公平的資料、無偏見的預測、和可解釋的系統

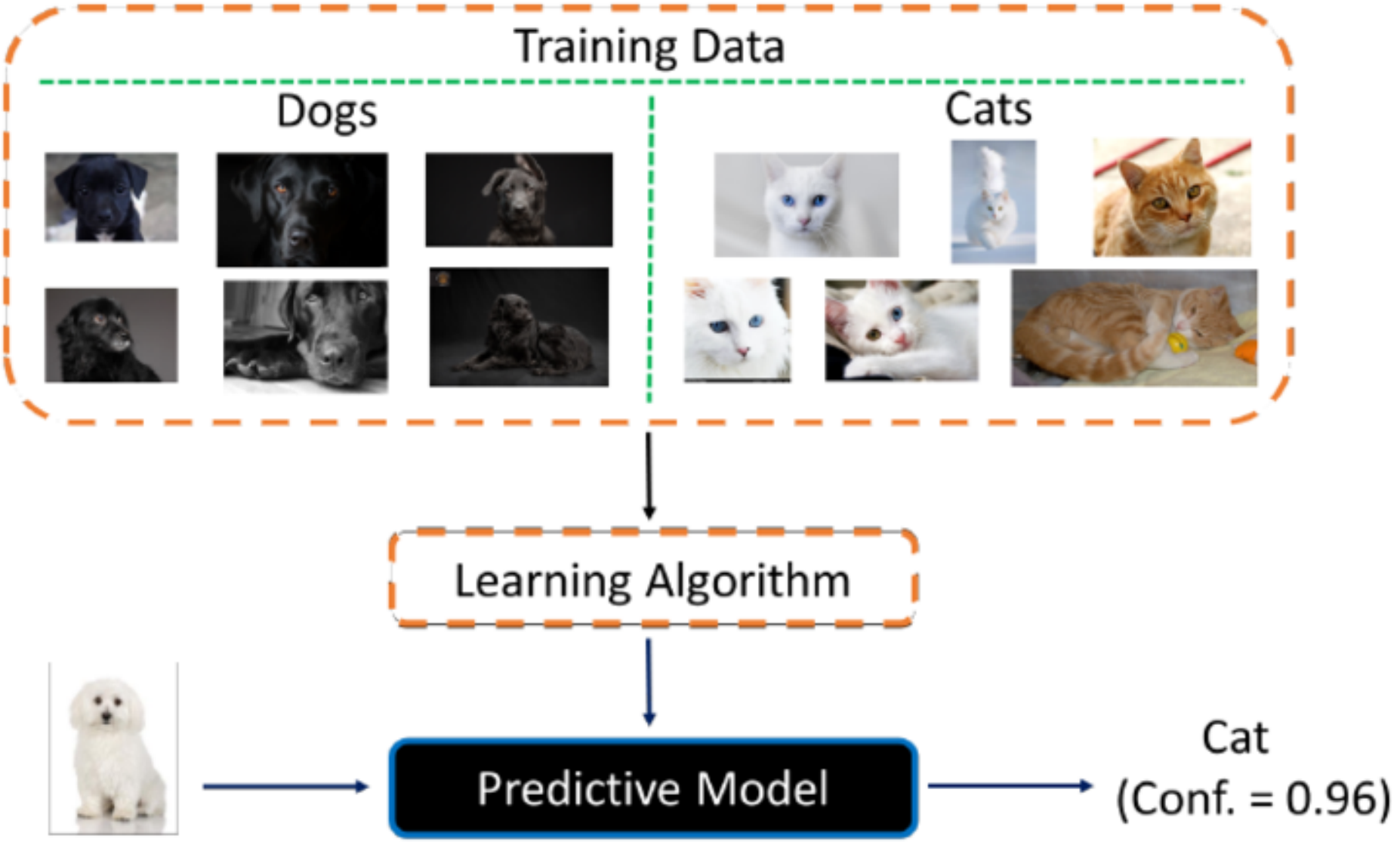
FAIR Data, Unbiased Predictions, and Explainable Systems

2018-03-24

「AI、醫療與法律」論壇
Forum on AI, Health Care, and the Law

莊庭瑞　Tyng-Ruey Chuang
中央研究院　Academia Sinica

trc@iis.sinica.edu.tw



- https://www.bloomberg.com/news/articles/2017-12-04/researchers-combat-gender-and-racial-bias-in-artificial-intelligence
- Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. *Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration,* in AAAI 2017. https://arxiv.org/abs/1610.09064

如何面對有偏見的預測？

- 機器學習（統計和經濟理性的 “AI” ）
  - 著重資料處理效率、預測精準度與速度、應用領域效益
- 機器學習的侷限
  - 資料：具代表性的大量資料如何取得？
  - 當整體預測的表現合乎期許時，能否解釋原因？
  - 當個別預測的結果無法接受時，原因出在哪裡？
- 對於當下的這筆新資料，請問是
  - 歸屬於合乎期許的整體預測表現？
  - 還是會得到無法接受的預測結果？

「公平的資料」 “FAIR Data”

- 資料的代表性（？）如何能被獨立、分別的檢驗？

- 資料找得到 – Data is **F**indable.
- 資料拿得到 – Data is **A**ccessible.
- 資料看得懂 – Data is **I**nteroperable.
- 資料可再用 – Data is **R**eusable.

- “The FAIR Data Principles”
  - https://www.force11.org/group/fairgroup/fairprinciples

可再次使用的資料 Reusable Data

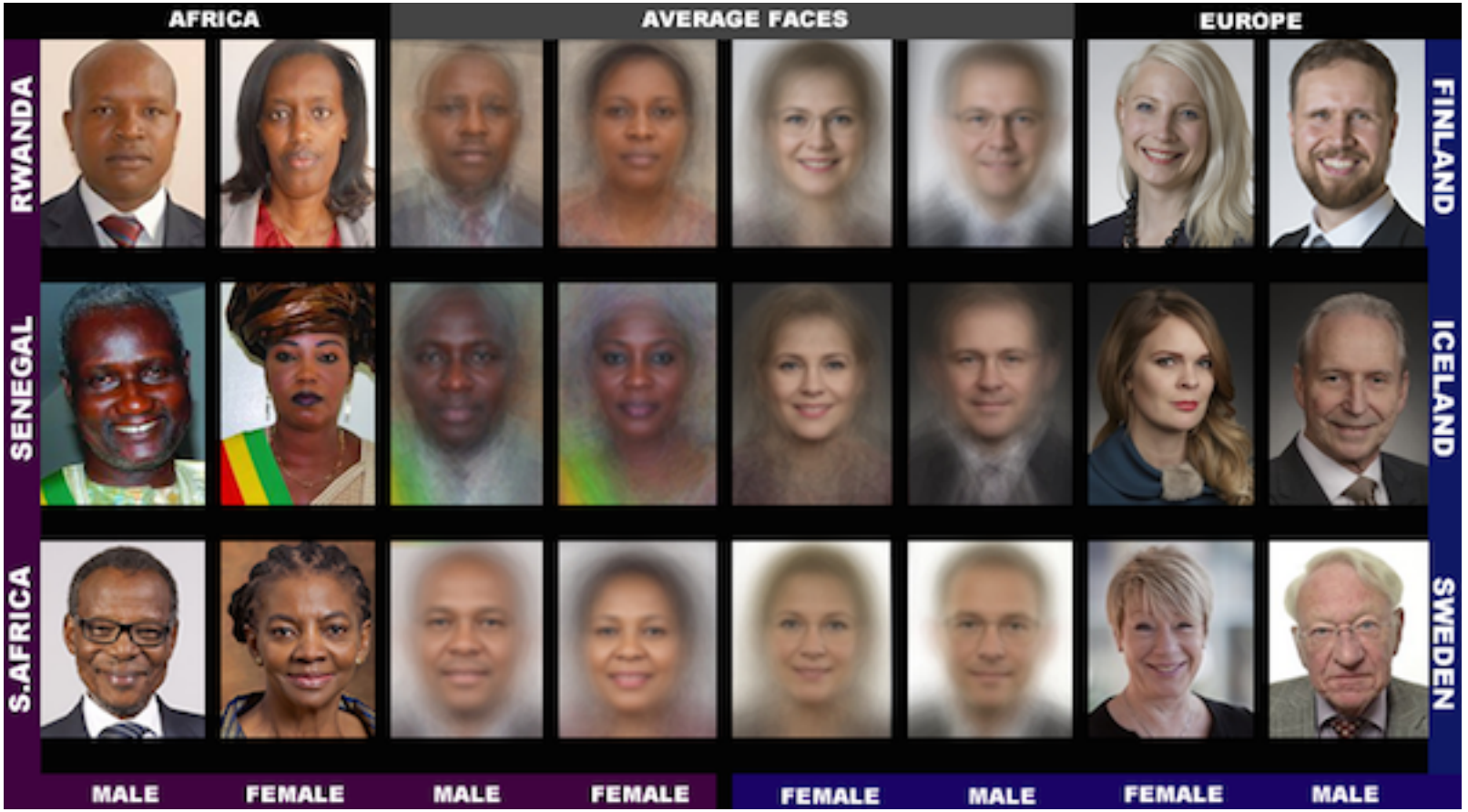
- 後設資料 (關於資料的資料) 應有多項精確與相關的屬性
  - 資料釋出時應搭配清楚與可親近的資料使用授權方式
  - 資料應結合其來源資訊
  - 資料應符合領域相關的社群標準

- Metadata have a plurality of accurate and relevant attributes
  - Data are released with a clear and accessible data usage license
  - Data are associated with their provenance
  - Data meet domain-relevant community standards

- “The FAIR Data Principles”
  - https://www.force11.org/group/fairgroup/fairprinciples

如何走向可解釋的系統？

- “AI” 系統不能是黑盒子
  - 機器學習所得的模型可有錯誤？模型可否能被理解？
  - 「無偏見的預測」需要「公平的資料」以及「可解釋的系統」
- 「可解釋的系統」需要「可重現的研究」的幫助
- （計算上）可重現的研究 Reproducible Research
  - 資料可取得、可再次使用
  - 程式可取得、可再次使用
  - 過程可複製、結果可重現
  - 人人可以做、不限您和我
- 典範： Gender Shades, by Joy Buolamwini & Timnit Gebru.



"At the time of evaluation, none of the [three] companies tested reported how well their computer vision products perform across gender, skin type, ethnicity, age or other attributes."

"The description of classification methodology lacked detail and there was no mention of what training data was used."

- "How well do IBM, Microsoft, and Face++ AI services guess the gender of a face?" http://gendershades.org
- Joy Buolamwini and Timnit Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. In Proc. of Machine Learning Research, 2018. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

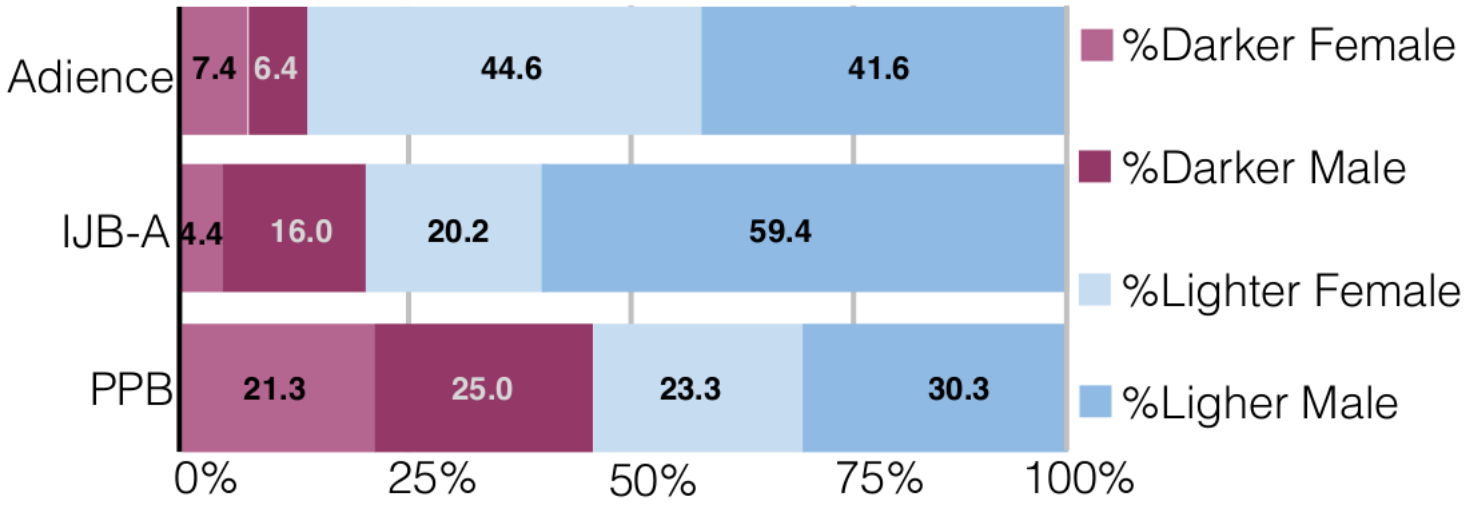


Figure 3: The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB.

- "How well do IBM, Microsoft, and Face++ AI services guess the gender of a face?" http://gendershades.org
- Joy Buolamwini and Timnit Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. In Proc. of Machine Learning Research, 2018. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

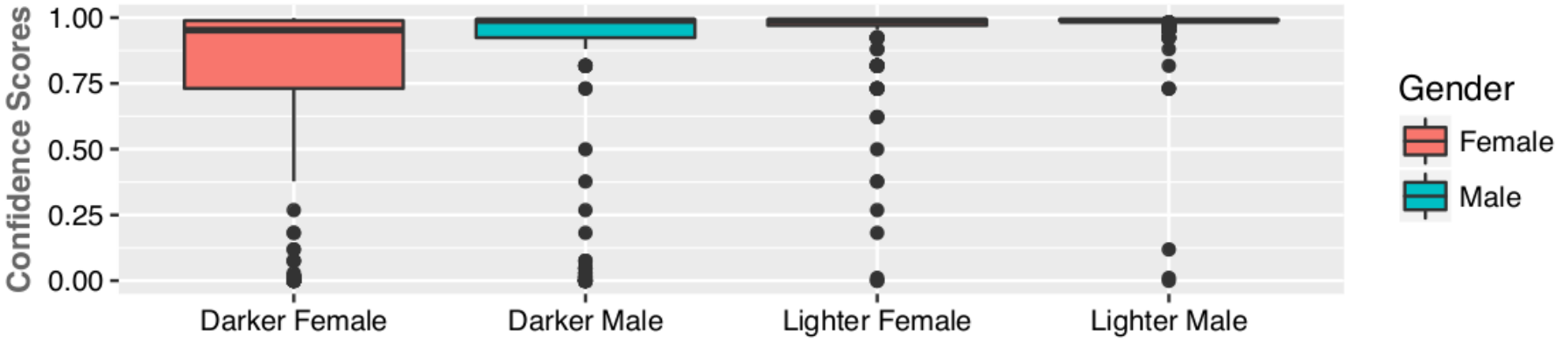


Figure 4: Gender classification confidence scores from IBM (IBM). Scores are near 1 for lighter male and female subjects while they range from $\sim 0.75 - 1$ for darker females.

- "How well do IBM, Microsoft, and Face++ AI services guess the gender of a face?" http://gendershades.org
- Joy Buolamwini and Timnit Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. In Proc. of Machine Learning Research, 2018. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf