

**Shapley values**

# **Aprendizaje Automático**

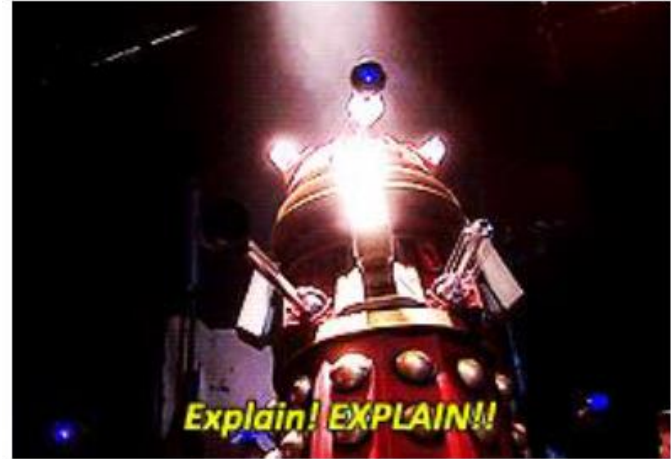
Juan David Martínez

[jdmartinev@eafit.edu.co](mailto:jdmartinev@eafit.edu.co)

# Agenda

- ML interpretable
- SHapley Additive exPlanations

# Estado actual de ML



# Usos



<https://www.tesla.com/videos/autopilot-self-driving-hardware-neighborhood-long>



NYPPost



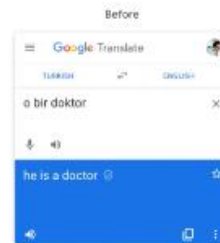
MIT Technology Review



DeepMind

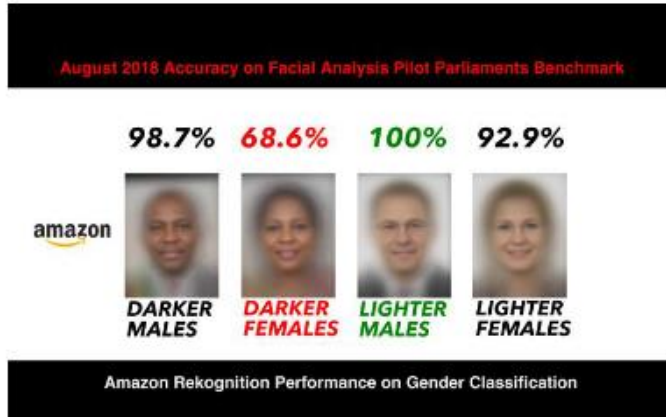


DeepMind



# Problemas

## Sesgos de los algoritmos



<https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced>

## Machine Learning can amplify bias.



- Data set: 67% of people cooking are women
- Algorithm predicts: 84% of people cooking are women

<https://www.infoq.com/presentations/unconscious-bias-machine-learning/>

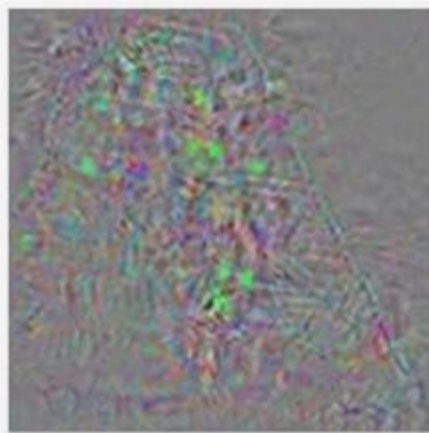
# Problemas

## Ejemplos adversarios

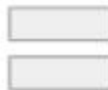


**Original image**

Temple (97%)



**Perturbations**



**Adversarial example**

Ostrich (98%)

# ML interpretable

## Tenemos varios problemas:

- No confiamos en los modelos
- No sabemos qué pasa en casos extremos
- Los errores pueden ser costosos/nocivos
- ¿Los modelos cometen errores similares a los de los humanos?
- ¿Cómo cambiamos el modelo si no da los resultados esperados?

**Una forma de lidiar con estos problemas es a través de la interpretabilidad**

# Shapley Additive exPlanations - SHAP

---

## A Unified Approach to Interpreting Model Predictions

---

**Scott M. Lundberg**

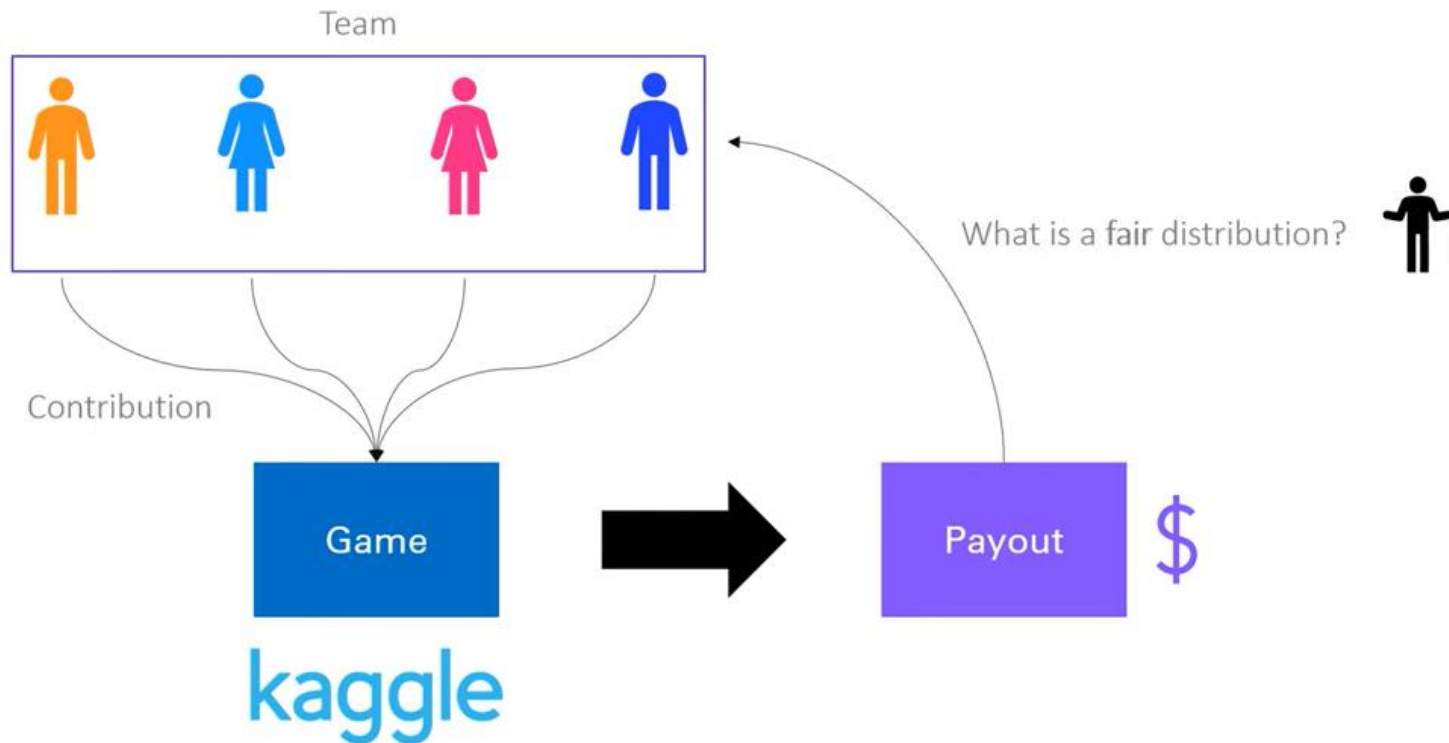
Paul G. Allen School of Computer Science  
University of Washington  
Seattle, WA 98105  
slund1@cs.washington.edu

**Su-In Lee**

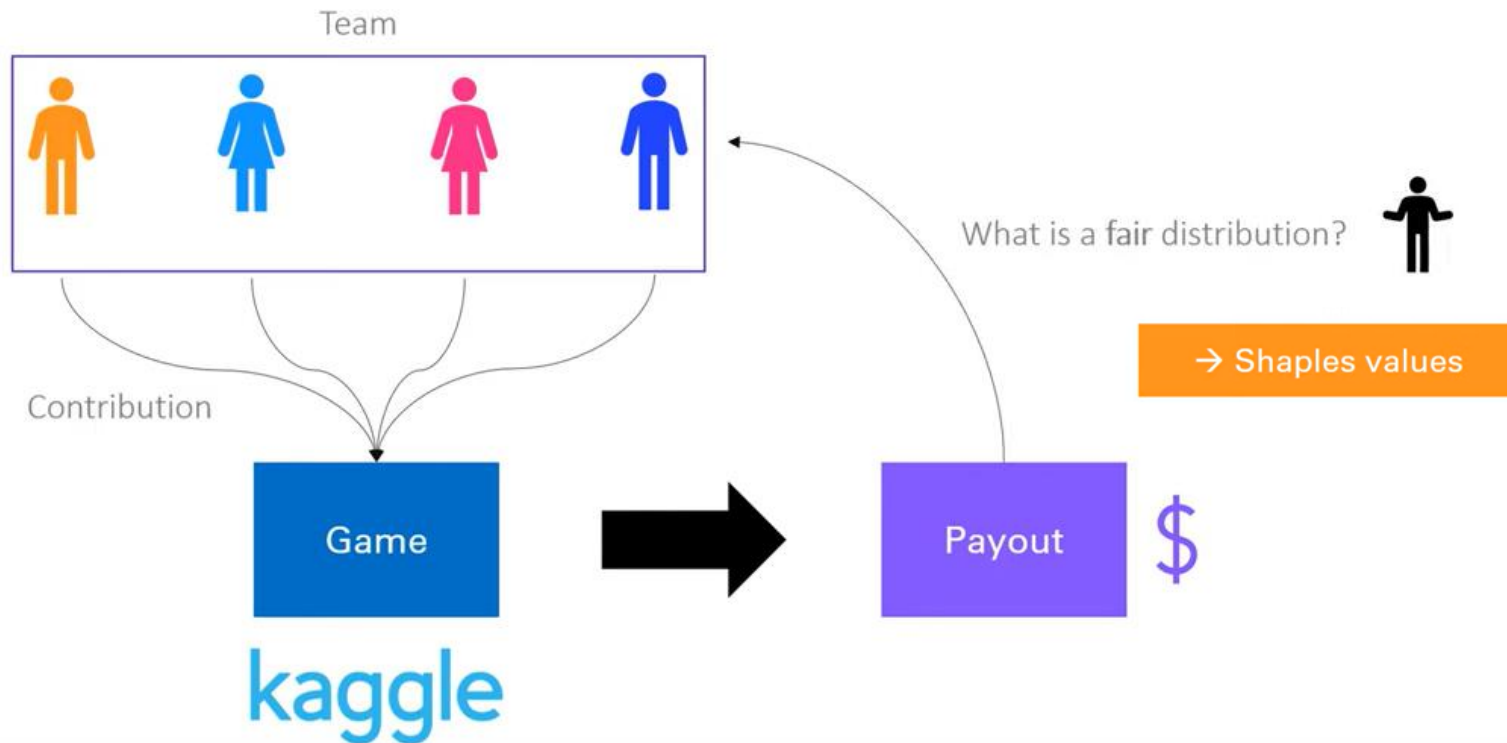
Paul G. Allen School of Computer Science  
Department of Genome Sciences  
University of Washington  
Seattle, WA 98105  
suinlee@cs.washington.edu



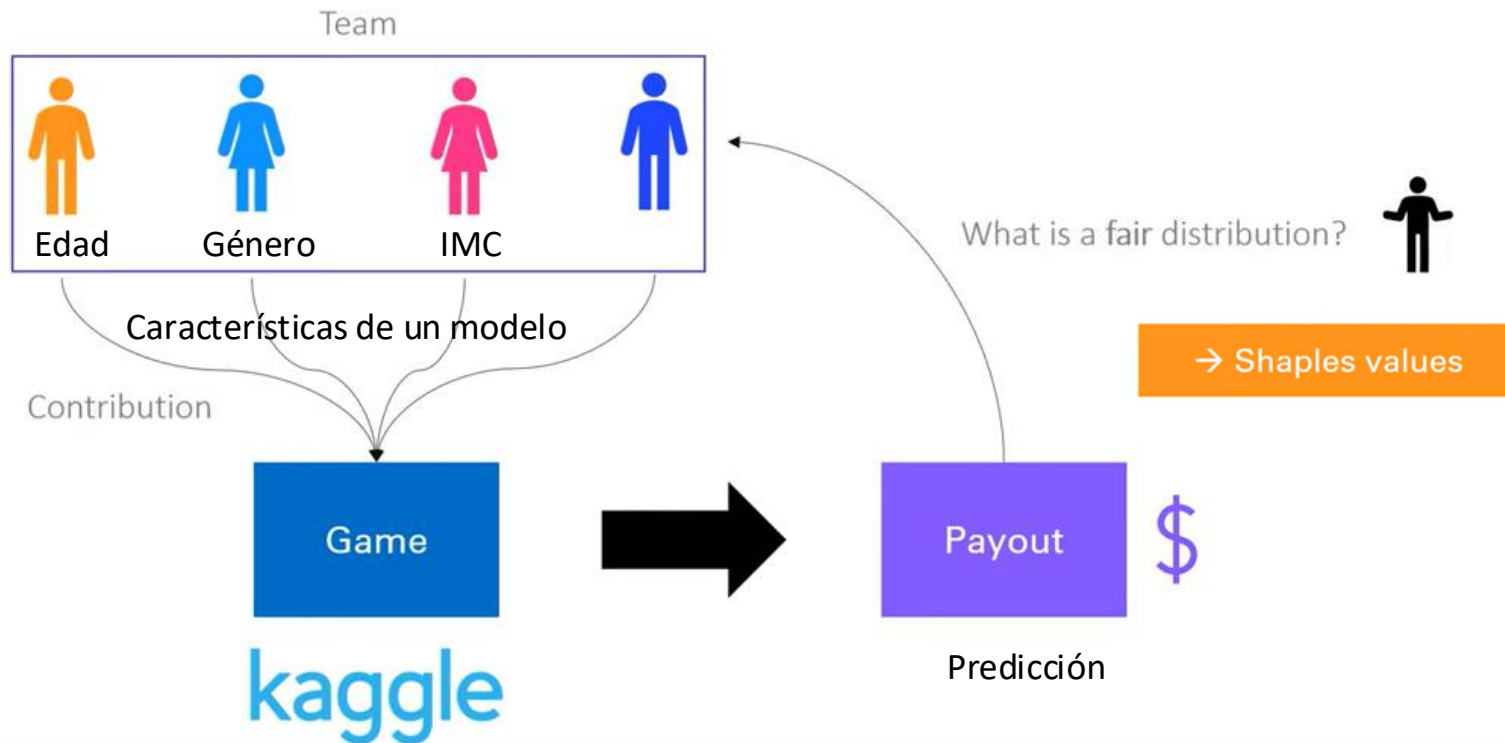
# Shapley Additive exPlanations - SHAP



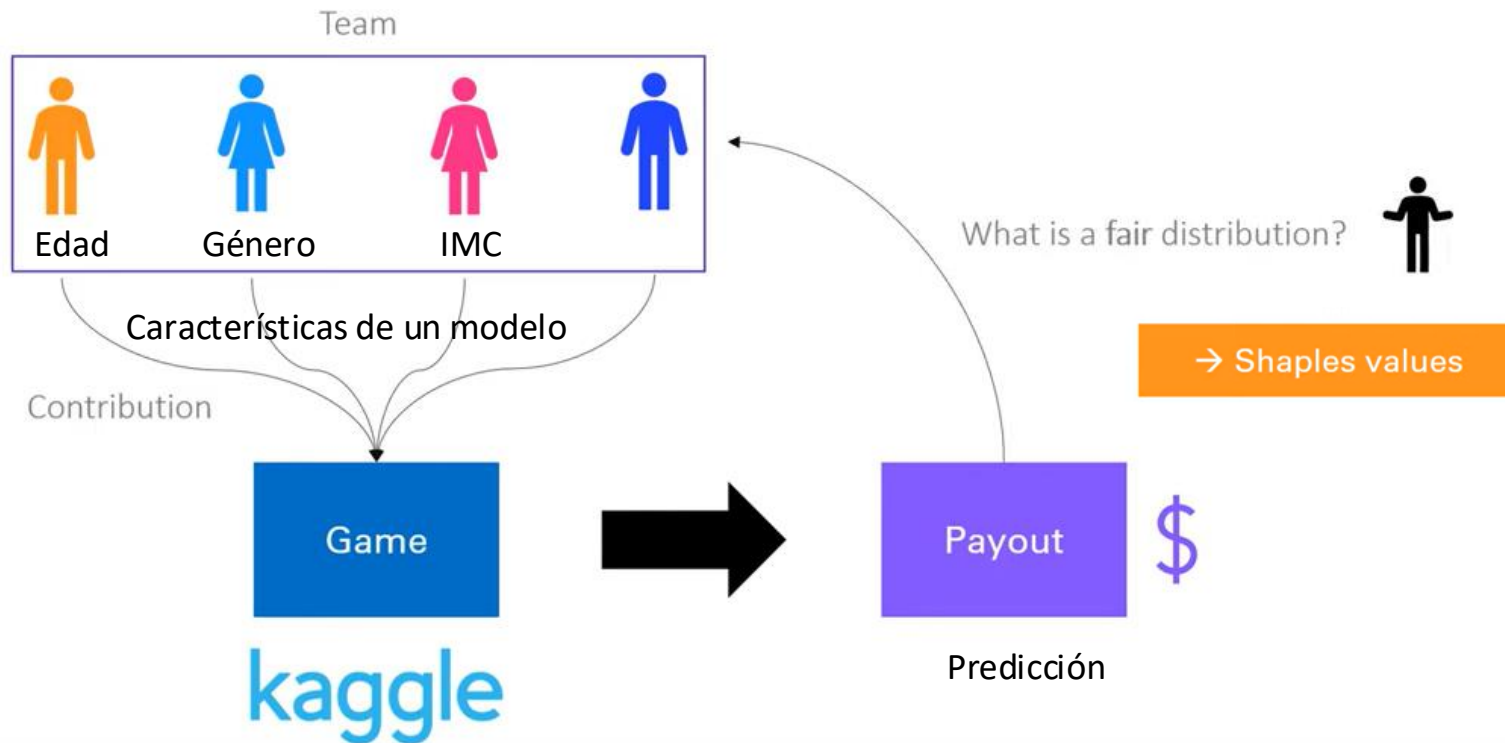
# Shapley Additive exPlanations - SHAP



# Shapley Additive exPlanations - SHAP



# Shapley Additive exPlanations - SHAP



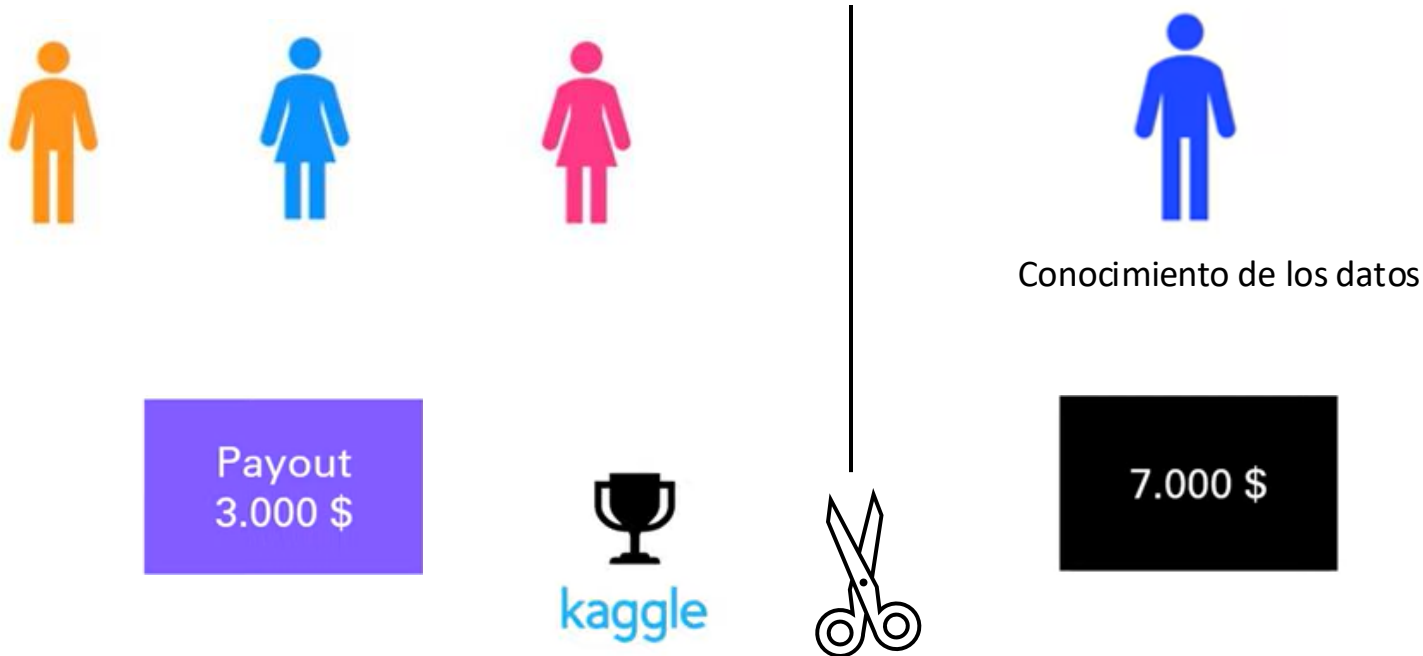
# Shapley Additive exPlanations - SHAP



Payout  
10.000 \$



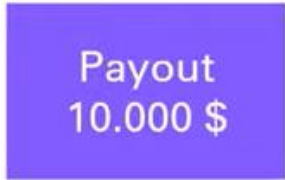
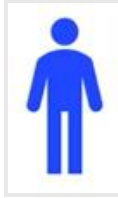
# Shapley Additive exPlanations - SHAP



# Shapley Additive exPlanations - SHAP



Experto ML



# Shapley Additive exPlanations - SHAP



Payout  
10.000 \$



Todas las combinaciones posibles: Contribución marginal de cada participante



# Shapley Additive exPlanations - SHAP

$$\phi_i(f, \mathbf{x}) = \sum_{\mathbf{z}' \subseteq \mathbf{x}'} \frac{|\mathbf{z}'|!(M - |\mathbf{z}'| - 1)!}{M!} [f(\mathbf{z}') - f(\mathbf{z}' \setminus i)]$$

$\mathbf{z}'$ : Subconjunto de características

$\mathbf{z}' \setminus i$ : Subconjunto de características sin la característica  $i$

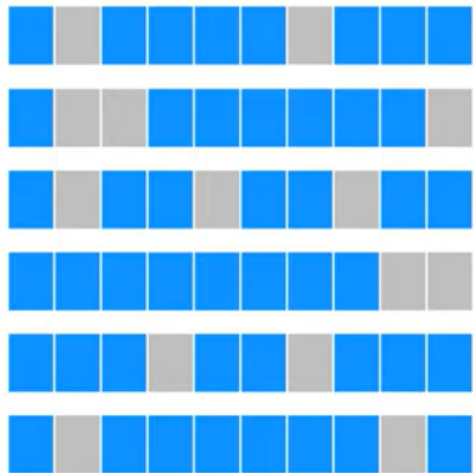
$\mathbf{x}'$	Age = 56	Gender = F	Body Mass Index = 30	Heart disease = yes	...
$\mathbf{z}'$	Age = 56	Gender = F	Body Mass Index = 30	Heart disease = yes	...

# Shapley Additive exPlanations - SHAP



# Shapley Additive exPlanations - SHAP

[https://github.com/deepfindr/xai-series/blob/master/03\\_shap.py](https://github.com/deepfindr/xai-series/blob/master/03_shap.py)



...

$$2^{10} = 1024$$

$2^n$  = total number of  
subsets of a set