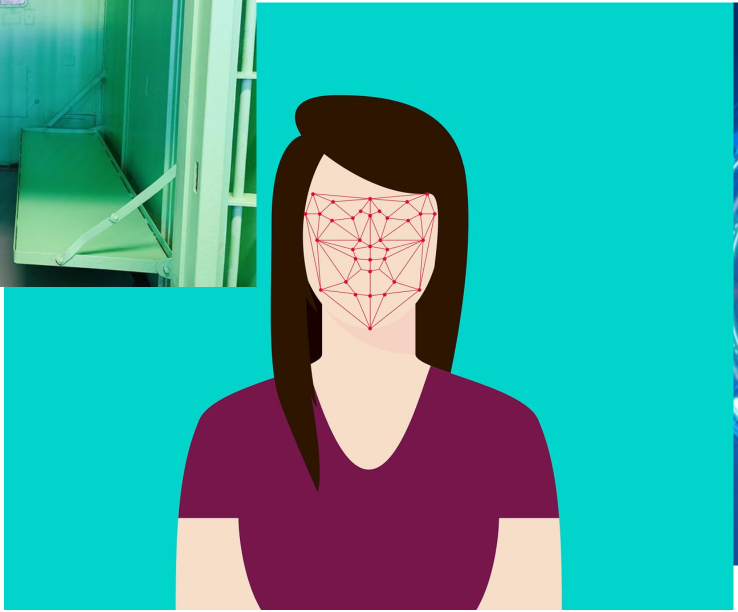
- Fairness in machine learning (Marius)
- Transparency/Explicability in MIR
- Discussion

Human bias in decision making



Danziger et al, Extraneous factors in judicial decisions (2011)

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 et al., Automated underwriting in mortgage lending: Good news for the underserved? (2002)

[..] 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 et al., Human decisions and machine predictions (2006)

Bias may affect formal assessments and leave room for discrimination

McKay and McDaniel, A reexamination of black-white mean differences in work performance: More data, more moderators (2006)

Bias vs 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.

Metcalf, Disambiguating Bias and Unfairness in Algorithmic Products (2017)

Bias vs Fairness

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

Barocas et al., Fairness in Machine Learning (2017)

Bias vs Fairness

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

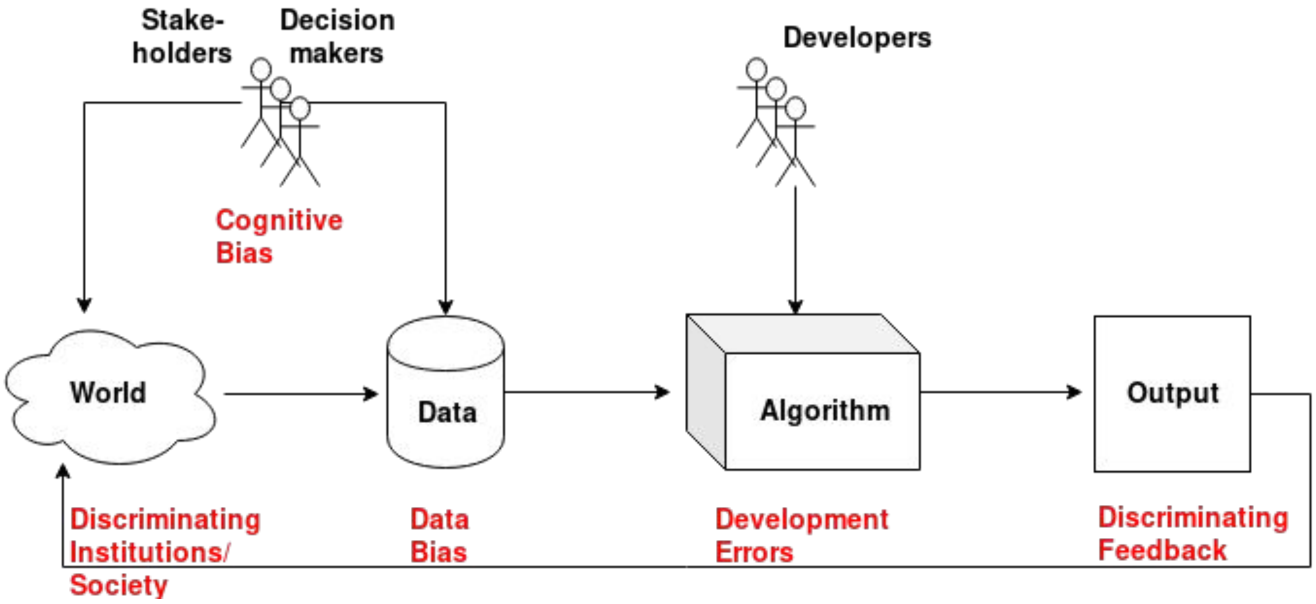
Surprising view of computer scientists:

“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.

Narayanan, 21 fairness definitions and their politics (2018)

Bias vs Fairness



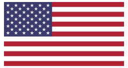
Tolan, Discrimination in Algorithmic Justice (2018)

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)



Fairness

Real challenge

Design systems that support human values.

Narayanan, 21 fairness definitions and their politics (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 et al, Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment (2018)

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)

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 and Selbst, Big data's disparate impact (2016)

Disparate impact

Unjustified discrimination resulted from facially neutral practices

Outcome depends on group membership

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

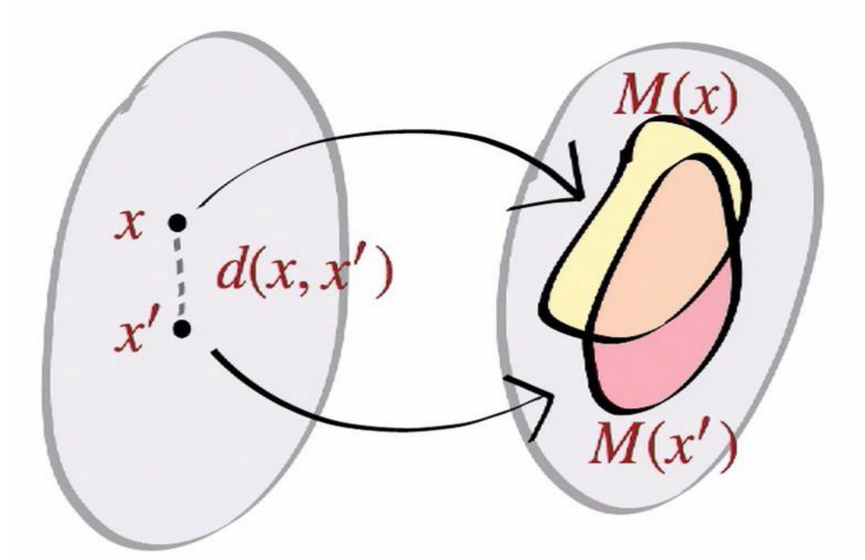
Must come with rigorous proof - account for confounders, exogenous effects

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

Barocas and Selbst, Big data's disparate impact (2016)

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)



Dwork et al, Fairness through awareness (2011)

Group Fairness

Protected features

***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)

Example: face recognition

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini, The Safe Face Pledge (2019)

Example: face recognition

Stakeholders

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Stakeholder

Buolamwini, The Safe Face Pledge (2019)

Example: binary classification

	Labeled high-risk	Labeled low-risk
Recidivated	TP	FN
Did not recidivate	FP	TN

Buolamwini, The Safe Face Pledge (2019)

Example: binary classification

	Labeled high-risk	Labeled low-risk
Recidivated	TP	FN
Did not recidivate	FP	TN

Stakeholder

Example: binary classification - group metrics

A protected feature has two categories: A and B (can be race A and race B)

	Labeled high-risk	Labeled low-risk
Recidivated	TP_A	FN_A
Did not recidivate	FP_A	TN_A

	Labeled high-risk	Labeled low-risk
Recidivated	TP_B	FN_B
Did not recidivate	FP_B	TN_B

Metrics A: FPR_A, FNR_A,...

Metrics B: FPR_B, FNR_B,...

Example: binary classification - group metrics

A protected feature has two categories: A and B (can be race A and race B)

	Labeled high-risk	Labeled low-risk
Recidivated	TP_A	FN_A
Did not recidivate	FP_A	TN_A

	Labeled high-risk	Labeled low-risk
Recidivated	TP_B	FN_B
Did not recidivate	FP_B	TN_B

Metrics A: FPR_A, FNR_A,...

Metrics B: FPR_B, FNR_B,...

Stakeholder **FPR_A/ FPR_B**

Trade-offs - Impossibility theorems

	Labeled high-risk	Labeled low-risk
Recidivated	TP	FN
Did not recidivate	FP	TN

There are at least 21 definitions of “fairness” which may contradict each other.

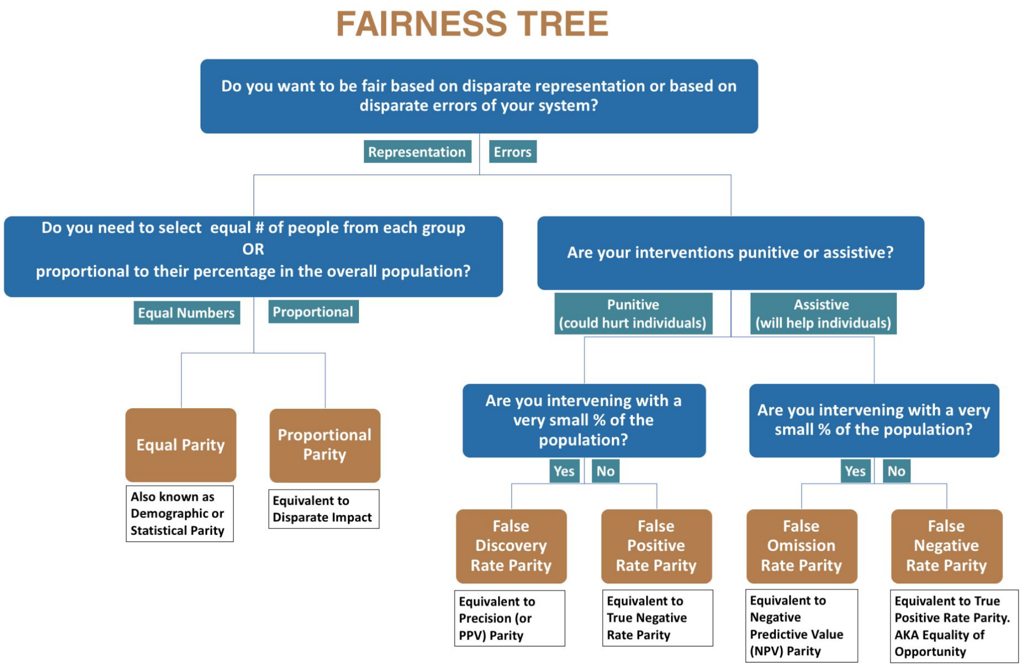
Many of these definitions do not match legal or social definitions of equality.

In reality we have many ways to measure discrimination.

Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments (2017)

Fairness - domain specific

Machine learning is domain-specific: understand legal and social context



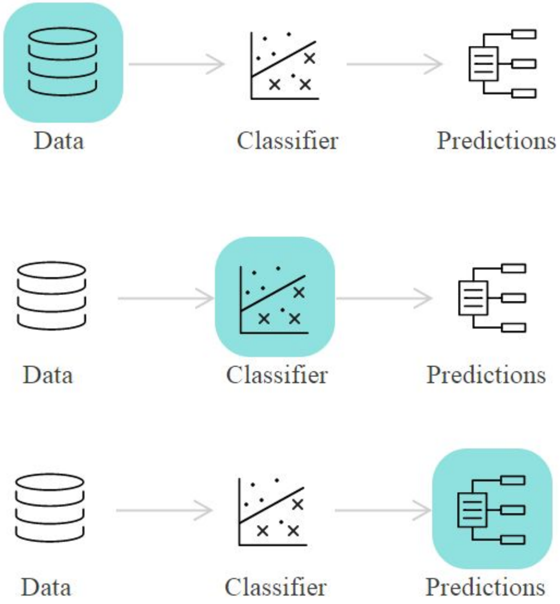
http://aequitas.dssg.io/

Mitigation

- Pre-processing

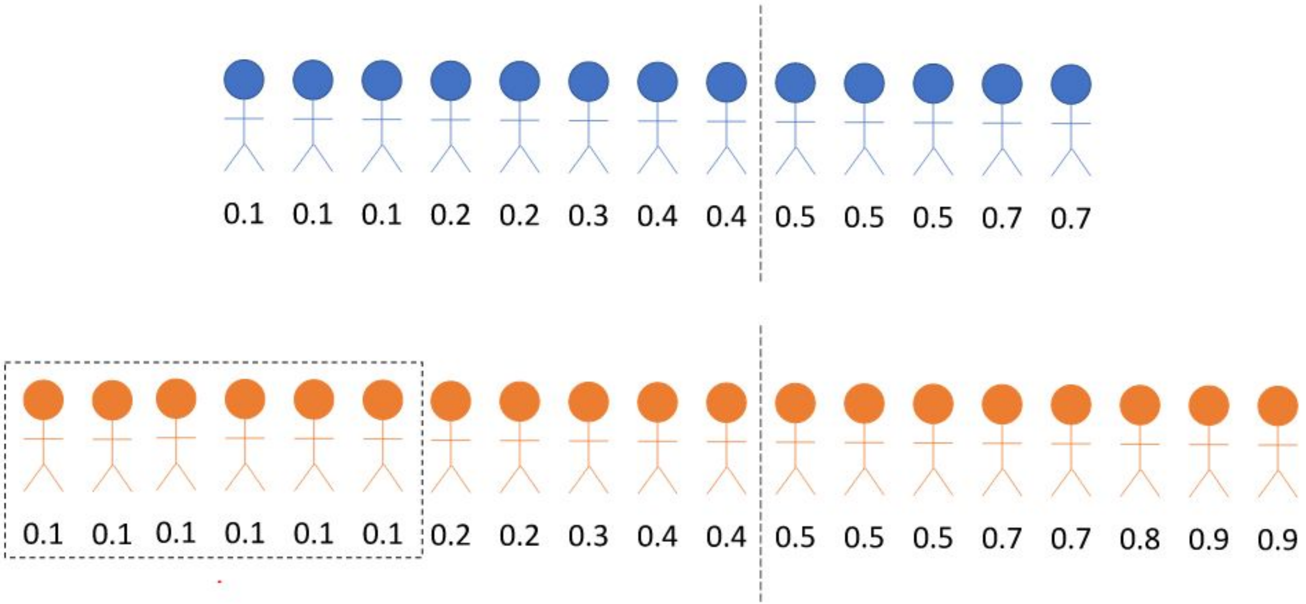
- In-processing

- Post-processing



http://aif360.mybluemix.net/data

Deceptive equalization of False Positive Rates



Detention rate	False positive rate
38%	25%
~~62%~~ **42%**	~~42%~~ **22%**

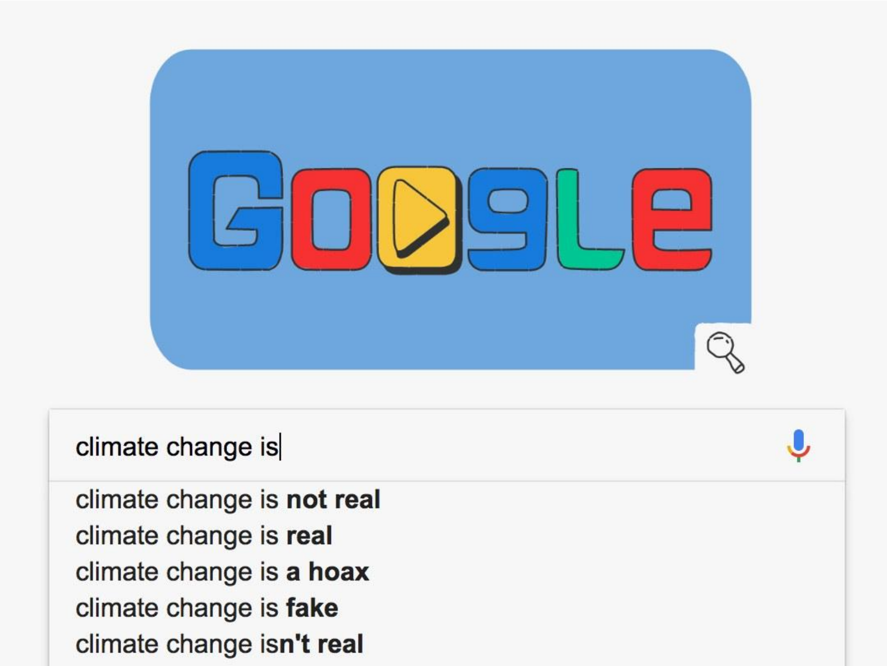
Castillo, Discrimination in Supervised Classification, (2019)

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**



Castillo, Fairness and Transparency in Ranking, (2018)

Fairness in recommendation

Multisided (Group) Fairness

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

CP-fairness

Burke, Multisided fairness for recommendation, (2017)

FAT-MIR in a tutorial

- Introduction
- Ethical principles in practical MIR scenarios

BREAK

- Fairness in machine learning
- Transparency/Explicability in MIR (Bob)
- Discussion

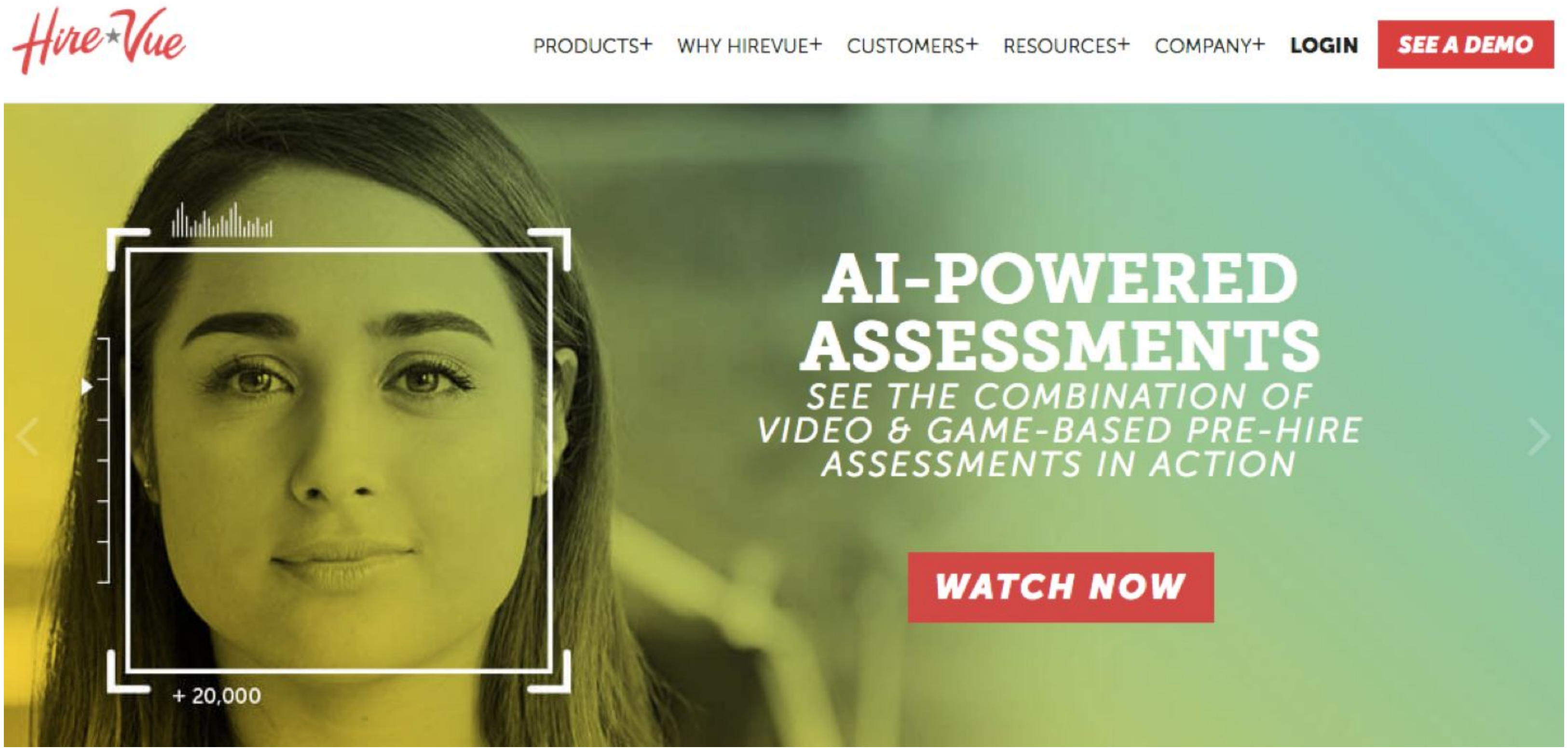
Ethical principles (III)

4. HLEGAI Key Requirement: **Transparency**

- the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Consider *HireVue*

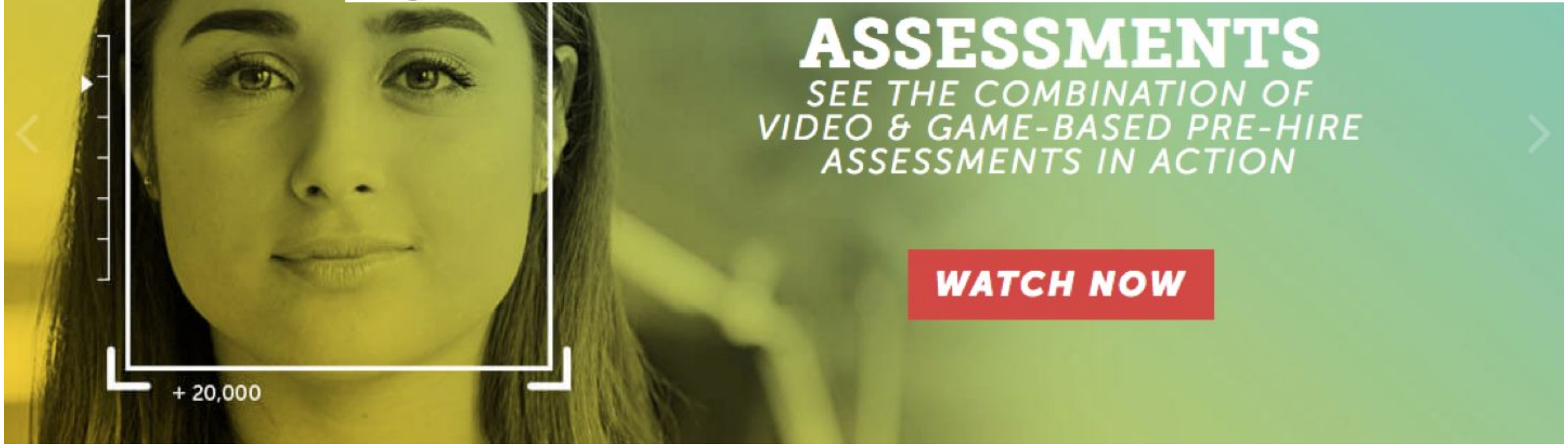


Consider *Hi*

HireVue said its system dissects the tiniest details of candidates' responses — their facial expressions, their eye contact and perceived "enthusiasm" — and compiles reports companies can use in deciding whom to hire or disregard.

The system, HireVue said, employs superhuman precision and impartiality to zero in on an ideal employee, picking up on telltale clues a recruiter might miss.

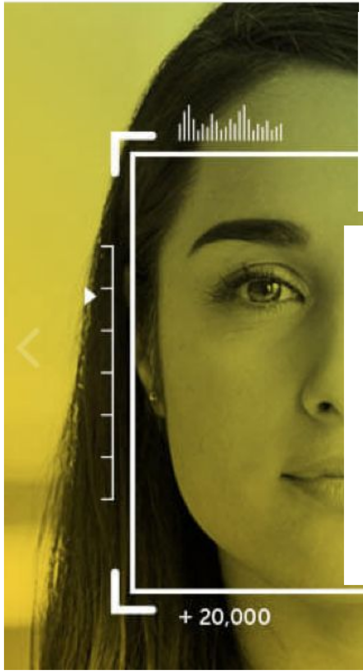
Hire★Vue



https://wapo.st/335o35O

Consider *Hi*

HireVue



HireVue said its system dissects the tiniest details of candidates' responses — their facial expressions, their eye contact and perceived "enthusiasm" — and compiles reports companies can use in deciding whom to hire or disregard.

The system, HireVue said, employs superhuman precision and impartiality to zero in on an ideal employee, picking up on telltale clues a recruiter might miss.

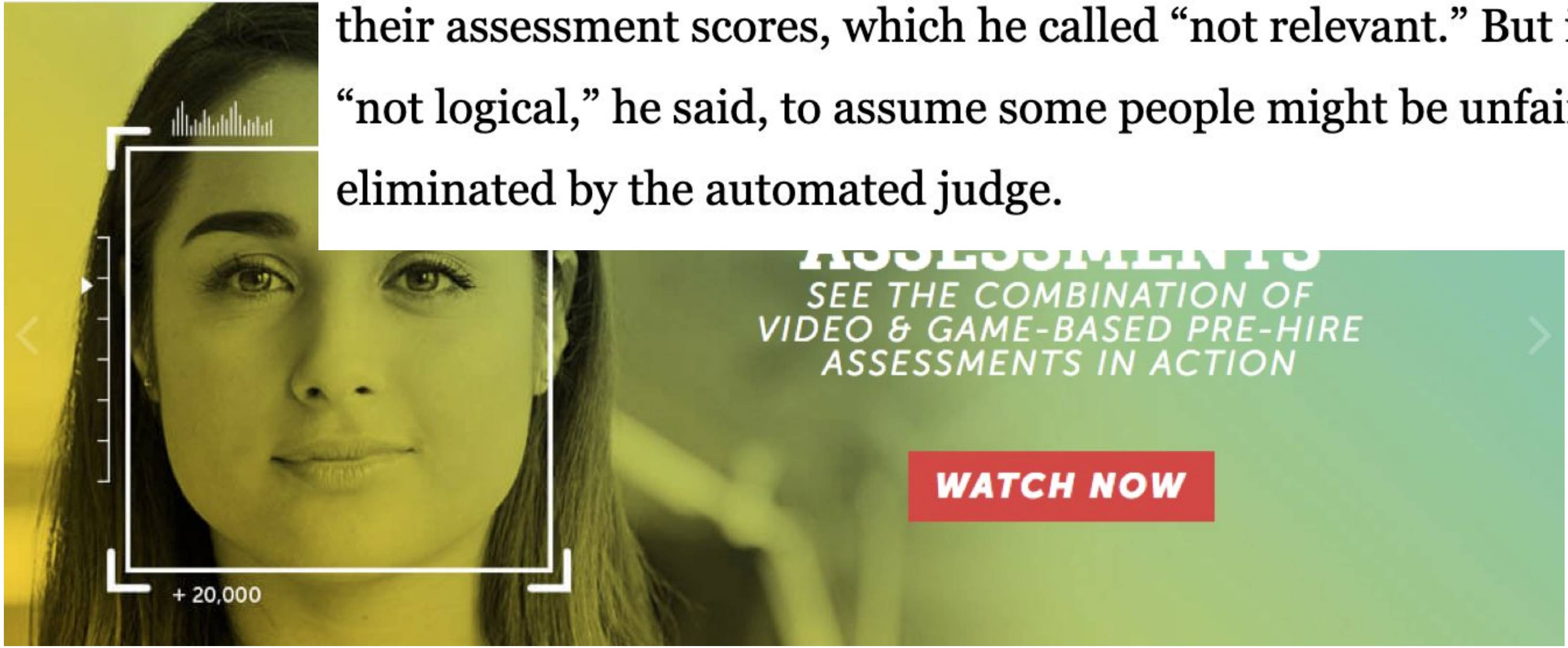
After a new candidate takes the HireVue test, the system generates a report card on their "competencies and behaviors," including their "willingness to learn," "conscientiousness & responsibility" and "personal stability," the latter of which is defined by how well they can cope with "irritable customers or co-workers."

https://wapo.st/335o35O

Consider *HireVue*

Loren Larsen, HireVue's chief technology officer,

The AI, he said, doesn't explain its decisions or give candidates their assessment scores, which he called "not relevant." But it is "not logical," he said, to assume some people might be unfairly eliminated by the automated judge.



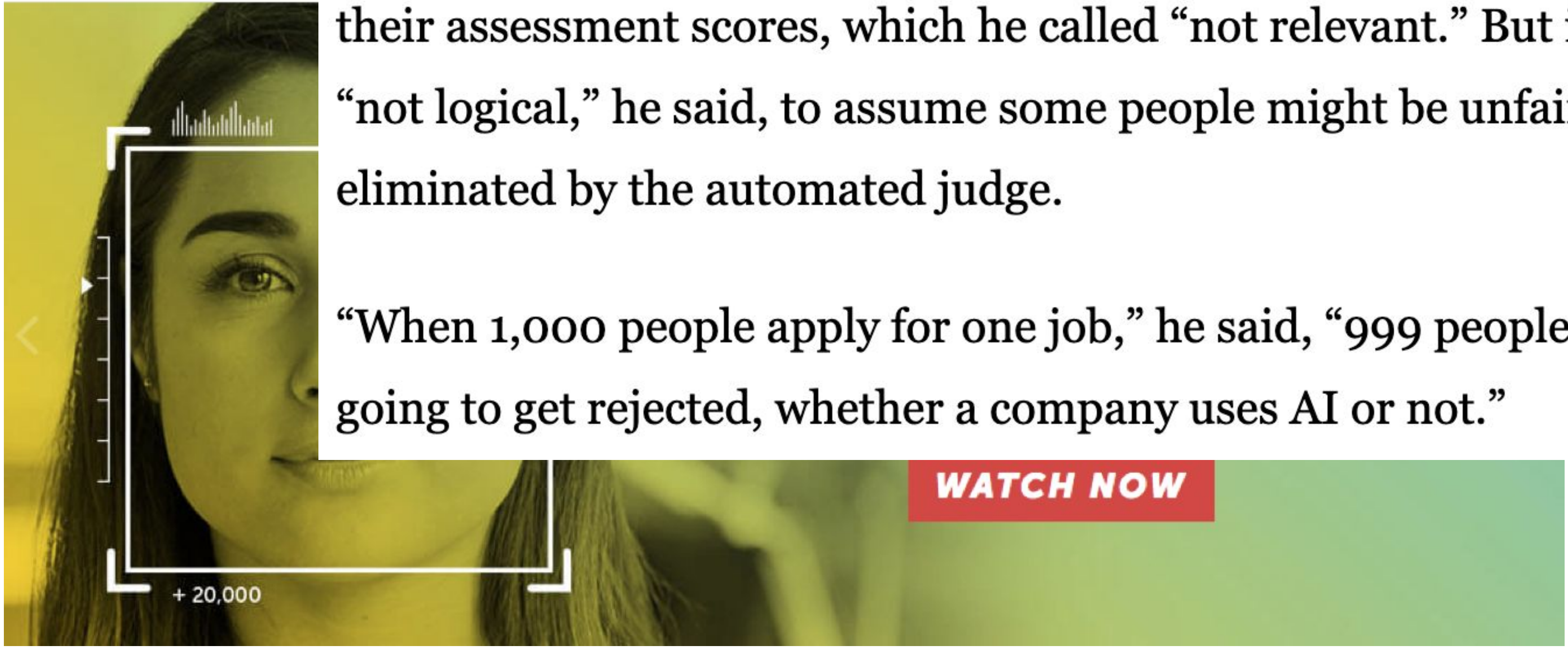
https://wapo.st/335o35O

Consider *HireVue*

Loren Larsen, HireVue's chief technology officer,

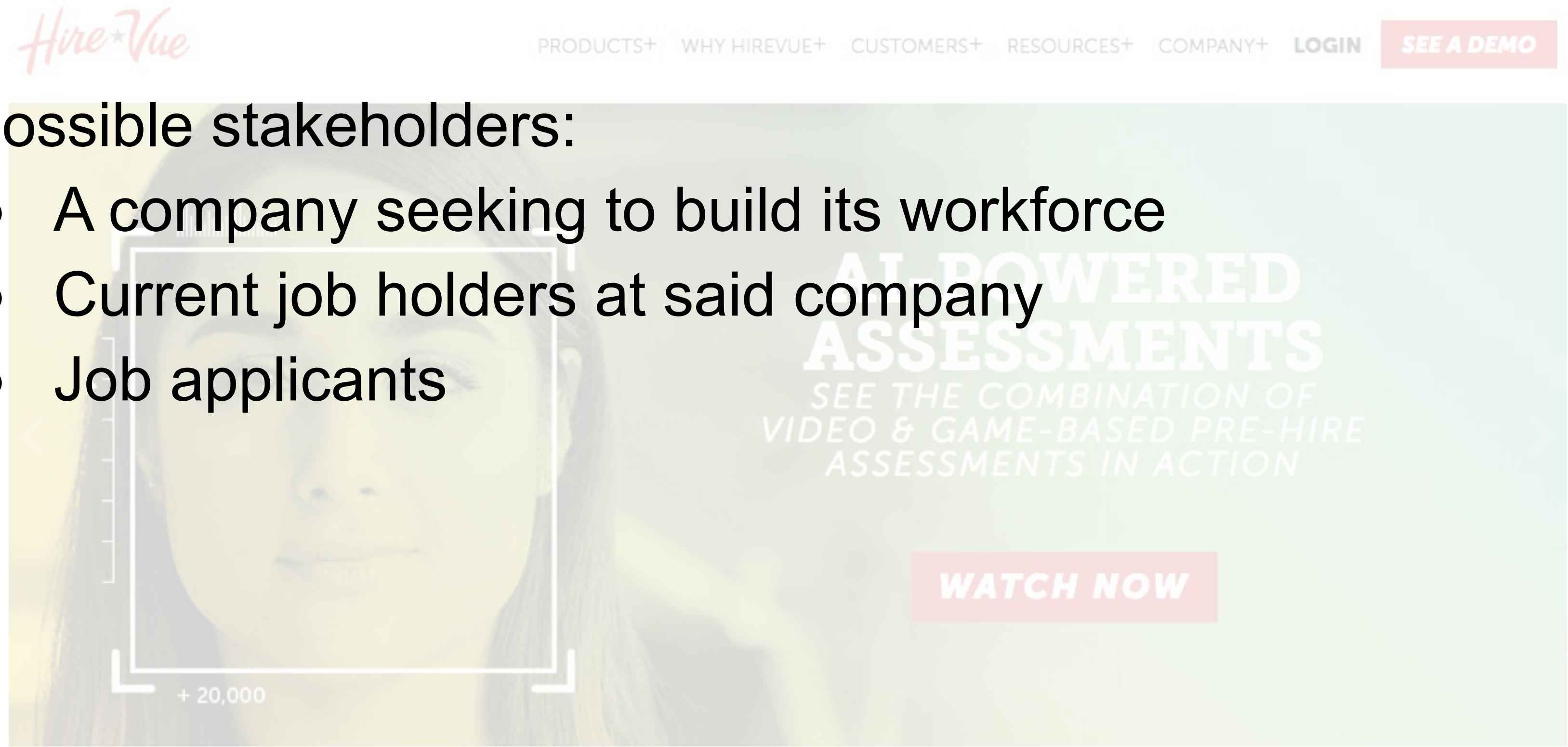
The AI, he said, doesn't explain its decisions or give candidates their assessment scores, which he called "not relevant." But it is "not logical," he said, to assume some people might be unfairly eliminated by the automated judge.

"When 1,000 people apply for one job," he said, "999 people are going to get rejected, whether a company uses AI or not."



https://wapo.st/335o35O

Consider *HireVue*



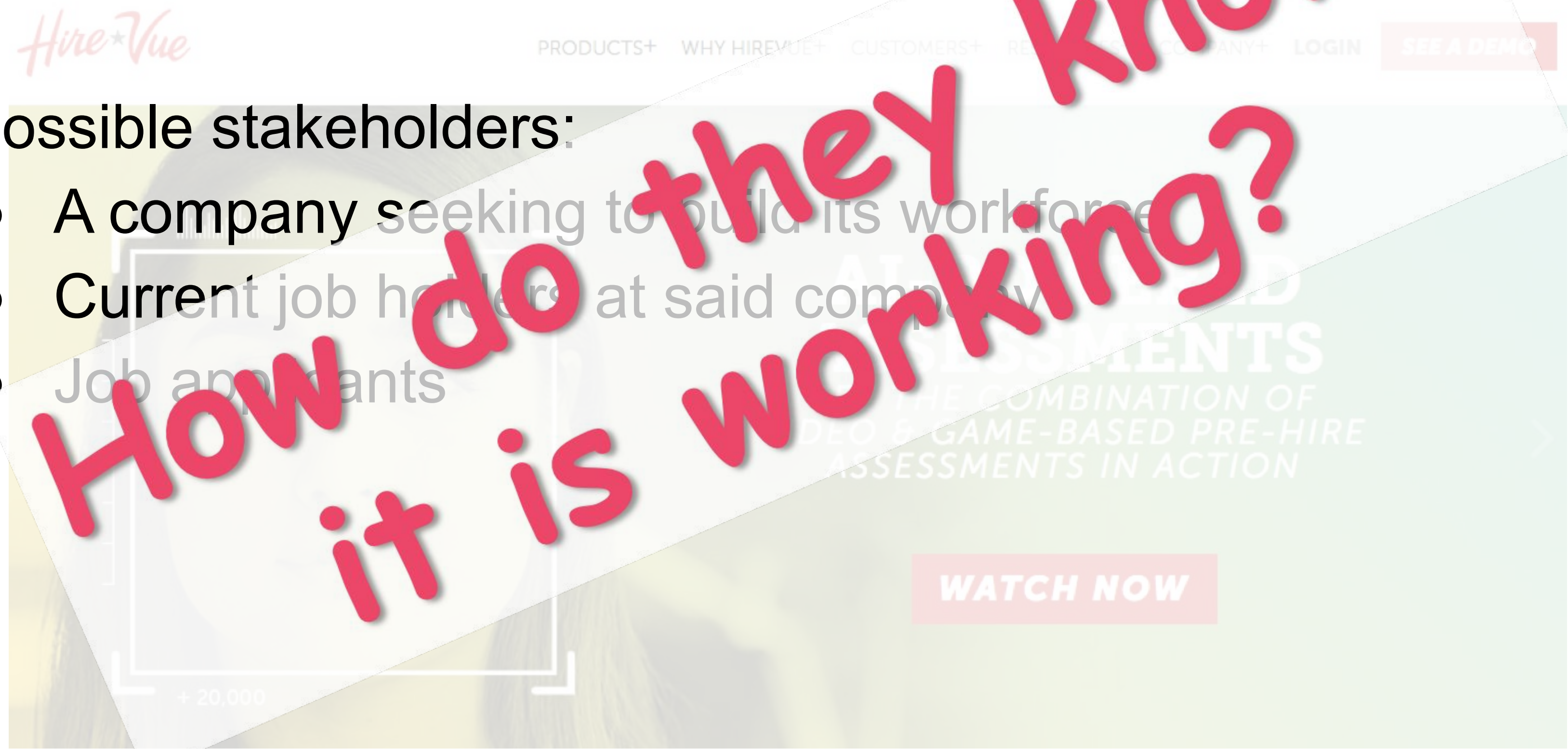
Possible stakeholders:

- A company seeking to build its workforce
- Current job holders at said company
- Job applicants

Consider *HireVue*

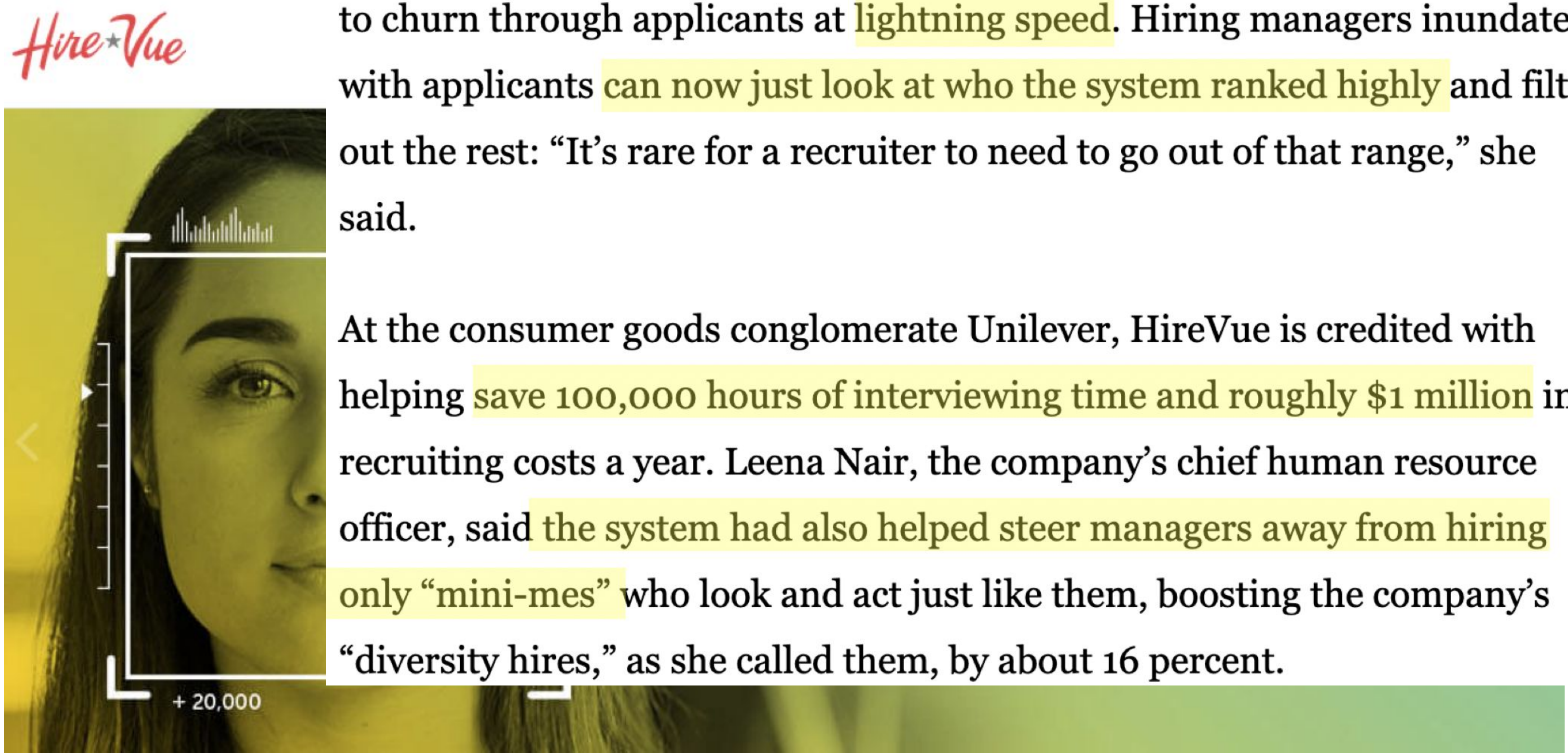
Possible stakeholders:

- A company seeking to build its workforce
- Current job holders at said company
- Job applicants



How do they know
it is working?

Consider *Hi*

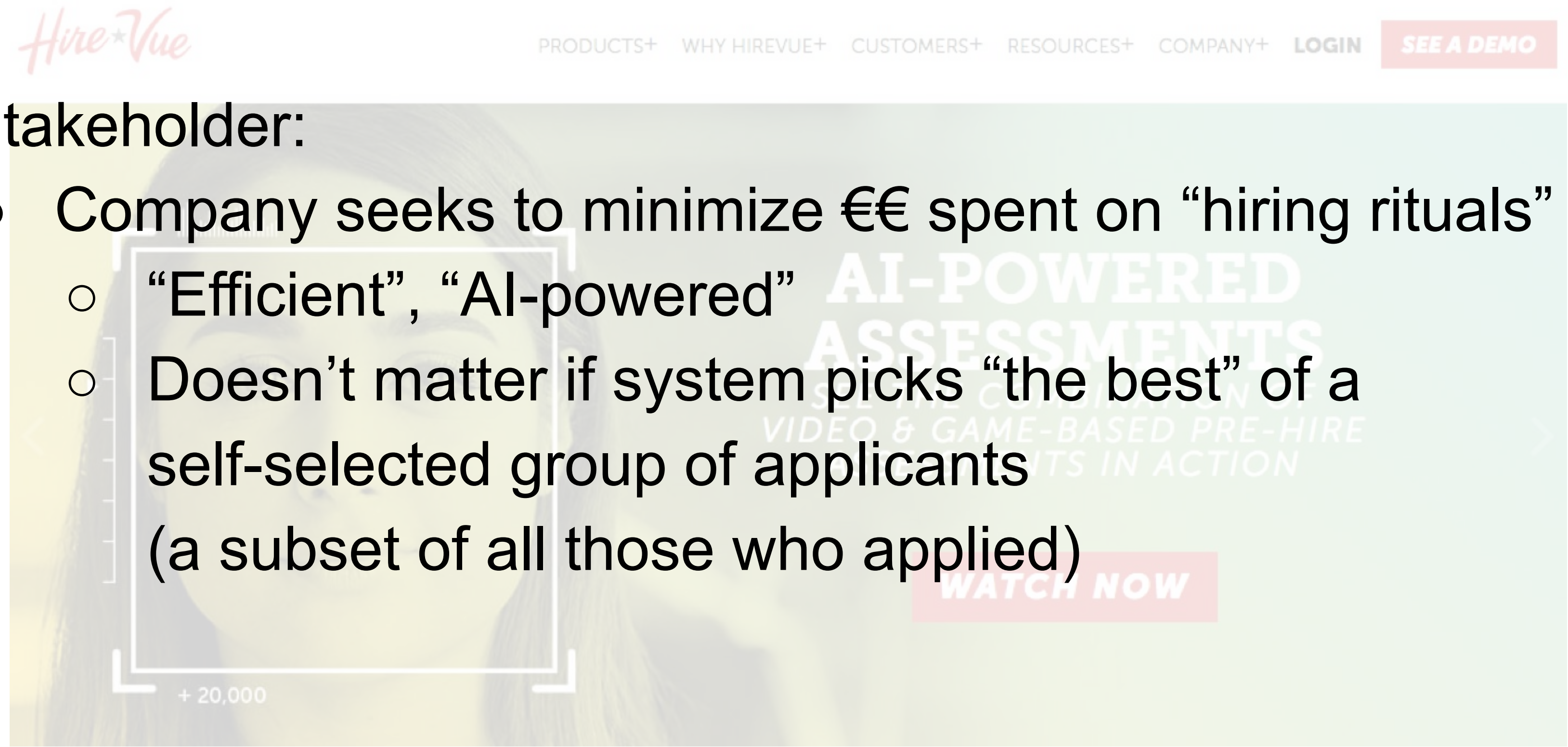


Sarah Smart, the company's vice president of global recruitment, said the system has radically redrawn Hilton's hiring rituals, allowing the company to churn through applicants at lightning speed. Hiring managers inundated with applicants can now just look at who the system ranked highly and filter out the rest: "It's rare for a recruiter to need to go out of that range," she said.

At the consumer goods conglomerate Unilever, HireVue is credited with helping save 100,000 hours of interviewing time and roughly $1 million in recruiting costs a year. Leena Nair, the company's chief human resource officer, said the system had also helped steer managers away from hiring only "mini-mes" who look and act just like them, boosting the company's "diversity hires," as she called them, by about 16 percent.

https://wapo.st/335o35O

Consider *HireVue*



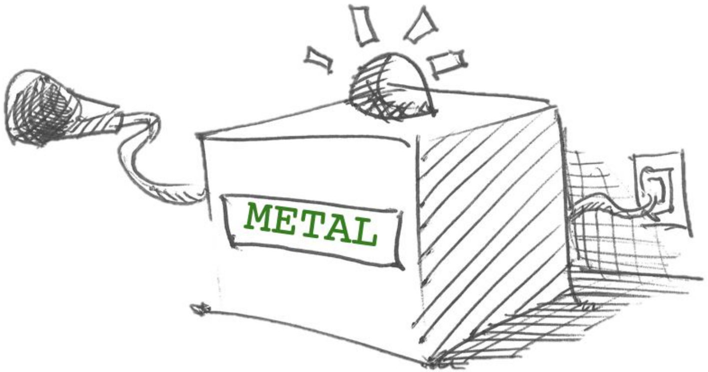
Stakeholder:

- Company seeks to minimize €€ spent on "hiring rituals"
  - "Efficient", "AI-powered"
  - Doesn't matter if system picks "the best" of a self-selected group of applicants (a subset of all those who applied)

Ethical principles (III)

4. HLEGAI Key Requirement: **Transparency**

- the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.



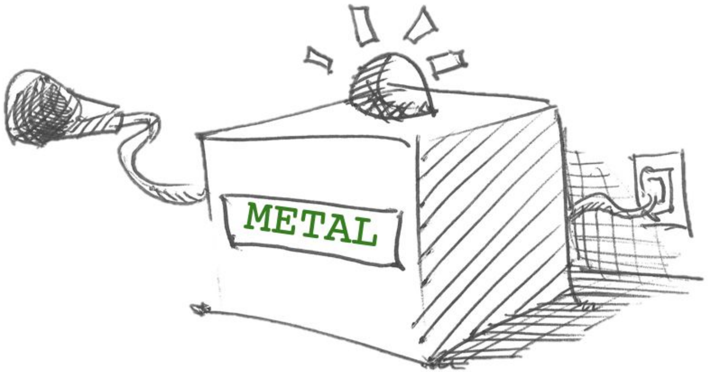
https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Ethical principles (III)

4. HLEGAI Key Requirement: **Transparency**

- the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

HOW DO YOU KNOW IT IS WORKING?



https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

How do you know it is working?

First you have to define what “to work” means (suitcase terms):

1. What is the intended mode of operation?
2. What are the **success criteria**?

These help define relevant
and reliable evaluation.


15th International Society for Music Information Retrieval Conference (ISMIR 2014)

FORMALIZING THE PROBLEM OF MUSIC DESCRIPTION

Bob L. Sturm Aalborg University Denmark bst@create.aau.dk

Rolf Bardeli Fraunhofer IAIS Germany rolf.bardeli@iais.fraunhofer.de

Thibault Langlois Lisbon University Portugal tl@di.fc.ul.pt

Valentin Emiya Aix-Marseille Université CNRS UMR 7279 LIF valentin.emiya@lif.univ-mrs.

ABSTRACT

of a formalism for “the problem of music descrip-
other things: ambiguity in what
address, how it

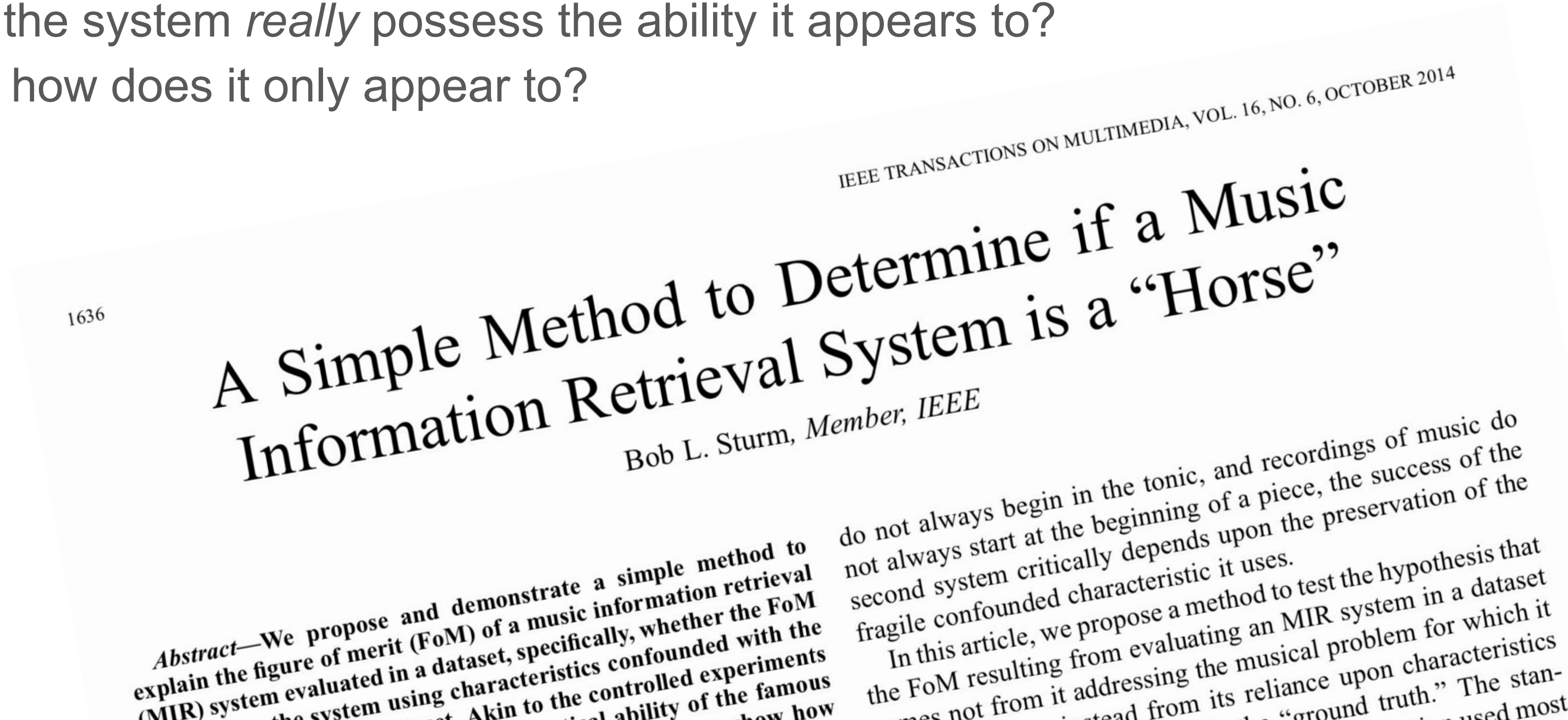
problems and questions that have major repercussions on the design and evaluation of any proposed system. For example, What is “genre”? What is “useful”? How is “feeling” related to “listening”? “Similar” in what respects?

With respect to the problem of music description, som
in MIR discusses the meaningfulness, worth, and fu
artificial systems to describe music [28
“ground truth” [3, 6, 1

How do you know it is working?

Work to answer these questions:

1. Does the system *really* possess the ability it appears to?
2. If not, how does it only appear to?


IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 16, NO. 6, OCTOBER 2014

1636

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

Bob L. Sturm, *Member, IEEE*

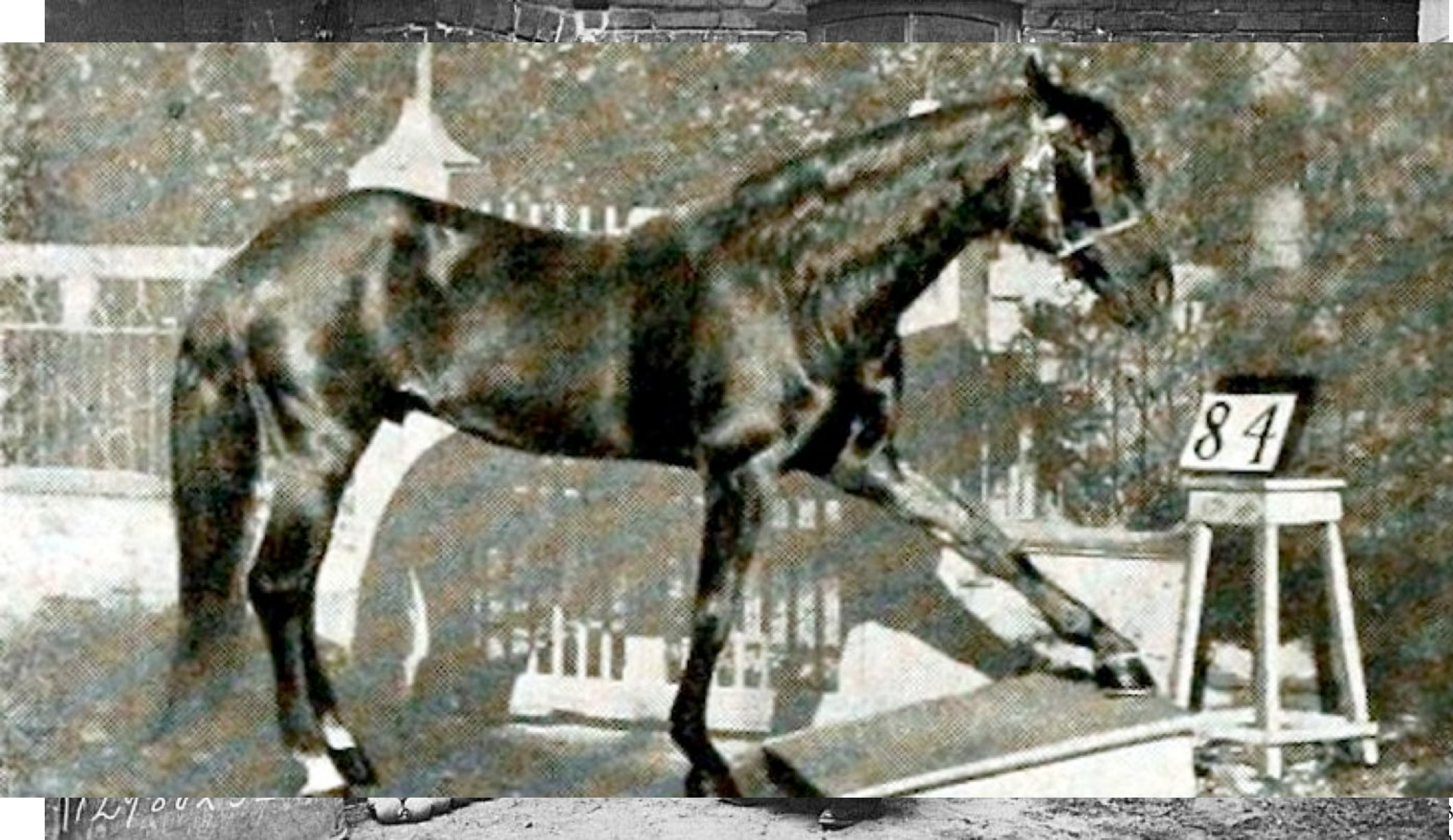
***Abstract*—We propose and demonstrate a simple method to explain the figure of merit (FoM) of a music information retrieval (MIR) system evaluated in a dataset, specifically, whether the FoM ... the system using characteristics confounded with the ... Akin to the controlled experiments ... ability of the famous ... how**

do not always begin in the tonic, and recordings of music do not always start at the beginning of a piece, the success of the second system critically depends upon the preservation of the fragile confounded characteristic it uses.

In this article, we propose a method to test the hypothesis that the FoM resulting from evaluating an MIR system in a dataset ... not from it addressing the musical problem for which it ... instead from its reliance upon characteristics ... "ground truth." The stan- ... used most

c do
f the
f the

is tha
datase
hich
teristi
he sta
sed m
“gro
3]) sin
sing


84

3]) sin
sing

How do you know it is working?

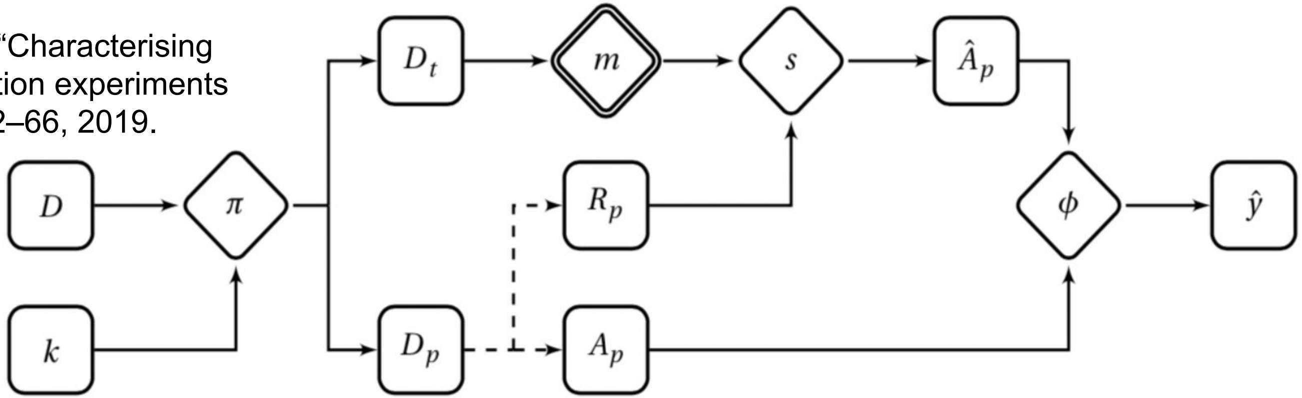
Work to answer these questions:

1. Does the system *really* possess the ability it appears to?
2. If not, how does it only appear to?

Work to explain *all* its answers, not just incorrect ones.

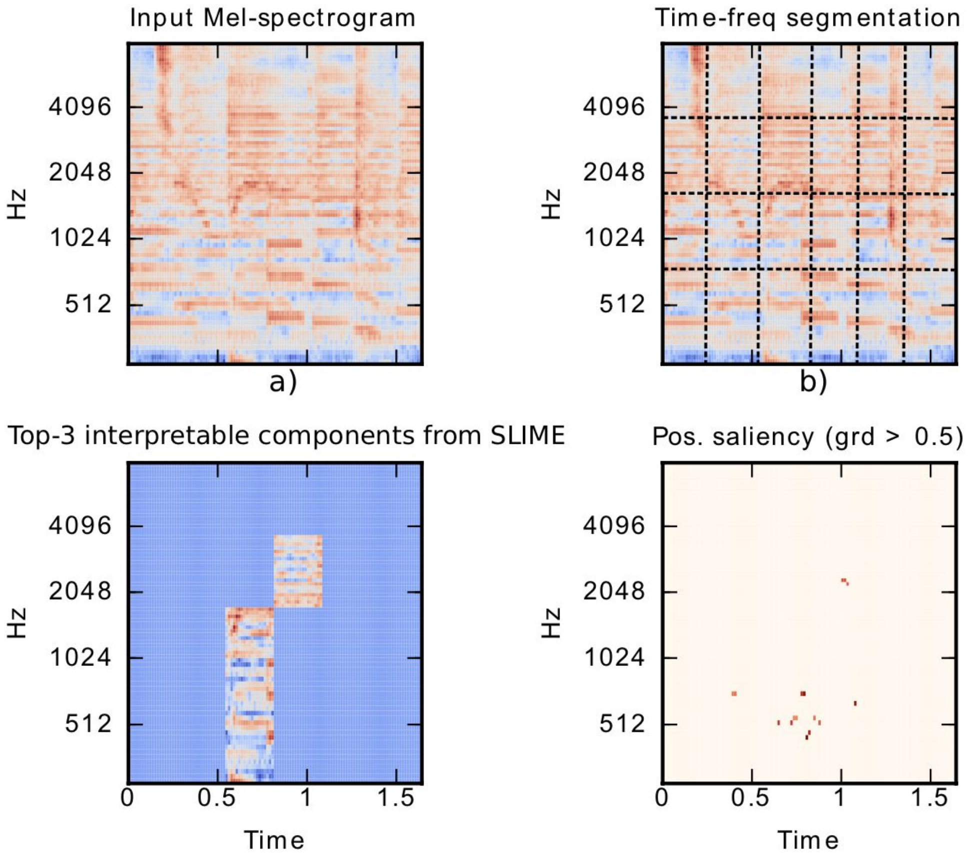
Don't just speculate. Use interventional experiments!

Rodríguez-Algarra, Sturm, and Dixon, "Characterising confounding effects in music classification experiments through interventions," TISMIR 2(1): 52–66, 2019.



How do you know it is working?

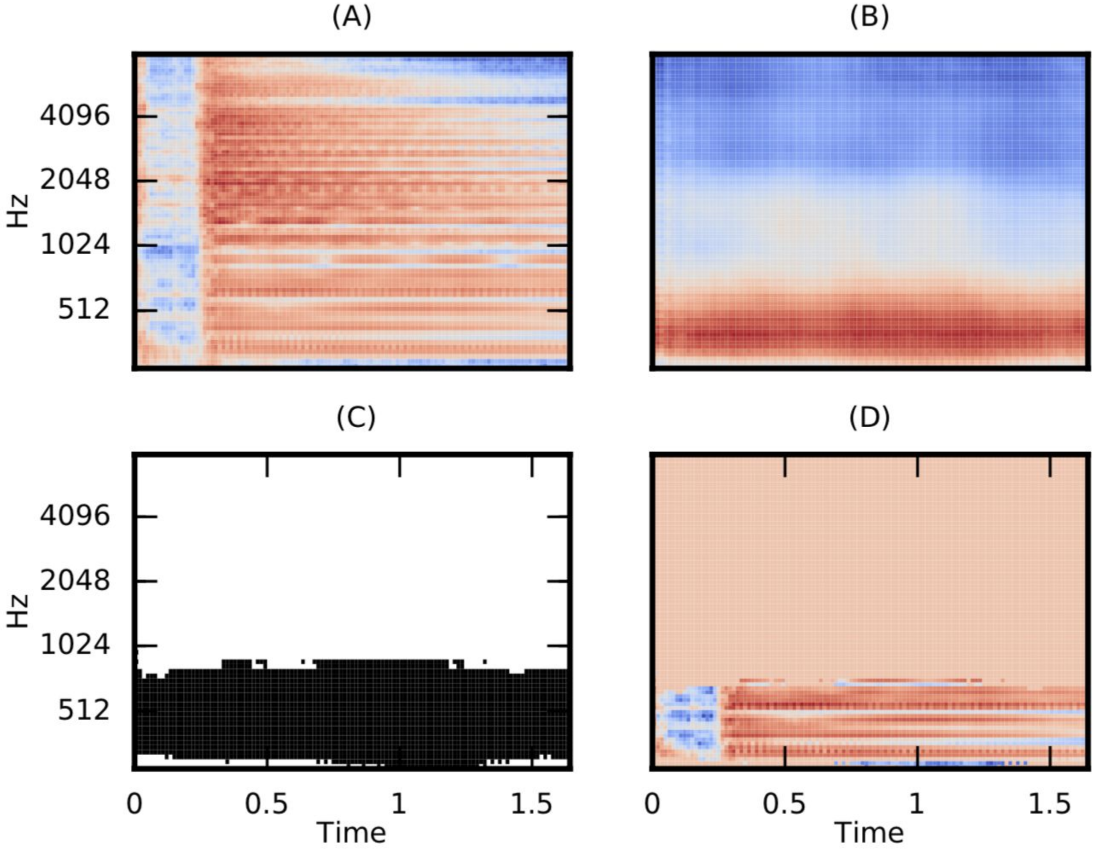
Mishra, Sturm, and Dixon, “Local interpretable model-agnostic explanations for music content analysis,” in *Proc. ISMIR*, 2017.



How do you know it is working?

Mishra, Sturm, and Dixon, "'What are you listening to?' Explaining predictions of deep machine listening systems," in *Proc. EUSIPCO*, 2018.

Mishra, Sturm, and Dixon, "Understanding a deep machine listening model through feature inversion," in *Proc. ISMIR*, 2018.



How do you know it is working?

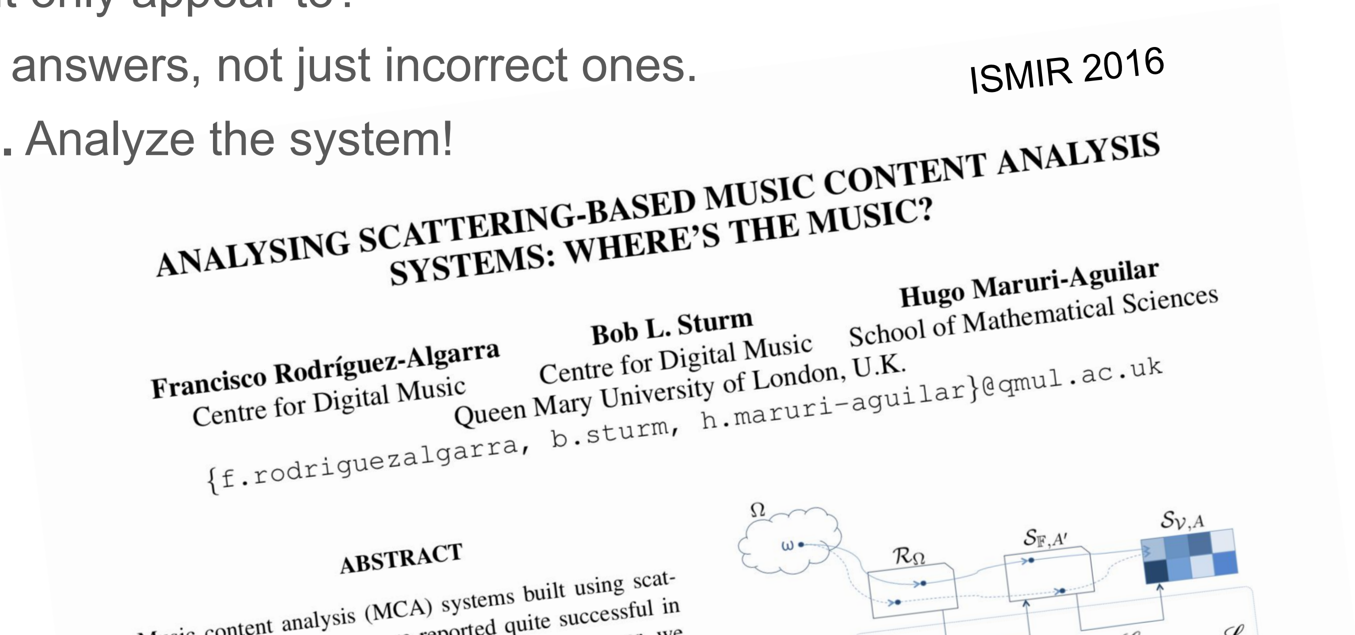
Work to answer these questions:

1. Does the system *really* possess the ability it appears to?
2. If not, how does it only appear to?

Work to explain *all* its answers, not just incorrect ones.

Don't just speculate. Analyze the system!

ISMIR 2016



ANALYSING SCATTERING-BASED MUSIC CONTENT ANALYSIS SYSTEMS: WHERE'S THE MUSIC?

Francisco Rodríguez-Algarra
Centre for Digital Music

Bob L. Sturm
Centre for Digital Music

Hugo Maruri-Aguilar
School of Mathematical Sciences

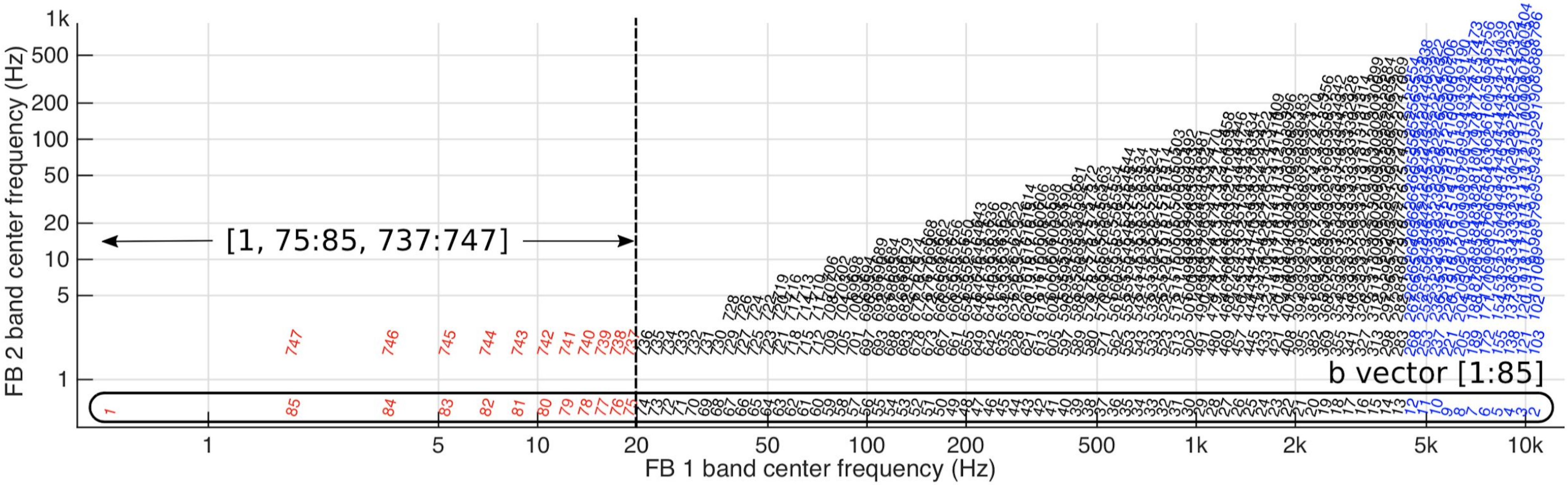
Queen Mary University of London, U.K.

{f.rodriguezalgarra, b.sturm, h.maruri-aguilar}@qmul.ac.uk

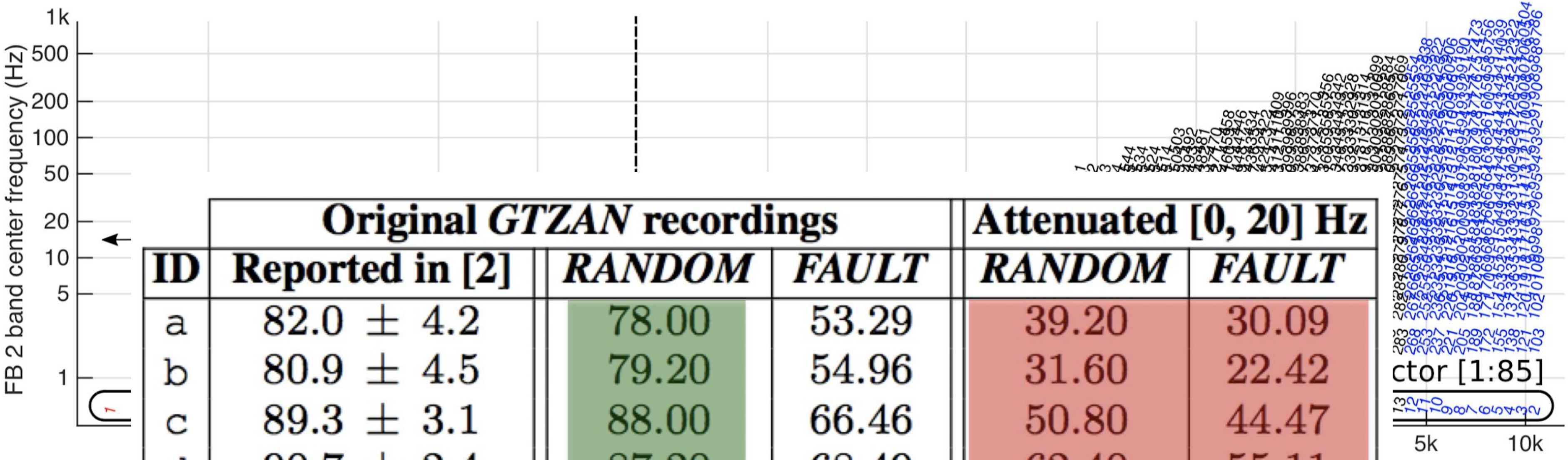
ABSTRACT

content analysis (MCA) systems built using scat-
reported quite successful in

How do you know it is working?



How do you know it is working?



	Original *GTZAN* recordings			Attenuated [0, 20] Hz	
ID	**Reported in [2]**	***RANDOM***	***FAULT***	***RANDOM***	***FAULT***
a	82.0 ± 4.2	78.00	53.29	39.20	30.09
b	80.9 ± 4.5	79.20	54.96	31.60	22.42
c	89.3 ± 3.1	88.00	66.46	50.80	44.47
d	90.7 ± 2.4	87.20	68.49	62.40	55.11
e	91.4 ± 2.2	85.60	68.61	64.80	44.52
f	89.4 ± 2.5	83.60	68.32	64.80	53.16

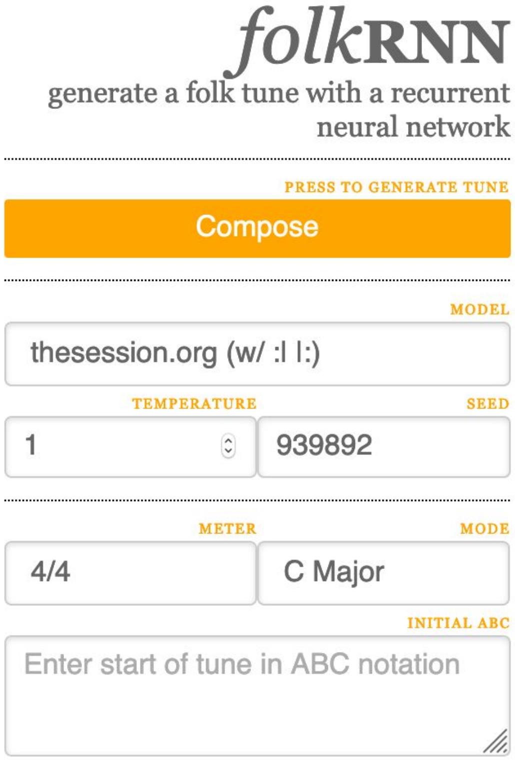
How do you know it is working?

Work to answer these questions:

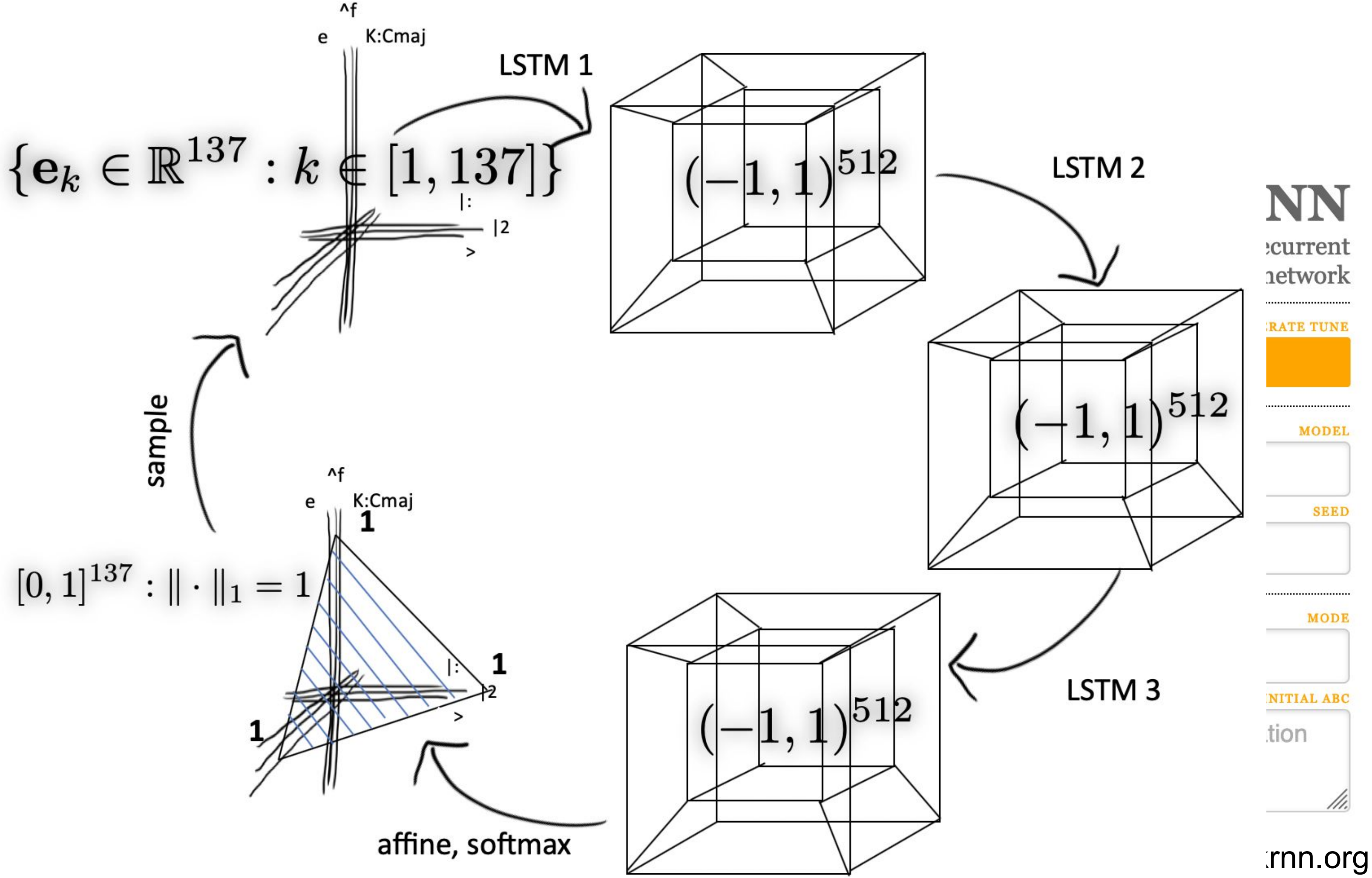
1. Does the system *really* possess the ability it appe
2. If not, how does it only appear to?

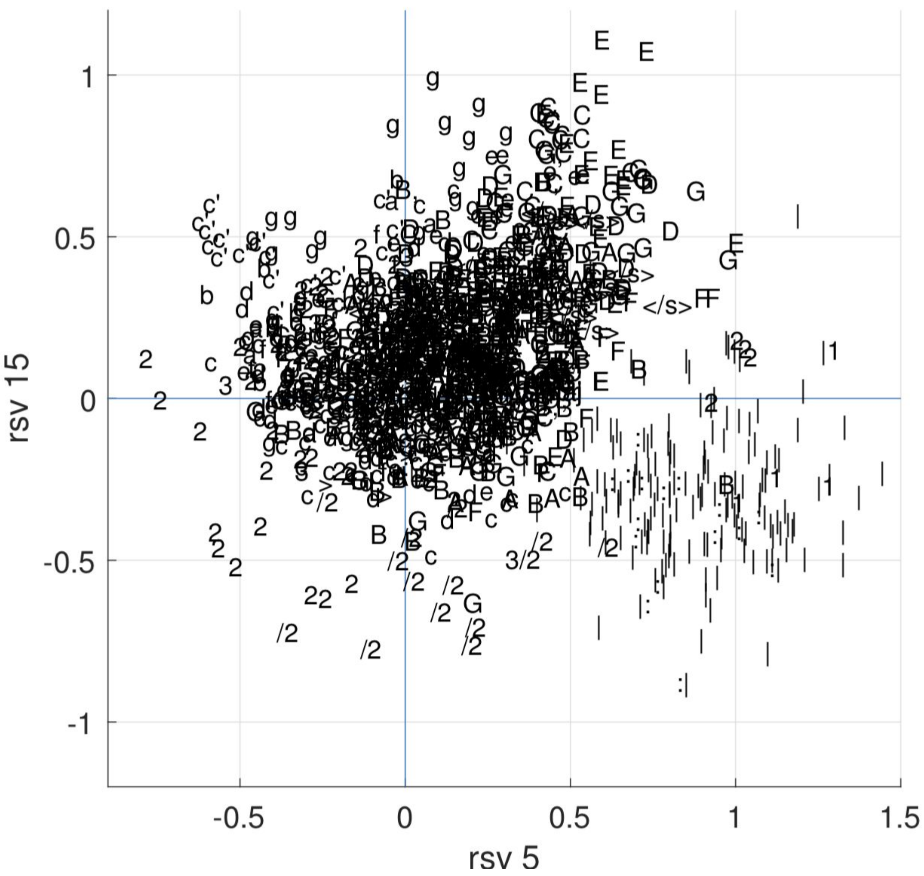
Work to explain *all* its answers, not just incorrect ones

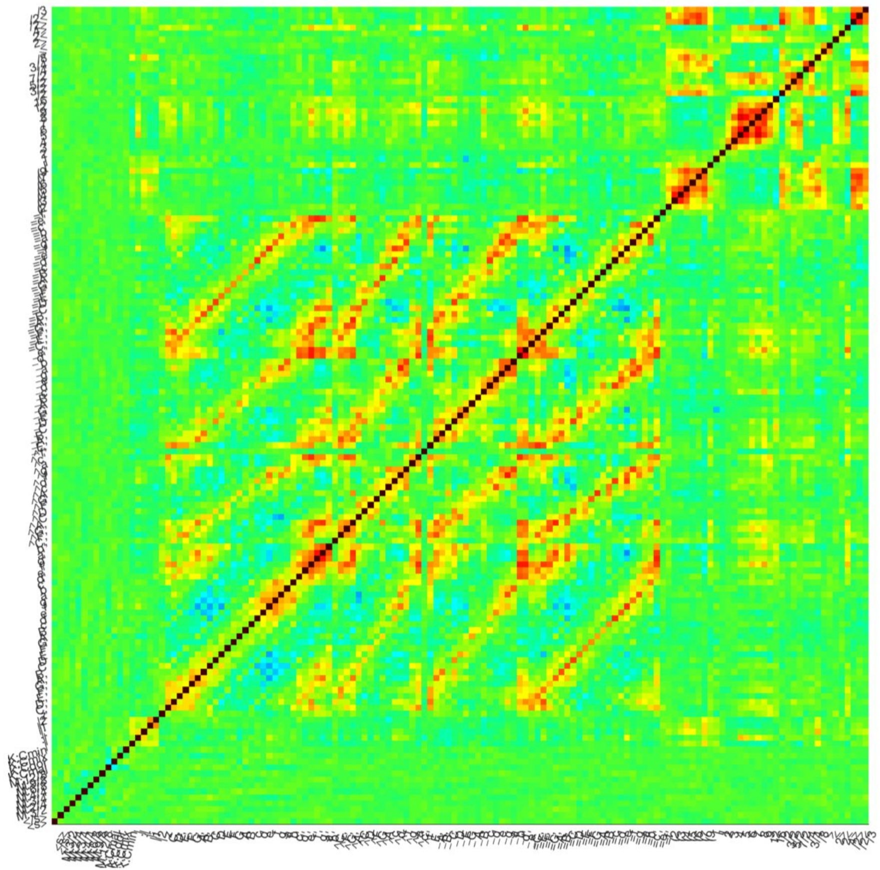
Don't just speculate. Analyze the system!



folkrnn.org


LSTM 1
LSTM 2
LSTM 3
$\{\mathbf{e}_k \in \mathbb{R}^{137} : k \in [1, 137]\}$
$(-1, 1)^{512}$
$(-1, 1)^{512}$
$(-1, 1)^{512}$
$[0, 1]^{137} : \| \cdot \|_1 = 1$
sample
affine, softmax
^f
e
K:Cmaj
|:
|2
>
1
1
1
NN
MODEL
SEED
MODE





1. Sturm, “What do these 5,599,881 parameters mean? An analysis of a specific LSTM music transcription model, starting with the 70,281 parameters of its softmax layer,” in *Proc. Music Metacreation*, 2018.
2. Sturm, “How stuff works: LSTM model of folk music transcriptions,” in *Proc. Workshop ML for Music*, ICML, 2018.

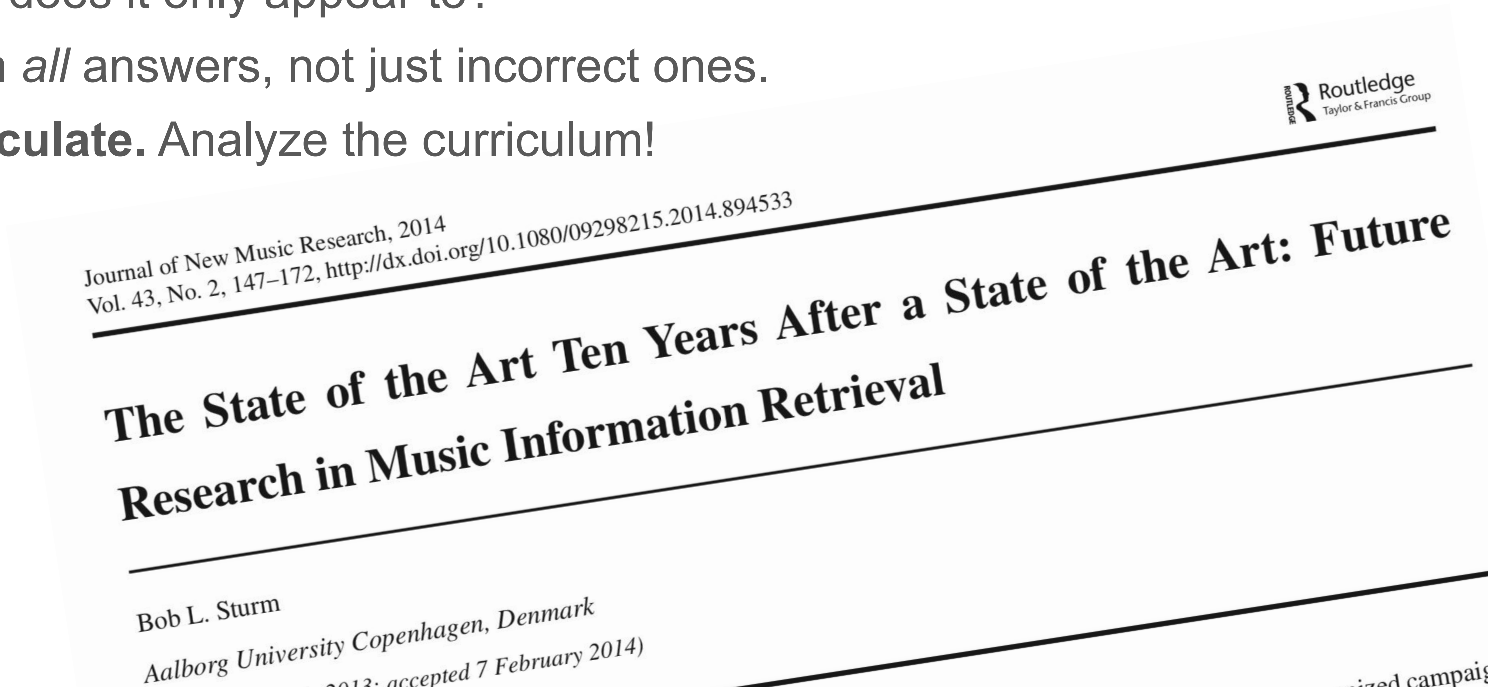
How do you know it is working?

Work to answer these questions:

1. Does the system *really* possess the ability it appears to?
2. If not, how does it only appear to?

Work to explain *all* answers, not just incorrect ones.

Don’t just speculate. Analyze the curriculum!


Routledge
Taylor & Francis Group

Journal of New Music Research, 2014
Vol. 43, No. 2, 147–172, http://dx.doi.org/10.1080/09298215.2014.894533

The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval

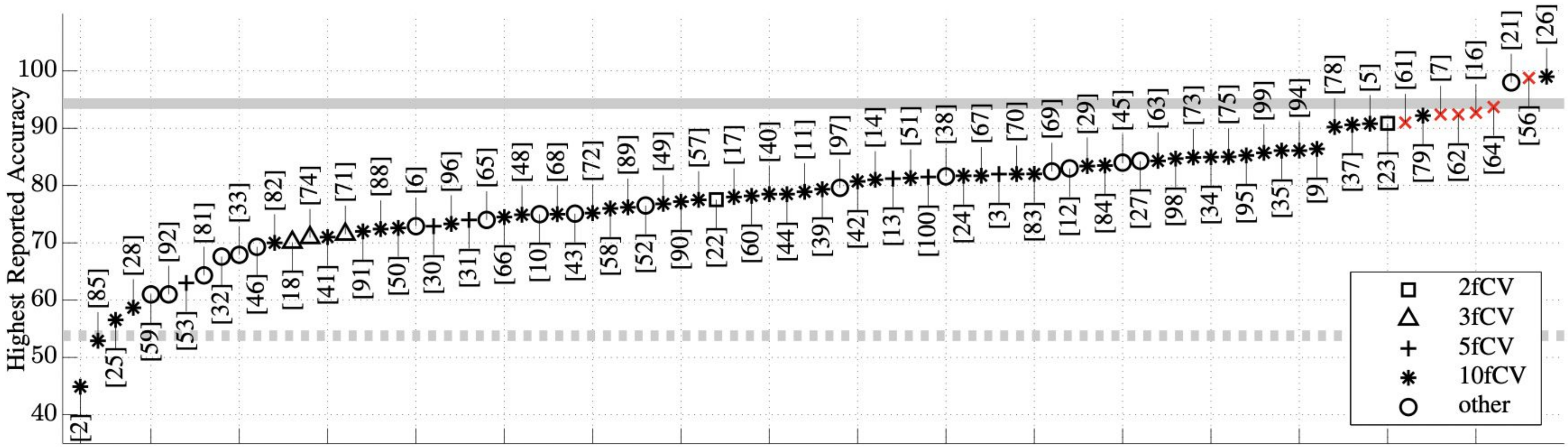
Bob L. Sturm

Aalborg University Copenhagen, Denmark

accepted 7 February 2014)

How do you know it is working?

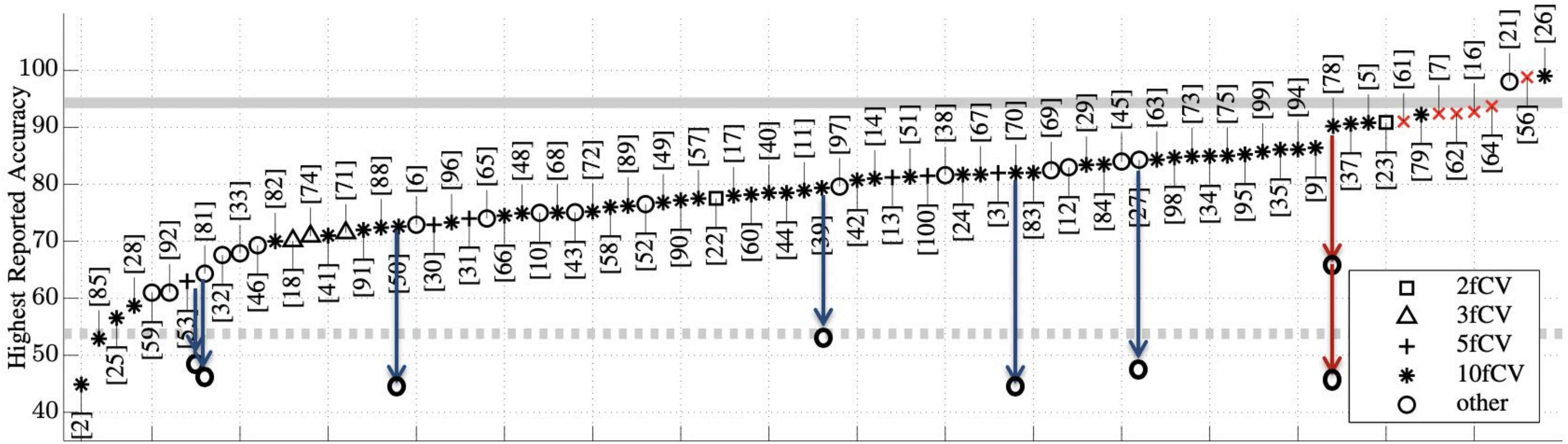
Got confounds?



Sturm, “The state of the art ten years after a state of the art: Future research in music information retrieval,” J. New Music Research 43(2): 147–172, 2014.

How do you know it is working?

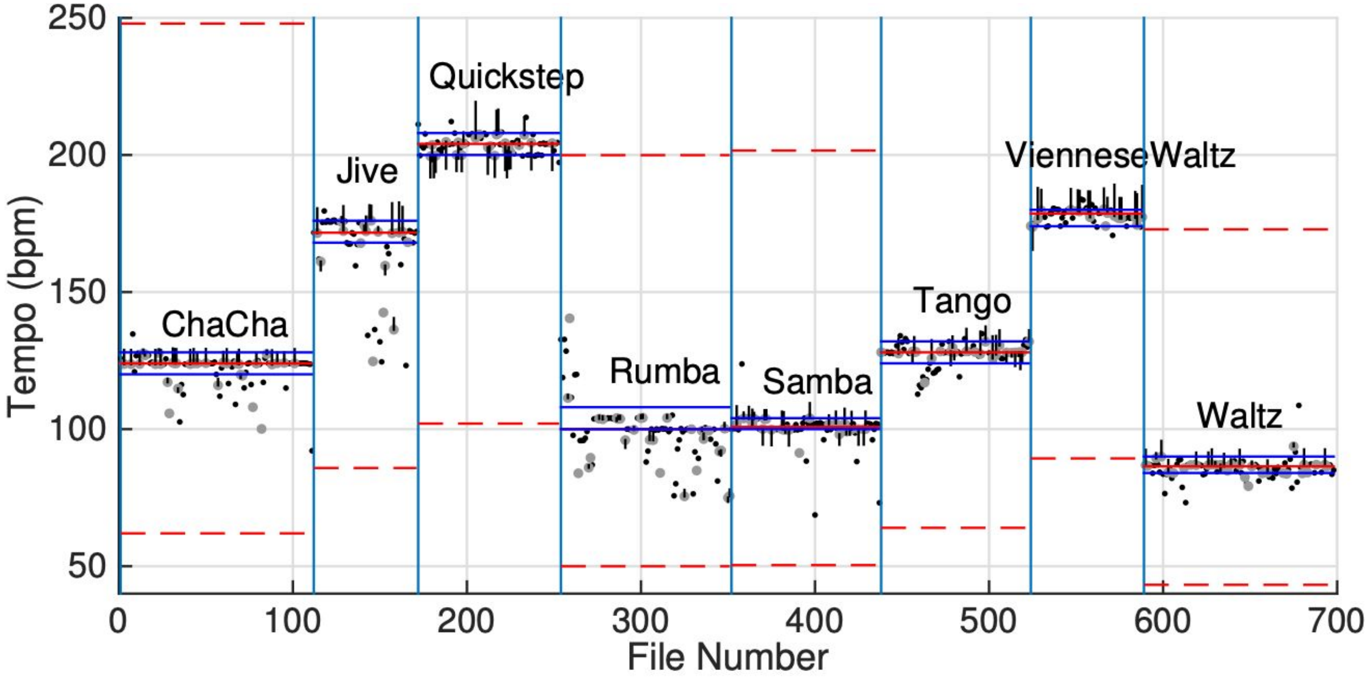
Got confounds?



Sturm, “The state of the art ten years after a state of the art: Future research in music information retrieval,” J. New Music Research 43(2): 147–172, 2014.

How do you know it is working?

Got confounds?



Sturm, “The “horse” inside: Seeking causes behind the behaviors of music content analysis systems,” *ACM Computers in Entertainment* 14(2) 2016.

How do you know it is working?

Work to answer these questions:

1. Does the system *really* possess the ability it appears to?
2. If not, how does it only appear to?

Work to explain *all* answers, not just incorrect ones.

Don't just speculate.

Get creative!

J. Schlüter, "Learning to pinpoint singing voice from weakly labeled examples," in Proc. ISMIR, 2016.

How do you know it is working?

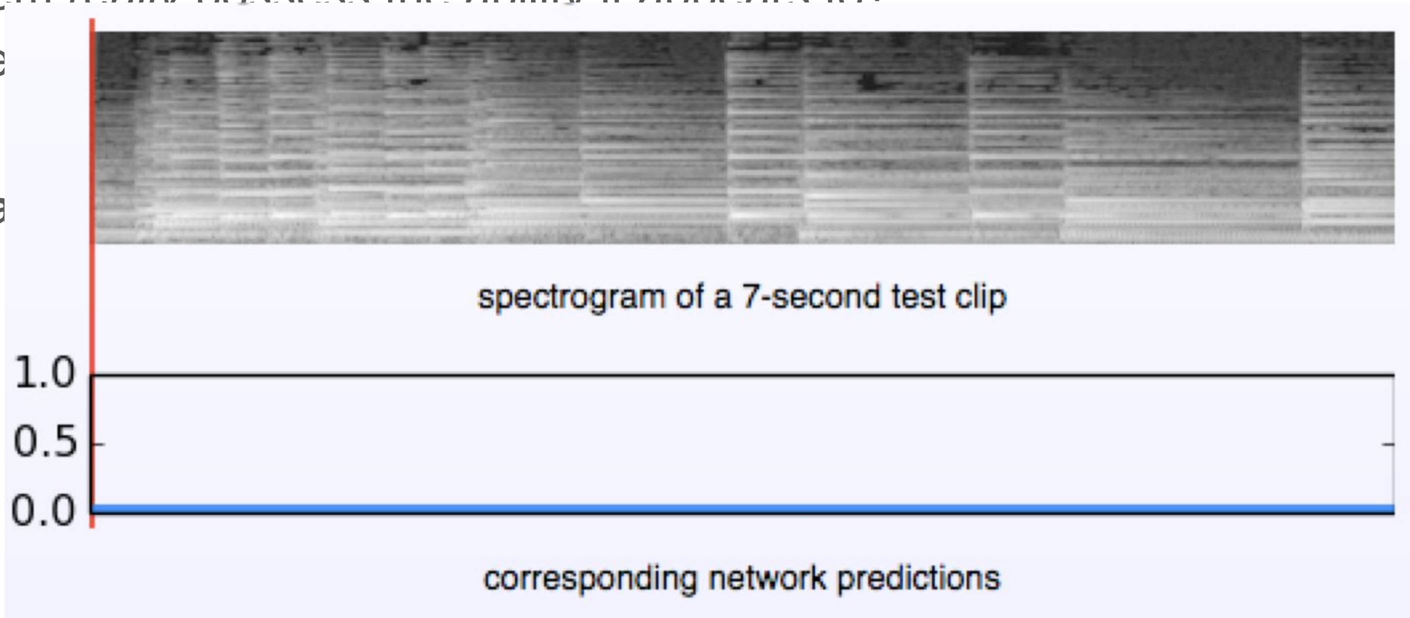
Work to answer these questions:

1. Does the system *really* possess the ability it appears to?
2. If not, how doe

Work to explain *all*

Don't just specula

Get creative!



Jan Schlüter, ISMIR 2016

How do you know it is working?

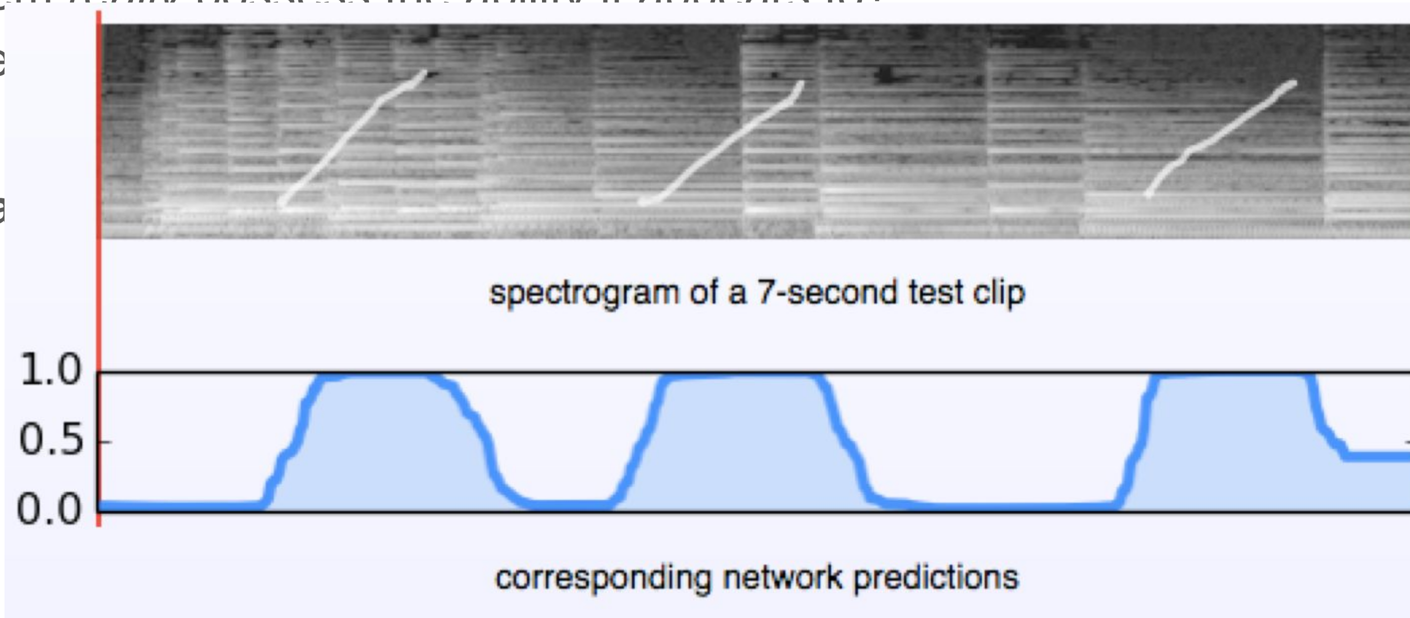
Work to answer these questions:

1. Does the system *really* possess the ability it appears to?
2. If not, how doe

Work to explain *all*

Don’t just specula

Get creative!



Jan Schlüter, ISMIR 2016

How do you know it is working?

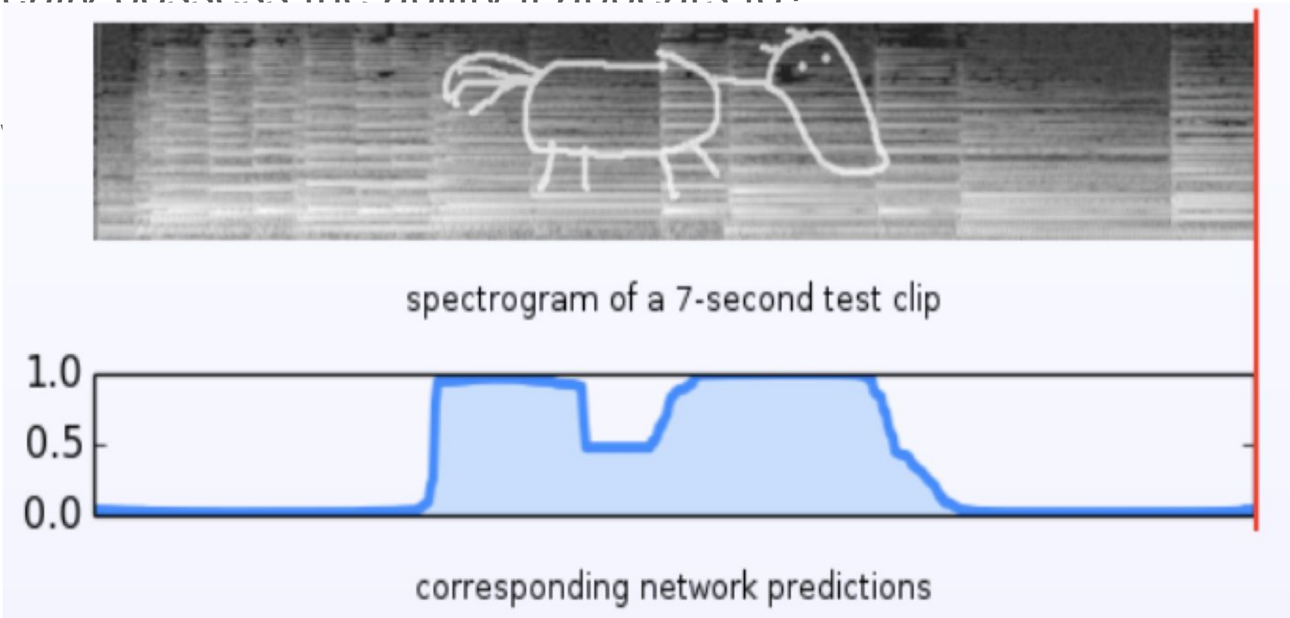
Work to answer these questions:

1. Does the system *really* possess the ability it appears to?
2. If not, how does it

Work to explain *all* ans

Don't just speculate.

Get creative!



spectrogram of a 7-second test clip

corresponding network predictions

Jan Schlüter, ISMIR 2016

How do you know it is working?

Work to answer these questions:

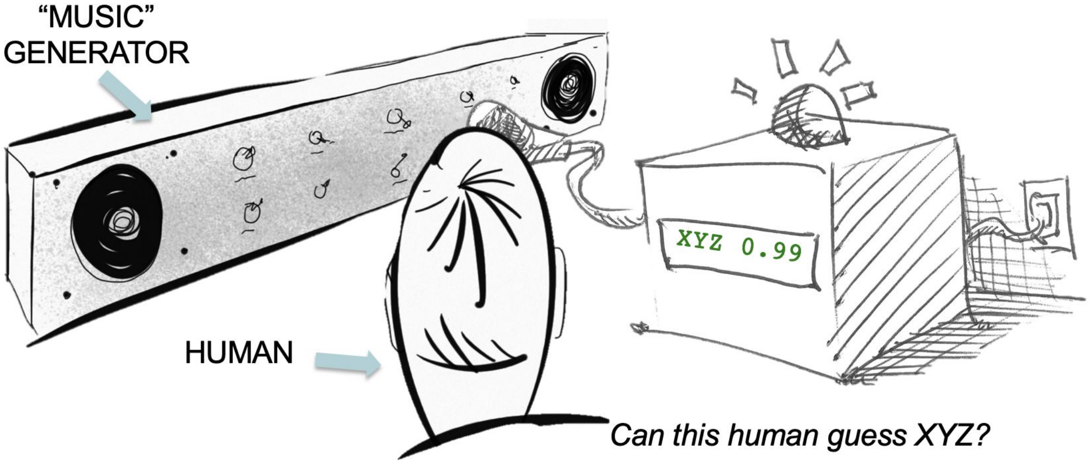
1. Does the system *really* possess the ability it appears to?
2. If not, how does it only appear to?

Work to explain *all* answers, not just incorrect ones.

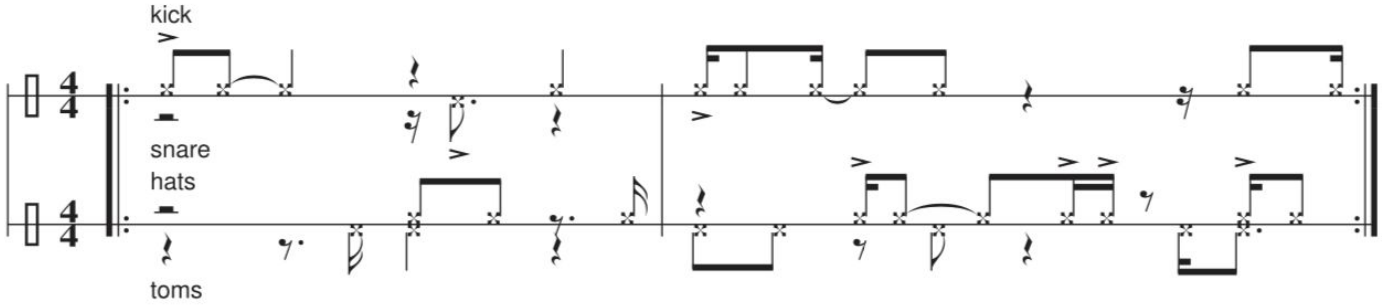
Don’t just speculate.

Get creative!

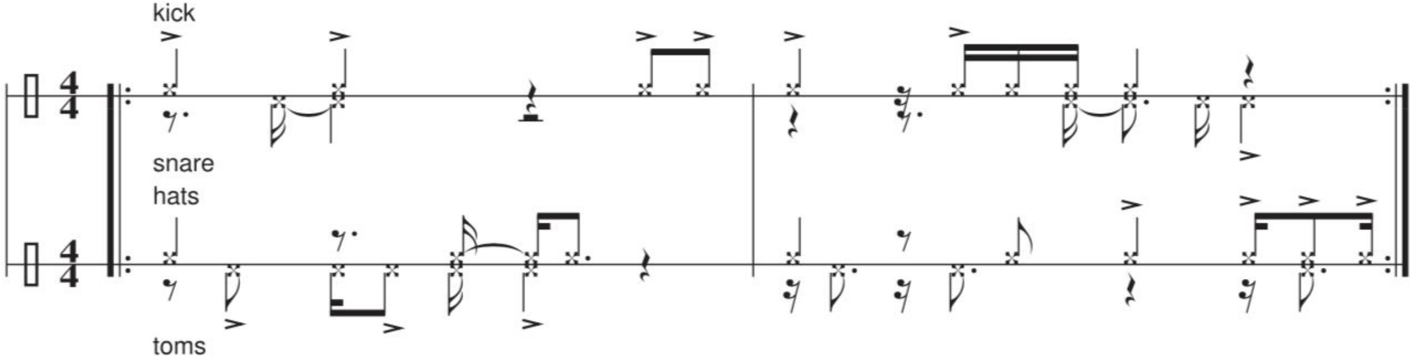
Sturm, “Two systems for automatic music genre recognition: What are they really recognizing?,” in *Proc. ACM MIRUM Workshop*, 2012.



How do you know it is working?



(c) Pattern 5519, classified with maximum confidence as Tango



(d) Pattern 2684, classified with maximum confidence as Waltz

Sturm, “The “horse” inside: Seeking causes behind the behaviors of music content analysis systems,” *ACM Computers in Entertainment* 14(2) 2016.

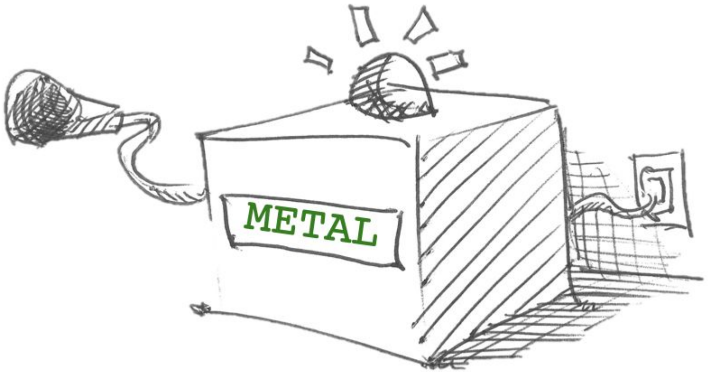
Ethical principles (III)

4. HLEGAI Key Requirement: **Transparency**

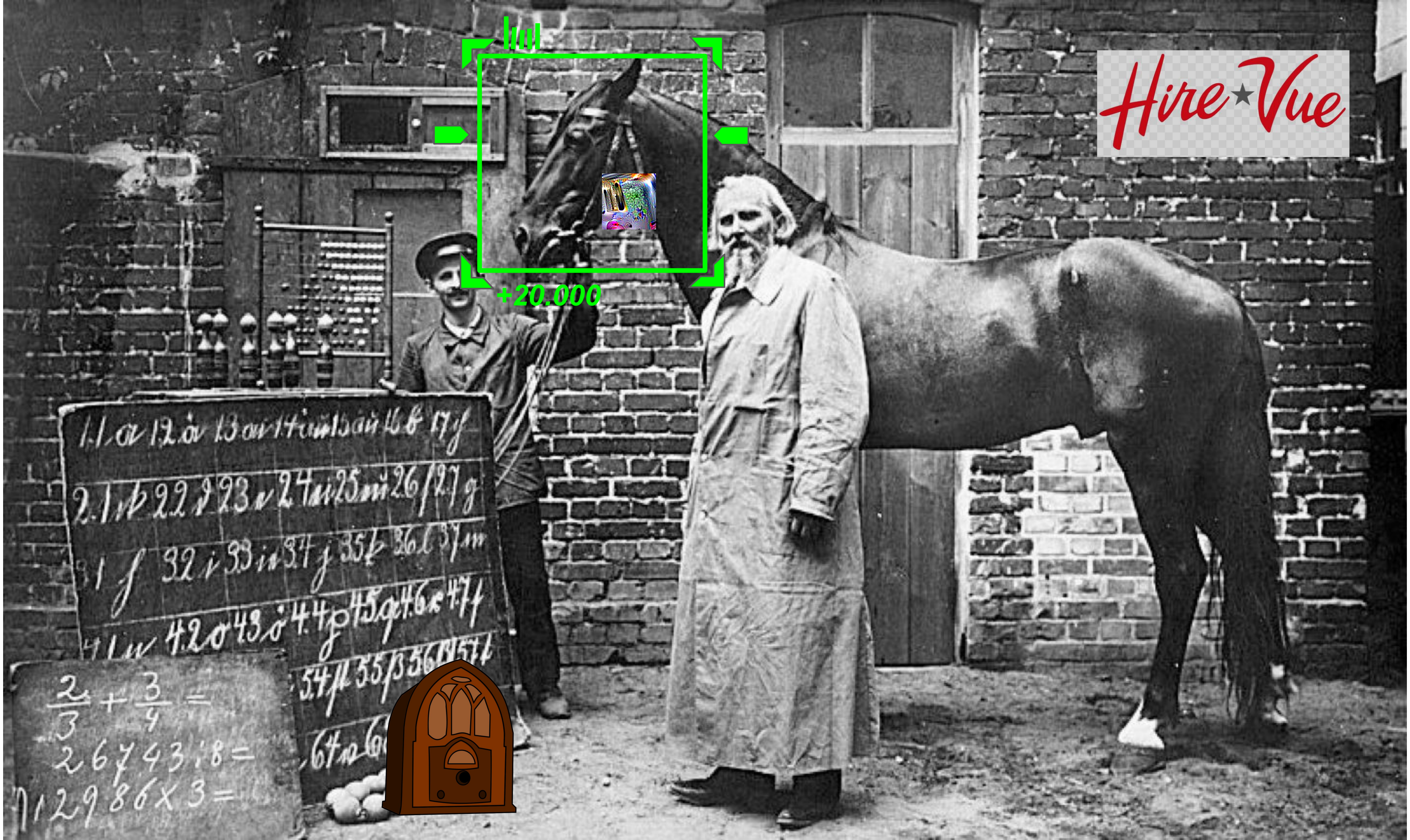
- the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

HOW DO YOU KNOW IT IS WORKING?

Be brave: Release your code and invite others to break it!



https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai


Hire★Vue
+20.000
2/3 + 3/4 =
26743 : 8 =
12986 × 3 =

Questions?

Fairness, Accountability and Transparency in Music Information Research (FAT-MIR)

Tutorial

Emilia Gomez, Andre Holzapfel, Marius Miron, Bob L. Sturm

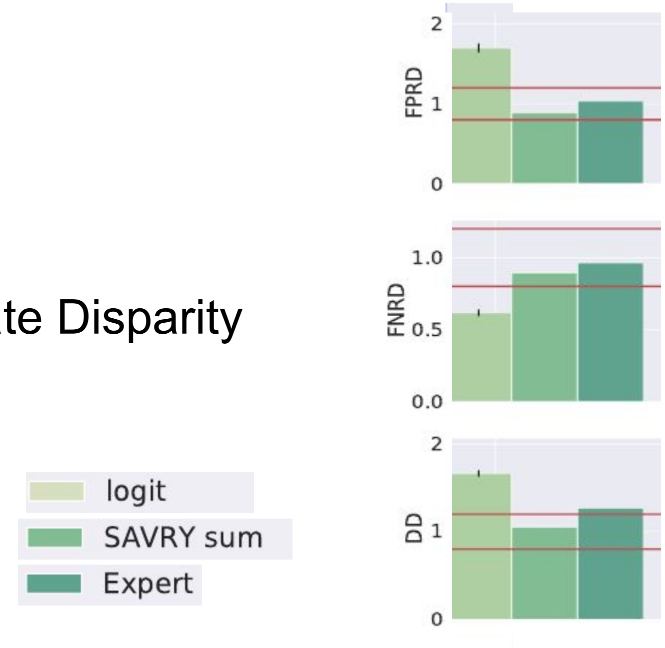
#fat-mir

20th International Society for Music Information Retrieval Conference
Delft, The Netherlands, November 4-8, 2019

Trade-offs - Fairness vs Predictive power

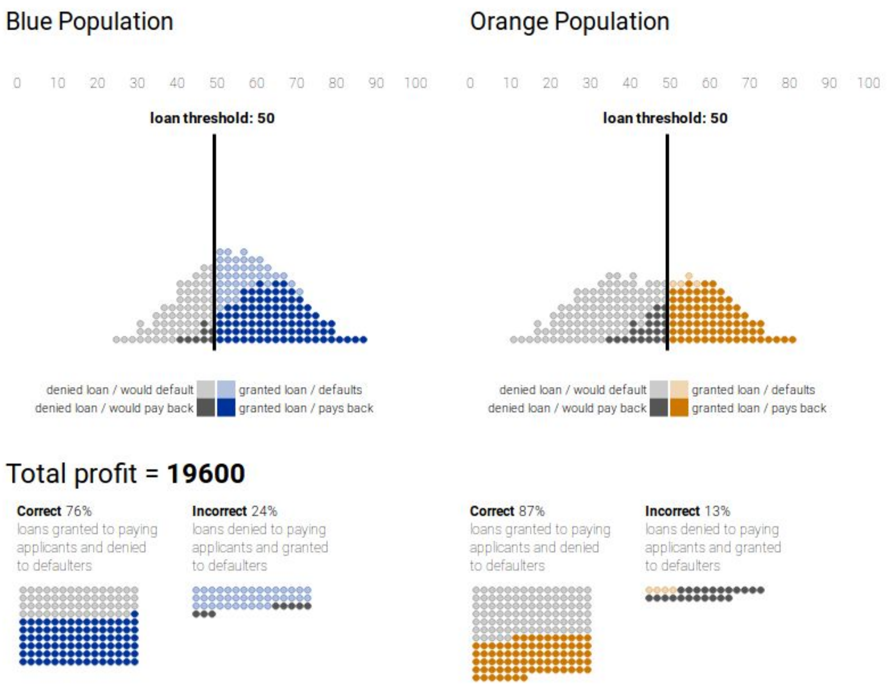
AUC of SAVRY sum = 0.64
AUC of expert = 0.66
Logistic regression: AUC = 0.71

However it also increases False Positive Rate Disparity (FPRD) between foreigners and nationals



Tolan et al., Why machine learning may lead to unfairness (2019)

Trade-offs - thresholds



https://research.google.com/bigpicture/attacking-discrimination-in-ml/