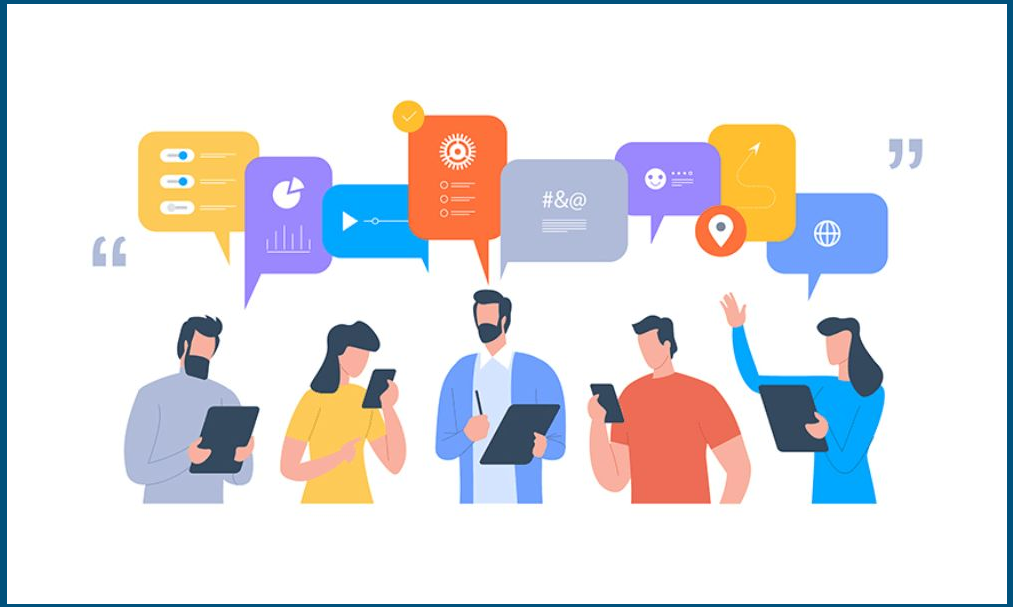
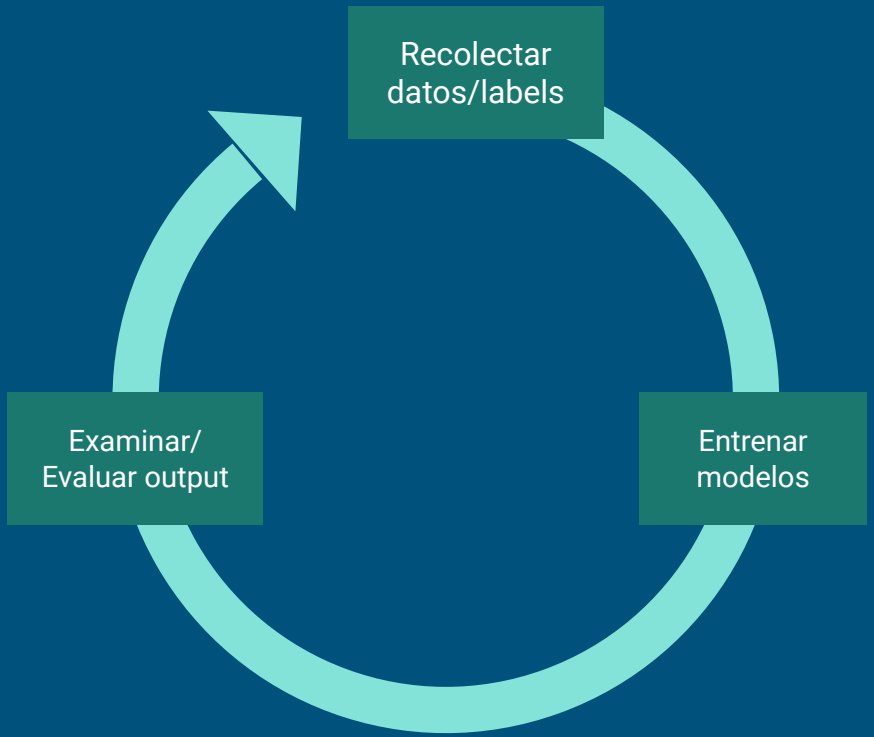
Taller básico de PLN:
Sesgos e Interpretabilidad

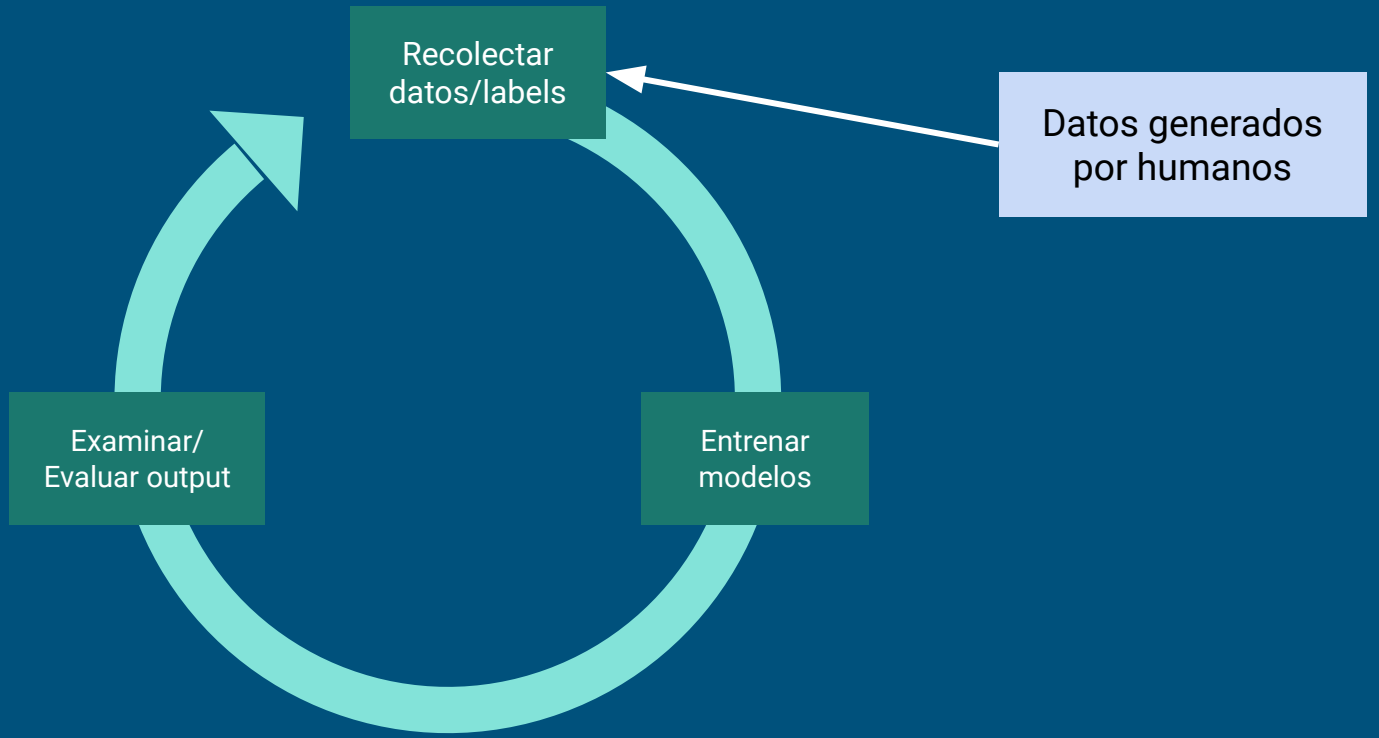
Ana Valeria González
Ximena Gutiérrez-Vasques

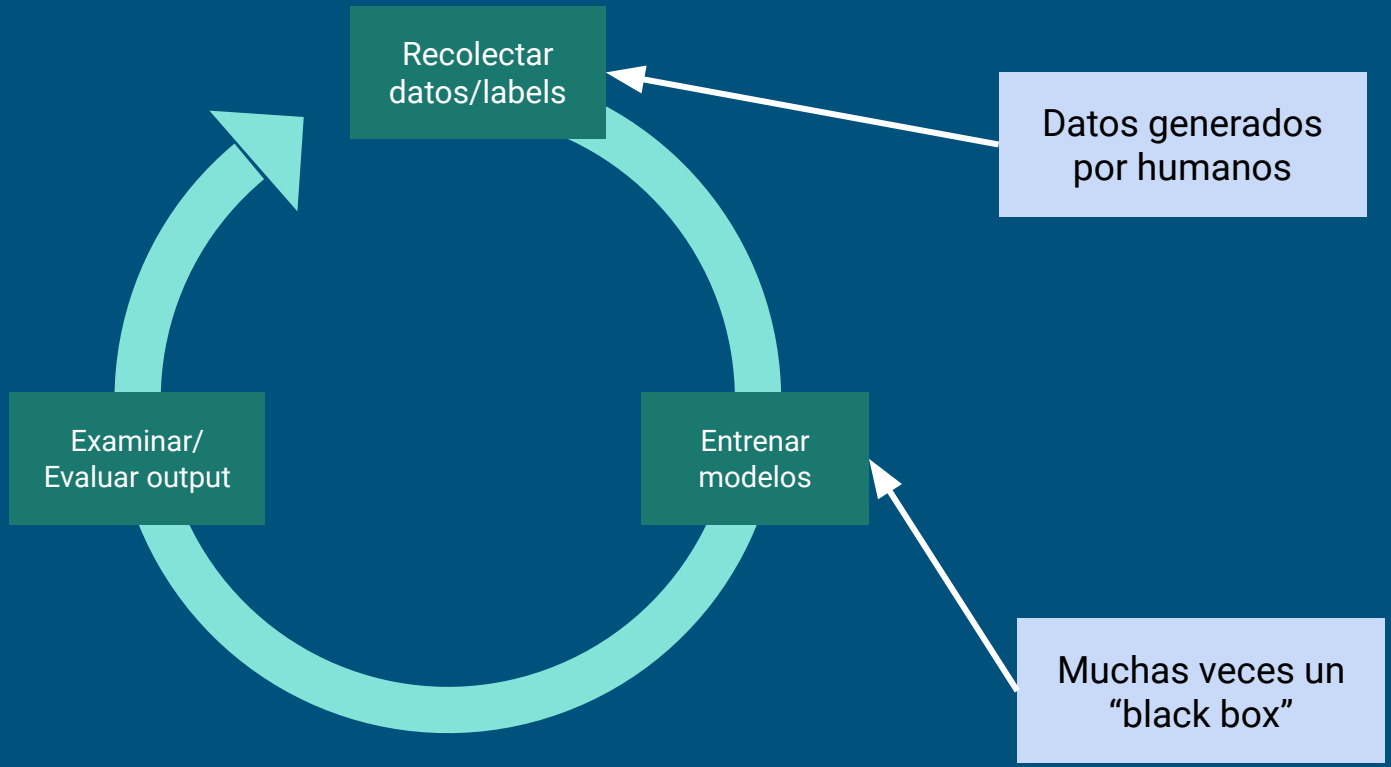
Sistemas de aprendizaje automático utilizan datos generados por humanos

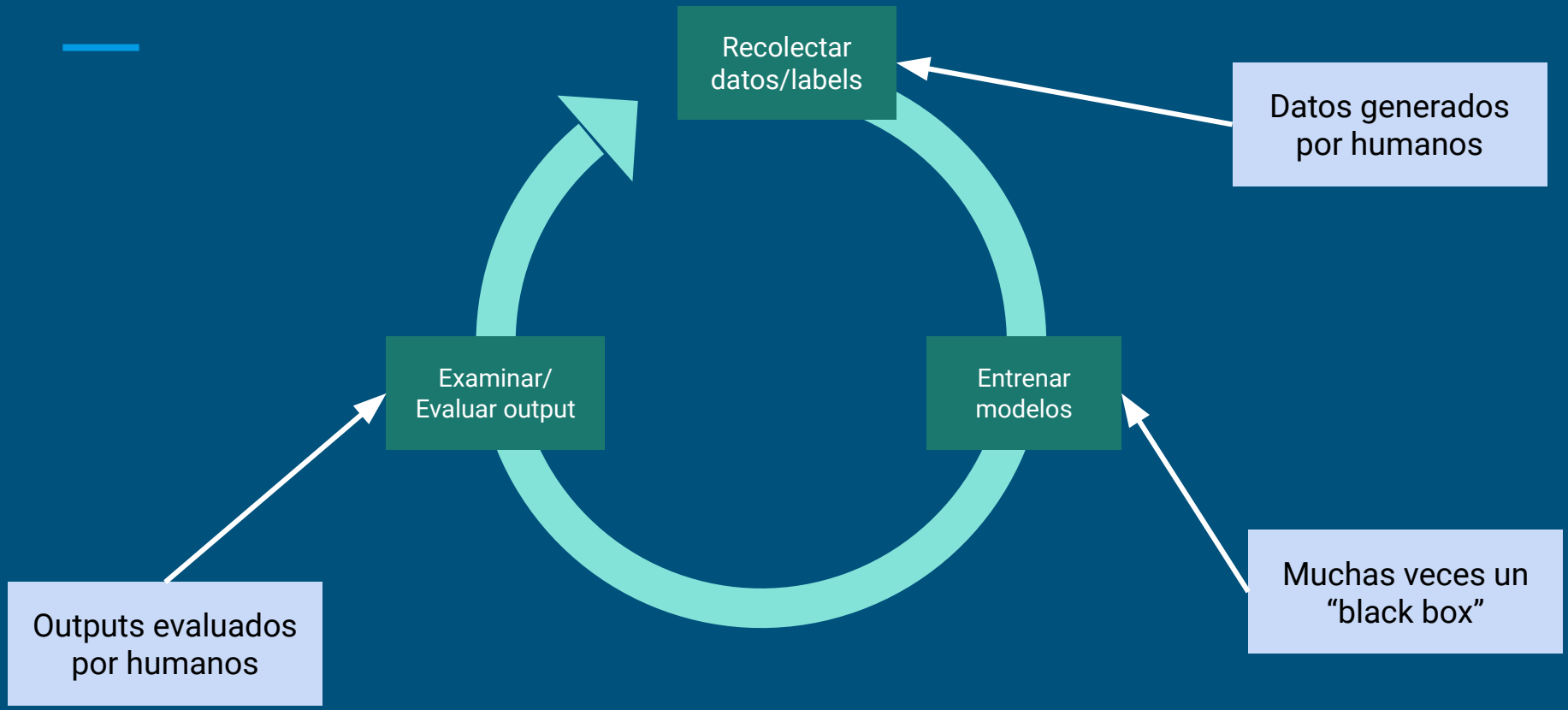


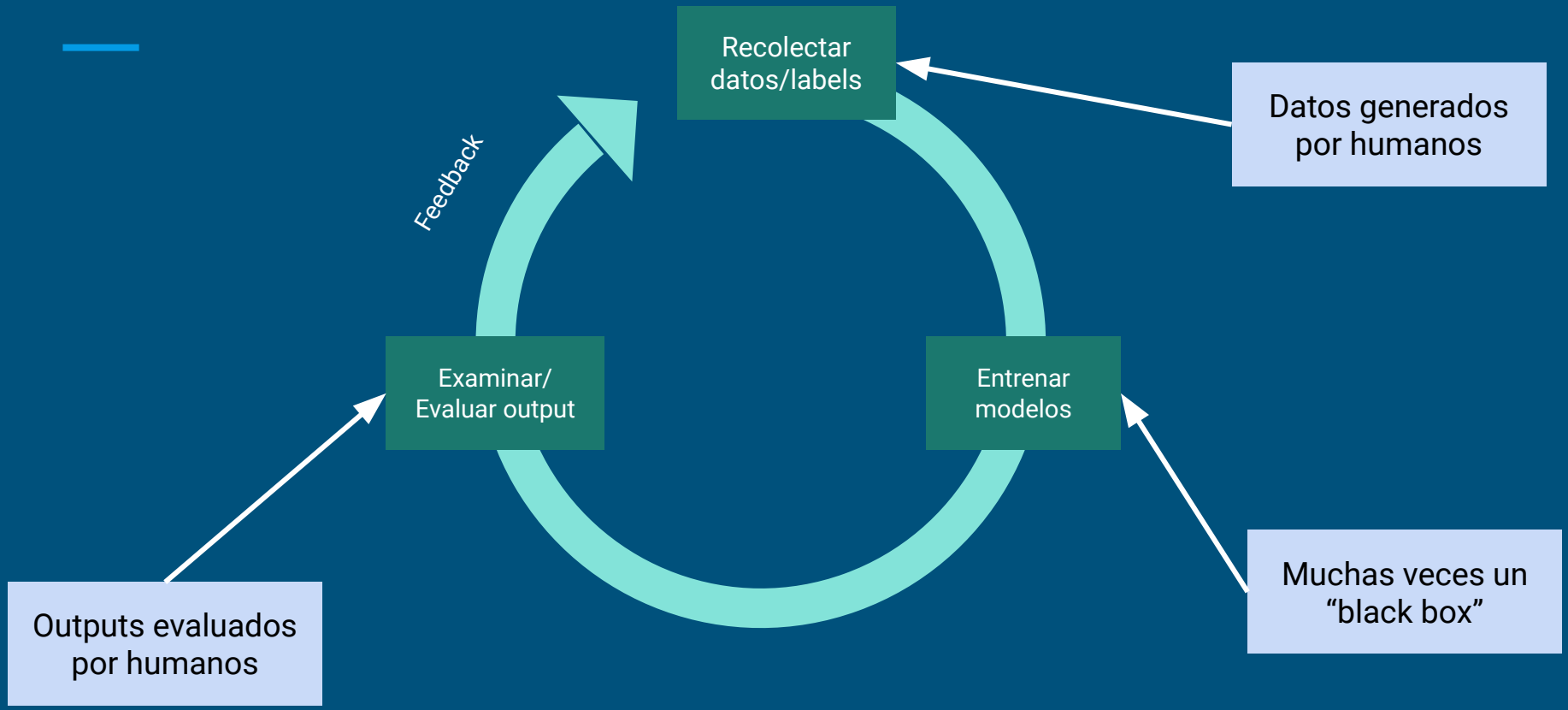
¿Cuáles son las implicaciones éticas ?


Recolectar datos/labels
Entrenar modelos
Examinar/ Evaluar output

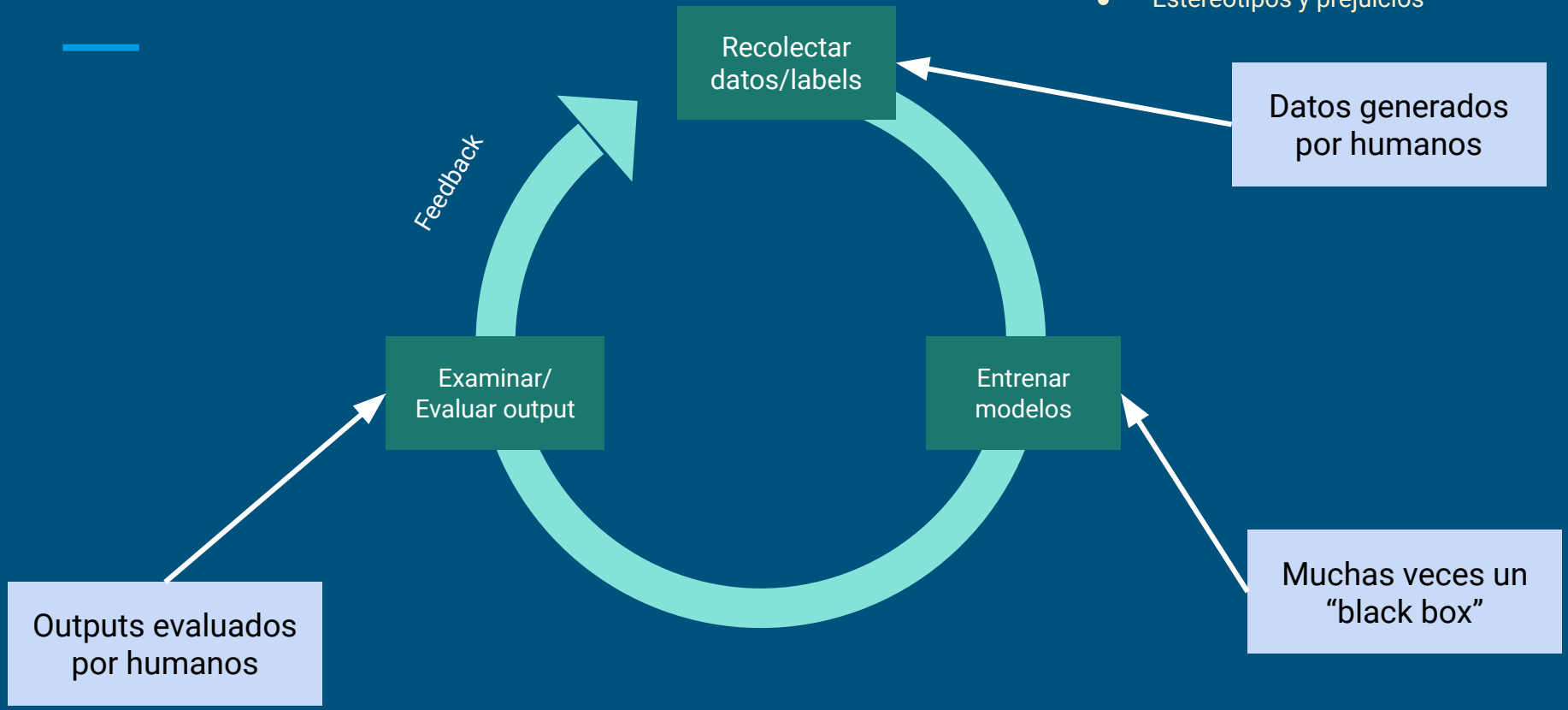

Recolectar datos/labels
Datos generados por humanos
Entrenar modelos
Examinar/ Evaluar output

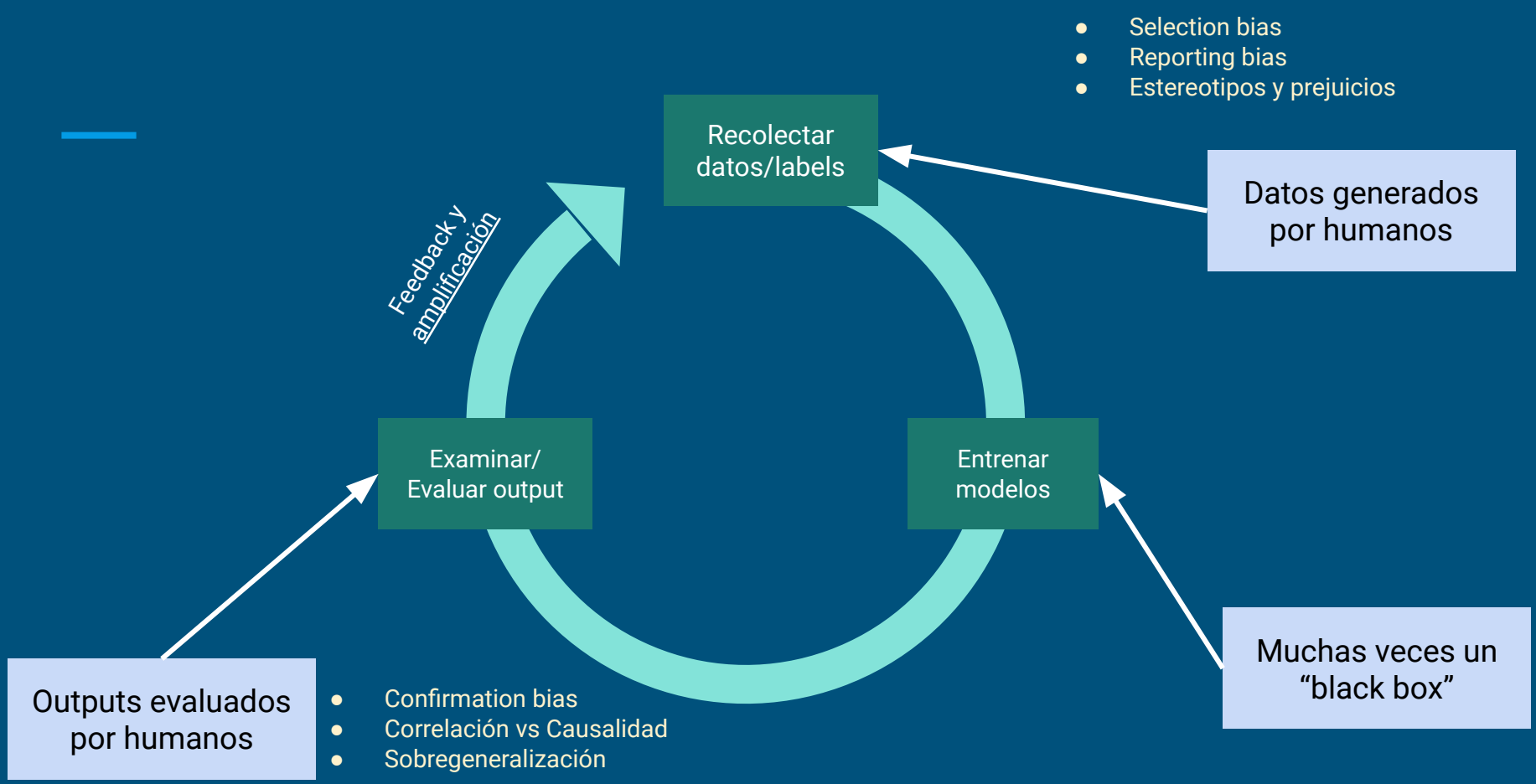

Recolectar datos/labels
Entrenar modelos
Examinar/ Evaluar output
Datos generados por humanos
Muchas veces un “black box”


Recolectar datos/labels
Datos generados por humanos
Entrenar modelos
Muchas veces un “black box”
Examinar/ Evaluar output
Outputs evaluados por humanos


Recolectar datos/labels
Datos generados por humanos
Feedback
Examinar/ Evaluar output
Entrenar modelos
Outputs evaluados por humanos
Muchas veces un “black box”

- Selection bias
- Reporting bias
- Estereotipos y prejuicios




Recolectar datos/labels
Entrenar modelos
Examinar/ Evaluar output
Feedback y amplificación
Datos generados por humanos
Selection bias
Reporting bias
Estereotipos y prejuicios
Muchas veces un “black box”
Outputs evaluados por humanos
Confirmation bias
Correlación vs Causalidad
Sobregeneralización

Outline de esta plática

- Sesgos en la recolección de datos
- Sesgos en la interpretación de los modelos
- La importancia de la interpretabilidad en NLP/Machine learning

Bias en la recolección de datos

Reporting Bias

(Gordon and Van Durme, 2013)

Palabra	Frecuencia
Head	18 millones
Eyes	18 millones
Arms	6 millones
Brain	3 millones
Liver	250 mil
Kidney	180 mil
Spleen	47 mil
Pancreas	24 mil
Gallbladder	17 mil

Reporting Bias

(Gordon and Van Durme, 2013)

Palabra	Frecuencia
Head	18 millones
Eyes	18 millones
Arms	6 millones
Brain	3 millones
Liver	250 mil
Kidney	180 mil
Spleen	47 mil
Pancreas	24 mil
Gallbladder	17 mil

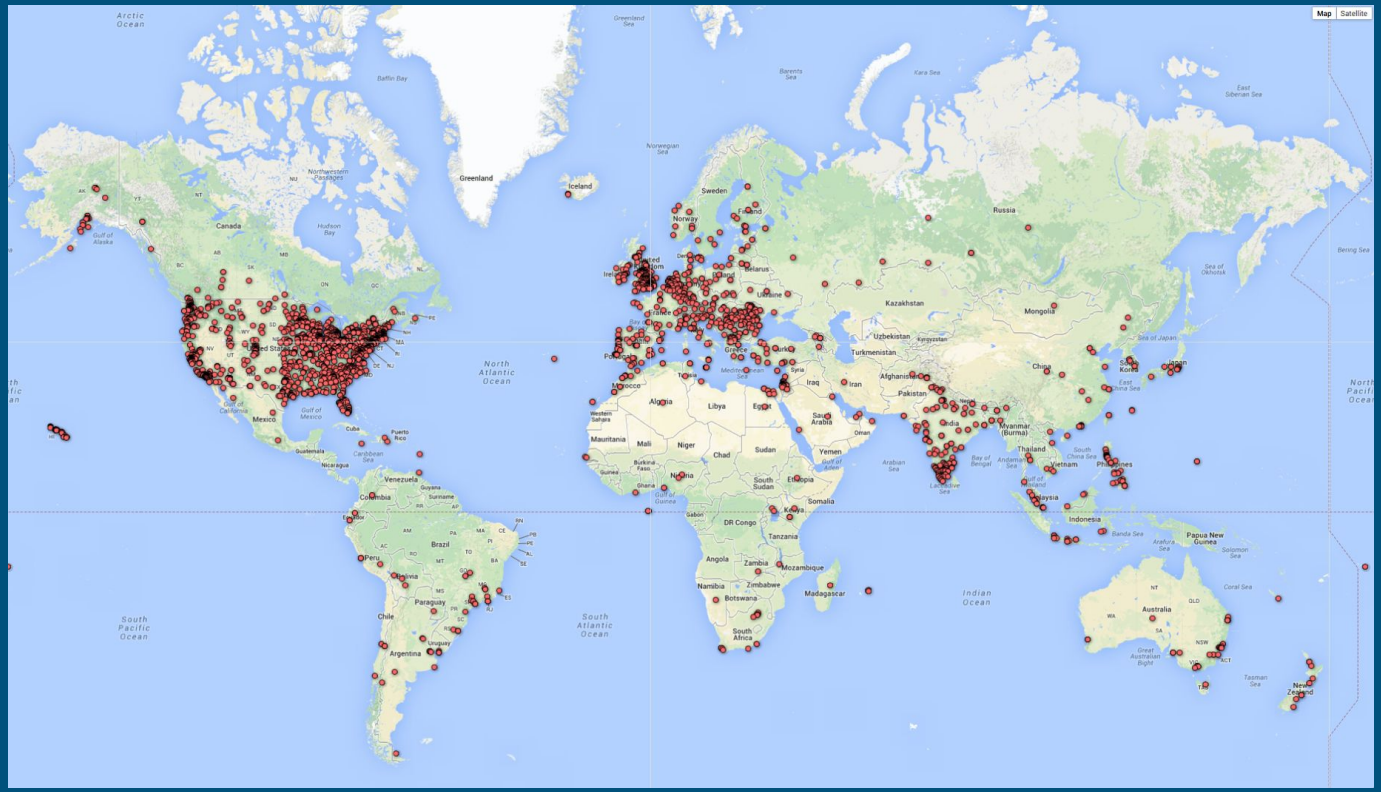
Reporting Bias

(Gordon and Van Durme, 2013)

La cabeza es más común que el páncreas?

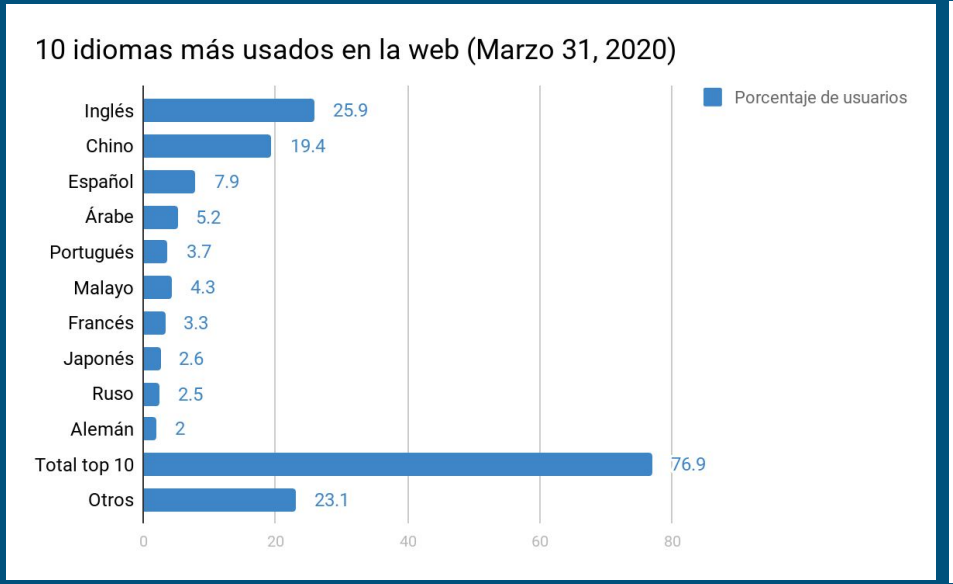
Palabra	Frecuencia
Head	18 millones
Eyes	18 millones
Arms	6 millones
Brain	3 millones
Liver	250 mil
Kidney	180 mil
Spleen	47 mil
Pancreas	24 mil
Gallbladder	17 mil

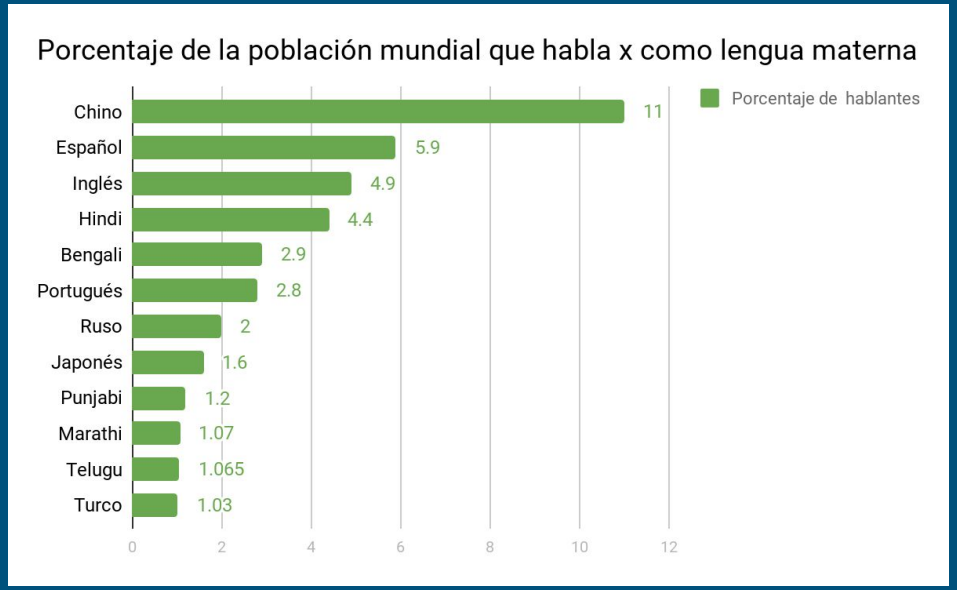
Selection Bias: la selección de grupos no refleja a la población



Mechanical turkers: http://turktools.net/crowdsourcing/

Selection Bias





https://www.ethnologue.com/

https://www.internetworldstats.com/stats7.htm

Selection Bias: ejemplos

- Los hombres son sobrerrepresentados en conversaciones de twitter
- La mayoría de los contribuidores de Wikipedia son hombres
- Muy pocas biografías sobre mujeres en Wikipedia

https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia

Estereotipos y representaciones negativas

Podemos tener la misma cantidad de datos para cada grupo, pero ciertos grupos pueden ser representados de forma negativa o estereotípica

- Páginas de wikipedia sobre mujeres con tonos sexistas
- Los pocos artículos sobre historia de África tienden a perpetuar imágenes negativas

https://en.wikipedia.org/wiki/Racial_bias_on_Wikipedia

https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia

Bias en los datos = Bias en las representaciones vectoriales

Bolukbasi et al. ***Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.*** NIPS (2016)

Garg et al. ***Word embeddings quantify 100 years of gender and ethnic stereotypes.*** PNAS. (2018)

Manzini et al. ***Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings.*** NAACL (2019).

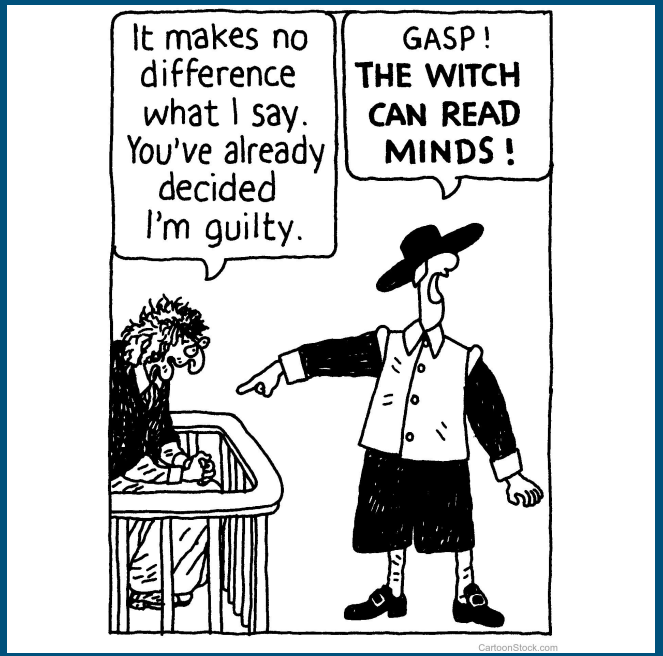
Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

Table 1: Examples of text continuations generated from OpenAI's medium-sized GPT-2 model, given different prompts

(Sheng et al. 2019)

Bias al interpretar modelos

Confirmation Bias



“Tendencia a favorecer, buscar, interpretar, y recordar la información que confirma las propias creencias o hipótesis, dando desproporcionadamente menos consideración a posibles alternativas.”

Ejemplo: confirmation bias

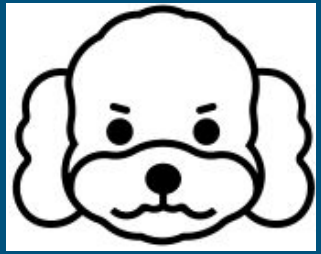
Una ingeniera está creando un modelo que predice la agresividad de los perros en función de distintas características (altura, peso, raza, entorno).

La ingeniera fue agredida por un poodle cuando era niña y desde ese entonces está segura que los poodles son agresivo.

La predicción del modelo indica que los poodles son dóciles.

La ingeniera "*mejora*" el procesamiento de los datos y los parámetros del modelo hasta que este arroja el resultado "*correcto*": los poodles son agresivos.

Creencias previas



Correlación vs Causalidad

La correlación implica asociación, pero no causalidad...

A la inversa, la causalidad implica asociación, pero no correlación.

Muchísimos más tipos de sesgos

https://developers.google.com/machine-learning/glossary

Interpretabilidad

Varias definiciones:

1. El grado en el que los humanos entienden la causa de cierta decisión
2. El grado en el que los humanos pueden predecir el resultado final del modelo

Hoy solo hablaremos de la importancia de la interpretabilidad

https://christophm.github.io/interpretable-ml-book/

Diagnosticar biases

Podemos mejorar sistemas automáticos y la formulación de ciertos problemas si tenemos más información sobre cómo toma decisiones el sistema

The Guardian, 2015

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

▲ Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates. Photograph: Brian Snyder/Reuters

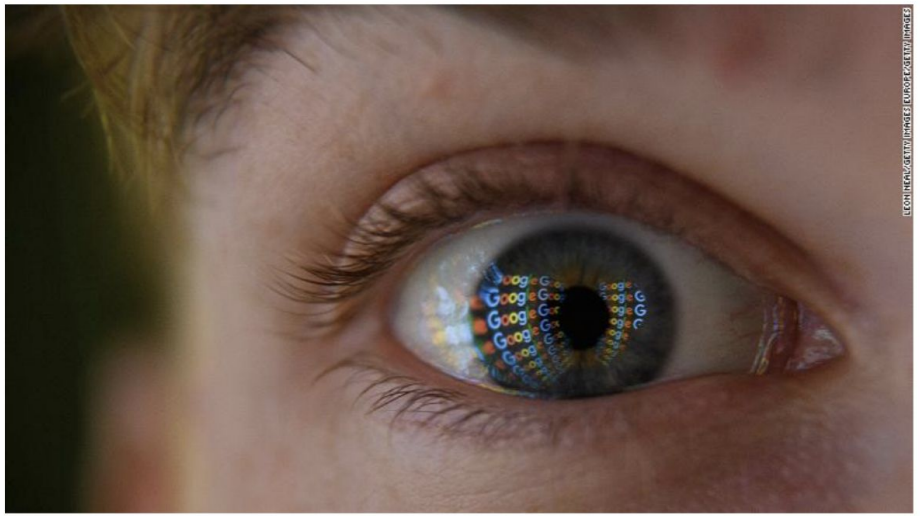
Confianza en los sistemas de aprendizaje automático

Para confiar en una decisión, los humanos necesitan entender cómo el sistema tomó esa decisión

Would you trust an algorithm to diagnose an illness?

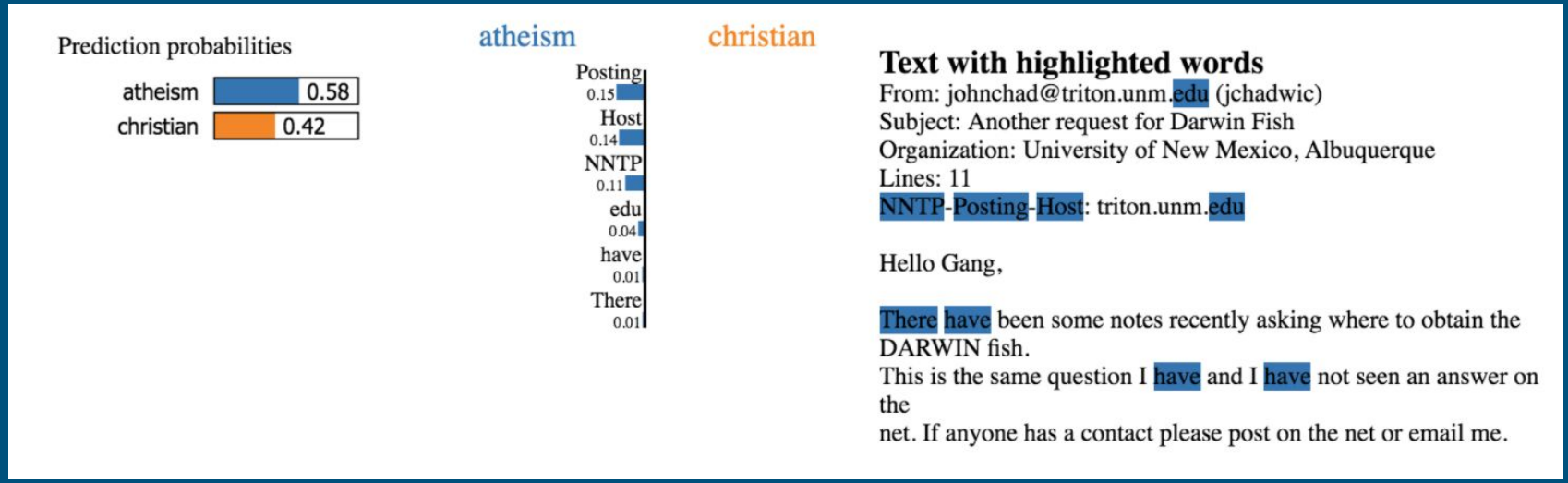
By Nell Lewis, CNN Business

Updated 0903 GMT (1703 HKT) July 15, 2019

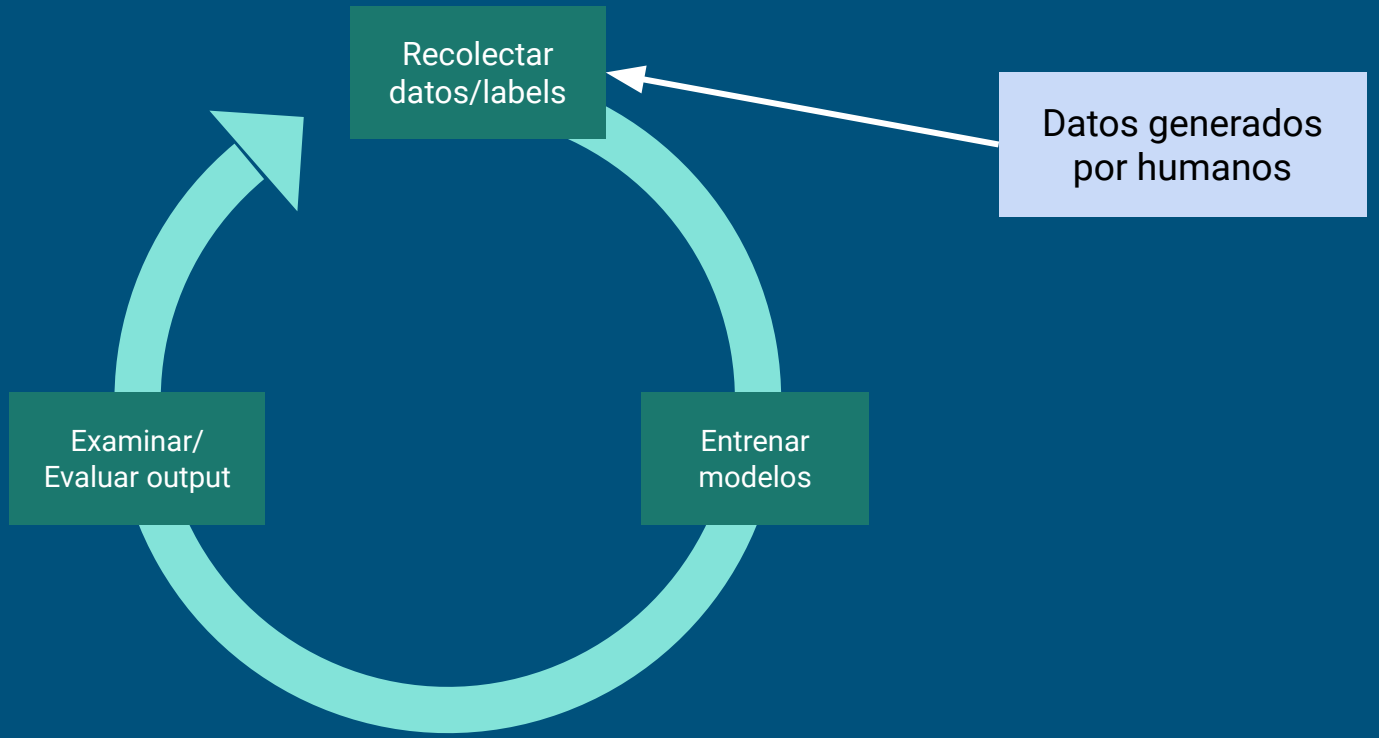


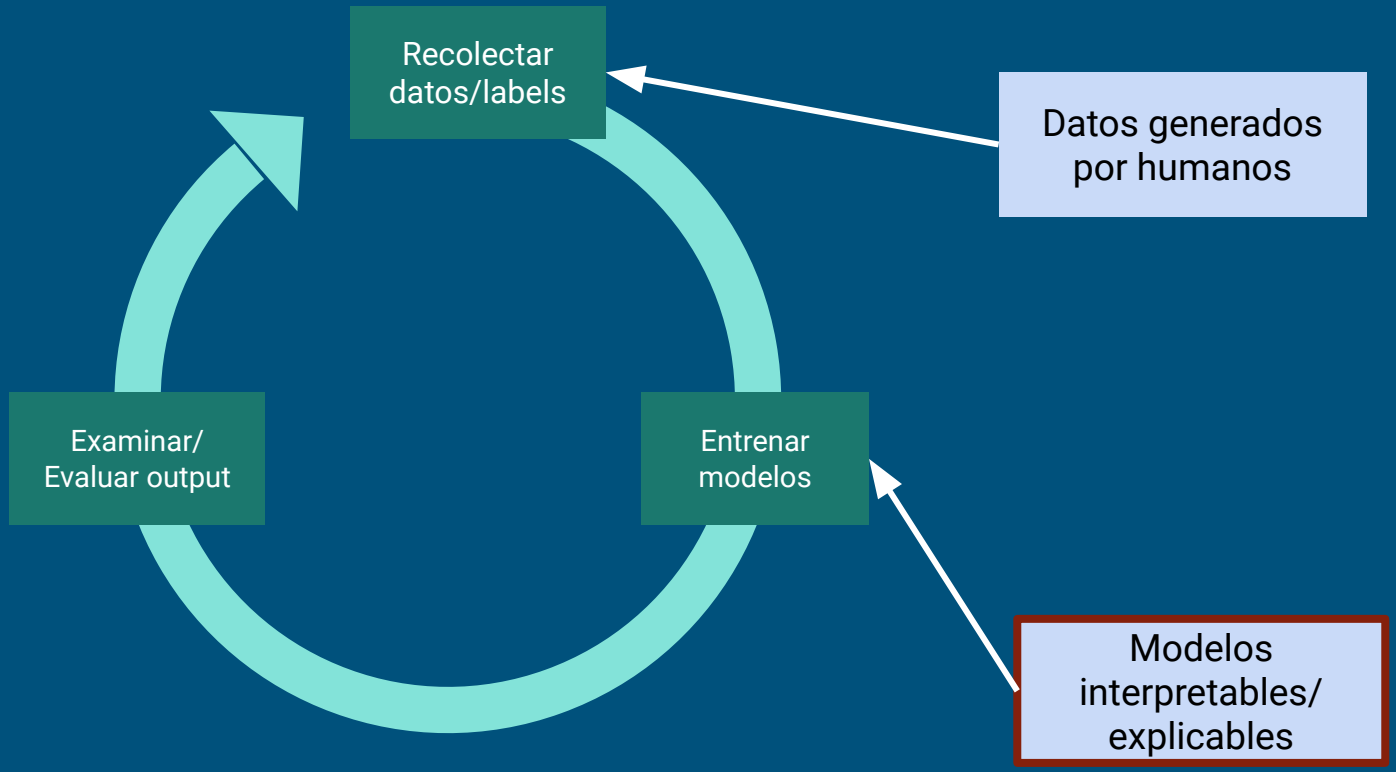
Deep learning algorithms are excellent at pattern-matching in images. This technique could streamline the process of reading medical scans

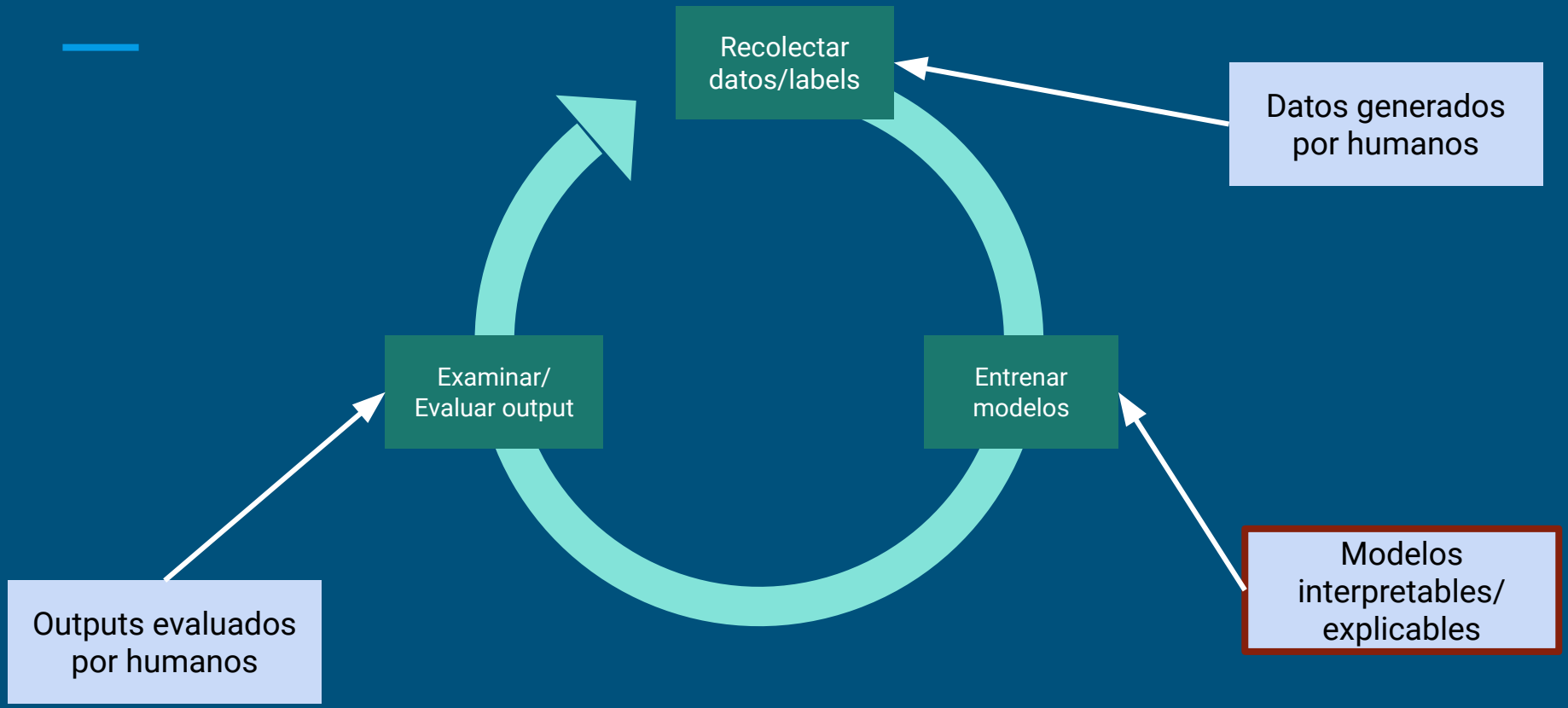
Métodos de interpretabilidad nos permiten examinar lo que el sistema está aprendiendo...

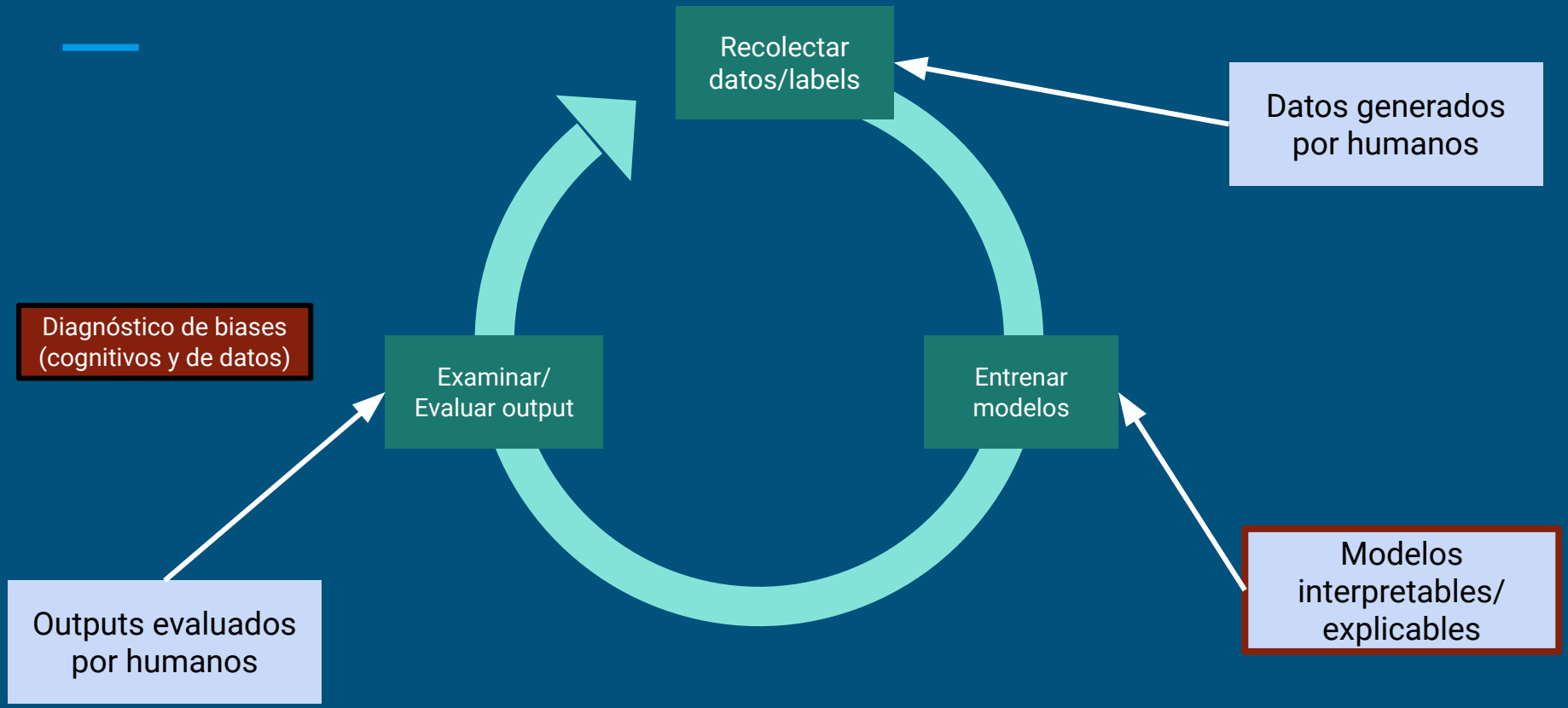


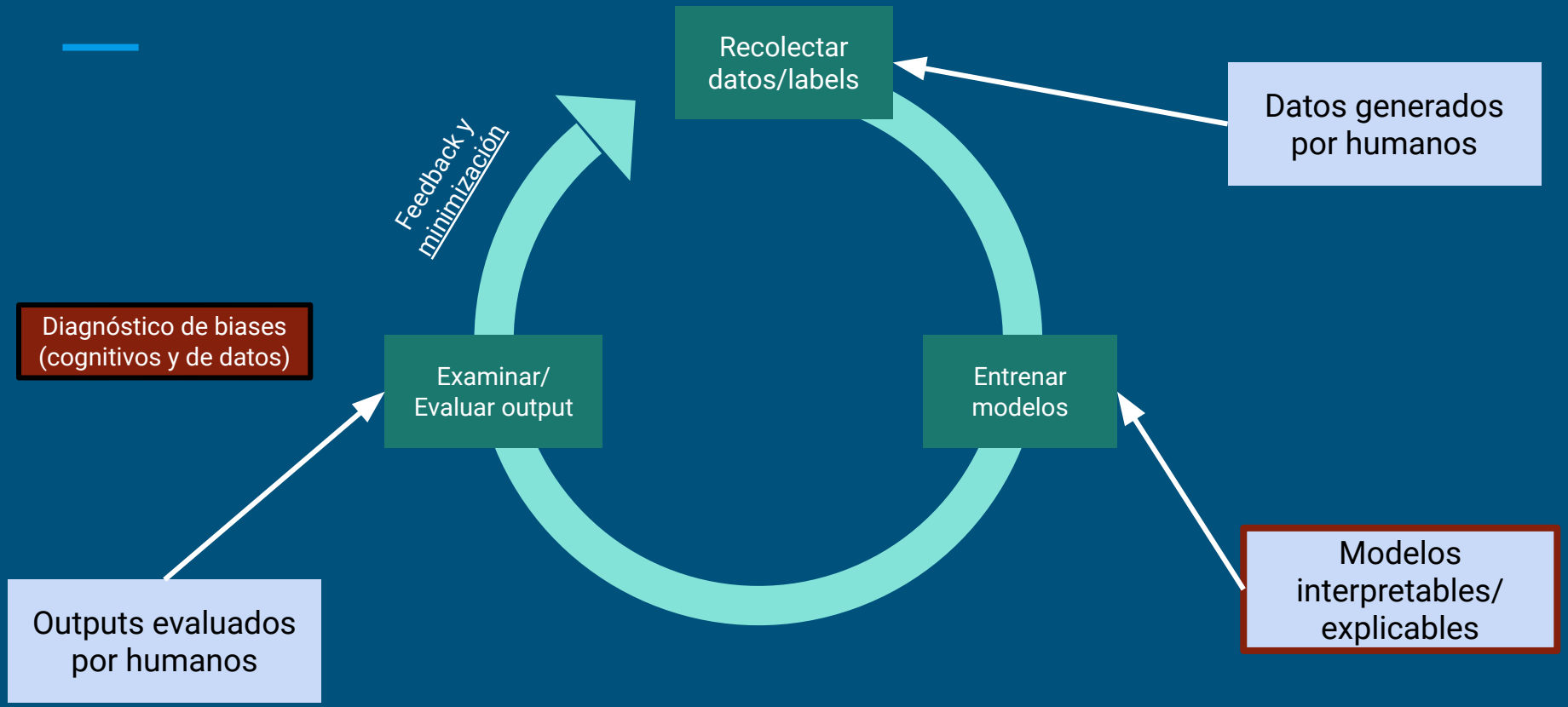
En este caso, el sistema da la respuesta correcta, por las razones incorrectas


Recolectar datos/labels
Datos generados por humanos
Entrenar modelos
Examinar/ Evaluar output


Recolectar datos/labels
Datos generados por humanos
Entrenar modelos
Modelos interpretables/ explicables
Examinar/ Evaluar output


Recolectar datos/labels
Datos generados por humanos
Entrenar modelos
Modelos interpretables/ explicables
Examinar/ Evaluar output
Outputs evaluados por humanos


Recolectar datos/labels
Datos generados por humanos
Diagnóstico de biases (cognitivos y de datos)
Examinar/ Evaluar output
Entrenar modelos
Outputs evaluados por humanos
Modelos interpretables/ explicables


Recolectar datos/labels
Datos generados por humanos
Feedback y minimización
Diagnóstico de biases (cognitivos y de datos)
Examinar/ Evaluar output
Entrenar modelos
Outputs evaluados por humanos
Modelos interpretables/ explicables

Comentarios finales

- **PLN** es un área creciendo **aceleradamente** y con gran relevancia en el panorama de la **inteligencia artificial**

- Sin embargo… No solo es importante la ingeniería detrás de estas tecnologías del lenguaje sino las **implicaciones en la sociedad** que las usa

Comentarios finales

- Al estar trabajando con lenguas, tenemos que pensar en las comunidades de personas usuarias de esas **lenguas** y sus **necesidades**
- En el caso de lenguas **minorizadas/indígenas** es importante incluir a los hablantes para el desarrollo de **tecnologías** para sus propias **comunidades**

“Technology is never neutral, it's made by humans. If we don't assure truly diverse work groups, we are not really creating technology for all”

Dorothy Gordon, Ghana (Technology activist)

Comentarios finales

- **El PLN y la lingüística computacional** constituyen también una herramienta para tener una visión más **profunda** y general del **lenguaje humano**

Algunos recursos

- **HuggingFace transformers** https://huggingface.co/transformers/. Diversas arquitecturas (Bert, GPT-2,...) con modelos pre-entrenados para muchas lenguas
- **Pytorch, TensorFlow** (útiles para construir redes neuronales)
- **Scikit-learn, Pandas, NLTK, Spacy** (paquetería de Python útil para procesar textos)
- Lista de trabajos de PLN enfocados a las lenguas indígenas de América: https://github.com/pywirrarika/naki
- **LIME**, https://github.com/marcotcr/lime, Local Interpretable Model-Agnostic Explanations. Método post-hoc de interpretabilidad.
- **Captum** https://captum.ai/. Librería de python para interpretar modelos utilizando diversos métodos (integrated gradients, Layerwise Relevance Propagation, DeepLift)

Gracias

Escuela de verano 2020

Contacto:

Ana Valeria González:

ana@di.ku.dk

Ximena Gutiérrez-Vasques:

xim@unam.mx @ximgutierrez