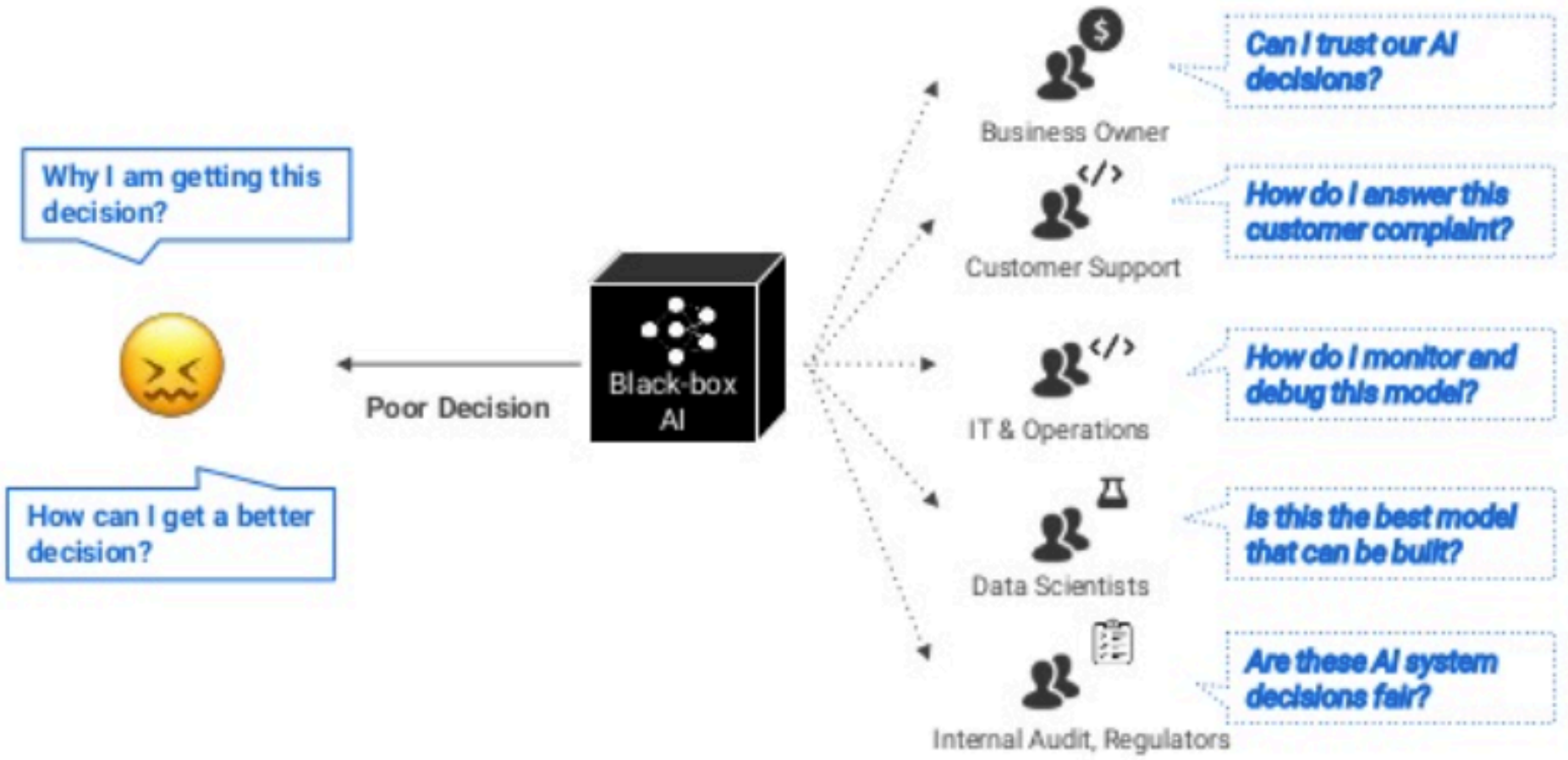
# Interpretabilidade em Modelos de ML

Luciano Barbosa

# Modelos Black Box



# Viés em Modelos

# Viés em Modelos

Bloomberg Businessweek

Apple Card's Gender-Bias Claims Look Familiar to Old-School Banks

Updated on November 12, 2019, 4:23 AM

MIT News

Study finds gender and skin-type bias in commercial AI systems

Feb 12, 2018

BBC NEWS

Tay: Microsoft issues apology over racist chatbot fiasco

Sep 22, 2017

Missouri S&T News and Research

After Uber, Tesla incidents, can artificial intelligence be trusted?

Apr 10, 2018

Guilty! AI Is Found to Perpetuate Biases in Jailing

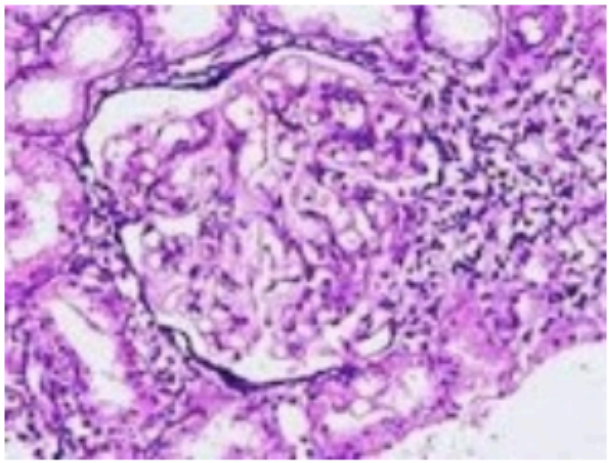
1 day ago

# Para que Interpretabilidade?

- Verifcar se o modelo funciona como o esperado

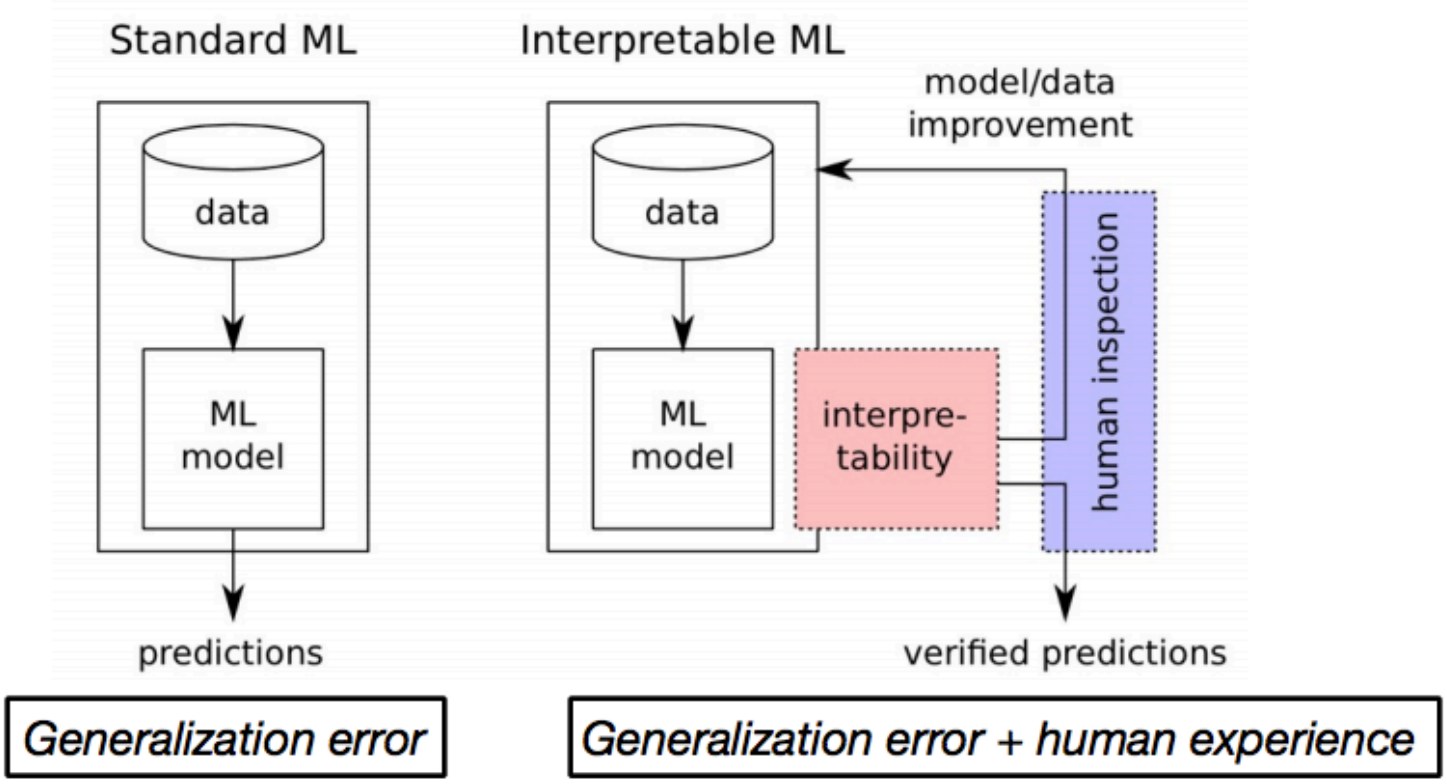
*"Autonomous car crashes, because it wrongly recognizes …"*

*"AI medical diagnosis system misclassifies patient's disease …"*



# Para que Interpretabilidade?

- Verifcar problemas e melhorar modelos



# Para que Interpretabilidade?

- Verifcar problemas e melhorar modelos

Top label: **"clog"**

Why did the network label this image as **"clog"**?

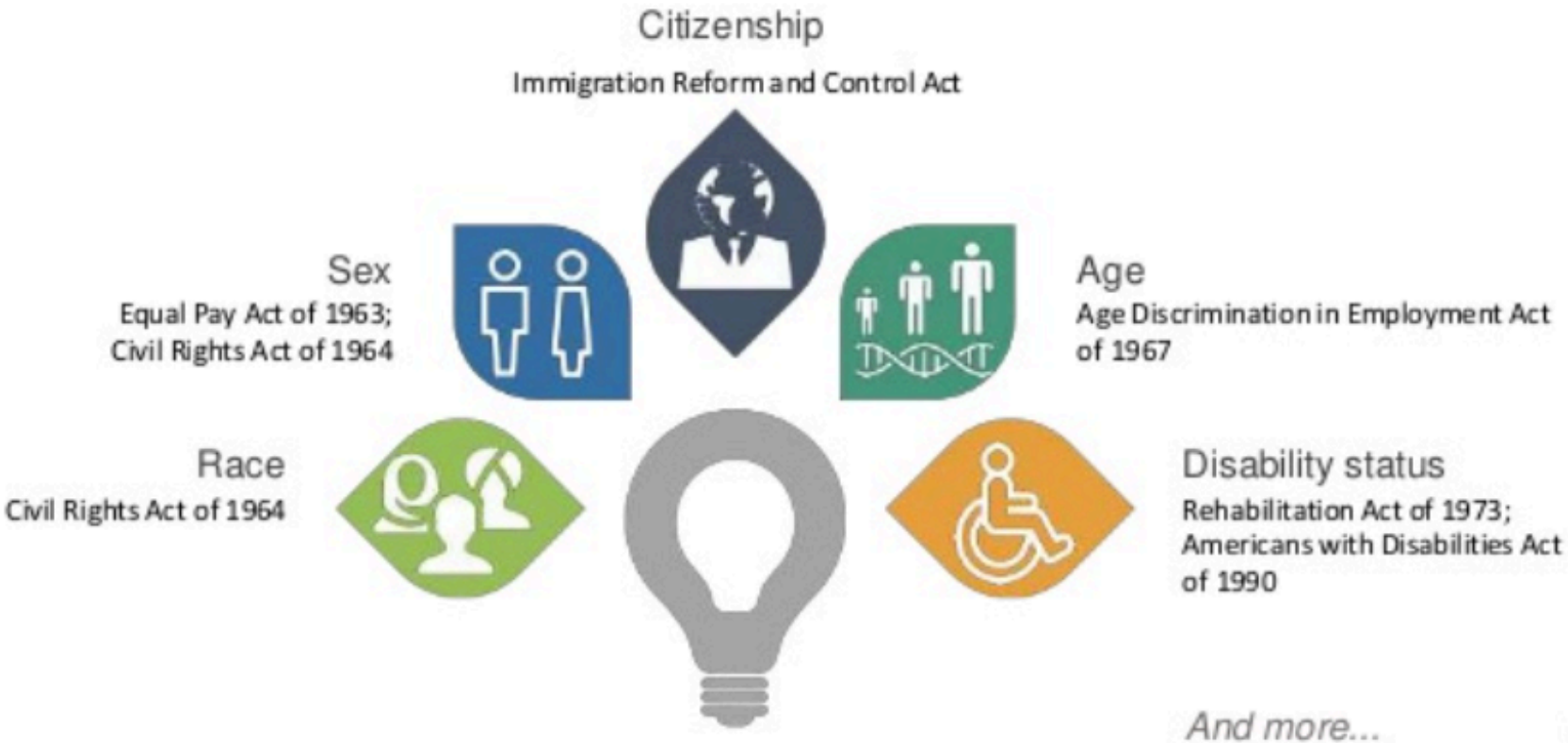
# Para que Interpretabilidade?

- Aprender novos insights

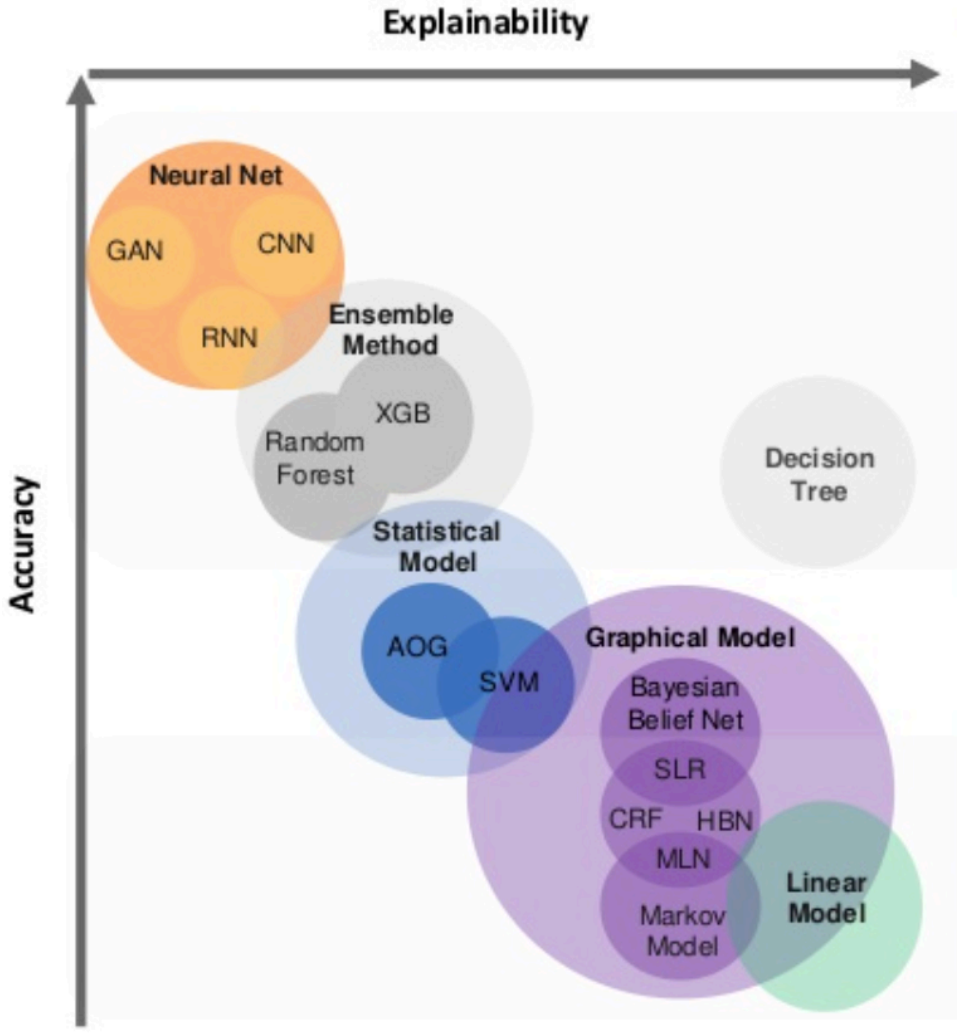
*“It's not a human move. I've never seen a human play this move.” (Fan Hui)*

# Para que Interpretabilidade?

- Confirmade com legislação



# Acurácia X Explicabilidade



# Estratégias

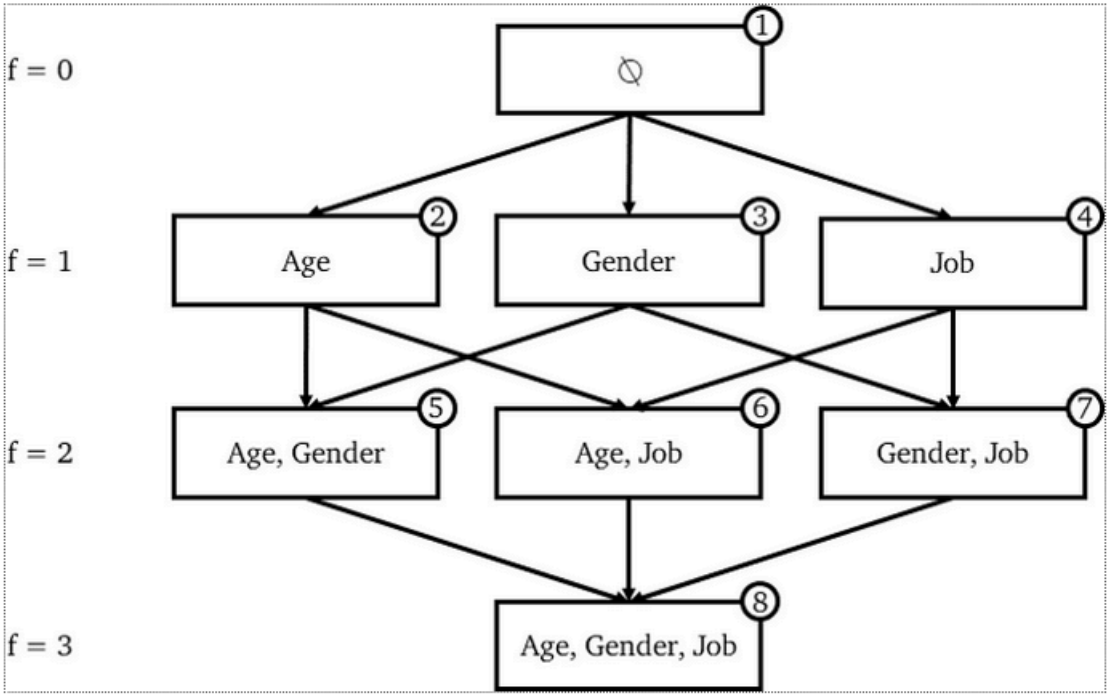
- Explicação dado um modelo de ML
  - Predição individual
  - Predição global
- Construção de um modelo interpretável
  - Decision trees, regressão linear etc

# SHAP: SHapley Additive exPlanations

- Predição individual
- Utiliza teoria dos jogos
- Jogo: saída do modelo
- Jogadores: features do modelo
- Contribuição de cada feature (jogador)para o jogo (predição)

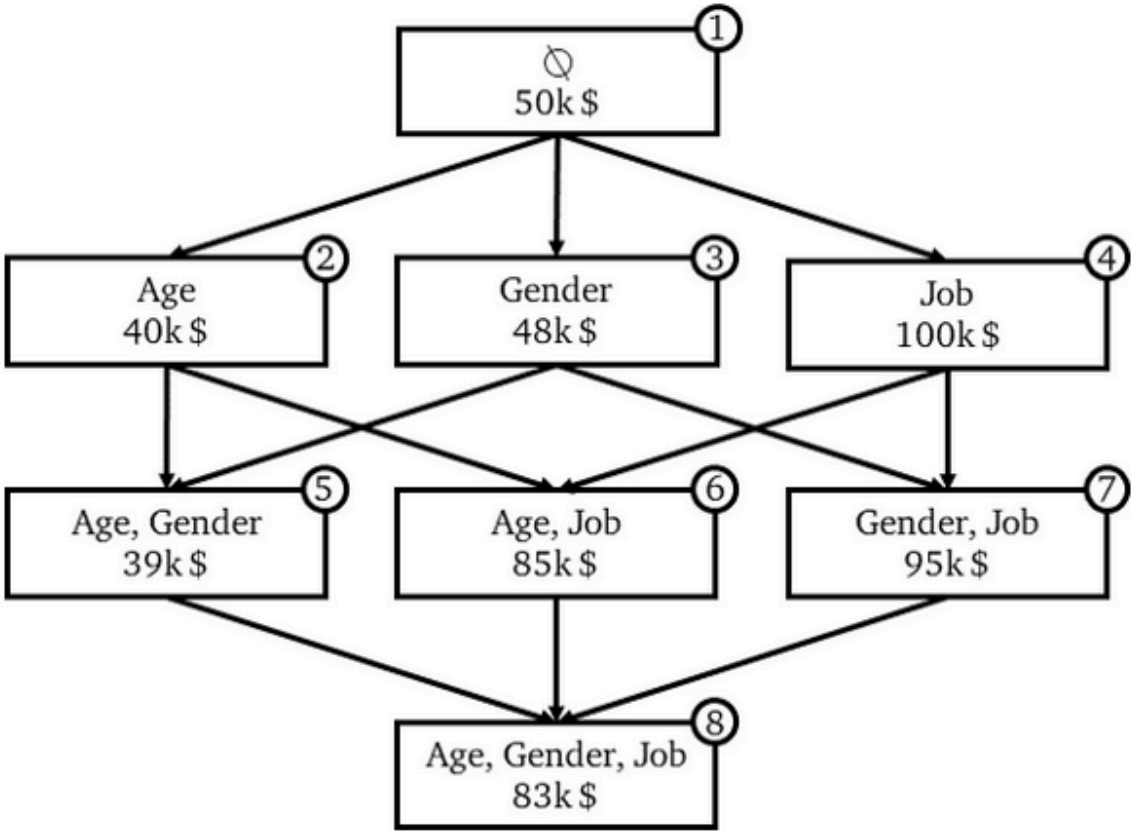
# SHAP: Exemplo

- Cada possível coalizão de jogadores determina a importância de um jogador
- Predizer renda a partir de idade, sexo e emprego



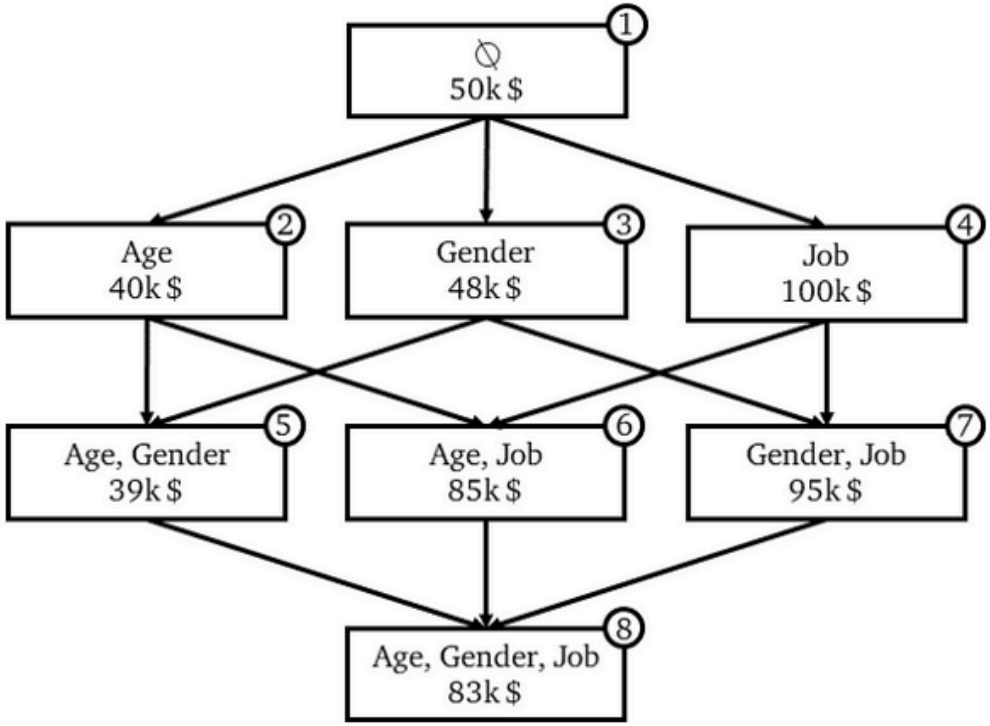
# SHAP: Exemplo

- Treina um modelo para cada coalizão (mesmo conjunto de treinamento e hiper-parâmetros)

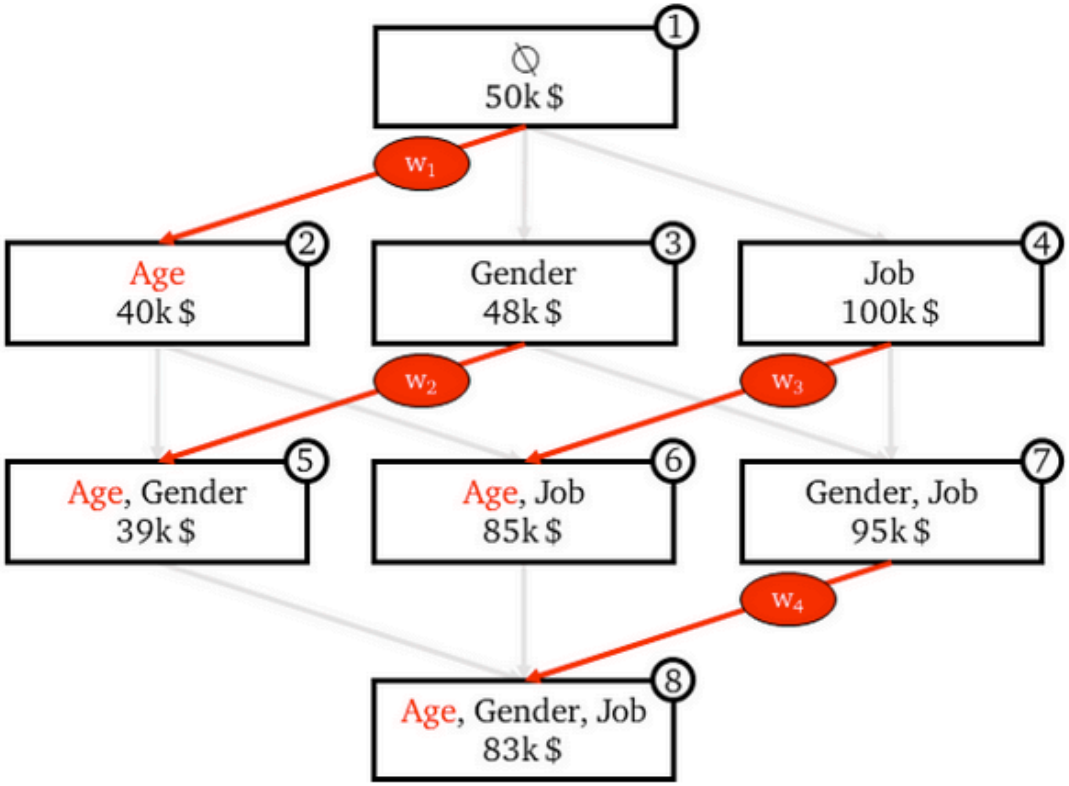


# SHAP: Exemplo

- Contribuição marginal (aresta): diferença entre predições de dois nós conectados é o efeito da feature adicional



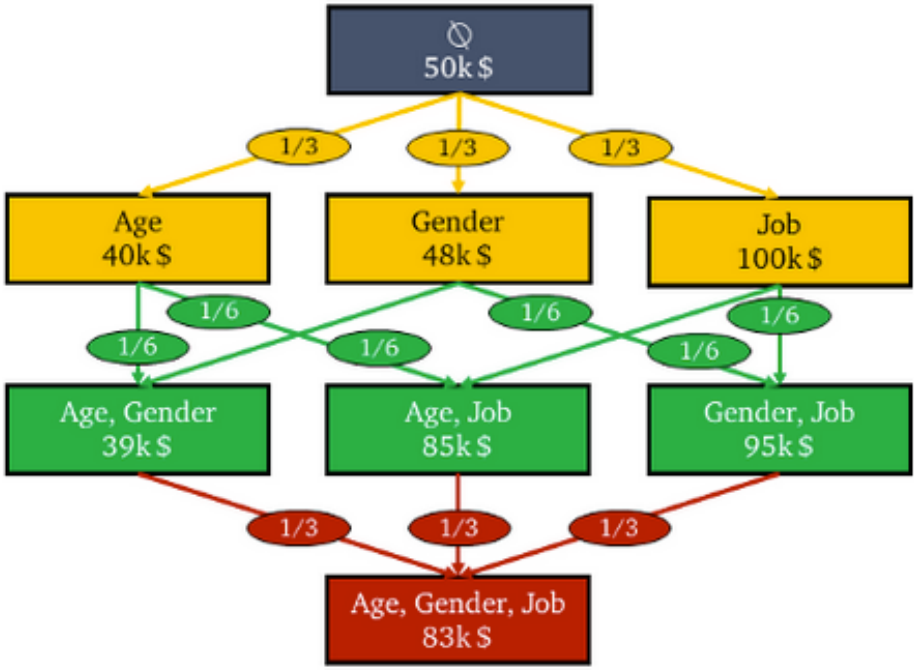
# Contribuição da idade



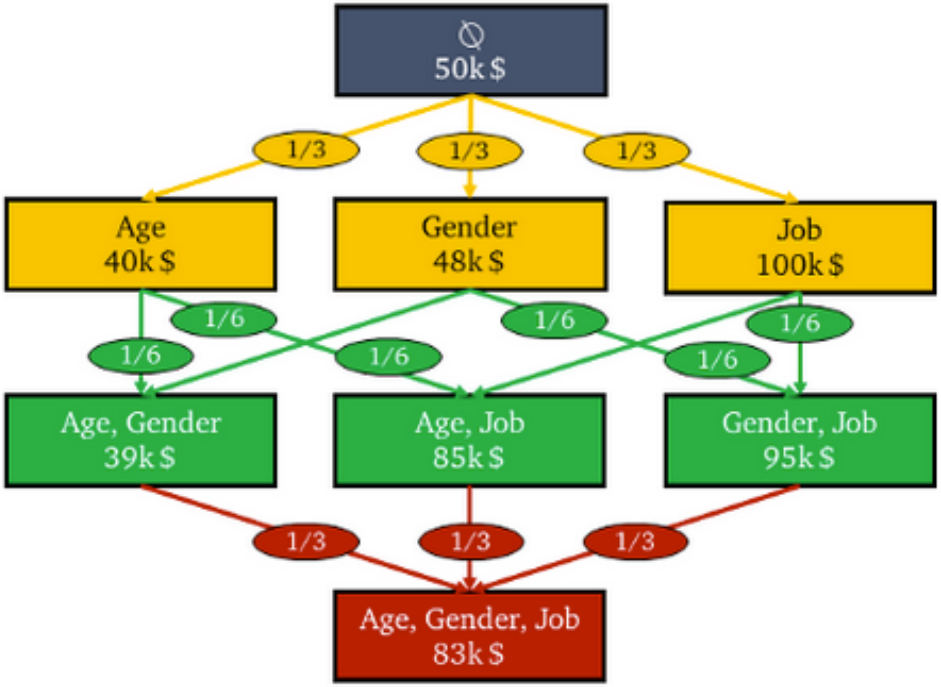

$$SHAP_{Age}(x_0) = w_1 \times MC_{Age,\{Age\}}(x_0) + w_2 \times MC_{Age,\{Age,Gender\}}(x_0) + w_3 \times MC_{Age,\{Age,Job\}}(x_0) + w_4 \times MC_{Age,\{Age,Gender,Job\}}(x_0)$$

# Pesos

- Soma dos pesos para cada linha tem que ser igual: $w_1 = w_2 + w_3 = w_4$
- Os pesos na mesma linha são iguais: $w_1 = 1/3$; $w_2 = 1/6$;



# Resultado para Idade




$$SHAP_{Age}(x_0) = \frac{1}{3} \times (-10k\$) + \frac{1}{6} \times (-9k\$) + \frac{1}{6} \times (-15k\$) + \frac{1}{3} \times (-12k\$)$$

$$= -11.33k\$$$

# Resultado Final

$$SHAP\_Age(xo) = -11.33k\,\$$$

$$SHAP\_Gender(xo) = -2.33k\,\$$$

$$SHAP\_Job(xo) = +46.66k\,\$$$