Working as intended?
Perceptions of success in AI Systems

Cynthia C. S. Liem

Multimedia Computing Group

Delft University of Technology

About me

Foto: Marco Borggreve

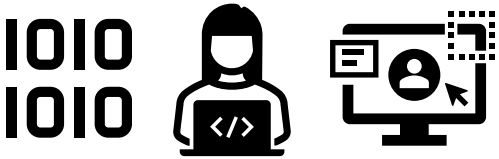
Foto: Marcel Krijger

https://www.trouw.nl/politiek/hoe-de-belastingdienst-lage-inkomens-profileerde-in-de-jacht-op-fraude~bbb66add/

Methodological schools of thought

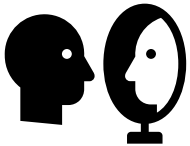
Design

Science



The computer scientist

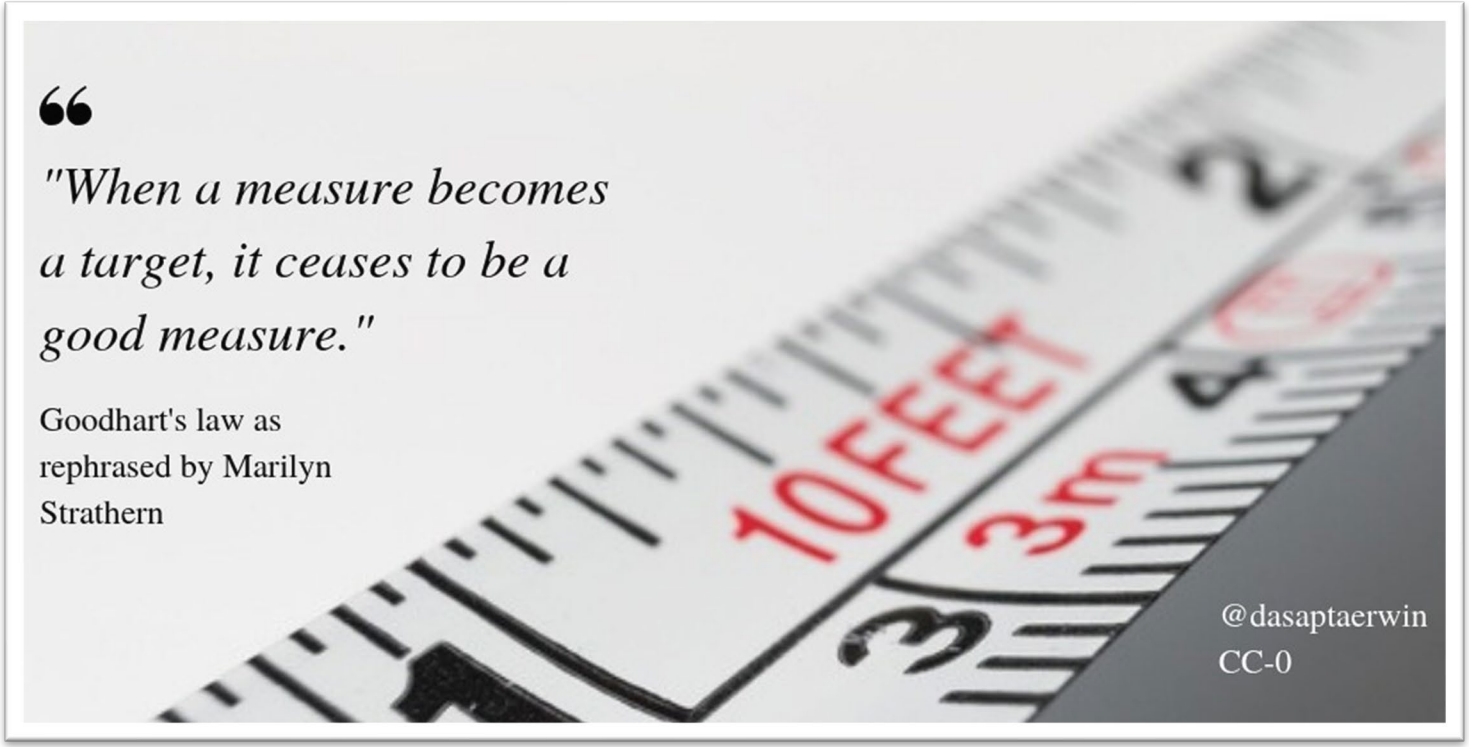
Engineering



Social sciences / humanities

Arts

A world seeking to measure



Babylonic confusion

Algorithms

Exercise

- Write down your favorite recipe

A classical algorithmic problem: sorting

- Can be done in stupid and smart ways
- What is efficient for large numbers?
- What is applicable to any problem in which one can agree on a way of ordering?

Corresponding values

- Efficiency
- Generalization
- Scaling up
- Using the exact same procedure

Corresponding values

- Efficiency
- Generalization
- Scaling up
- Using the exact same procedure

- This may turn people into numbers

Corresponding values

- Efficiency
- Generalization
- Scaling up
- Using the exact same procedure

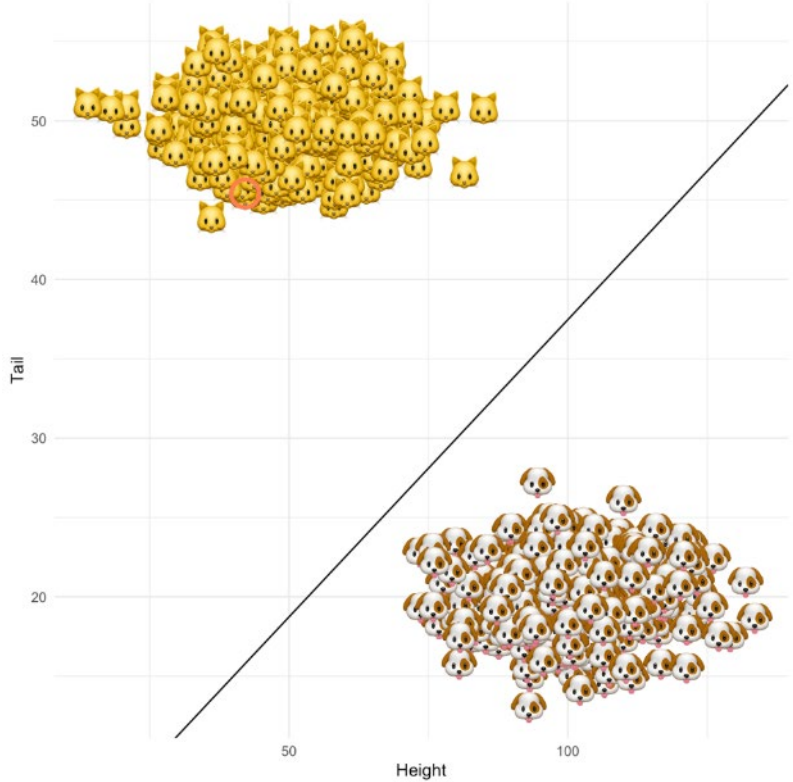
- They may be useful ‘mirrors’ to our thinking

Being precise is difficult!

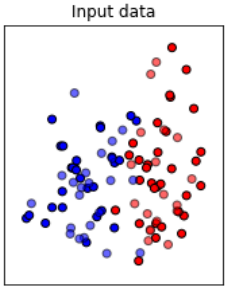
https://www.youtube.com/watch?v=cDA3_5982h8

Machine learning

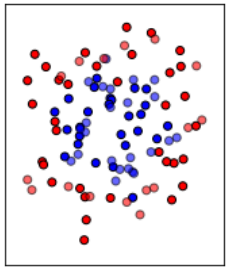
Learning patterns from data and labels

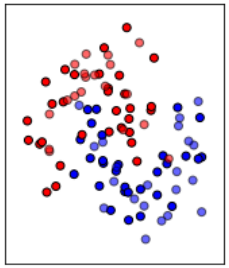


Learning patterns from data and labels

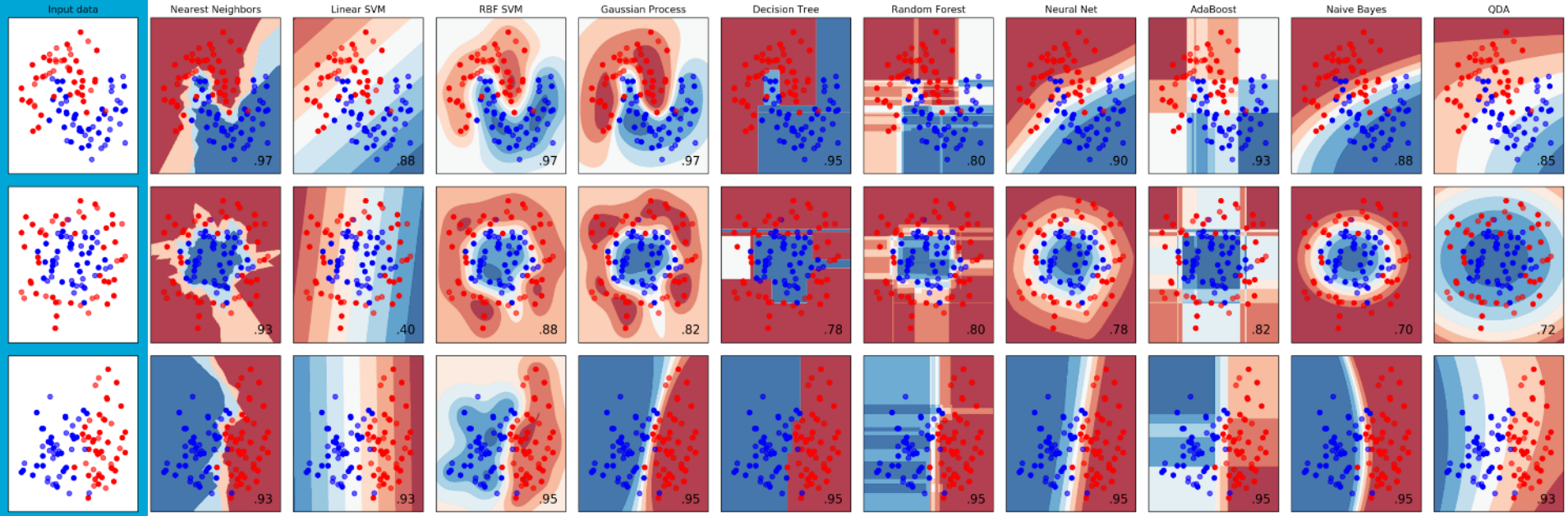


Learning patterns from data and labels

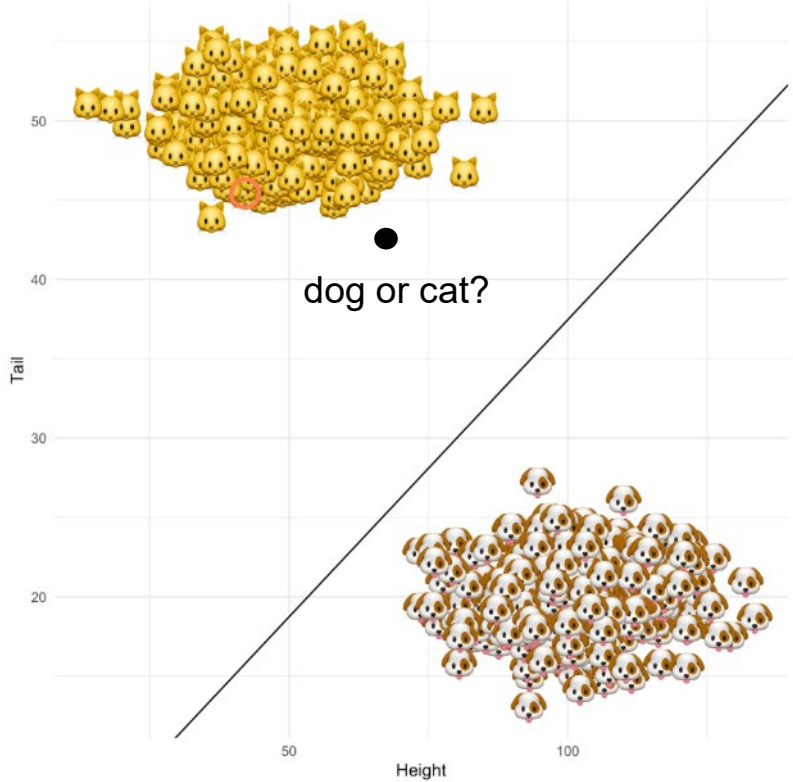




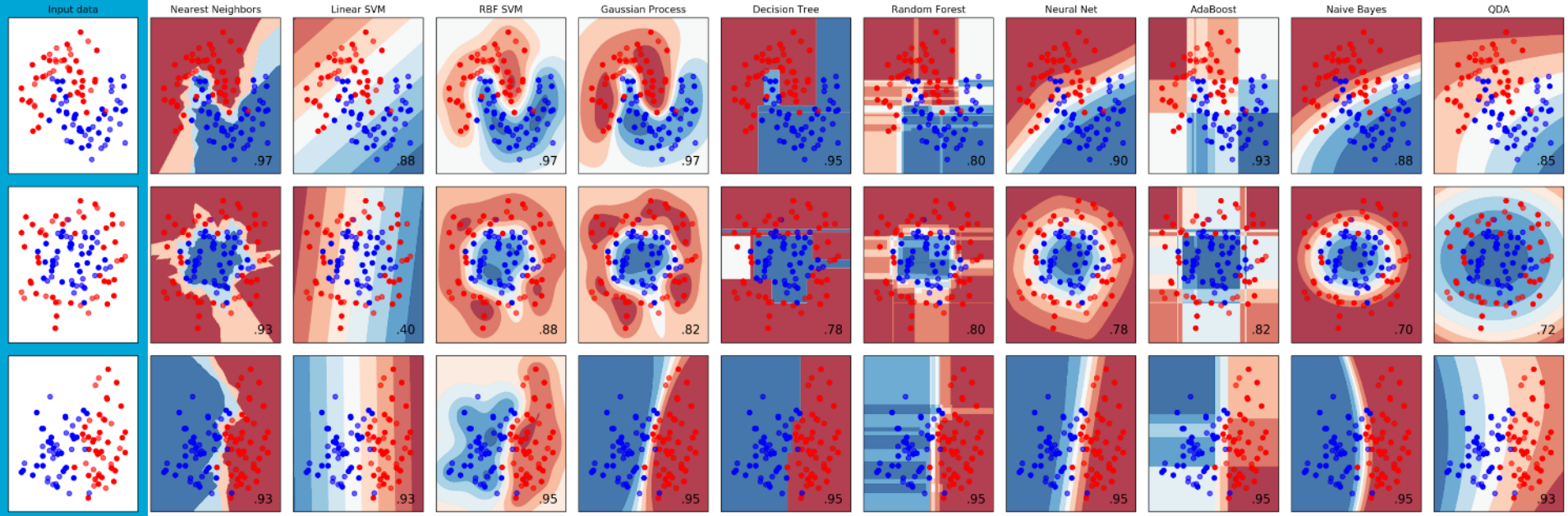
How we like to see the world



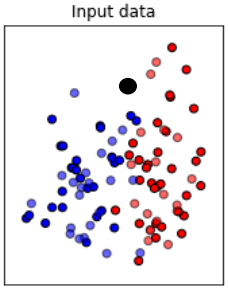
Predicting unseen data points



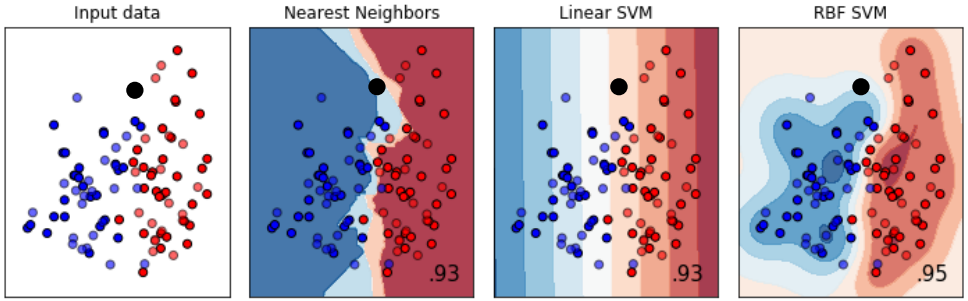
Prediction or priorities?



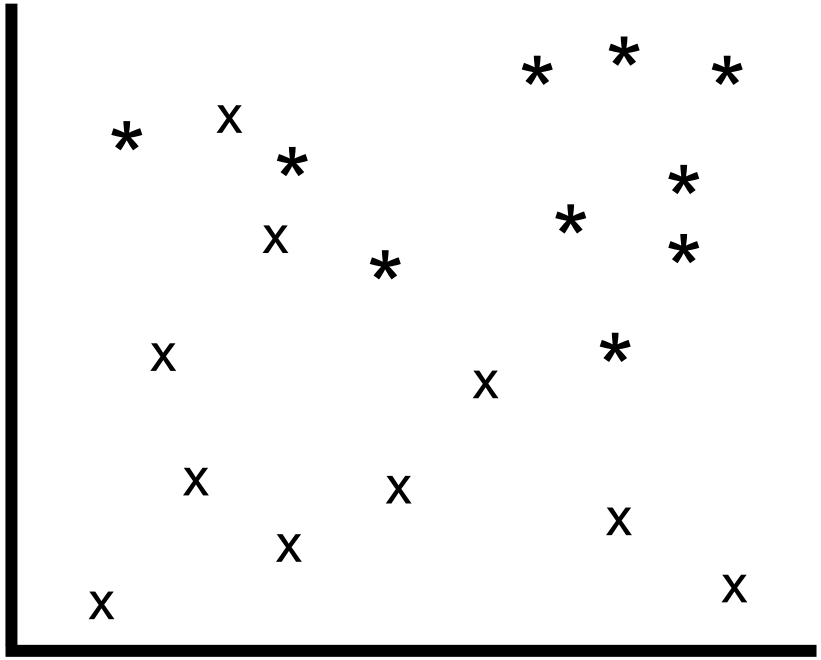
Prediction or priorities?



Prediction or priorities?

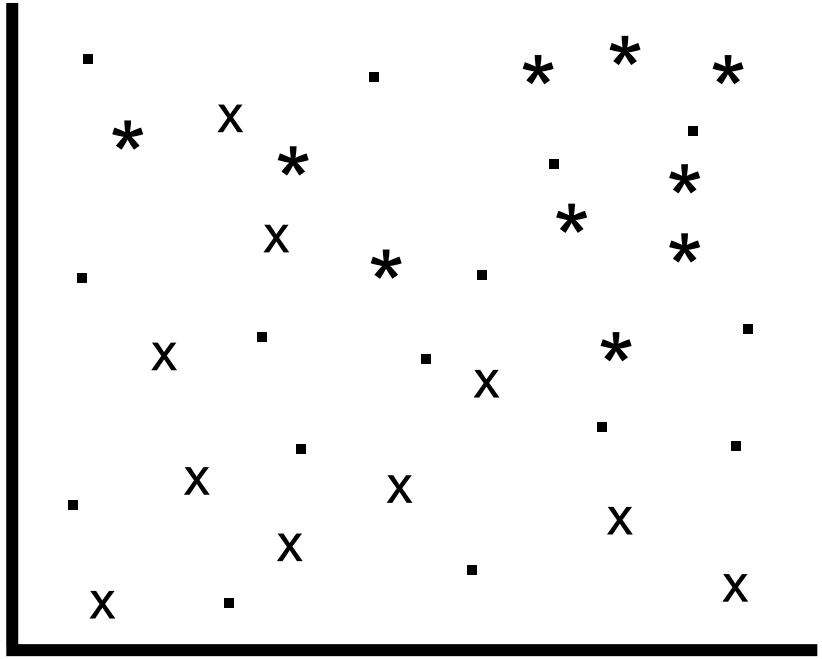


We may find patterns in datasets



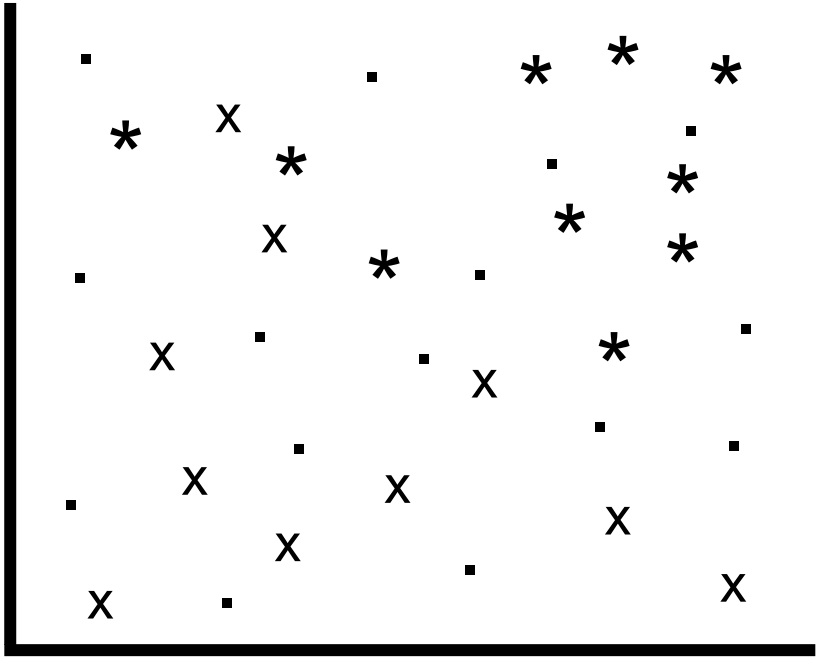
And often have limited resources

- Choose 5 data points



Who gets prioritized?

- Choose 5 (fictional) human beings

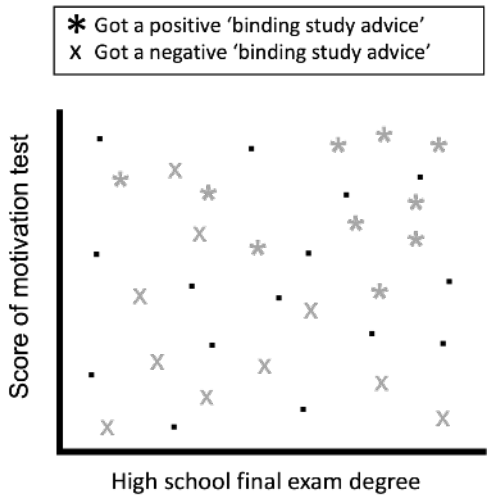


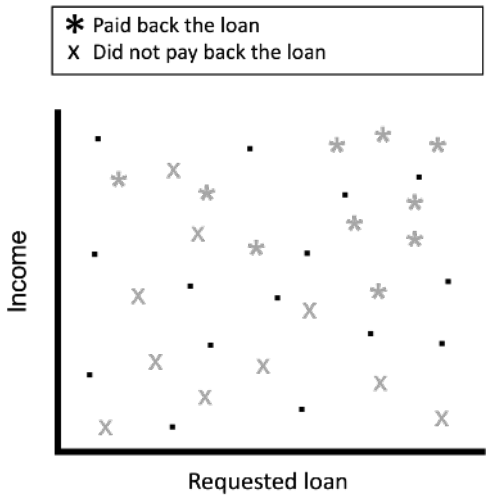
Who gets prioritized?

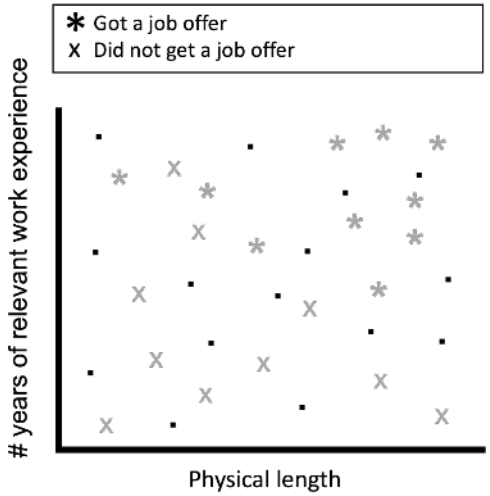
- Choose 5 (fictional) human beings

https://tudelft.fra1.qualtrics.com/jfe/form/SV_6VuOKlmn7QgDt5A

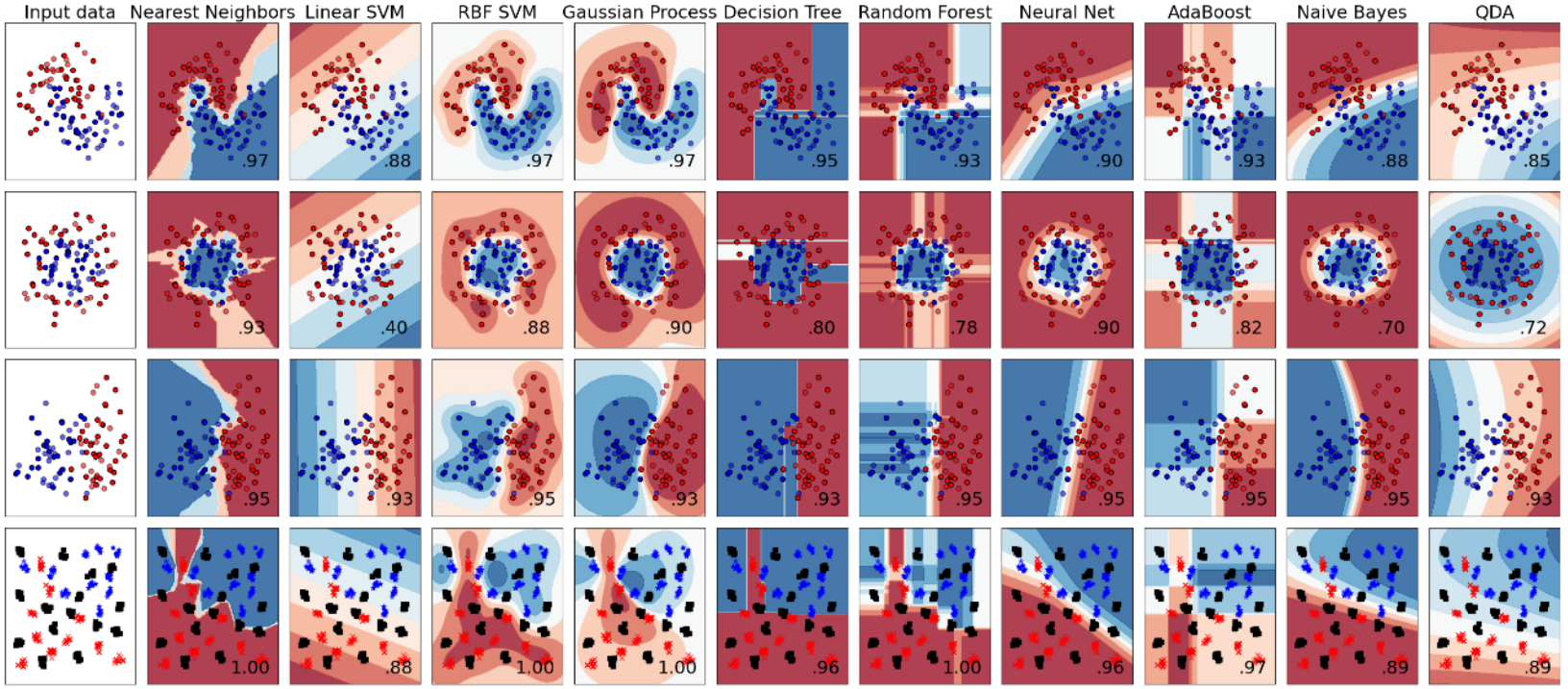
Same or different?



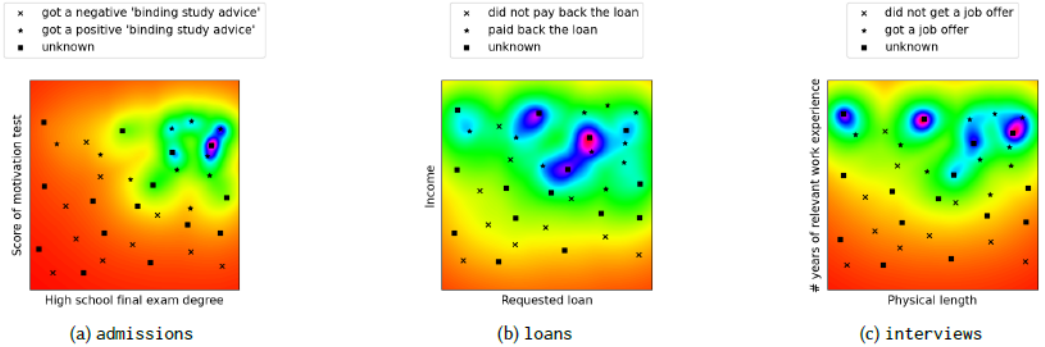




Classifiers would not distinguish…

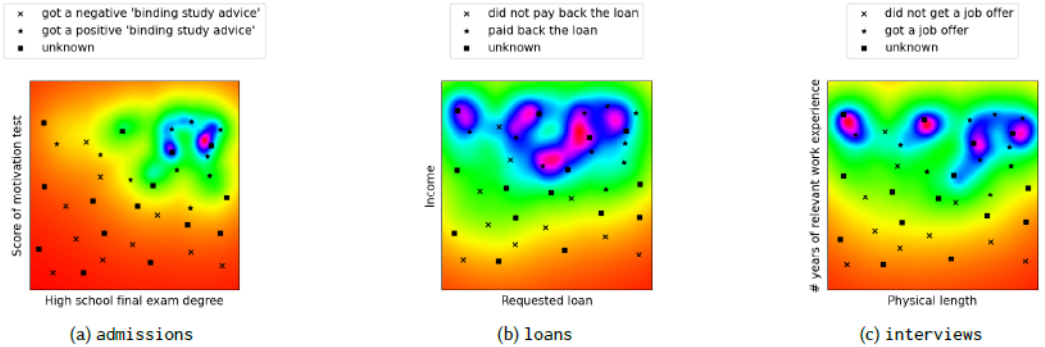


Humans actually do!



(a) admissions (b) loans (c) interviews

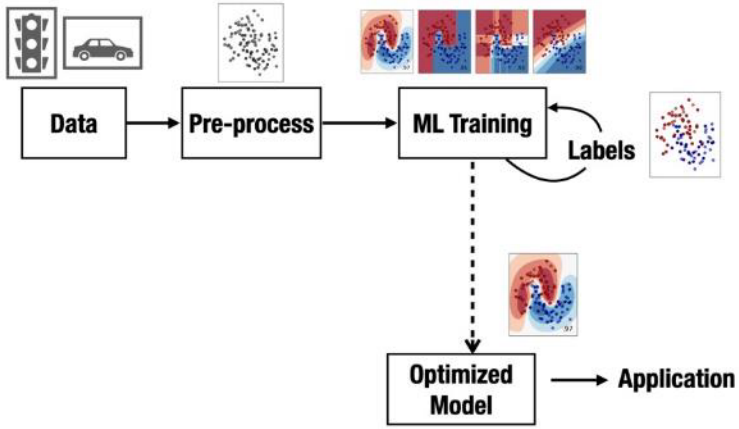
Fig. 6. Aggregated click heatmaps from audiences highly likely to have technical machine learning experience (python, dataScience, CSstudents, testingStudents and AICongress combined).



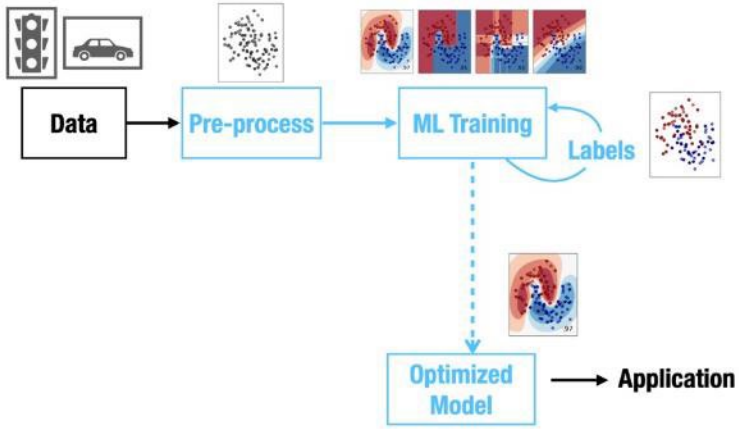
(a) admissions (b) loans (c) interviews

Fig. 7. Aggregated click heatmaps from audiences highly unlikely to have technical machine learning experience (rotary, policy, librarians, interim, professionalEdu, privacyCongress and ruleOfLaw combined).

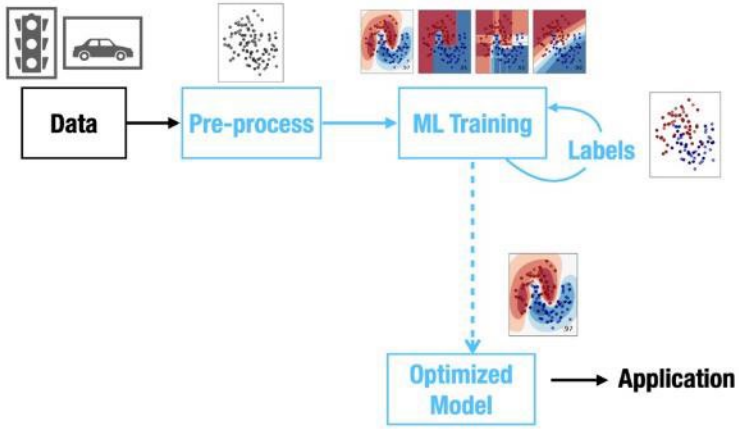
Applying machine learning



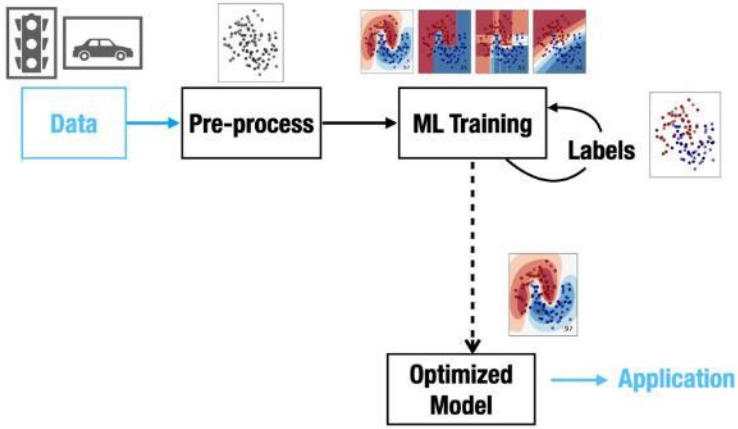
Our usual focus



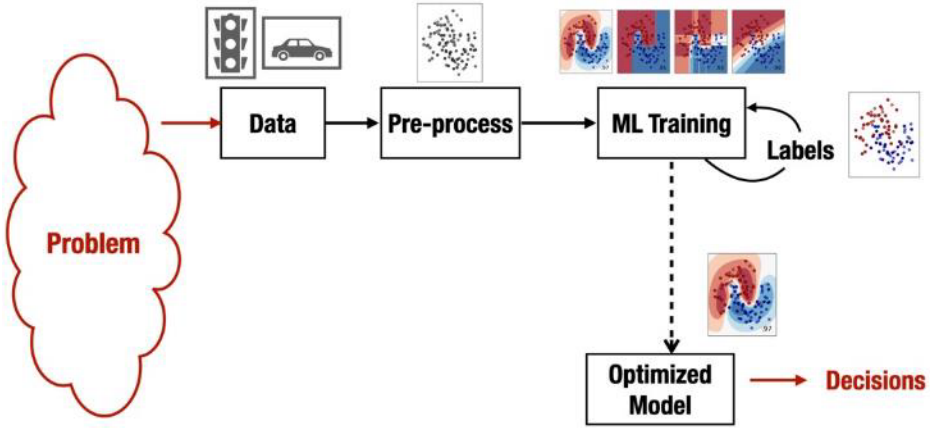
This focus often is too narrow



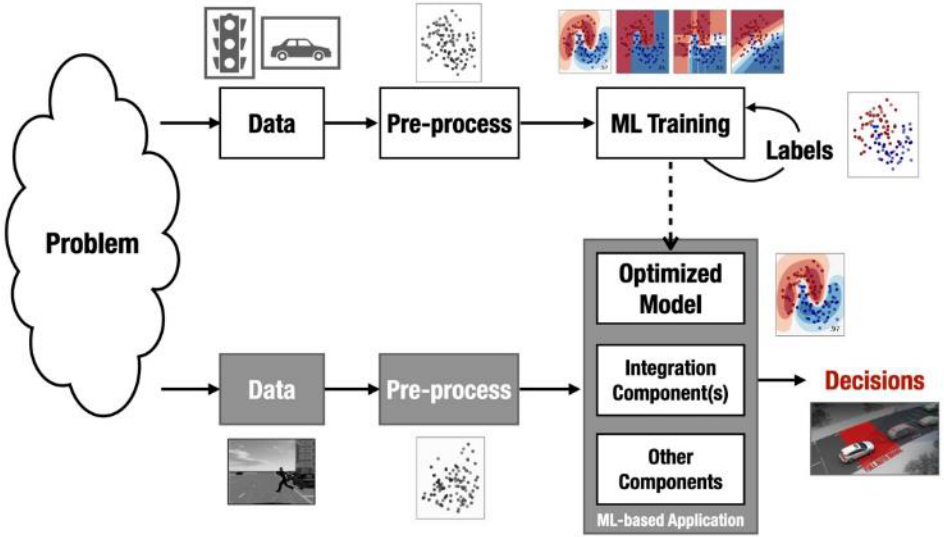
The domain expert's focus



What is it really about?



What may happen in practice?

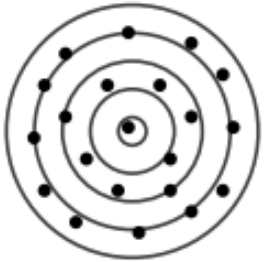


Interdisciplinarity: AI and hiring

Cynthia C. S. Liem et al., Psychology meets Machine Learning: Interdisciplinary perspectives on algorithmic job candidate screening, in Explainable and Interpretable Models in Computer Vision and Machine Learning (pp. 197-253), 2018

Validity and reliability

Not Valid but Reliable

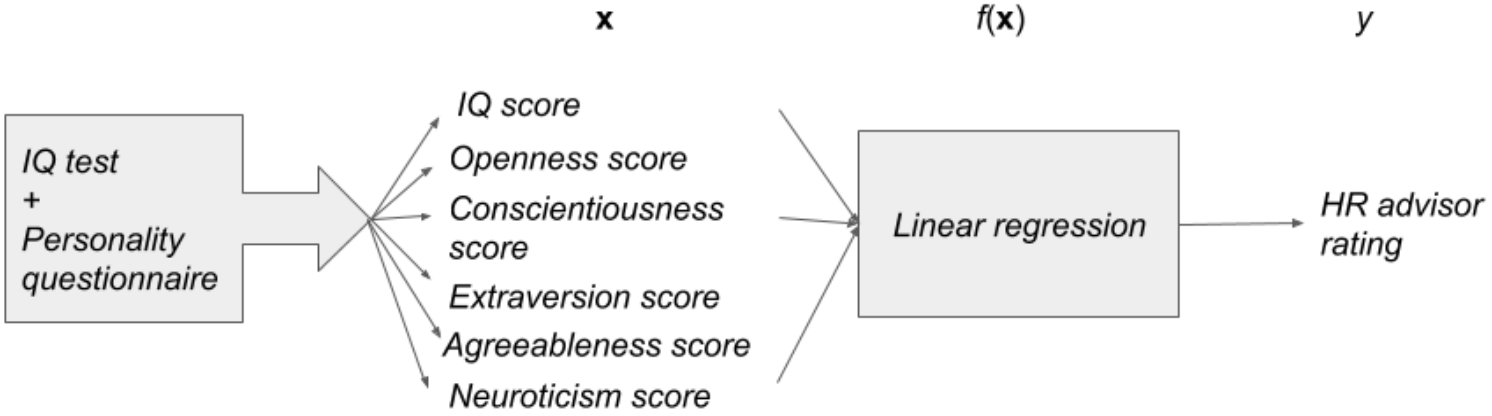


Valid but Not Reliable

Neither Valid Nor Reliable

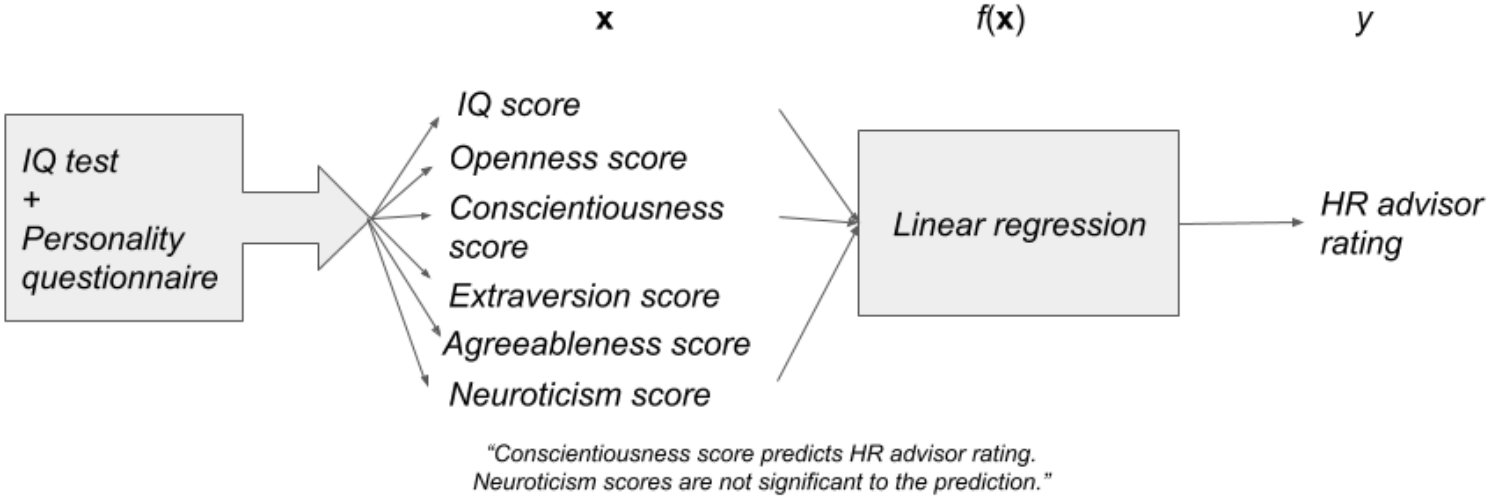
Both Valid and Reliable

Relevant pipelines to my colleagues



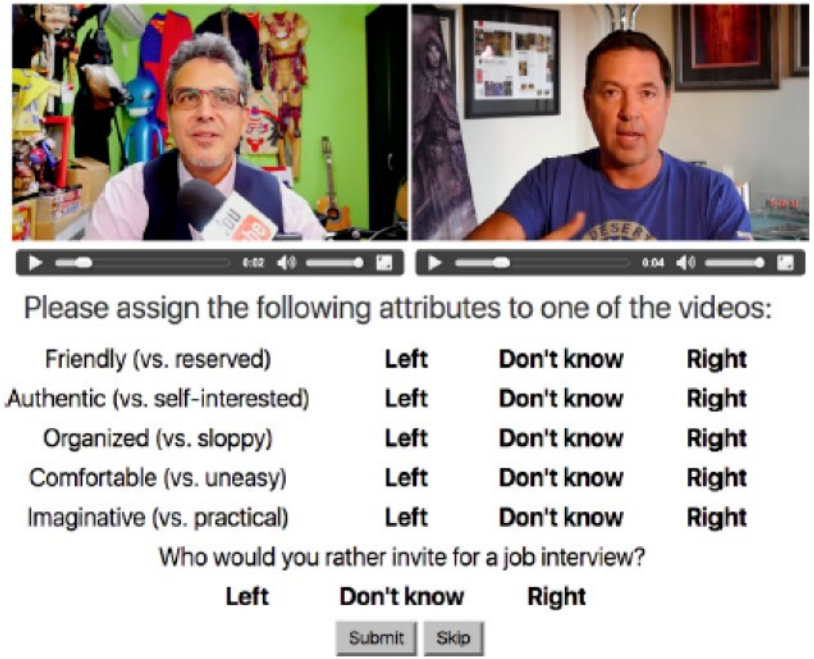
- looks familiar to what we do?

Relevant pipelines to my colleagues



- they really focus on **x** and *y*, where we focus on *f*(**x**)
- our **x** is much more low-level than theirs
- their **x** and *y* often came from **validated instruments**

This is not a validated instrument…



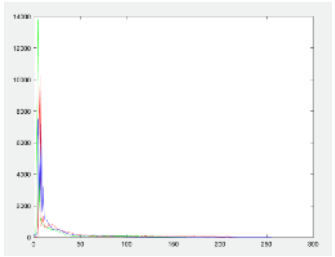
More on measurement:

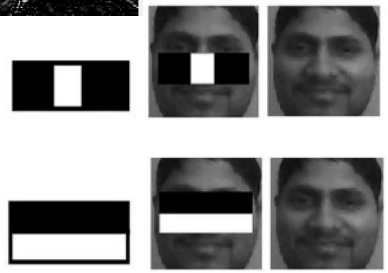
Chris Welty, Praveen Paritosh and Lora Aroyo, “Metrology for AI: From Benchmarks to Instruments”, arXiv:1911.01875
Abigail Z. Jacobs and Hanna Wallach, “Measurement and Fairness”, Proc. FAccT 2021.

And what are x and *y* ?

The ‘semantic gap’

WHAT WE SEE





from [Viola & Jones, 2001]

WHAT THE COMPUTER SEES

Raw, unstructured data

Feature extraction

[0.5, 2.1, 7.8, 0.2….]

[[3.2, 1.7, 9.1],
[1.2, 0.3, 9.2], ….]

Model

The ‘semantic gap’

WHAT WE SEE

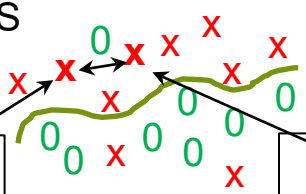
Same topic?

Similar image?






WHAT THE COMPUTER SEES



[0.5, 2.1, 7.8, 0.2….]

[[3.2, 1.7, 9.1],
[1.2, 0.3, 9.2], ….]

Some distance in some high-dimensional space?

[4.5, 3.3, 0.1, 3.8….]

[[0.1, 2.2, 9.0],
[2.2, 4.3, 1.5], ….]

How do we represent ‘the right answer’, and what does this imply?

When does my system perform well?

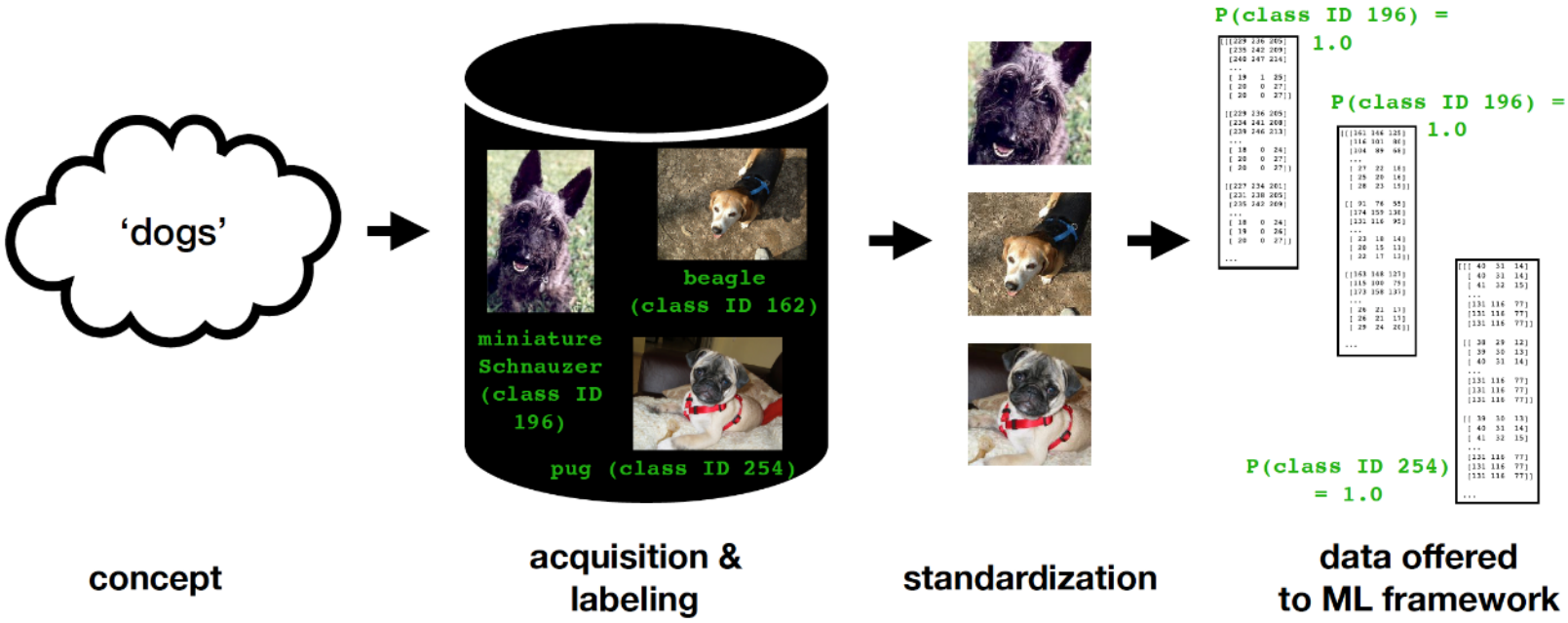
- Our usual focus: how often are we **right**?
- Trivial measure: accuracy
  - # correctly classified samples / # total samples

When mathematical translation and natural vs. social measurements get blurry:
examples from ImageNet

ImageNet

- Visual hierarchical ontology
- Large scale visual recognition challenges: automatic class recognition
- ILSVRC 2012: 1000 object classes

ImageNet


'dogs'
beagle (class ID 162)
miniature Schnauzer (class ID 196)
pug (class ID 254)
P(class ID 196) = 1.0
P(class ID 196) = 1.0
P(class ID 254) = 1.0
concept
acquisition & labeling
standardization
data offered to ML framework

ImageNet labeling

- Amazon Mechanical Turk
- “Is there an [insert class name] in the image?”

Is there a bucket in this image?

ImageNet labeling

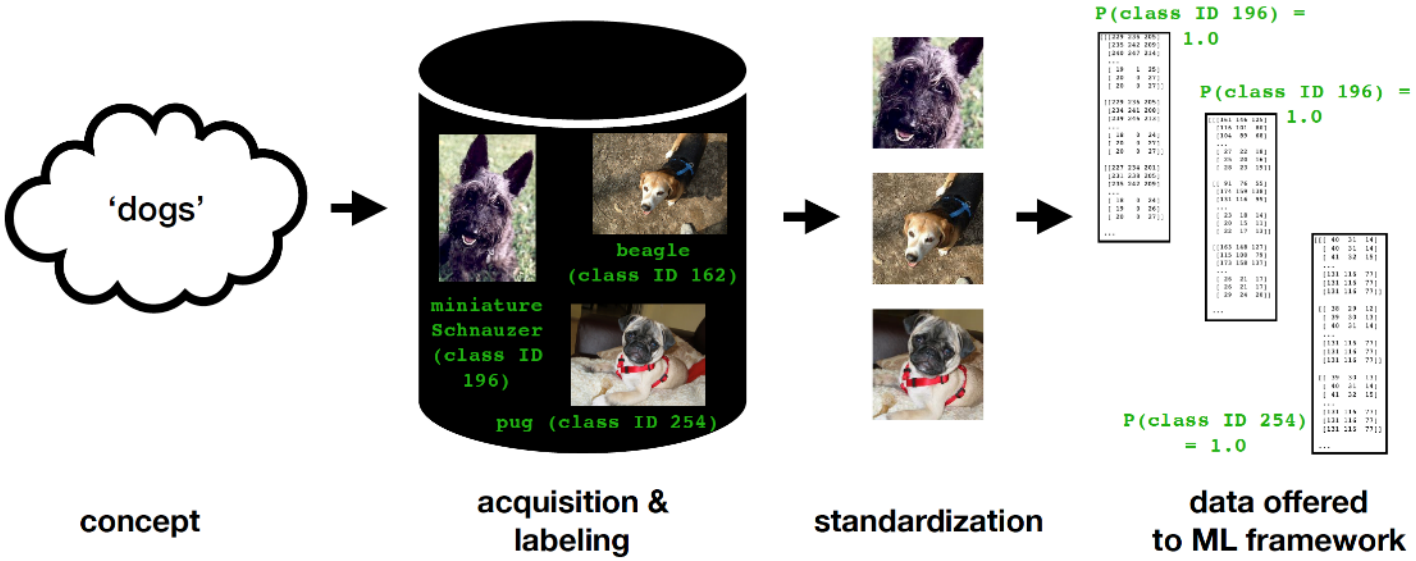
- Single label per image
- Top-5 accuracy

Preparation for mathematical batch-processing

Item ID	…	baseball	…	Miniature schnauzer	…	bucket	…
…	…	…	…	…	…	…	…
172839	…	0	…	0	…	1	…
…	…	…	…	…	…	…	…

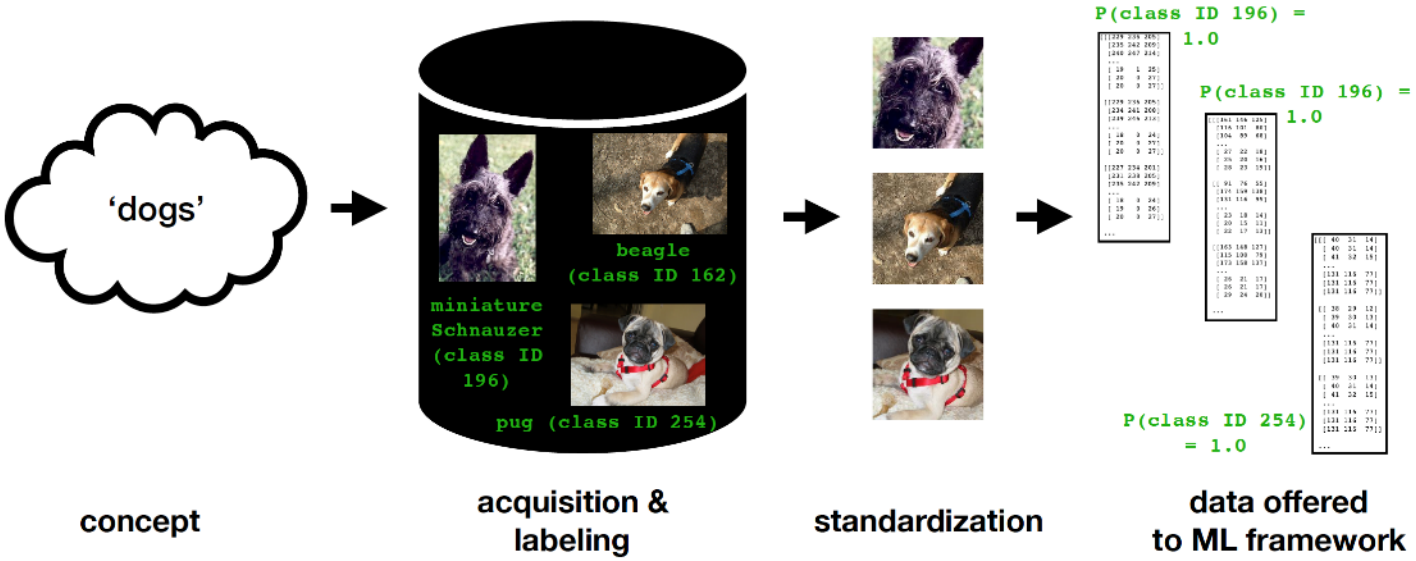
Some ILSVRC 2012 issues

- No true real-world sample (>100 sub-species of dogs)



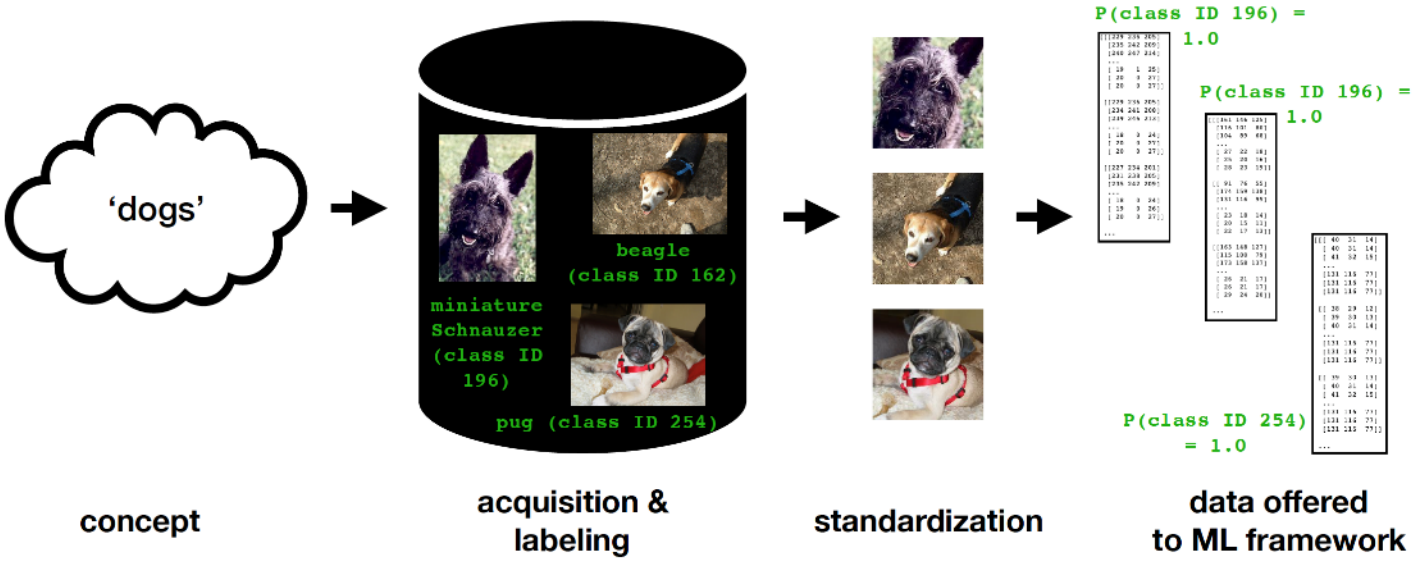
Some ILSVRC 2012 issues

- With a single class label per image, classes mathematically appear independent



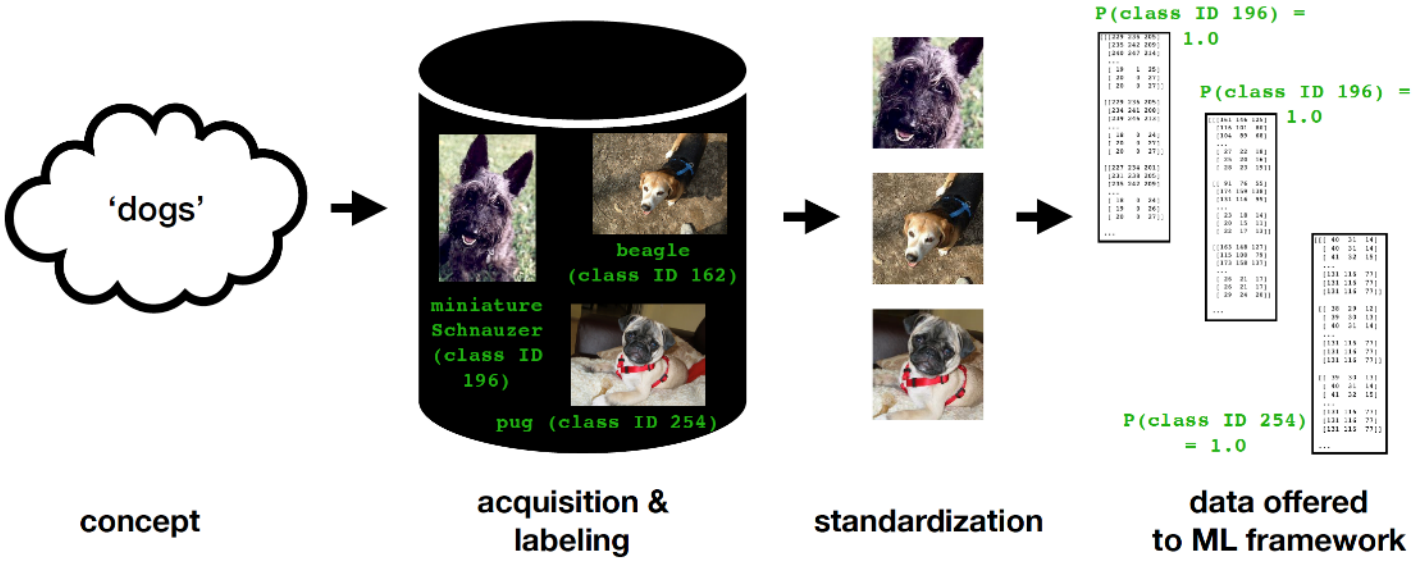
Some ILSVRC 2012 issues

- Learning typically optimizes for maximum likelihood of the target class



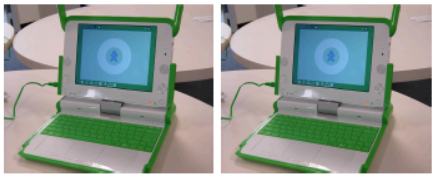
Some ILSVRC 2012 issues

- Standardized input: crop from the image



Some ILSVRC 2012 issues

(a) Original (b) Cropped



(a) Original (b) Cropped

vgg16	vgg19	ResNet50	ResNet101
laptop (0.9592)	laptop (0.9796)	laptop (0.9954)	laptop (0.9984)
notebook (0.0346)	notebook (0.0191)	notebook (0.0042)	notebook (0.0015)
iPod (0.0024)	iPod (0.0004)	space bar (0.0002)	space bar (0.0000)
hand-held computer (0.0011)	desktop computer (0.0002)	computer keyboard (0.0000)	mouse (0.0000)
modem (0.0007)	space bar (0.0001)	mouse (0.0000)	computer keyboard (0.0000)

(c) Predictions

(a) Original (b) Cropped

vgg16	vgg19	ResNet50	ResNet101
notebook (0.7222)	notebook (0.7327)	notebook (0.7230)	notebook (0.8161)
laptop (0.1866)	laptop (0.1178)	laptop (0.1689)	laptop (0.1492)
desktop computer (0.0244)	desktop computer (0.0459)	desktop computer (0.0420)	modem (0.0100)
space bar (0.0097)	space bar (0.0243)	space bar (0.0239)	space bar (0.0091)
solar dish (0.0092)	hand-held computer (0.0152)	mouse (0.0059)	desktop computer (0.0041)

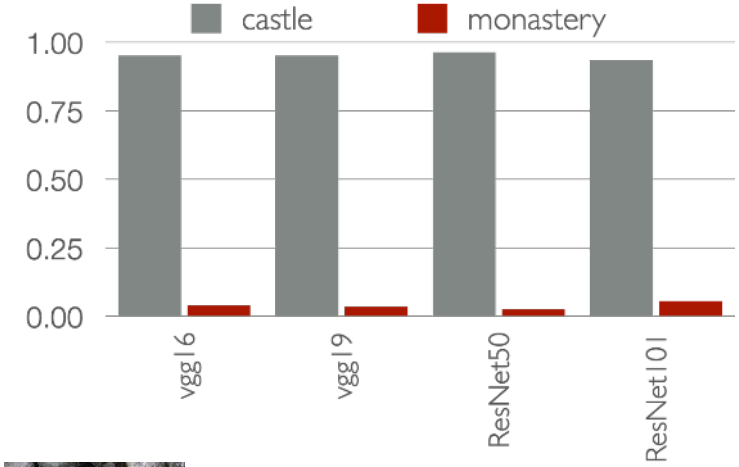
(c) Predictions

‘Ground truth’ and its implications

- Is there a castle in this image?

Cynthia C. S. Liem and Annibale Panichella, “Oracle Issues in Machine Learning and Where to Find Them,” Proc. RAISE 2020.

‘Ground truth’ and its implications



Cynthia C. S. Liem and Annibale Panichella, “Oracle Issues in Machine Learning and Where to Find Them,” Proc. RAISE 2020.

‘ILSVRC 2012 has been solved’

- In terms of %, few ’true’ errors
  - If ground truth class is not in top-5, model suggestions seem acceptable to humans
- For object class recognition, this setup is fine

- But this is **not** comprehensive visual understanding!

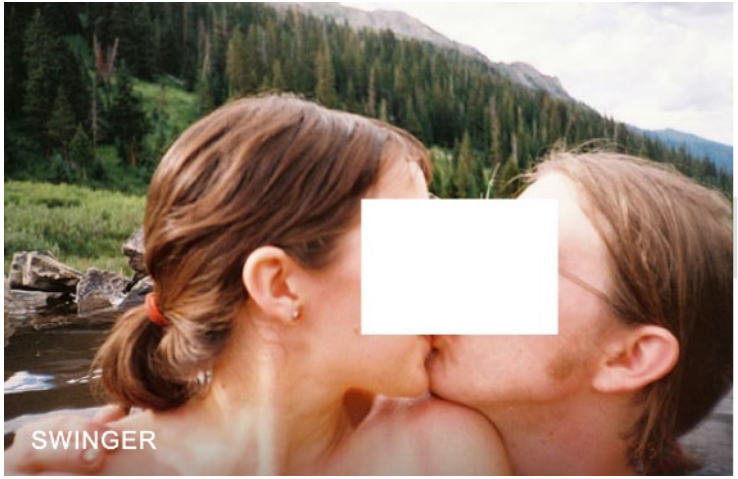
Other issues: what can one truly visually illustrate?

Excavating AI

The Politics of Images in Machine Learning Training Sets

By Kate Crawford and Trevor Paglen

You open up a database of pictures used to train artificial intelligence systems. At first, things seem straightforward. You're met with thousands of images: apples and oranges, birds, dogs, horses, mountains, clouds, houses, and street signs. But as you probe further into the dataset, people begin to appear: cheerleaders, scuba divers, welders, Boy Scouts, fire walkers, and flower girls. Things get strange: A photograph of a woman smiling in a bikini is labeled a "slattern, slut, slovenly woman, trollop." A young man drinking beer is categorized as an "alcoholic, alky, dipsomaniac, boozer, lush, soaker, souse." A child wearing sunglasses is classified as a "failure, loser, non-starter, unsuccessful person." You're looking at the "person" category in a dataset called ImageNet, one of the most widely used training sets for machine learning.



excavating.ai

Counting right vs. wrong

- Self-driving cars likely make less mistakes than humans---still, good reasons to not allow them on the road yet

False-positives vs. false-negatives

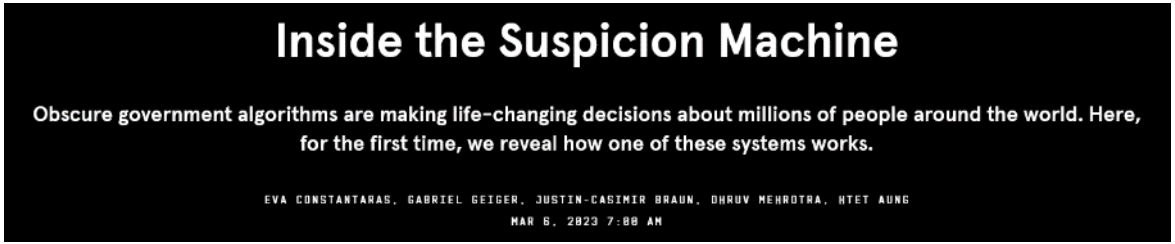
- Treat patient who may be ill in case of doubt?

False-positives vs. false-negatives

- Always halt similar people as ‘potentially high-risk’ if this may lead to feedback loops?

False-positives vs. false-negatives

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



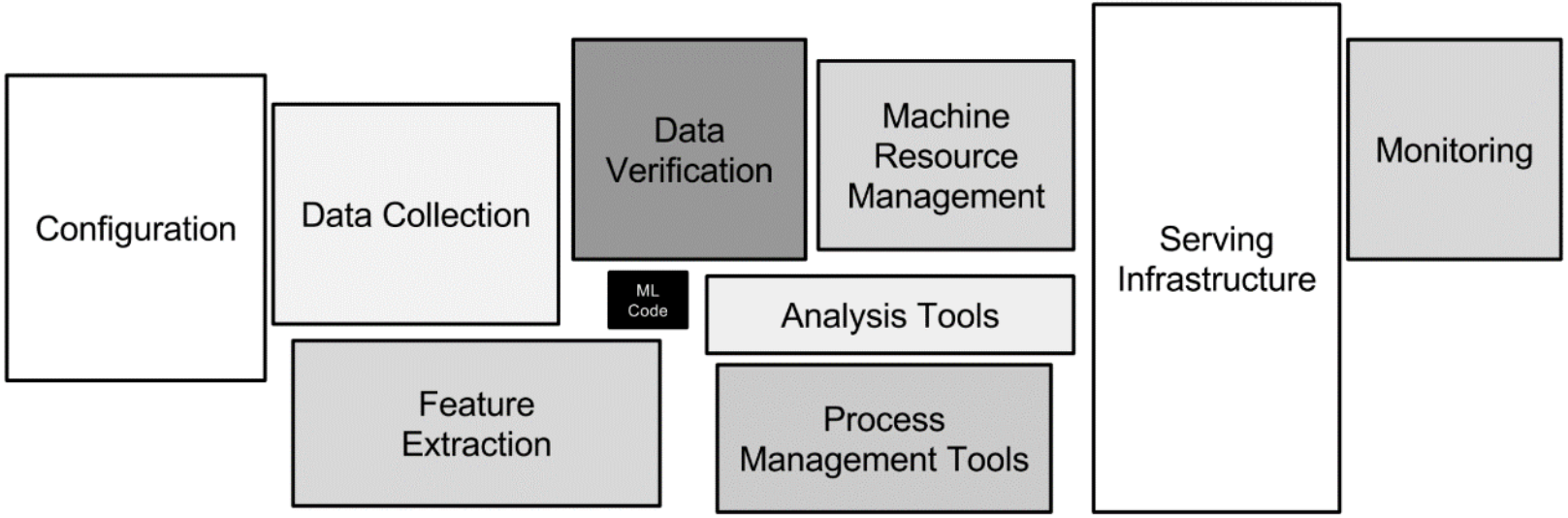
https://www.wired.com/story/welfare-state-algorithms/

Software and systems engineering perspectives

When does my system perform well?

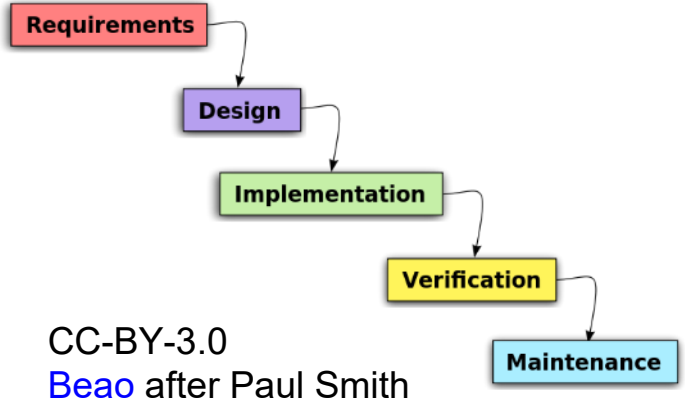
- Our usual focus: how often are we **right**?
- Often forgotten: how bad is **wrong**?

How the ML/AI part compares

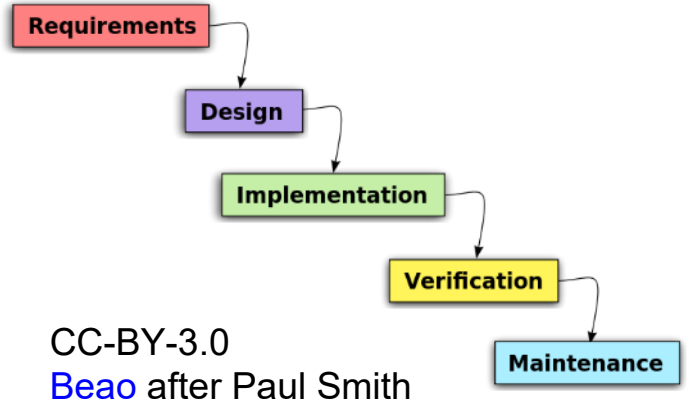


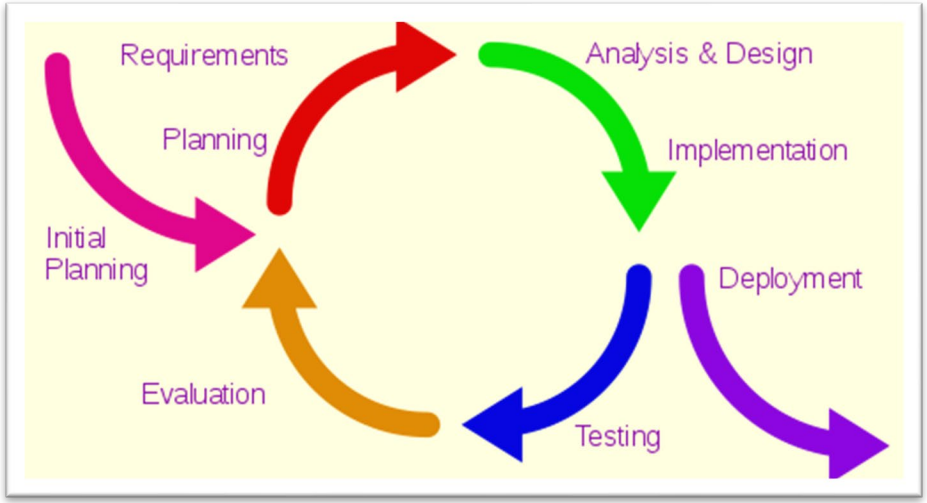
D. Sculley et al., “Hidden Technical Debt in Machine Learning Systems”, in proc. NIPS 2015.
Nithya Sambasivan et al., “Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI”, in proc. CHI 2021.

Ways of working



Ways of working





Questionable representations and reductions

World views

World views

Irrefutable patterns?

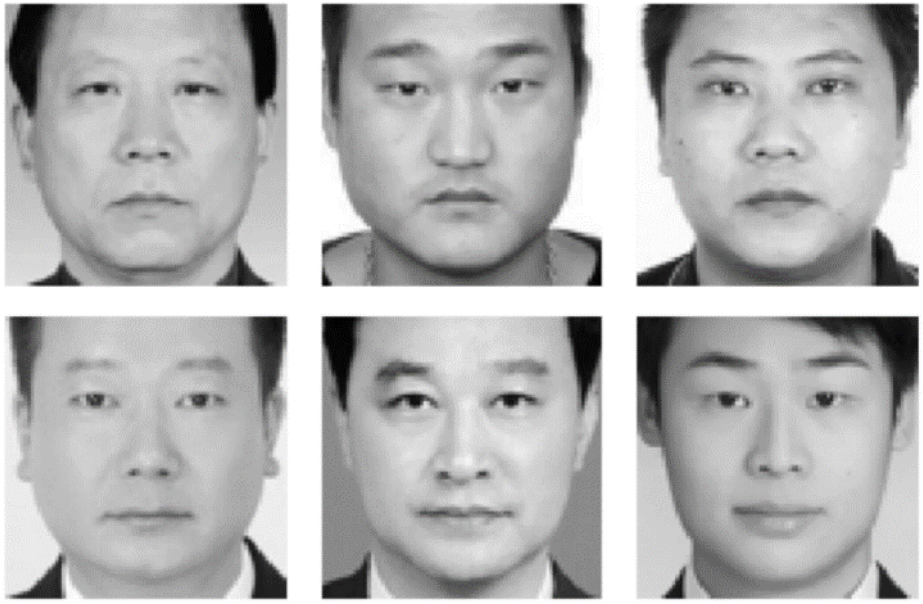


Figure 3. Wu and Zhang's "criminal" images (top) and "non-criminal" images (bottom). In the top images, the people are frowning. In the bottom, they are not. These types of superficial differences can be picked up by a deep learning system.

https://arxiv.org/abs/1611.04135

“More Professional”

‘Fairness’

- Fairness in AI currently a hot topic
- Be mindful: many mathematical definitions which cannot all be satisfied simultaneously (and link to differing ideological schools of thought)

Tutorial: 21 fairness definitions and their politics

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

Debiasing?

- Full debiasing is a myth, and in many cases undesired
- Still, important to explicitly consider human rights, especially those of those not commonly at the table

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence
https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Exercise: favorite and disliked food

- Think of your favorite food, as well as food that you really dislike

Exercise: favorite food commonalities

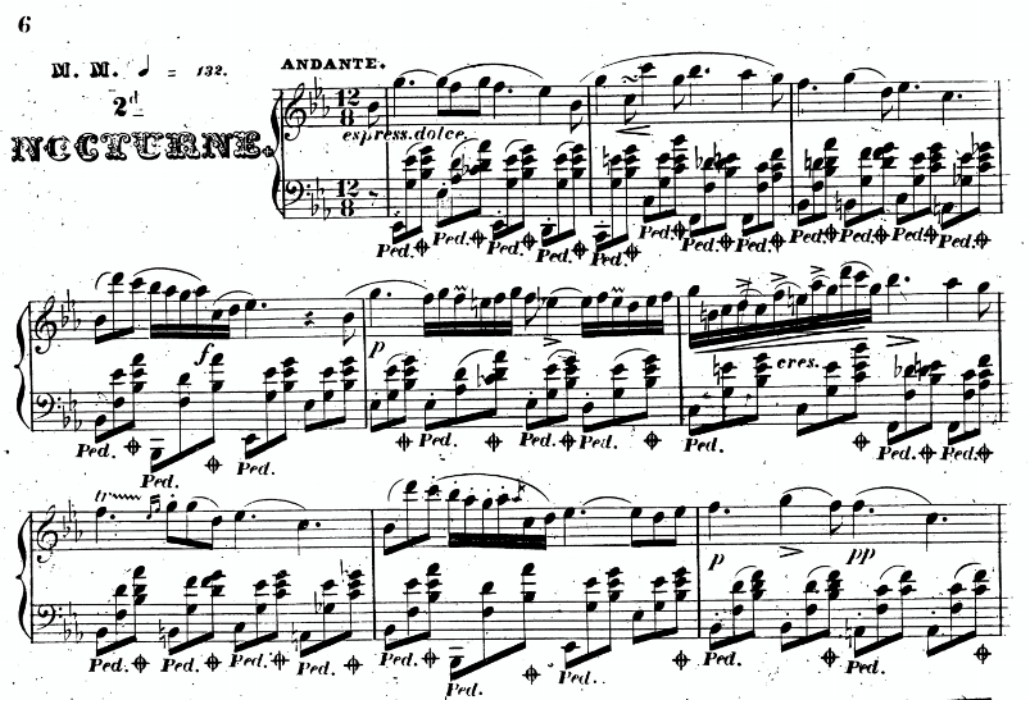
- Make 5 groups
- Find **3 common properties** in your favorite foods. One should be able to give a binary answer (e.g. ‘yes’ or ‘no’) on the property being present

Scoring

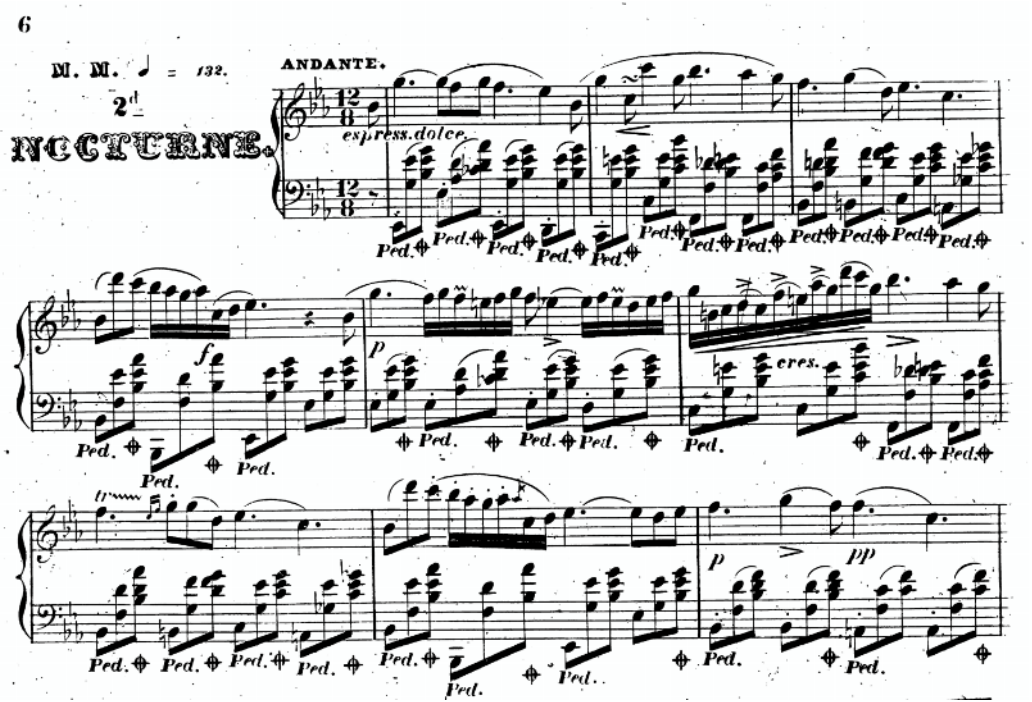
- What scores do we get for your favorite and disliked food?

Questions on ‘the truth’ in music: What do we seek?

The Composition



Is it about being as precise as possible?



uit ‘Chopin – Nocturne op. 9 no. 2’ https://www.youtube.com/watch?v=9E6b3swbnWg (Vadim Chaimovich)

World views


Showing Songs for chopin op 9 no 2
TITLE
ARTIST
ALBUM
Nocturnes, Op. 9: No. 2 in E-Flat Major. And... Frédéric Chopin, Bri... Chopin: The Essentials 4:34
Nocturnes, Op. 9: No. 2 in E-Flat Major. And... Frédéric Chopin, Bri... Chopin: Complete N... 4:34
Nocturnes, Op. 9: No. 2 in E-Flat Major Frédéric Chopin, Fra... Chopin: 21 Nocturnes 4:23
Nocturne Op. 9 No. 2 Frédéric Chopin, Olg... JUST THE BEST MU... 5:02
Nocturnes, Op. 9: No. 2 in E-Flat Major Frédéric Chopin, Art... Chopin Nocturnes 4:31
Nocturnes, Op. 9: No. 2 in E-Flat Major Frédéric Chopin, Art... Chopin: Nocturnes - ... 4:26
Nocturnes, Op. 9: No. 2 in E-Flat Major Frédéric Chopin, Art... 50 Masterworks - Ar... 4:28
Nocturnes, Op. 9: No. 2 in E-Flat Major Frédéric Chopin, Art... The Original Jacket ... 4:21
Nocturnes, Op. 9: No. 2 in E-Flat Major Frédéric Chopin, Ad... Classical Study Music 4:50
Nocturnes, Op. 9: No. 2 in E-Flat Major Frédéric Chopin, Elia... Chopin: Notturno 4:30

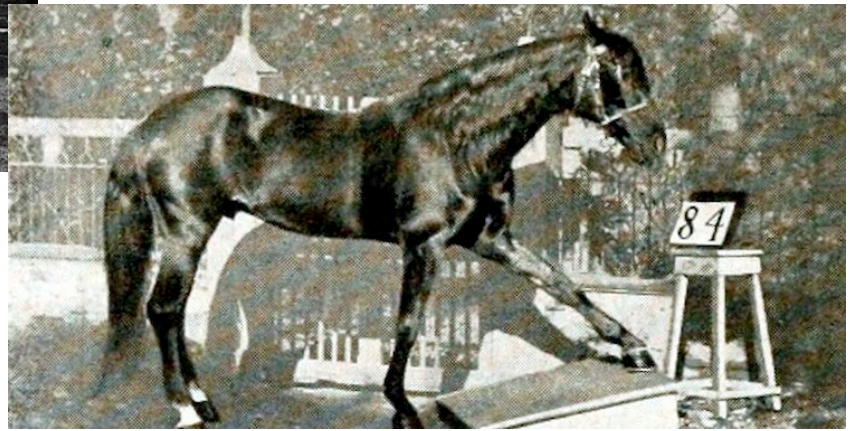
Musicians being puzzled

- Does music genre classification make sense?
- Do the datasets make sense?
- ‘Horse systems’, after Clever Hans

Bob Sturm

Bob Sturm. “A Simple Method to Determine if a Music Information Retrieval System is a “Horse””, IEEE Trans. Multimedia 16 (6), 1636-1644, 2014.

Clever Hans



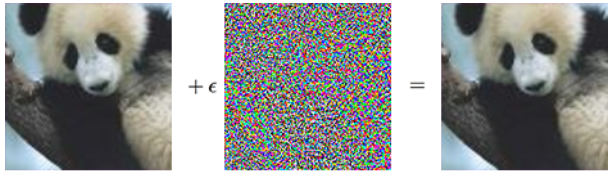
So what is a ‘horse’ system?

- A “horse” is a system that is not actually addressing the problem it appears to be solving.
- A system is a “horse” only in relation to a specific problem.
- A “horse” for one problem may not be a “horse” for another
  - “Reproduce ground truth by XYZ”
  - “Reproduce ground truth by any means”

Bob Sturm. “A Simple Method to Determine if a Music Information Retrieval System is a “Horse””, IEEE Trans. Multimedia 16 (6), 1636-1644, 2014.

How can we test whether a system is a ‘horse’?

- Apply ‘irrelevant transformations’
- See what your system actually will say
- cf. adversarial examples
- https://github.com/cleverhans-lab/cleverhans



irrelevant transformations in music genre recognition: https://www.youtube.com/watch?v=KFZH8gZMumU

Assessing something is off – even without a clear ground truth

Cynthia C. S. Liem and Chris Mostert, “Can’t trust the feeling? How open data reveals unexpected behavior of high-level music descriptors”, Proc. ISMIR 2020, https://program.ismir2020.net/poster_2-10.html

Meta-scientific trustworthiness analysis?

- Anyone can submit anything…so we don't know what the output should be?
- In software engineering and psychology, we saw '**testing**' can go **beyond 'known truths'**, exploiting **known relationships**.

Correlation between constructs

- Inspired by construct validity approaches in psychology
- Redundancy in constructs (e.g., multiple genre classifiers with overlapping labels)

Correlation between constructs

- `genre_rosamerica` classifier was **90.74 %** accurate on rock.
- `genre_tzanetakis` classifier was **60 %** accurate on rock.
- Pearson's *r* between the `genre_rosamerica` and `genre_tzanetakis` rock classifications in Acousticbrainz: **-0.07**

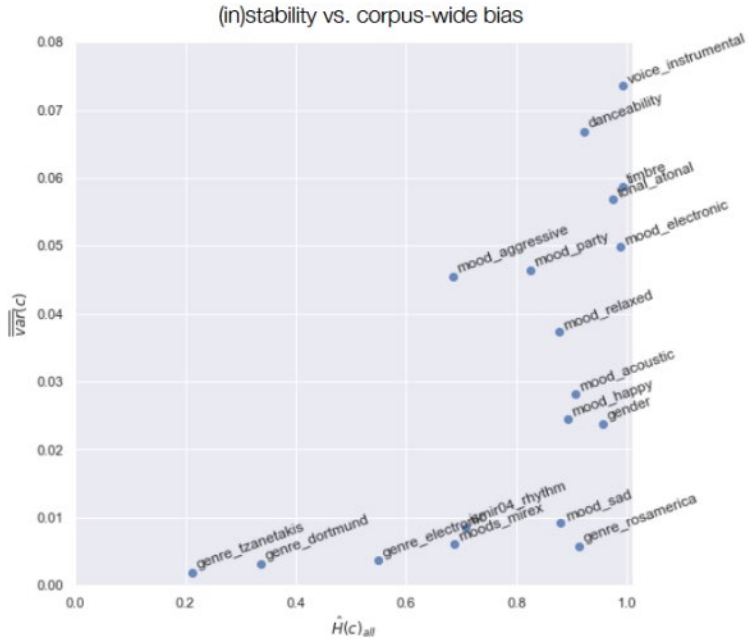
Classifier, label A	Classifier, label B	Pearson's r
genre_rosamerica, `cla`	genre_tzanetakis, `cla`	.29
genre_dortmund, `rock`	genre_rosamerica, `roc`	.24
genre_dortmund, `jazz`	genre_rosamerica, `jaz`	.22
genre_dortmund, `pop`	genre_rosamerica, `pop`	.11
genre_dortmund, `jazz`	genre_tzanetakis, `jaz`	.08
genre_rosamerica, `pop`	genre_tzanetakis, `pop`	.06
genre_rosamerica, `hip`	genre_tzanetakis, `hip`	.05
genre_rosamerica, `jaz`	genre_tzanetakis, `jaz`	.02
genre_dortmund, `blues`	genre_tzanetakis, `blu`	.01
genre_dortmund, `pop`	genre_tzanetakis, `pop`	-.05
genre_dortmund, `rock`	genre_tzanetakis, `roc`	-.06
genre_rosamerica, `roc`	genre_tzanetakis, `roc`	-.07
mood_aggressive, `aggressive`	mood_relaxed, `not_relaxed`	.59
mood_acoustic, `acoustic`	mood_electronic, `not_electronic`	.58
danceability, `danceable`	mood_party, `party`	.53
mood_electronic, `electronic`	genre_dortmund, `electronic`	.48
danceability, `danceable`	genre_rosamerica, `dan`	.33
mood_happy, `happy`	mood_party, `party`	.20
mood_happy, `happy`	mood_sad, `not_sad`	.13

Stability of resubmissions

- Inspired by derived oracles in software testing
- One would assume

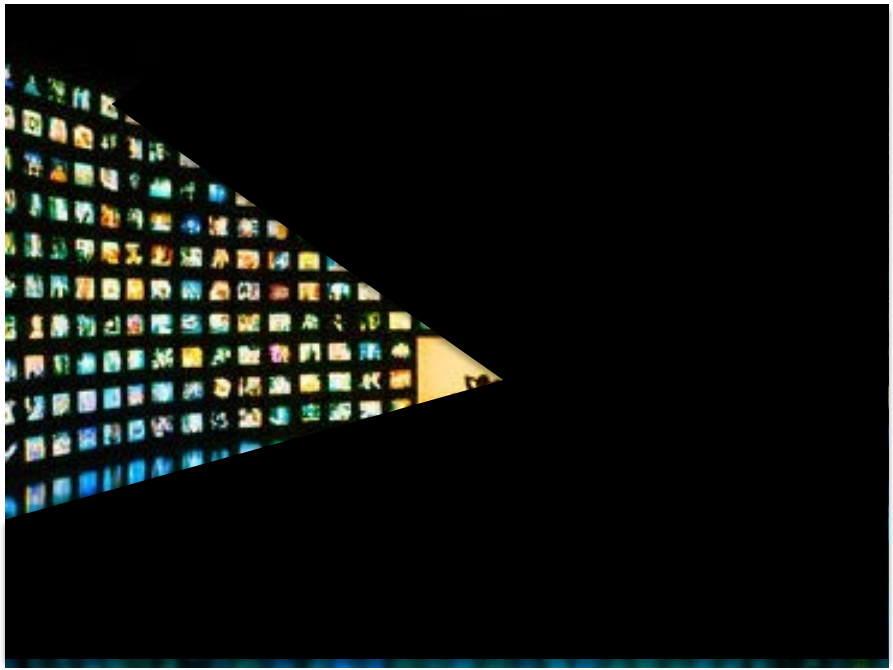
```
classify_c(my_preprocessing(m))
                ==
classify_c(your_preprocessing(m))
```

- In terms of classifier outputs, not necessarily…



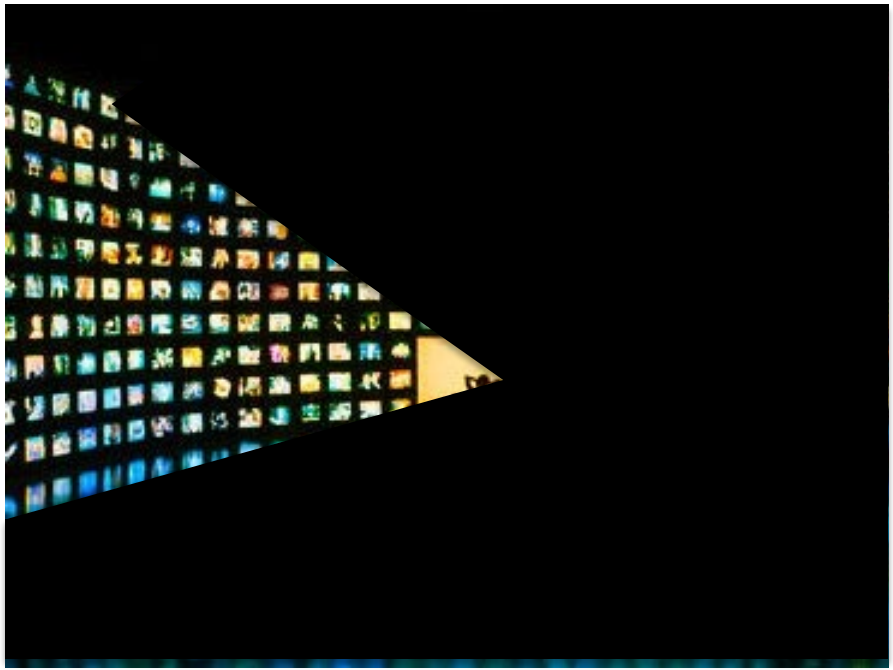
The long-standing ideal in multimedia consumption

The long-standing ideal in multimedia consumption



The long-standing ideal in multimedia consumption

relevant



Netflix 2017

“What’s more powerful: you telling me you would give five stars to the documentary about unrest in the Ukraine; that you’d give three stars to the latest Adam Sandler movie; or that you’d watch the Adam Sandler movie 10 times more frequently?” Yellin said. “What you do versus what you say you like are different things.”

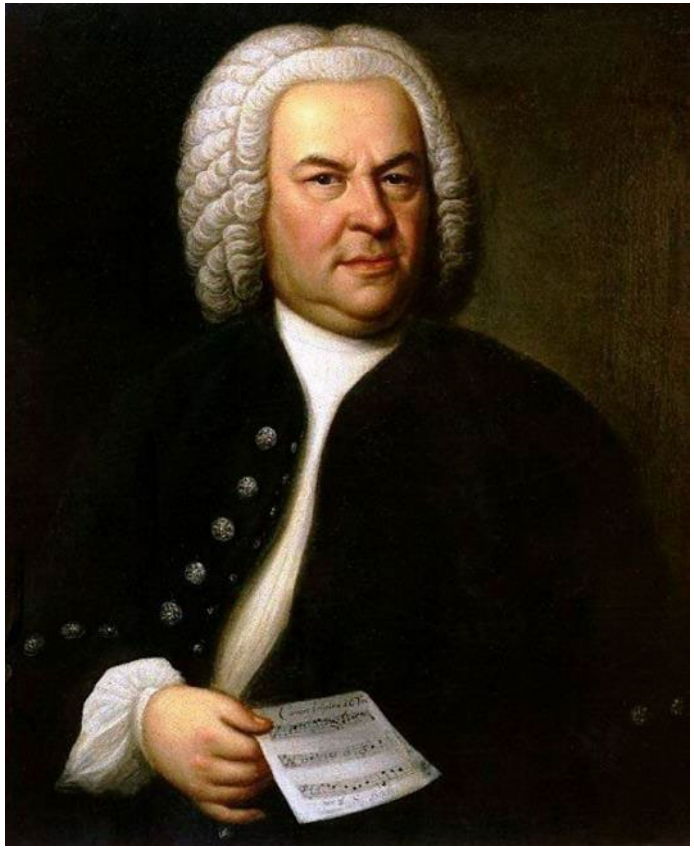
https://www.theverge.com/2017/3/16/14952434/netflix-five-star-ratings-going-away-thumbs-up-down

The long-standing ideal in multimedia consumption

The long-standing ideal in multimedia consumption

irrelevant?

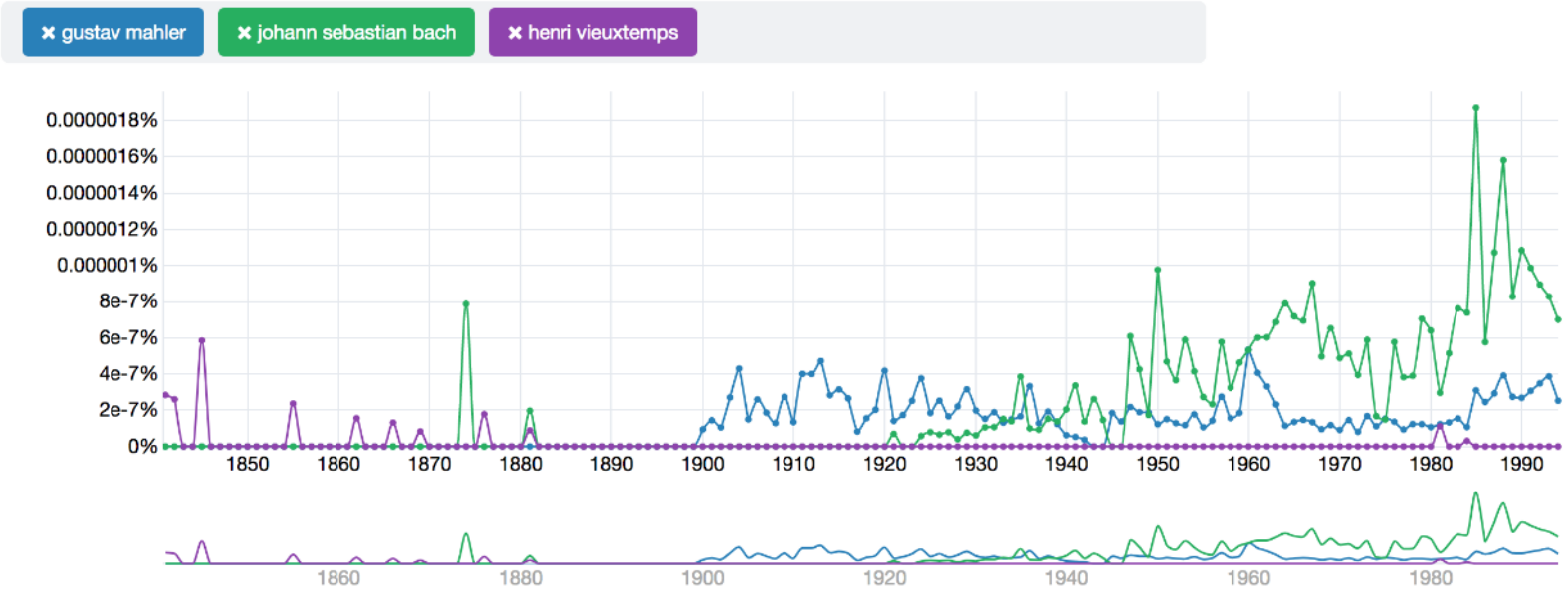
Relevant?



[Liem, 2018]

Trends over time

- KB Labs n-gram viewer



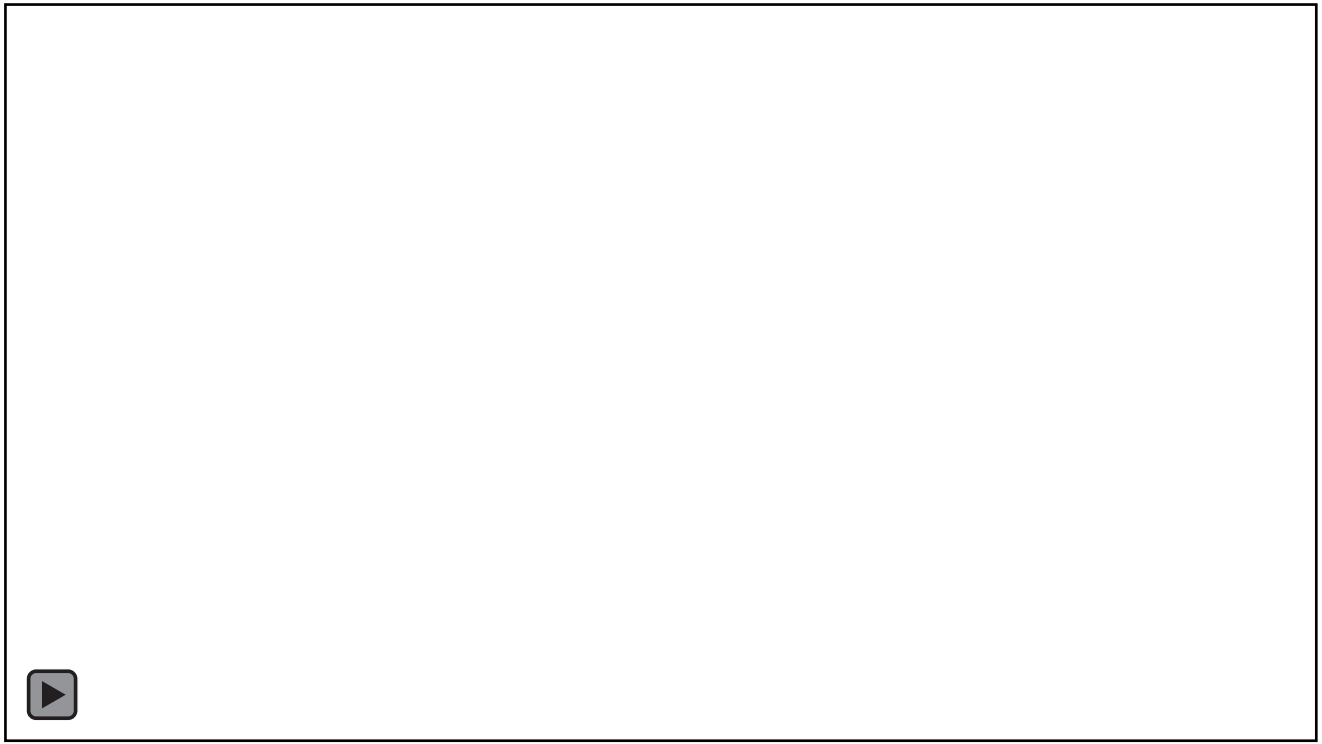
The champions

Felix Mendelssohn Bartholdy Franz Liszt

and several others…

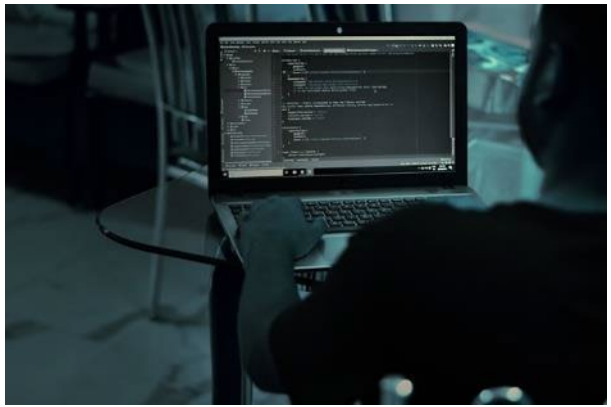
Being human in the age of Generative AI

The promise?



https://www.youtube.com/watch?v=YRb0XAnUpIk

Do we undervalue this?



While we already risked losing this?

- Archive Bookstore, London, March 16, 2019

And this?

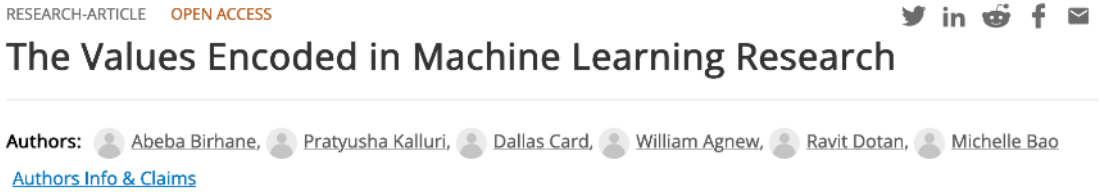
- Acquired taste
- First experience negative, repeated exposure needed

Photo by kiliweb for Open Food Facts

What fed the enthusiasm?

Who decides and describes?

Whose values and interests?



RESEARCH-ARTICLE OPEN ACCESS

The Values Encoded in Machine Learning Research

Authors: Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, Michelle Bao

Authors Info & Claims

"We annotate key features of papers which reveal their values: their justification for their choice of project, which attributes of their project they uplift, their consideration of potential negative consequences, and their institutional affiliations and funding sources. We find that few of the papers justify how their project connects to a societal need (15%) and far fewer discuss negative potential (1%). Through line-by-line content analysis, we identify 59 values that are uplifted in ML research, and, of these, we find that the papers most frequently justify and assess themselves based on Performance, Generalization, Quantitative evidence, Efficiency, Building on past work, and Novelty."

https://dl.acm.org/doi/10.1145/3531146.3533083

Global inequalities

https://www.technologyreview.com/2022/04/19/1049592/artificial-intelligence-colonialism/

Insights from museum practices

Huang & Liem, “Social Inclusion in Curated Contexts: Insights from Museum Practices”, proc. ACM FAccT, 2022, https://dl.acm.org/doi/abs/10.1145/3531146.3533095 .

Challenges of curation

- How can the selection process do justice to the original diversity in the broader collections?
- How can curation be performed in ways that respect, engage and include audiences beyond mainstream perspectives?

ICOM Museum definitions: 2007 → 2022

"A museum is a non-profit, permanent institution in the service of society and its development, open to the public, which acquires, conserves, researches, communicates and exhibits the tangible and intangible heritage of humanity and its environment for the purposes of education, study and enjoyment."

"A museum is a not-for-profit, permanent institution in the service of society that researches, collects, conserves, interprets and exhibits tangible and intangible heritage. Open to the public, accessible and inclusive, museums foster diversity and sustainability. They operate and communicate ethically, professionally and with the participation of communities, offering varied experiences for education, enjoyment, reflection and knowledge sharing."

Neutrality revisited & cultural humility

Simon Maris, ca. 1906

"East Indian type. Oriental girl sitting in an armchair" (1922)

"Little Negress" (1970)

"Young lady with a fan" (2015)

"Isabella" (2020)

Words Matter: https://www.tropenmuseum.nl/en/about-tropenmuseum/words-matter-publication

Situational interpretation

- Items situated in context
- Active collaboration across museum services
- Response to target group's needs

Community participation

Kamen's Minivan, Object 0001 of Authentic Rotterdam Heritage Collection (Echt Rotterdams Erfgoed)

Working as intended?
Perceptions of success in AI Systems

Cynthia C. S. Liem

c.c.s.liem@tudelft.nl

@cynthiacsliem@akademienl.social

@informusiccs