Why fairness can’t (and shouldn’t) be ‘solved’ by machine learning

Cynthia C. S. Liem

c.c.s.liem@tudelft.nl | @informusiccs

Multimedia Computing Group

Delft University of Technology

Fairness? Machine learning? Wait, weren’t you that music person?

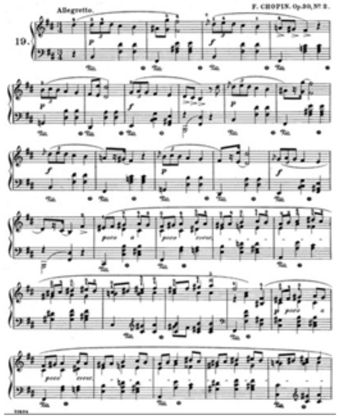
The classical music tradition

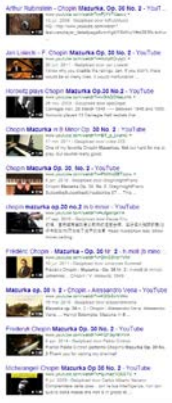
- A composer writes a composition

Allegretto.
F. CHOPIN. Op.30, No 2.
19.

The classical music tradition

- A composer writes a composition
- The composition gets performed by many interpreters





Arthur Rubinstein - Chopin Mazurka, Op. 30 No. 2 - YouT...

Jan Lisiecki - F. Chopin Mazurka Op. 30 No. 2 - YouTube

Horowitz plays Chopin Mazurka Op.30 No.2 - YouTube

Chopin Mazurka in B Minor Op. 30 No. 2 - YouTube

Chopin Mazurka Op. 30, No. 2 - YouTube

chopin mazurka op.30 no.2 in b minor - YouTube

Frédéric Chopin - Mazurka - Op. 30 Nr. 2 - h-moll (b mino...

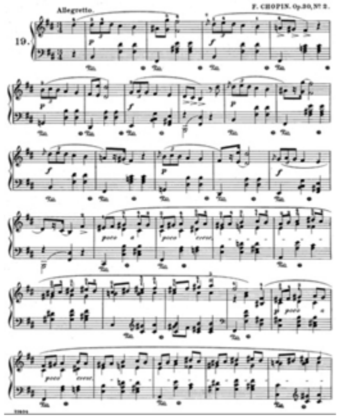
Mazurka op. 30 n. 2 - Chopin - Alessandro Vena - YouTube

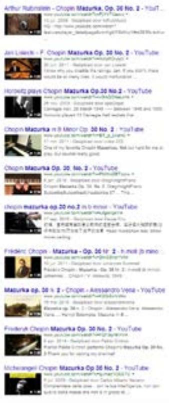
Frederik Chopin Mazurka Op. 30 No. 2 - YouTube

Michelangeli Chopin Mazurka Op 30 No. 2 - YouTube

The classical music tradition

- A composer writes a composition
- The composition gets performed by many interpreters
- Will interpreter n+1 just re-render the same content again?





Arthur Rubinstein - Chopin Mazurka, Op. 30 No. 2 - YouT...

Jan Lisiecki - F. Chopin Mazurka Op. 30 No. 2 - YouTube

Horowitz plays Chopin Mazurka Op.30 No.2 - YouTube

Chopin Mazurka in B Minor Op. 30 No. 2 - YouTube

Chopin Mazurka Op. 30, No. 2 - YouTube

chopin mazurka op.30 no.2 in b minor - YouTube

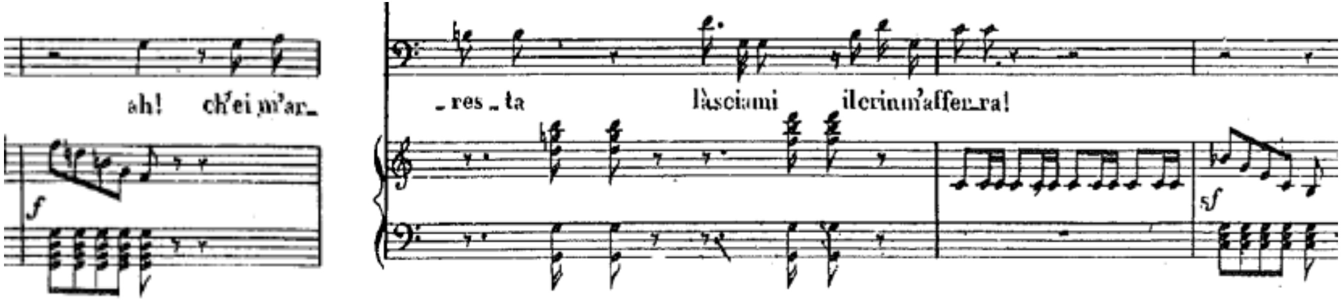
Frédéric Chopin - Mazurka - Op. 30 Nr. 2 - h-moll (b minor ...

Mazurka op. 30 n. 2 - Chopin - Alessandro Vena - YouTube

Frederyk Chopin Mazurka Op. 30 No. 2 - YouTube

Michelangeli Chopin Mazurka Op 30 No. 2 - YouTube

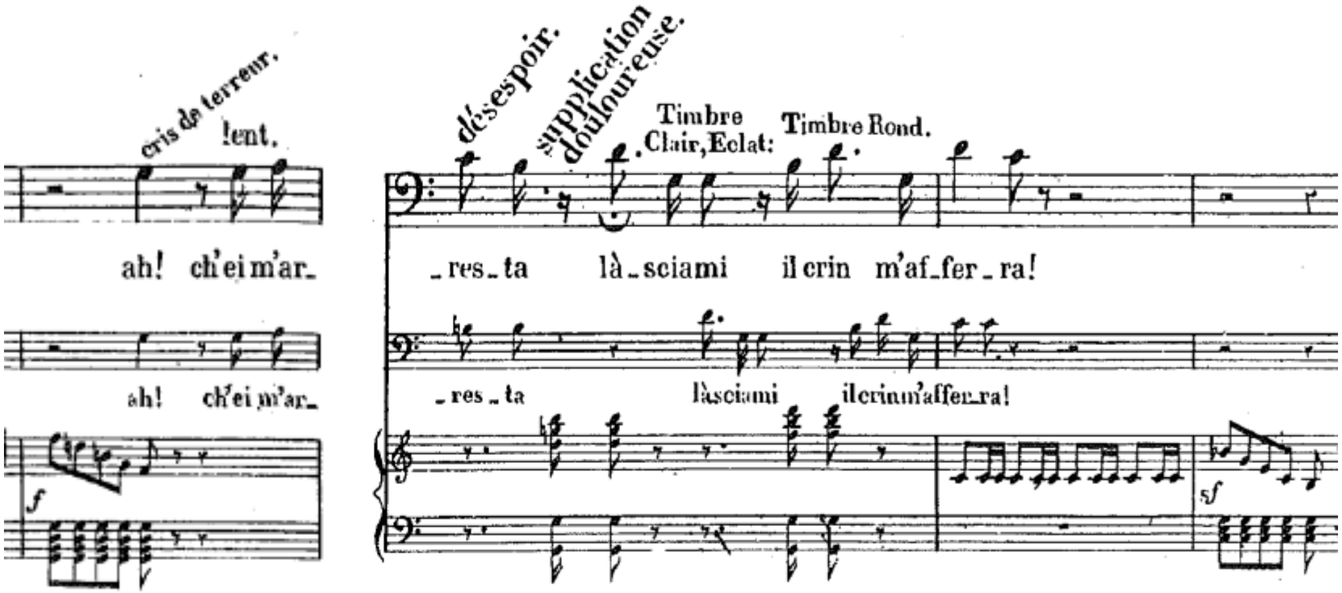
Expectation?



École de Garcia: Traité complet de l'art du chant, 11th edition, 1901
(1st edition: 1840/1847)

Reality!

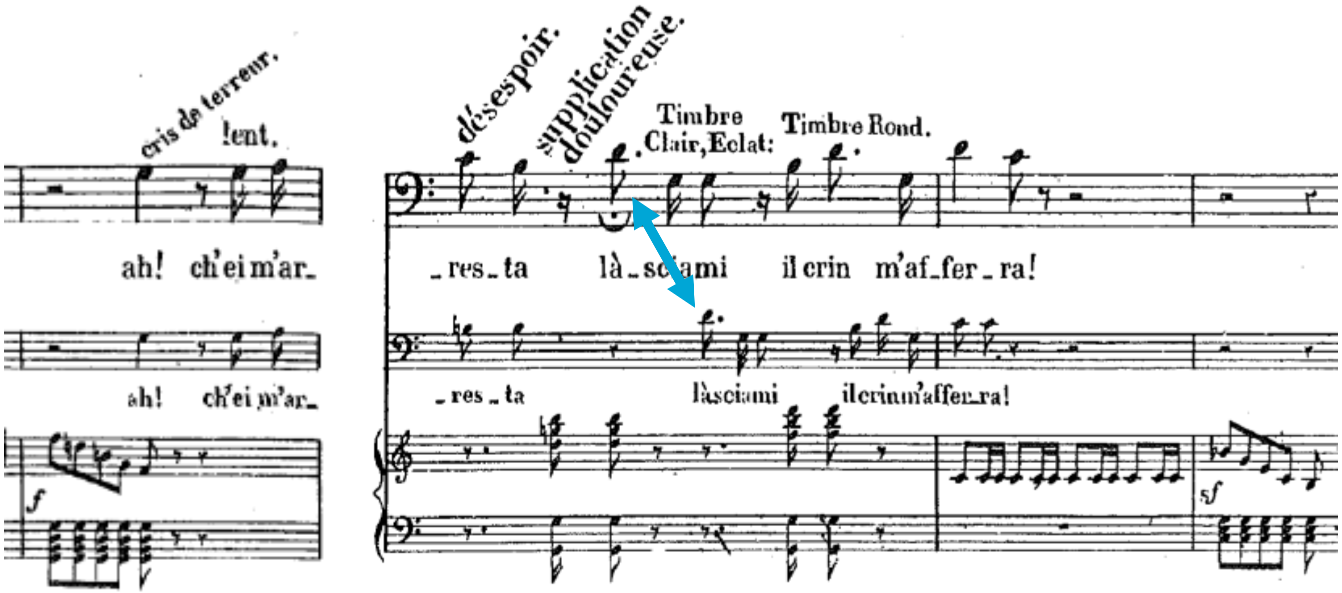
- You weren’t always expected to strictly follow the notes 😱


cris de terreur.
lent.
désespoir.
supplication douloureuse.
Timbre Clair, Eclat:
Timbre Rond.
ah! ch'ei m'ar_
_res_ta là_sciami il crin m'af_fer_ra!
ah! ch'ei m'ar_
_res_ta làsciami ilcrinm'affer_ra!
f
sf

École de Garcia: Traité complet de l’art du chant, 11th edition, 1901
(1st edition: 1840/1847)

Reality!

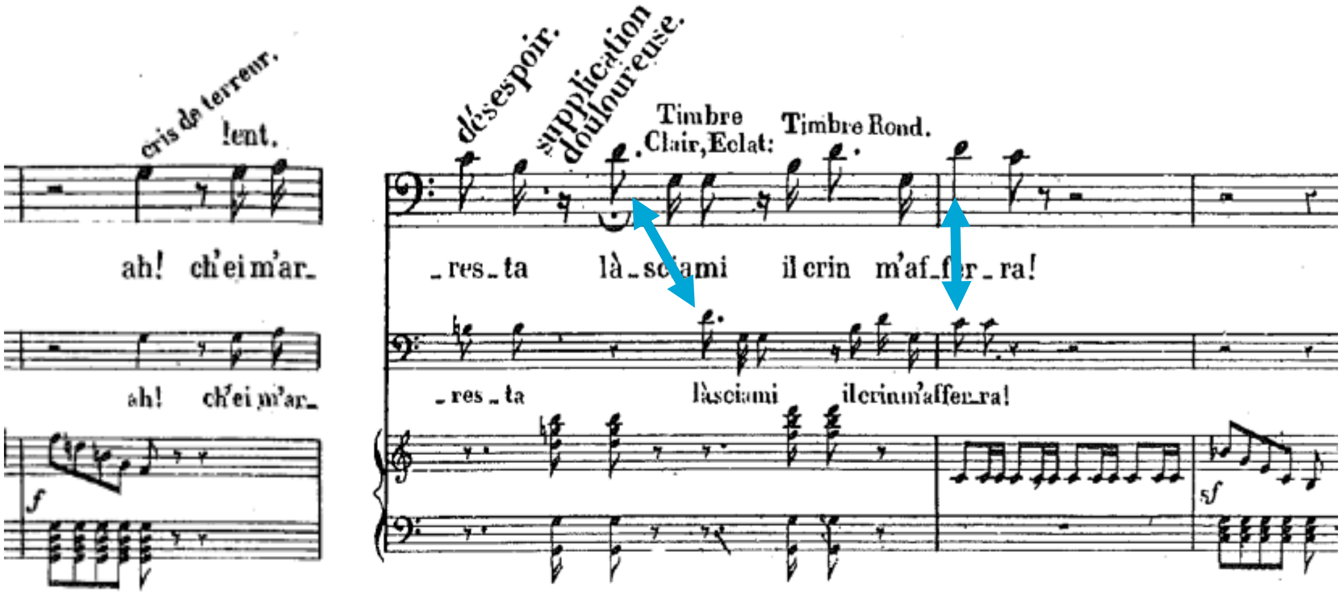
- You weren’t always expected to strictly follow the notes 😱



École de Garcia: Traité complet de l’art du chant, 11th edition, 1901
(1st edition: 1840/1847)

Reality!

- You weren’t always expected to strictly follow the notes 😱



cris de terreur.
lent.
désespoir.
supplication douloureuse.
Timbre Clair, Eclat:
Timbre Rond.
ah! ch'ei m'ar_
_res_ta là_sciami il crin m'af_fer_ra!
ah! ch'ei m'ar_
_res_ta làsciami ilcrinm'affer_ra!

École de Garcia: Traité complet de l’art du chant, 11th edition, 1901
(1st edition: 1840/1847)

Research interest 1: description

- Represent relevant information in multimedia objects more holistically and comprehensively

- Multiple parallel modalities (different signal domains)
- Multiple parallel perspectives (labels & ‘anomalies’)

- ‘Vague’ (but grounded) human concepts to be translated to mathematical frameworks. What can possibly go wrong?

So we wanted to perform lesser-known works…

- Concert halls: no way! That won’t sell!
- Gain visibility, change presentation strategies, champion
- (this is risky and expensive)

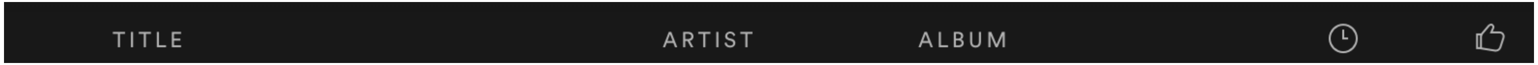
Research interest 2: exploration

- Make objects findable and retrievable, especially in cases they are not ‘on the radar’

- Algorithmic filtering: learn from behavioral data, but don’t necessarily literally re-predict it
  - ‘I truly liked this’ vs. ‘I clicked on it’
  - ‘no way’ vs. ‘this may work’

- User factors: present in accessible ways
- ‘Risky’ items need more effort. Find & create contexts in which this is acceptable and appreciated

Under-representation in music



TITLE ARTIST ALBUM 🕒 👍

- Title-artist-album ontology: library system for pop

- Many classical works do not map well into this
- Neither do works in the genre ‘world music’

- ‘Market is too small to fix this’
  - But if users won’t have a means to engage, accessibility is hampered and no interactions will be evidenced → self-fulfilling prophecy

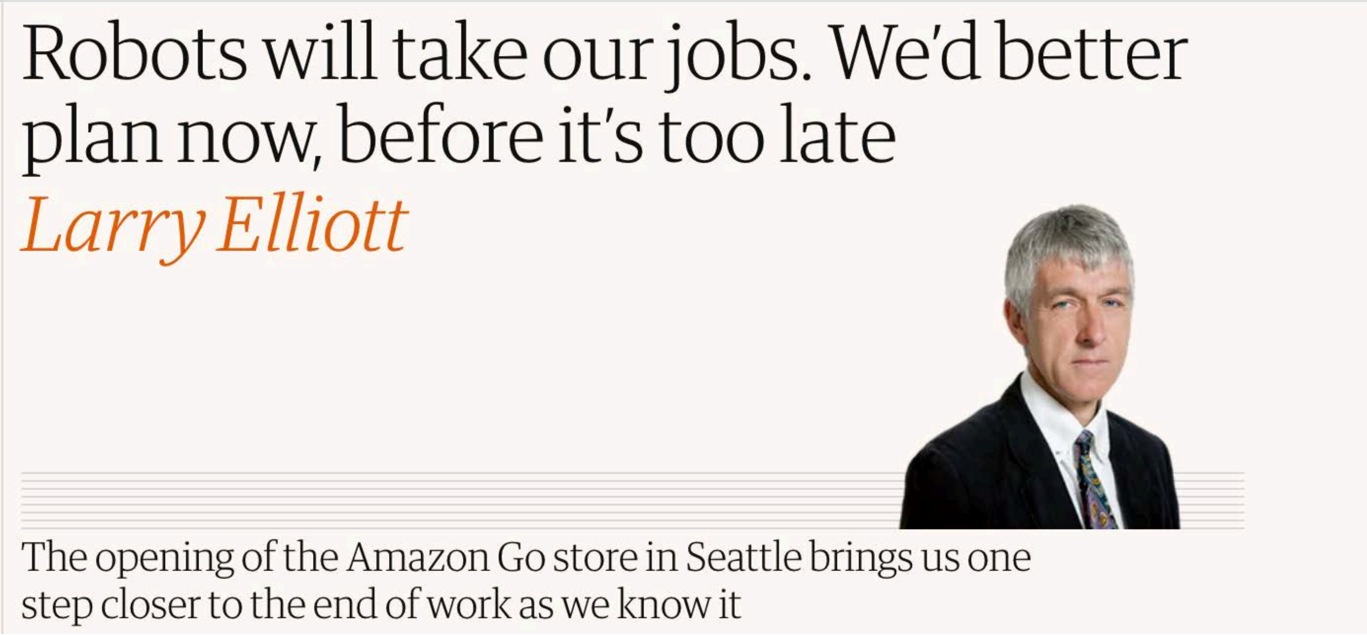
Use cases in job candidate screening

Big Data in Psychological Assessment

Erasmus+

Co-funded by the Erasmus+ Programme of the European Union

The future of work



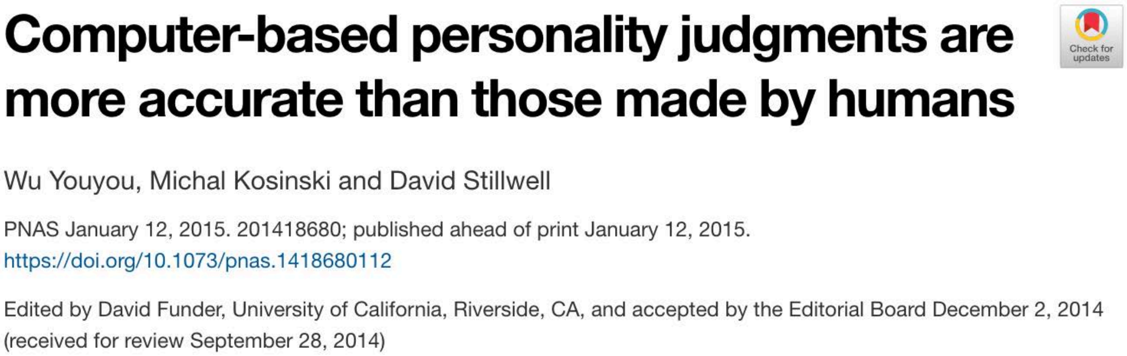
Robots will take our jobs. We'd better plan now, before it's too late

Larry Elliott

The opening of the Amazon Go store in Seattle brings us one step closer to the end of work as we know it

The digital promise

TIPS TO PREPARE AN IDEAL VIDEO RESUME TO GET YOU HIRED EASILY IN 2018
1/50
F1.8
±0.0
ISO AUTO



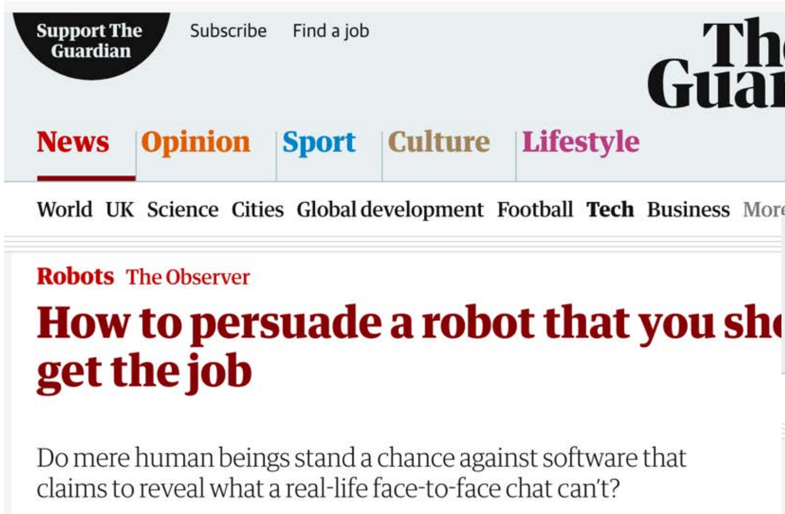
Computer-based personality judgments are more accurate than those made by humans

Wu Youyou, Michal Kosinski and David Stillwell

PNAS January 12, 2015. 201418680; published ahead of print January 12, 2015.
https://doi.org/10.1073/pnas.1418680112

Edited by David Funder, University of California, Riverside, CA, and accepted by the Editorial Board December 2, 2014 (received for review September 28, 2014)

The future of getting work



Support The Guardian

Subscribe Find a job

The Guardian

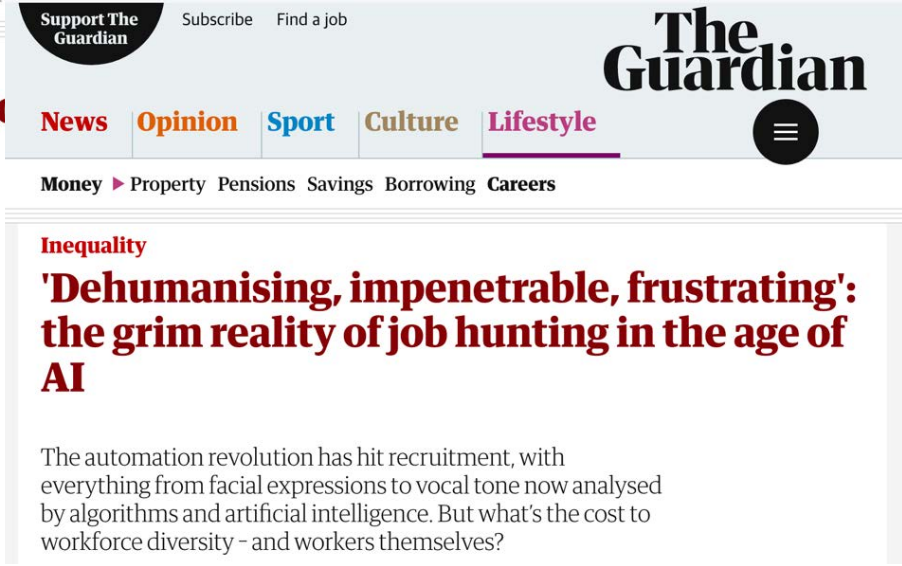
News Opinion Sport Culture Lifestyle

World UK Science Cities Global development Football Tech Business More

Robots The Observer

How to persuade a robot that you sh
get the job

Do mere human beings stand a chance against software that claims to reveal what a real-life face-to-face chat can't?



Support The Guardian

Subscribe Find a job

The Guardian

News Opinion Sport Culture Lifestyle

Money ▸ Property Pensions Savings Borrowing Careers

Inequality

'Dehumanising, impenetrable, frustrating': the grim reality of job hunting in the age of AI

The automation revolution has hit recruitment, with everything from facial expressions to vocal tone now analysed by algorithms and artificial intelligence. But what's the cost to workforce diversity - and workers themselves?

The big questions

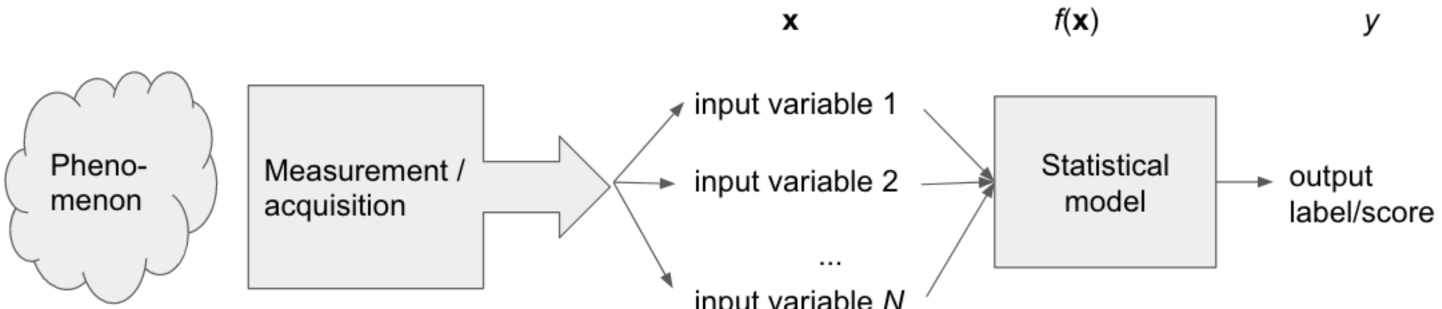
- In a digitized, data-rich world…

- …what novel skills do workers and hiring specialists need?
- …how can/should data-driven analysis methods and technological interventions be integrated in candidate screening?
- …what are major ethical risks?

A fundamental misunderstanding

[Liem et al., 2018]

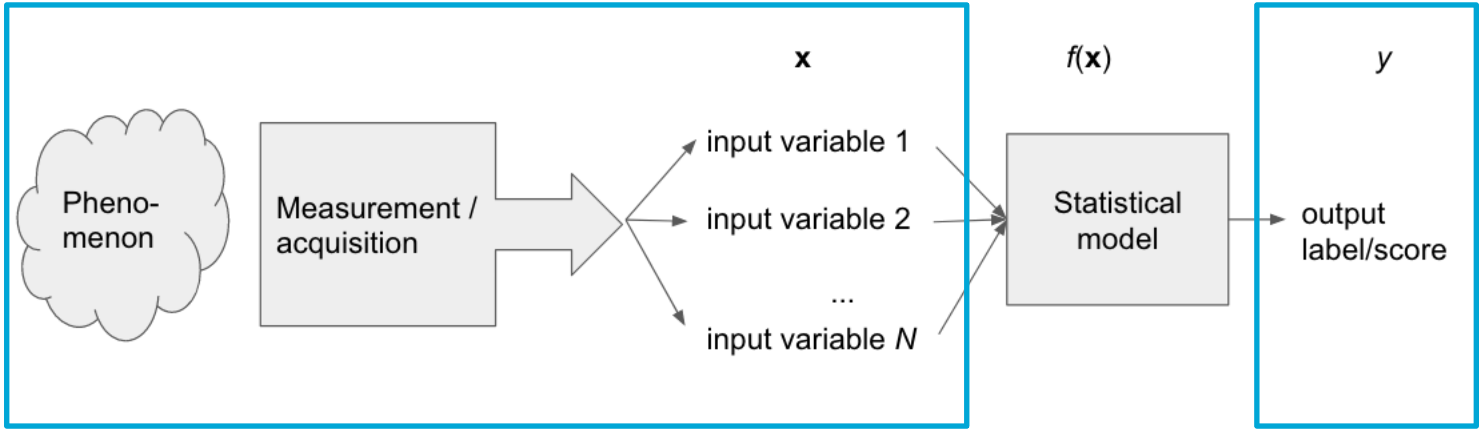
A common pipeline



Different focus areas, different perceptions of success

[Liem et al., 2018]

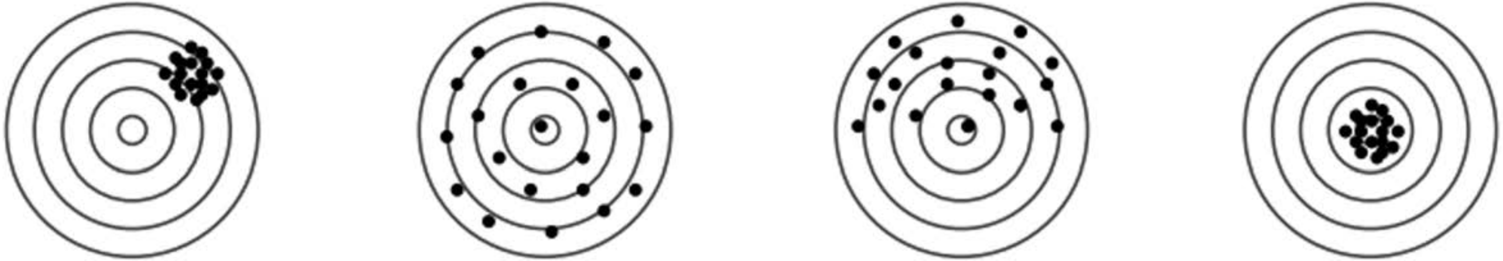
- A psychologist normally focuses on measuring and understanding **x** (and possibly *y*)



[Liem et al., 2018]

Psychometrics

- *Constructs* are not directly measurable, how can we trust them?
- *Instruments* need **validity** and **reliability**

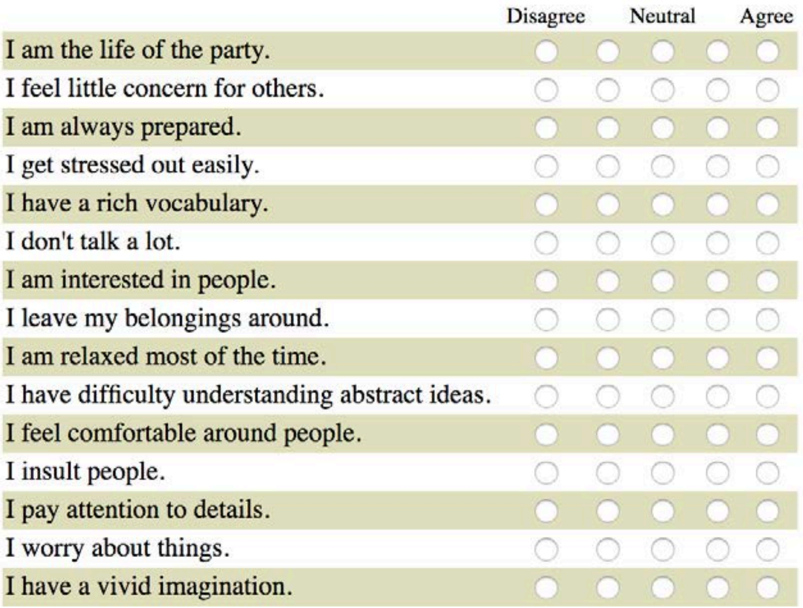


Not Valid but Reliable   Valid but Not Reliable   Neither Valid Nor Reliable   Both Valid and Reliable

[Liem et al., 2018]

Big Five (‘OCEAN’)

- Openness
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism
- Valid & reliable instruments exist



	Disagree		Neutral		Agree
I am the life of the party.	○	○	○	○	○
I feel little concern for others.	○	○	○	○	○
I am always prepared.	○	○	○	○	○
I get stressed out easily.	○	○	○	○	○
I have a rich vocabulary.	○	○	○	○	○
I don't talk a lot.	○	○	○	○	○
I am interested in people.	○	○	○	○	○
I leave my belongings around.	○	○	○	○	○
I am relaxed most of the time.	○	○	○	○	○
I have difficulty understanding abstract ideas.	○	○	○	○	○
I feel comfortable around people.	○	○	○	○	○
I insult people.	○	○	○	○	○
I pay attention to details.	○	○	○	○	○
I worry about things.	○	○	○	○	○
I have a vivid imagination.	○	○	○	○	○

[Liem et al., 2018]

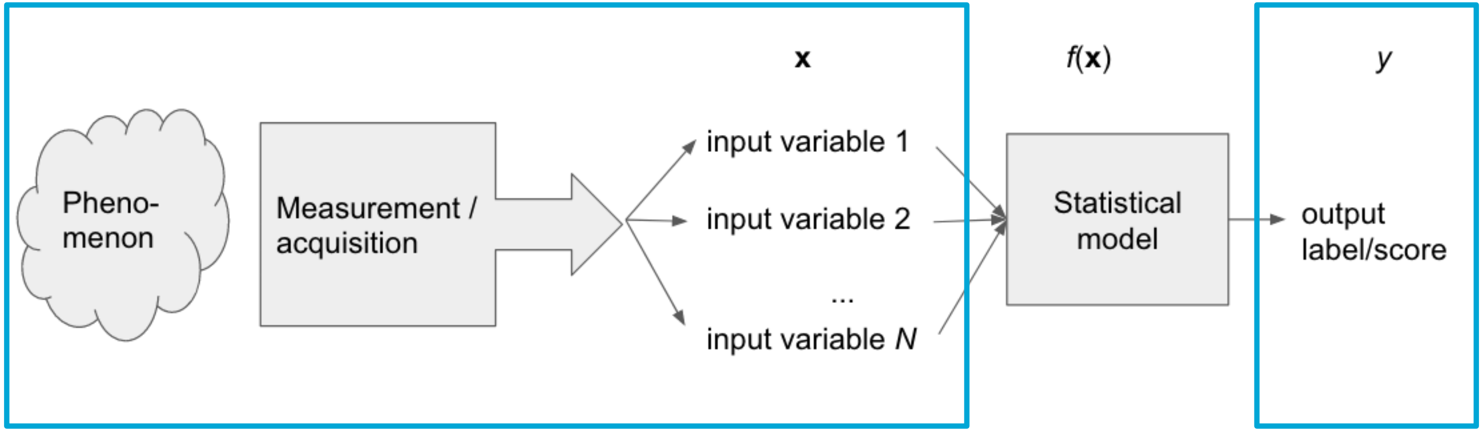
Myers-Briggs

- No valid & reliable instruments exist
- Yet, extremely popular, both in HR and social media

Different focus areas, different perceptions of success

[Liem et al., 2018]

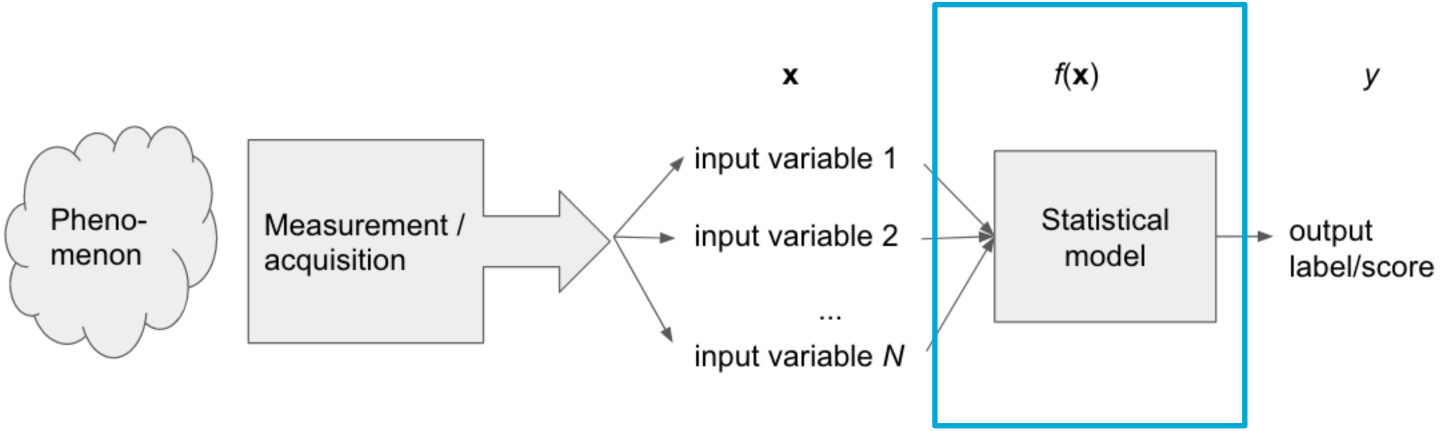
- A psychologist normally focuses on measuring and understanding **x** (and possibly *y*)



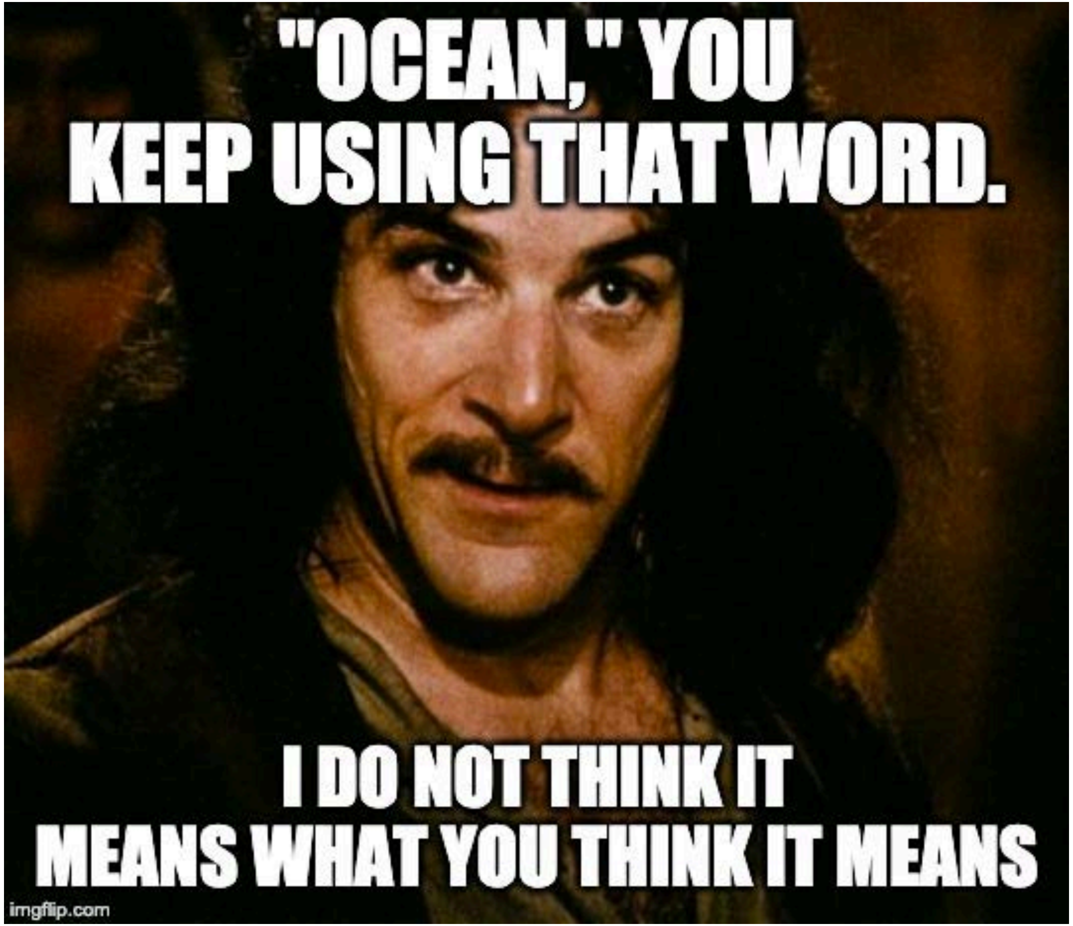
Different focus areas, different perceptions of success

[Liem et al., 2018]

- A machine learning expert normally focuses on optimizing *f*(**x**)
- The data is the responsibility of the domain expert



[Liem et al., 2018]


"OCEAN," YOU KEEP USING THAT WORD.
I DO NOT THINK IT MEANS WHAT YOU THINK IT MEANS
imgflip.com

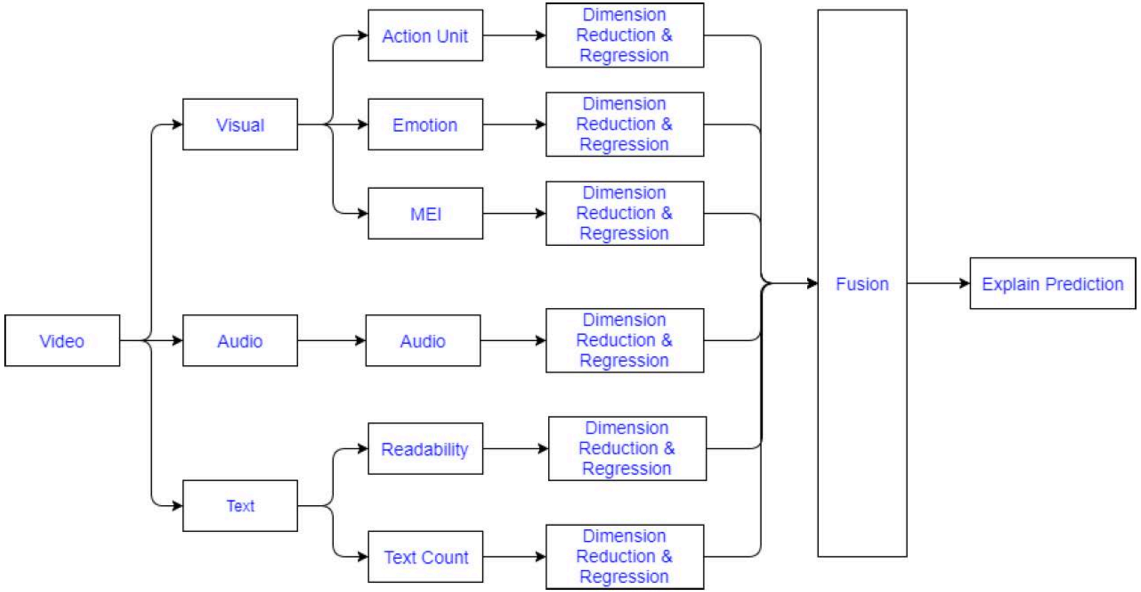
ChaLearn 'Looking at People'

- Driven by the Computer Vision community
- First impressions dataset v2:
  - 10,000 15-sec. vlog excerpts from YouTube
  - Transcriptions of speech
  - OCEAN & invite-to-interview labels
  - http://chalearnlap.cvc.uab.es/dataset/24/description/
- Qualitative & Quantitative challenges
  - http://chalearnlap.cvc.uab.es/challenge/23/description/

[Liem et al., 2018]

Quantitative performance

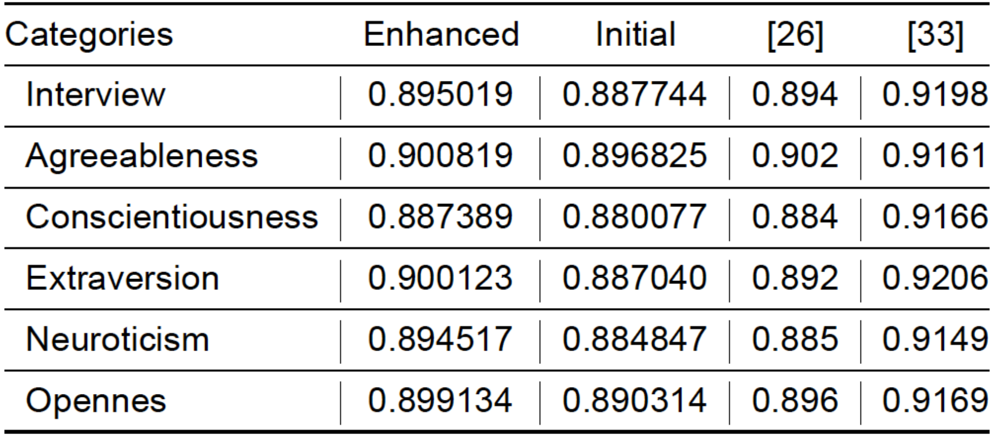
- My colleagues wanted to explicitly understand the data
- ‘Old-fashioned’ feature engineering: deep features would be meaningless


Video
Visual
Audio
Text
Action Unit
Emotion
MEI
Audio
Readability
Text Count
Dimension Reduction & Regression
Dimension Reduction & Regression
Dimension Reduction & Regression
Dimension Reduction & Regression
Dimension Reduction & Regression
Dimension Reduction & Regression
Fusion
Explain Prediction

[Liem et al., 2018]

Quantitative performance

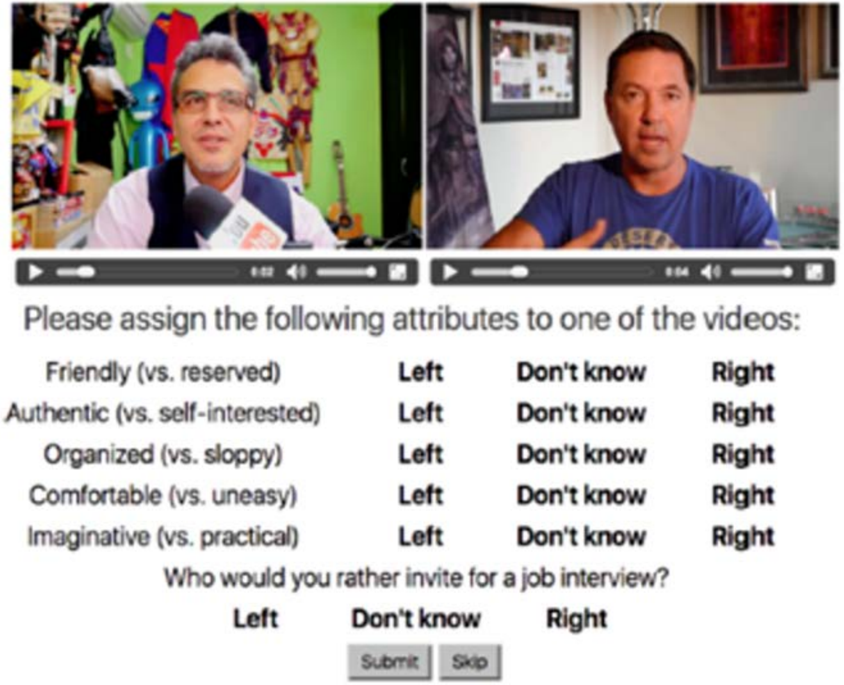
- My colleagues wanted to explicitly understand the data
- ‘Old-fashioned’ feature engineering: deep features would be meaningless



Categories	Enhanced	Initial	[26]	[33]
Interview	0.895019	0.887744	0.894	0.9198
Agreeableness	0.900819	0.896825	0.902	0.9161
Conscientiousness	0.887389	0.880077	0.884	0.9166
Extraversion	0.900123	0.887040	0.892	0.9206
Neuroticism	0.894517	0.884847	0.885	0.9149
Opennes	0.899134	0.890314	0.896	0.9169

[Liem et al., 2018]

Crowdsourced single-item scores



Please assign the following attributes to one of the videos:

Friendly (vs. reserved)	Left	Don't know	Right
Authentic (vs. self-interested)	Left	Don't know	Right
Organized (vs. sloppy)	Left	Don't know	Right
Comfortable (vs. uneasy)	Left	Don't know	Right
Imaginative (vs. practical)	Left	Don't know	Right

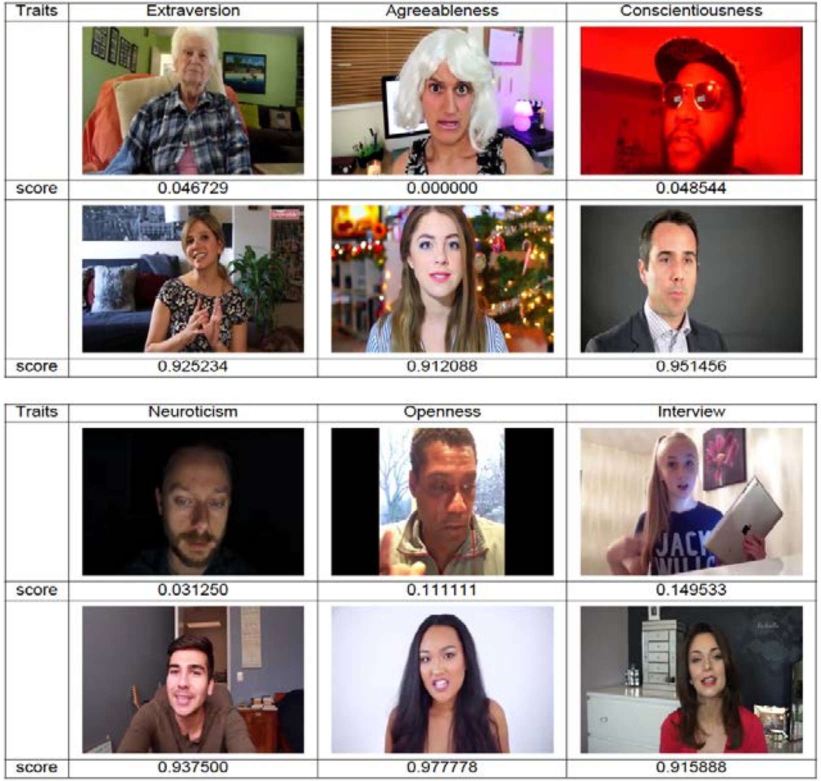
Who would you rather invite for a job interview?

Left Don't know Right

[Ponce-López et al., 2016]

[Liem et al., 2018]

Score maxima & minima



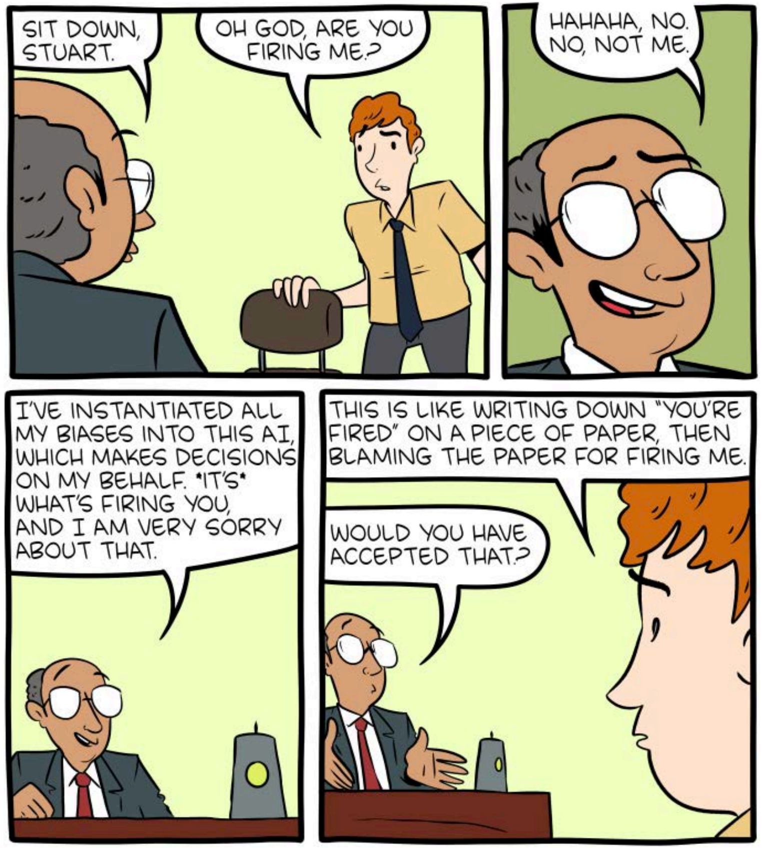
Traits	Extraversion	Agreeableness	Conscientiousness
score	0.046729	0.000000	0.048544
score	0.925234	0.912088	0.951456

Traits	Neuroticism	Openness	Interview
score	0.031250	0.111111	0.149533
score	0.937500	0.977778	0.915888

ML wasn’t a ‘solution’ in this case

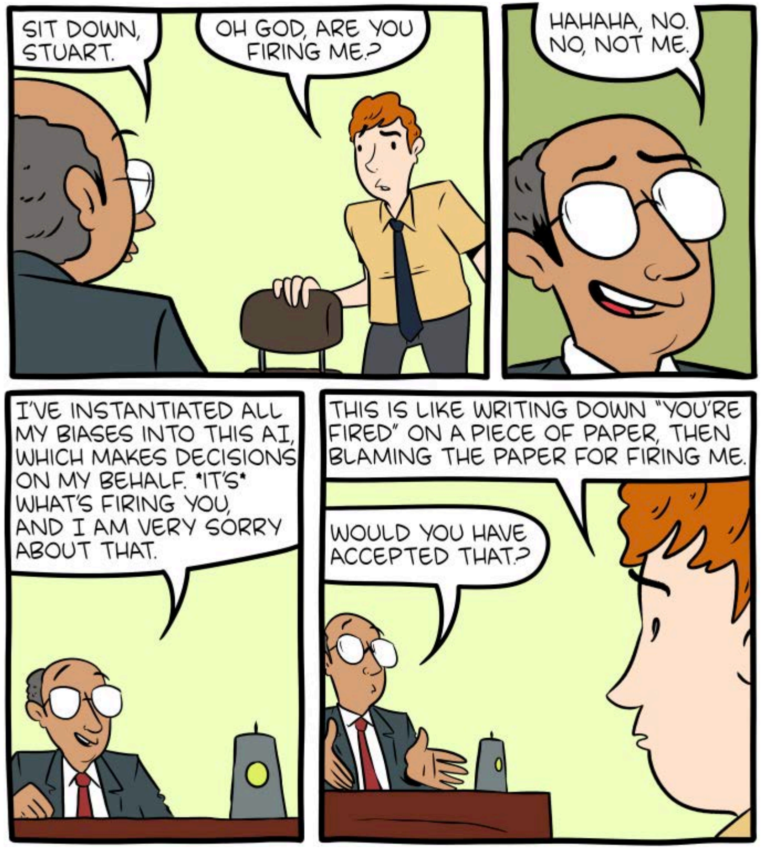
- My colleagues couldn’t handle the scale and complexity of multimedia input data. ML—when designed consciously—provided a useful tool
- But researching what could make for a better, interpretable **x** and *y* were the main interests
- If these are both unclear, you won’t gain insights by choosing a stronger *f*(**x**)
- I rather think my colleagues needed human-in-the-loop support to better reflect on their problem case

Let’s not be like this


SIT DOWN, STUART.
OH GOD, ARE YOU FIRING ME?
HAHAHA, NO. NO, NOT ME.
I'VE INSTANTIATED ALL MY BIASES INTO THIS AI, WHICH MAKES DECISIONS ON MY BEHALF. *IT'S* WHAT'S FIRING YOU, AND I AM VERY SORRY ABOUT THAT.
THIS IS LIKE WRITING DOWN "YOU'RE FIRED" ON A PIECE OF PAPER, THEN BLAMING THE PAPER FOR FIRING ME.
WOULD YOU HAVE ACCEPTED THAT?

https://www.smbc-comics.com/comic/ai-4
presented with permission by creator Zach Weinersmith

Let’s not be like this

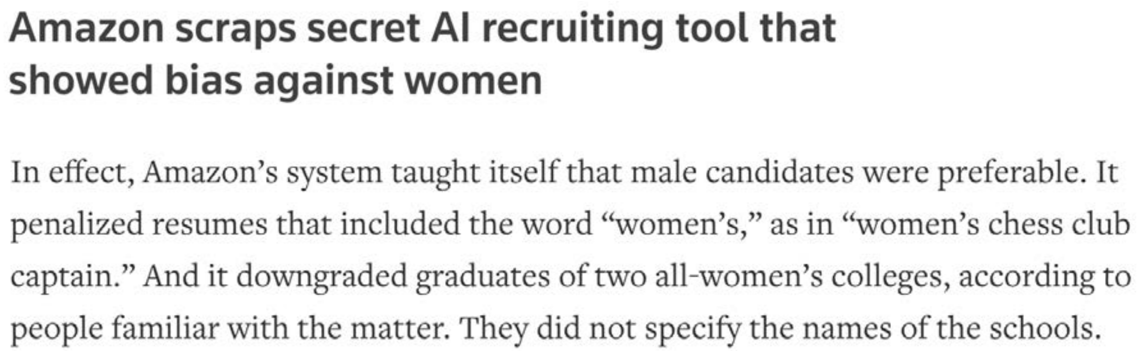

SIT DOWN, STUART.
OH GOD, ARE YOU FIRING ME?
HAHAHA, NO. NO, NOT ME.
I'VE INSTANTIATED ALL MY BIASES INTO THIS AI, WHICH MAKES DECISIONS ON MY BEHALF. *IT'S* WHAT'S FIRING YOU, AND I AM VERY SORRY ABOUT THAT.
THIS IS LIKE WRITING DOWN "YOU'RE FIRED" ON A PIECE OF PAPER, THEN BLAMING THE PAPER FOR FIRING ME.
WOULD YOU HAVE ACCEPTED THAT?

https://www.smbc-comics.com/comic/ai-4
presented with permission by creator Zach Weinersmith

- ML specialists tend to believe all information is in ‘the data’
- Academic narrative bias: my model is better than yours (watch the COVID-19 discussions…)
- Non-ML specialists tend to believe that ‘AI’ can help fixing problems they do not fully understand
- We should be careful with:
  - the ‘superhuman’ narrative
  - providing a ‘quick fix’
  - ‘outsourcing’ responsibility

Bias and fairness

- One of the big issues in hiring: handling & promoting diversity



Amazon scraps secret AI recruiting tool that showed bias against women

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Bias and fairness

- The minority group will not have been evidenced in historic data
- The majority group will have defined the historic example of 'what worked well'

- We can enforce 'more desired' balances between majority/minority groups
  - But optimization procedures are blind to 'meaningful minority' vs. noise: needs explicit human steering
  - You can't be fair to all. Under restricted resources, advantaging one group means disadvantaging the other
  - If the issue is systemic, it should be addressed at that level

Fairness is no fixed concept

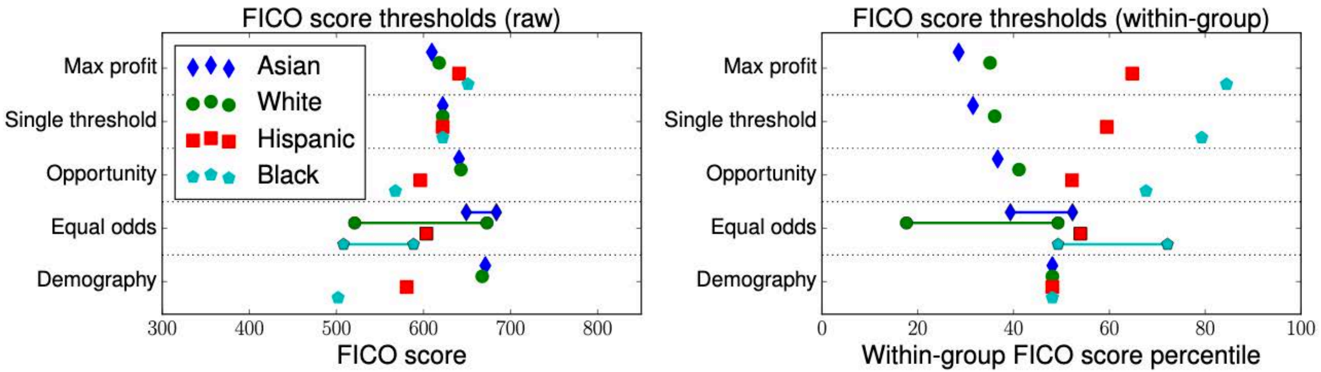
- Many (politically colored) definitions
  - see Arvind Narayanan’s tutorial linked below
- Within the same problem, different stakeholders will have different perceptions of what is fair
  - I don’t want to unrightfully be marked as a criminal (false positive)
  - Enforcers don’t want for too many true criminals to walk free (false negative)

Tutorial: 21 fairness definitions and their politics

https://www.youtube.com/watch?v=jIXIuYdnyyk

Trade-offs

- ‘Just’ blind the protected variable?
- Same treatment at individual or group level?
- True/False Positives/Negatives?



[Hardt et al., 2016]

So, no quick fixes. But there are things we can face

- Who are the stakeholders?
- What disagreements and trade-offs will happen?
- Does the data give room to alternative explanations?
- Should historic data (not) be replicated?

- Do we seek fairness?
  - Or rather accountability / explainability / transparency on decisions that necessarily will be controversial?
- Decision support rather than accuracy optimization?

Why fairness can’t (and shouldn’t) be ‘solved’ by machine learning

Cynthia C. S. Liem

c.c.s.liem@tudelft.nl | @informusiccs

Multimedia Computing Group

Delft University of Technology