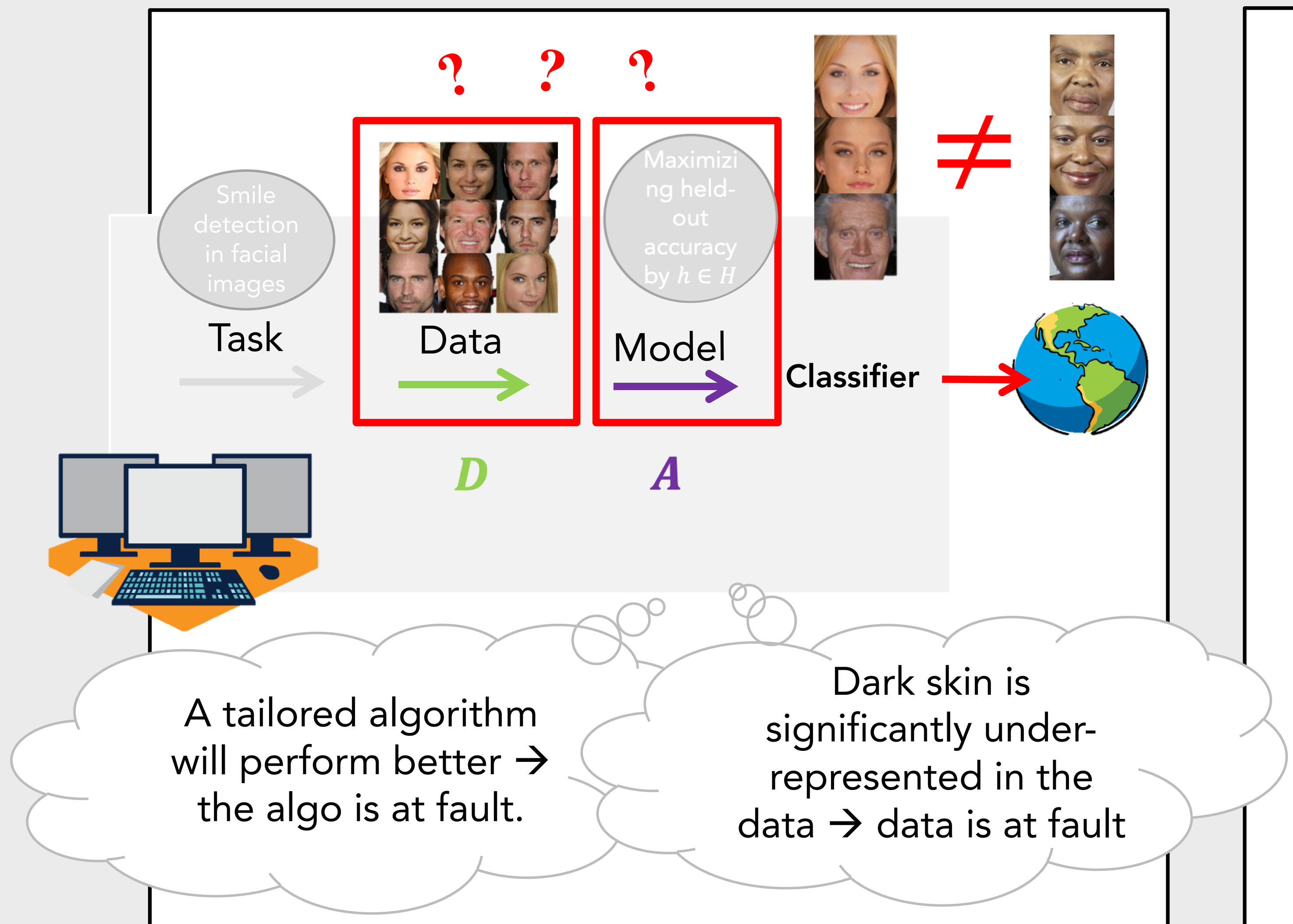
Who's Responsible? Jointly Quantifying the Contribution of the Learning Algorithm and Data

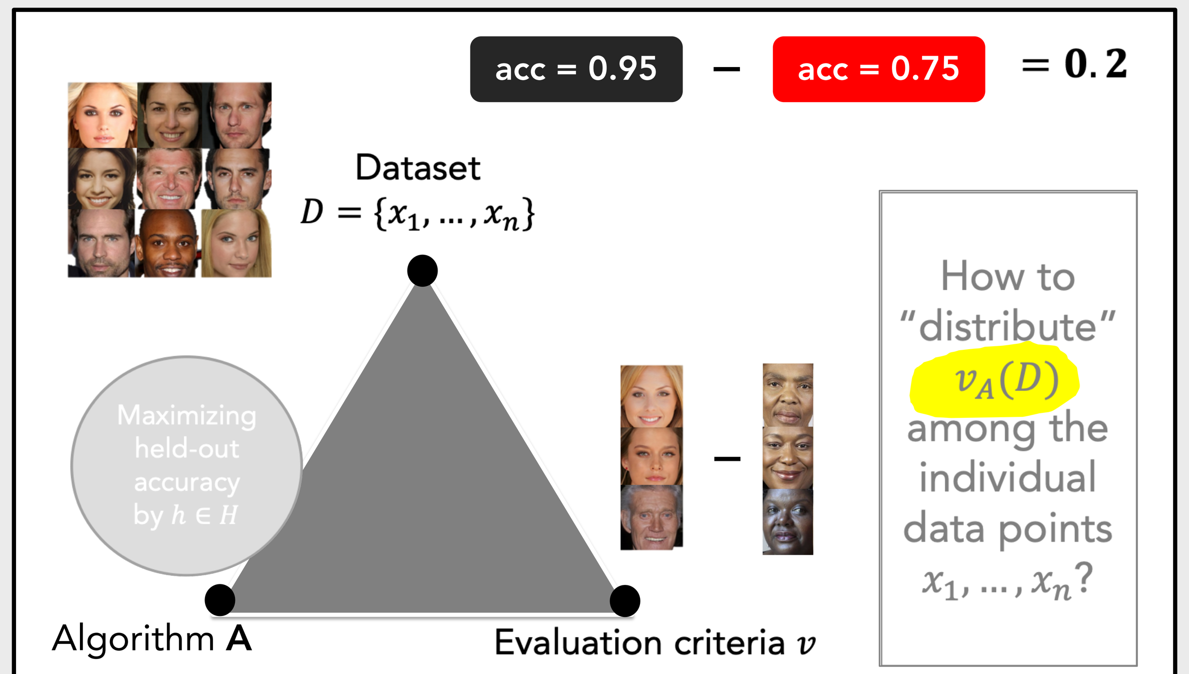
Gal Yona, Amirata Ghorbani & James Zou



Q: recognizing the interaction between existing biases in data and different (potentially subtle) modeling choices, can we disentangle their effect on the overall performance?

Joint data-algorithm valuation problem

A: A reduction to the **data valuation problem** (recently studied e.g. in [GZ19], [ADS19], [JDW+19])
Specifically the approach in [GZ19]

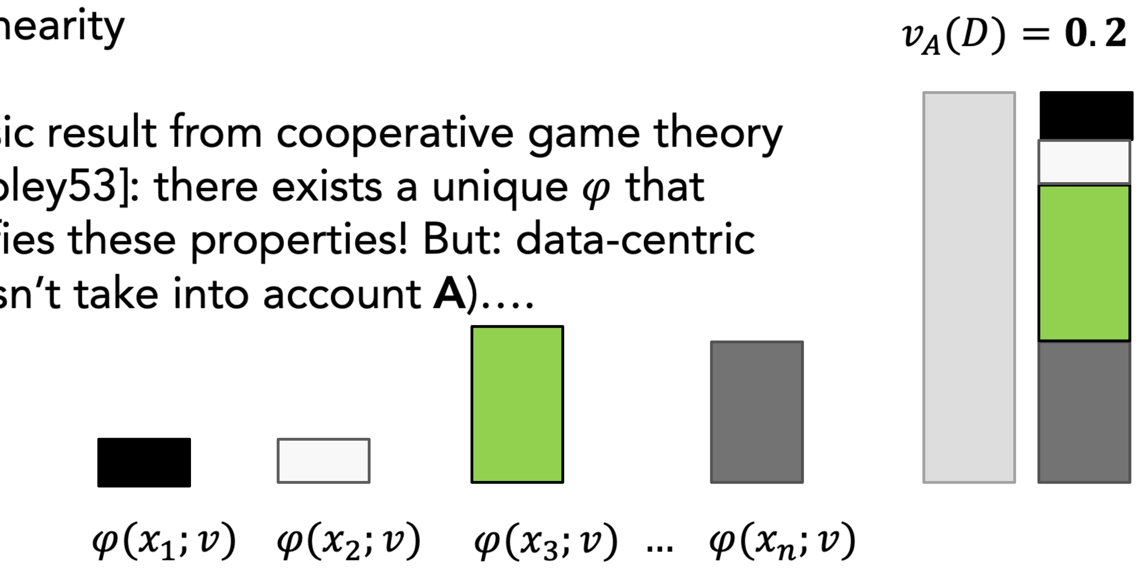


Data Shapley [GZ19]

Specifies *four natural conditions* for an equitable data valuation $\varphi: \{x_1, \dots, x_n\} \to \mathbb{R}^n$:

1. Null player receives *zero value*
2. Symmetric players receives *equal value*
3. *Sum of values is* $v_A(D)$
4. Linearity

Classic result from cooperative game theory [Shapley53]: there exists a unique φ that satisfies these properties! But: data-centric (doesn't take into account **A**)....



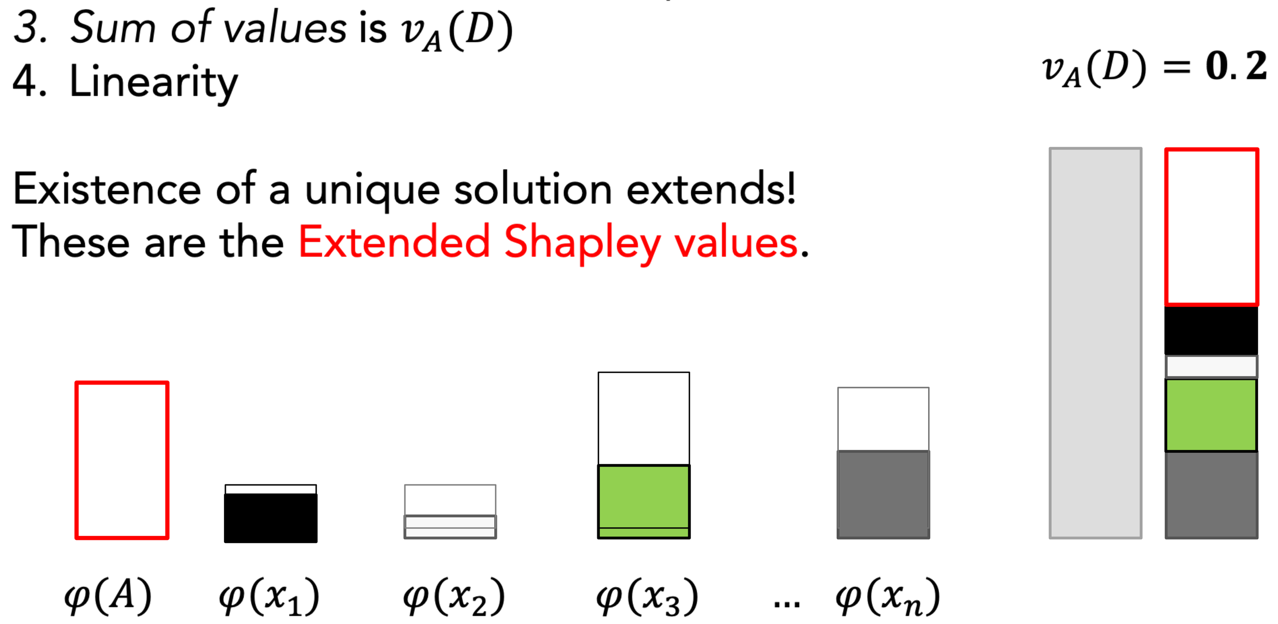
This work: Extended Shapley [YGZ19]

Fix a benchmark algorithm B and add the algorithm **A** as an "additional" $n+1$ player

Specify *five natural conditions* for an equitable data-algorithm valuation $\varphi: \{A, x_1, \dots, x_n\} \to \mathbb{R}^{n+1}$:

1. Null datum receives *zero value; if A is identical to B, algo receives zero value*
2. Symmetric players receives *equal value*
3. *Sum of values is* $v_A(D)$
4. Linearity

Existence of a unique solution extends!
These are the Extended Shapley values.



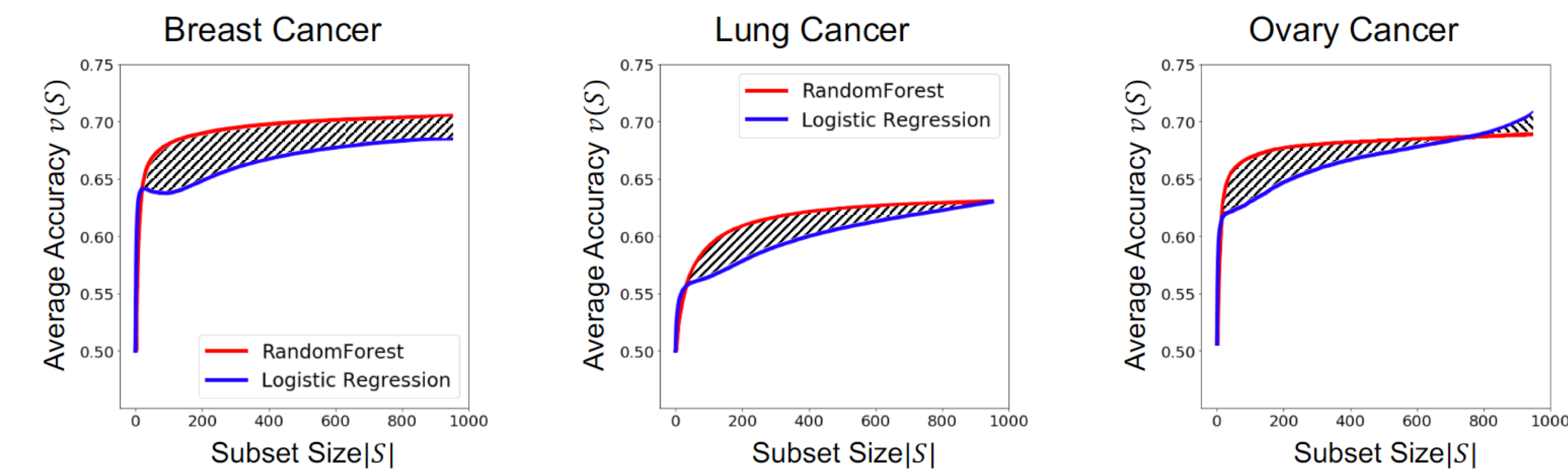
A practical framework!

#1: Closed-form expressions for the value of data and algorithm. E.g. value of algo is $E_S[v_A(S) - v_B(S)]$

Intuition:
If **A** is consistently doing worse off than baseline **B** (irrespective of the specific training set they are trained on), then **A** "deserves" a large portion of the blame.
If very specific combinations of datapoints contribute to A's failure, then these carry more of the "blame".

#2: Can re-use heuristics for efficient computation from [GZ19].

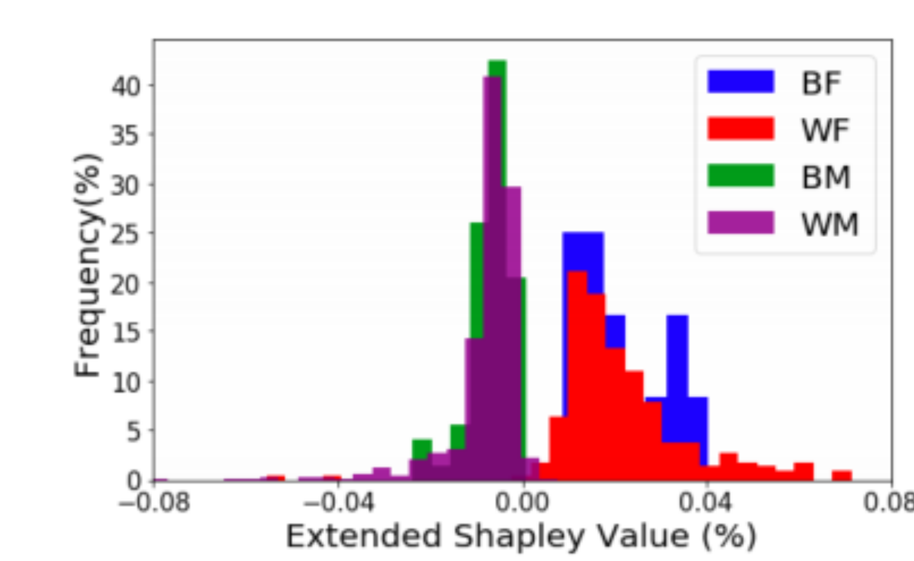
#1: Understanding Extended Shapley as a new performance metric



#2: Allocating responsibility for unfairness

- Training data: 1000 images from LFW+A dataset (imbalanced: 21% female, 5% black)
- Performance measure: maximal accuracy gap among groups {WM, WF, BM, BF} on the balanced PPB dataset
- Algorithm A: Logistic regression applied to 128-dimensional feature vectors obtained by passing the images through a ResNet-V1 pre-trained on CelebA.
- Benchmark B: a constant classifier (perfectly fair: $v_B(D) = 0$)

$v_A(D) = 22.9$ (WM-BF) $\varphi_A = 22.1$



arxiv: https://arxiv.org/abs/1910.04214. contact: gal.yona@gmail.com