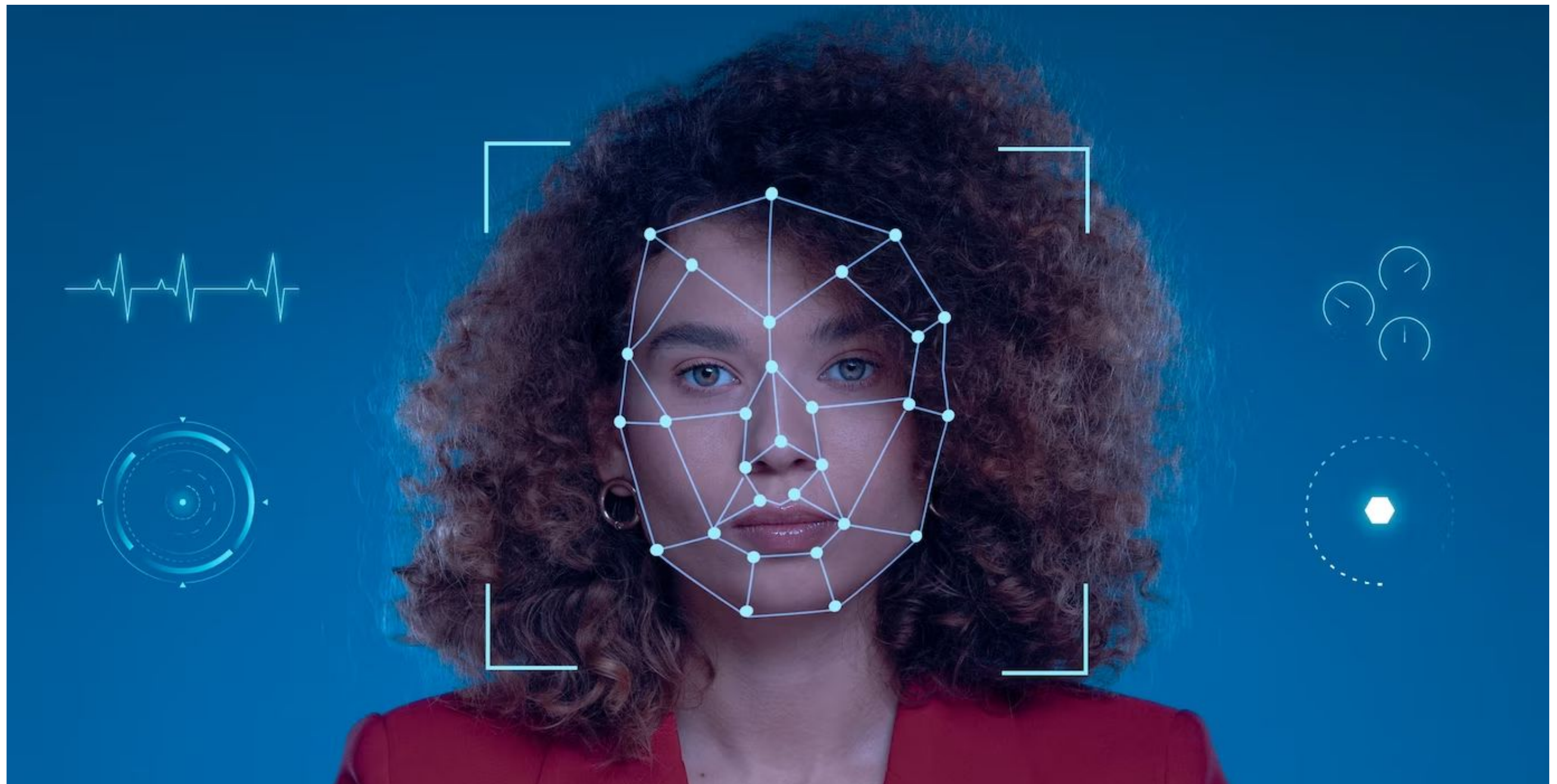




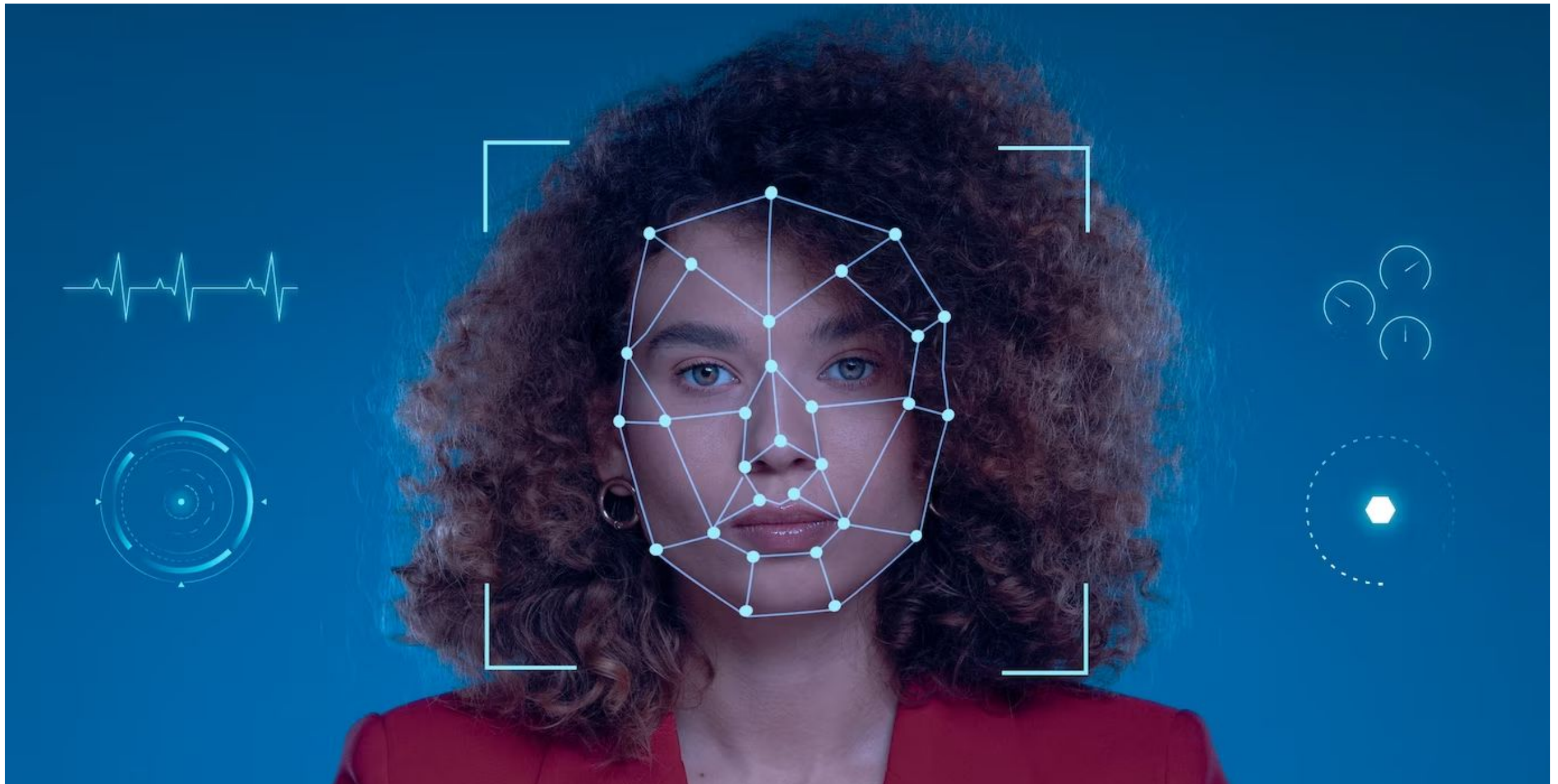
Ética en Machine Learning: Identificando riesgos y estrategias de mitigación

Primavera 2023

Jazmine Maldonado y Cinthia Sánchez



Importancia de la Ética en el desarrollo de soluciones basadas en datos



Objetivos de la Clase

1. Reconocer riesgos éticos
2. Medidas de mitigación

Nuestro Rol

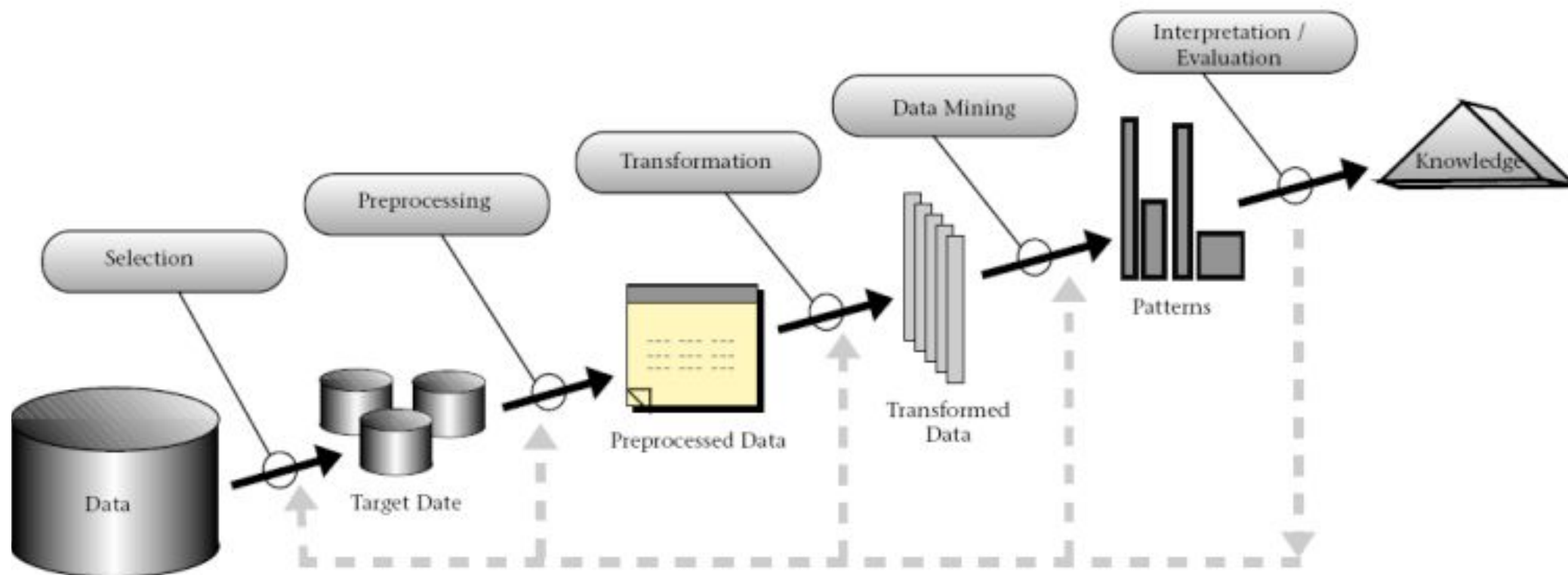
Como profesionales del sector tecnológico tendrán un rol clave en el desarrollo de nuevos productos, servicios y soluciones.



Además de la innovación técnica, tendrán la responsabilidad de considerar y abordar las implicancias éticas de sus creaciones.

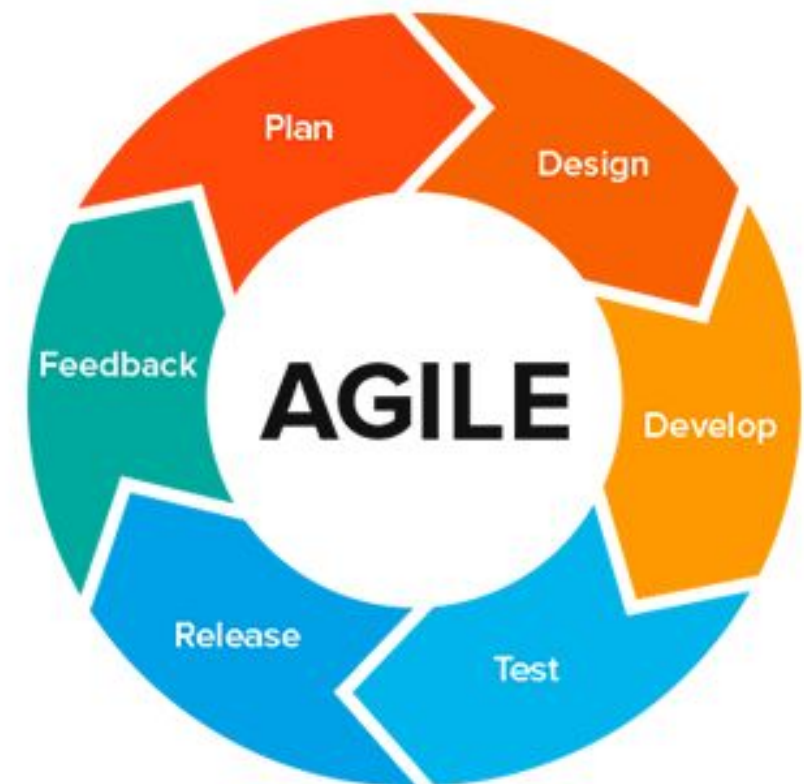
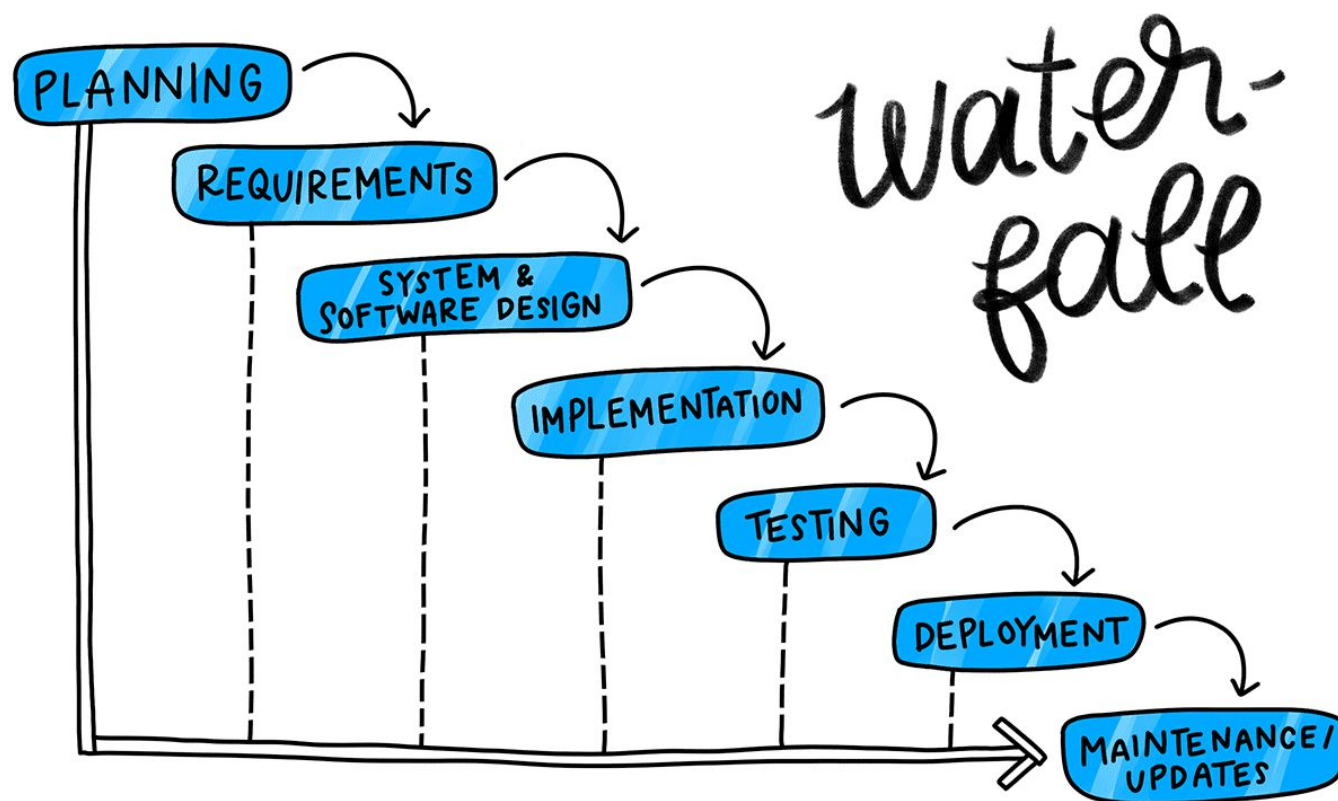
Nuestro Rol

Esta responsabilidad aplica tanto si estamos desarrollando análisis y minería de datos...



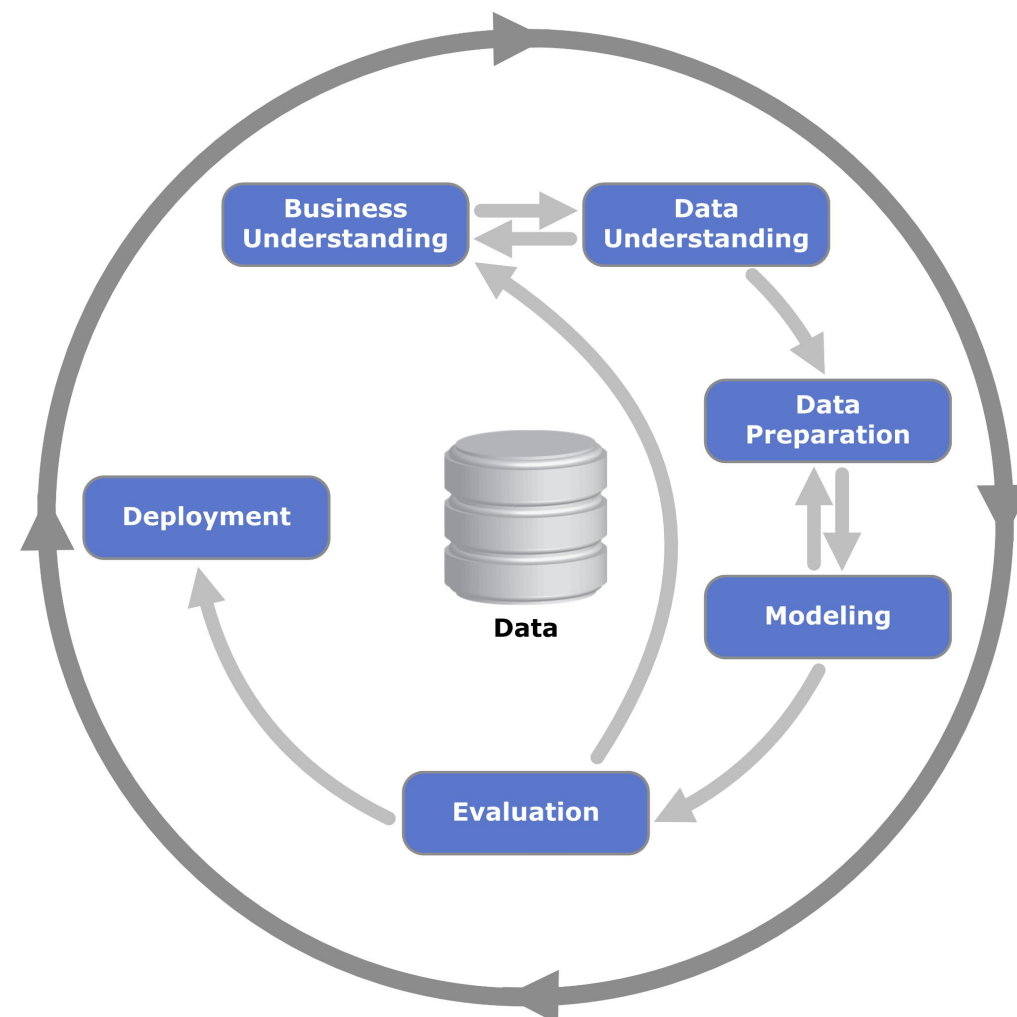
Nuestro Rol

... como si estamos desarrollando software...



Nuestro Rol

Y con mayor razón cuando estamos desarrollando modelos que toman decisiones autónomamente o que replican el comportamiento humano.



Ética

La ética, o filosofía moral, es la rama de la filosofía que estudia la conducta humana, lo correcto y lo incorrecto, lo bueno y lo malo, la moral, el buen vivir, la virtud, la felicidad y el deber.

- Se remonta a los griegos (Platón 427 A.C.)
- Permite crear un orden de convivencia humano y social.

Ética en Datos

Rama emergente de la ética aplicada

Objetivo: Promover el uso responsable y sostenible de los datos en beneficio de las personas y la sociedad y garantizar que el conocimiento adquirido a partir de los datos no se utilice en contra de intereses legítimos de un individuo o grupo.

(Source: Luciano Floridi, Mariarosaria Taddeo (2016) 'What is data ethics?'; Pernille Tranberg, Gry Hasselbalch, Birgitte Kofod Olsen & Catrine Søndergaard Byrne (2018) 'Data Ethics. Principles and Guidelines for Companies, Authorities & Organisations')

Ética Algorítmica

Objetivo: Se preocupa del impacto que tiene el desarrollo de soluciones algorítmicas en la sociedad.

Nuestros resultados y creaciones carecen de criterio ético, por lo que nosotros debemos identificar riesgos y aplicar criterios durante el proceso de creación desde el diseño.

Recordando En la 1ra Clase de Ética

Caso de la profesora Lisa Magrin que fue identificada a partir del análisis de datos superficialmente anónimos sobre ubicación en apps móviles.

<https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html>



Recordando

En la 1ra Clase de Ética

Caso de herramientas
policiales predictivas en
estados unidos que se
demostró eran
discriminatorias al amplificar
sesgos en los datos

<https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/>



Recordando

En la 1ra Clase de Ética

Caso de modelos de lenguaje que replican estereotipos, por ejemplo, generando analogías estereotipadas de género.

https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf



Valores

Para evaluar la ética de un algoritmo se trabaja sobre valores y principios que nos guían en la toma de decisiones.

- Valores y principios universales (dentro de lo posible):
 - Definidos por grupos diversos
 - Involucrando diferentes actores
 - Aprobados ampliamente

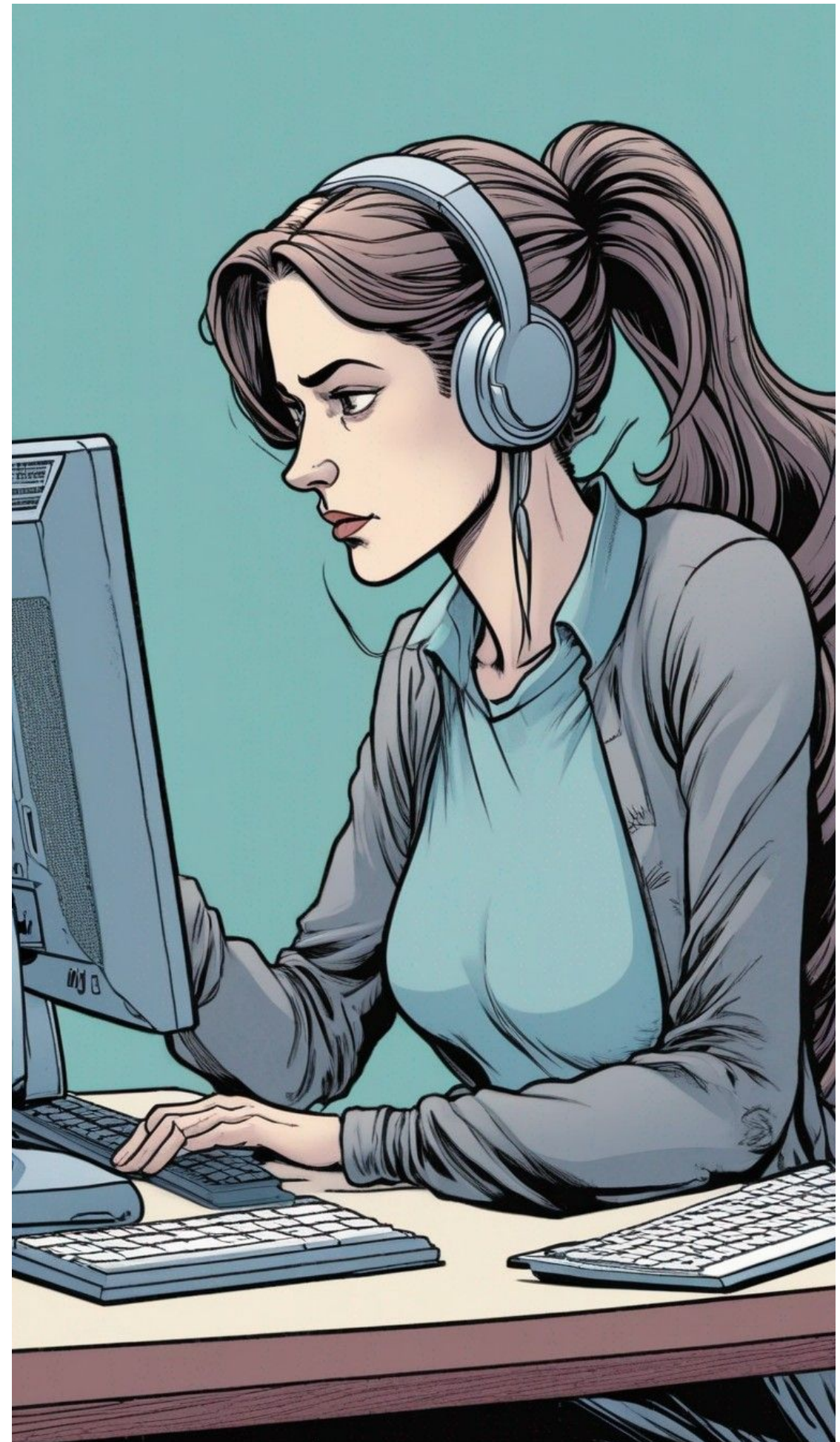
Por ejemplo, los valores que propone UNESCO.

Valores

Valores propuestos por UNESCO:

- Respeto, protección y promoción de los derechos humanos y libertades fundamentales, así como la dignidad humana.
- Prosperidad del medio ambiente y los ecosistemas.
- Garantizar la diversidad y la inclusión.
- Vivir en sociedades pacíficas, justas e interconectadas.

Parte 1. Reconocer Riesgos Éticos



Aspectos Éticos

- Sesgo y Justicia (Bias and Fairness)
- Privacidad
- Protección y Seguridad (safety and security)
- Transparencia
- Responsabilidad (Accountability)

Sesgo (Bias)

- Desviación sistemática de un valor real, ya sea positiva o negativa.
- Preocupación ética: "Inclinación sistemática en la toma de decisiones que resulta en resultados injustos" [CEDEI]
- La preocupación se acentúa cuando los resultados negativos afectan de manera desproporcionada a ciertos grupos.

Sesgo (Bias)

- En casos de trato injusto relacionado con características protegidas, como género o étnia, el sesgo **puede constituir discriminación ilegal** según las leyes pertinentes.

Sesgo (Bias)

¿Cuáles son las causas?

- Datos
- Diseño del algoritmo
- Percepción humana
- Proceso de toma de decisiones



Las causas pueden ser diversas

Algoritmos entrenados en datos históricos tienden a replicar y amplificar sesgos presentes en esos datos [SURESH]

Fairness (Justicia)

- No hay una definición única de Fairness
- Es contextual y varía según valores, perspectivas y sociedades ([Mehrabi] ofrece una buena encuesta de definiciones).



[Mehrabi] Ninareh Mehrabi; et al. [A Survey on Bias and Fairness in Machine Learning](#).

Fairness (Justicia)

- **Idea central:** Tratamiento igualitario a menos que exista una razón justificada.
- El criterio de Fairness se usa como marco ético para definir cuándo el sesgo (bias) es dañino.
- También se puede usar como guía para encontrar un balance entre diferentes principios éticos.

Fairness (Justicia)

- Enfoque en procesos y resultado
- **Resultados:** distribución justa de beneficios y costos, evitando sesgos injustos o decisiones arbitrarias.
- **Proceso:** Involucrar a las comunidades afectadas en decisiones, permitir la impugnación y búsqueda de reparación para decisiones tomadas por ML.

Privacidad

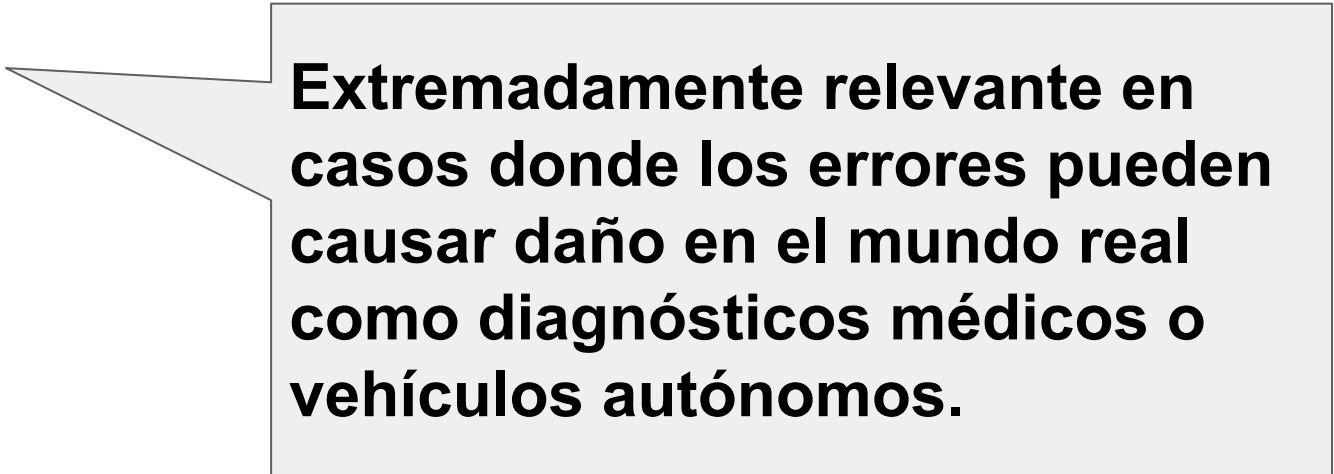
- Múltiples maneras en la que un sistema ML o software ponga en riesgo la privacidad.
- Tratamiento sin consentimiento informado
- Violación de privacidad en el acceso a datos de entrenamiento (ej. scrapping)
- Fugas de datos personales [Weidinger]

Privacidad

- La precisión de las predicciones de los sistemas de ML puede presentar riesgos a la privacidad.
- Inferencia de datos sensibles.
- Derecho al olvido (EU/GDPR), que podría incluir ser eliminado de los datos de entrenamiento de ML. [Bourtoule]

Protección y Seguridad

La seguridad en sistemas de ML implica precisión, confiabilidad, y funcionamiento constante a lo largo del tiempo.

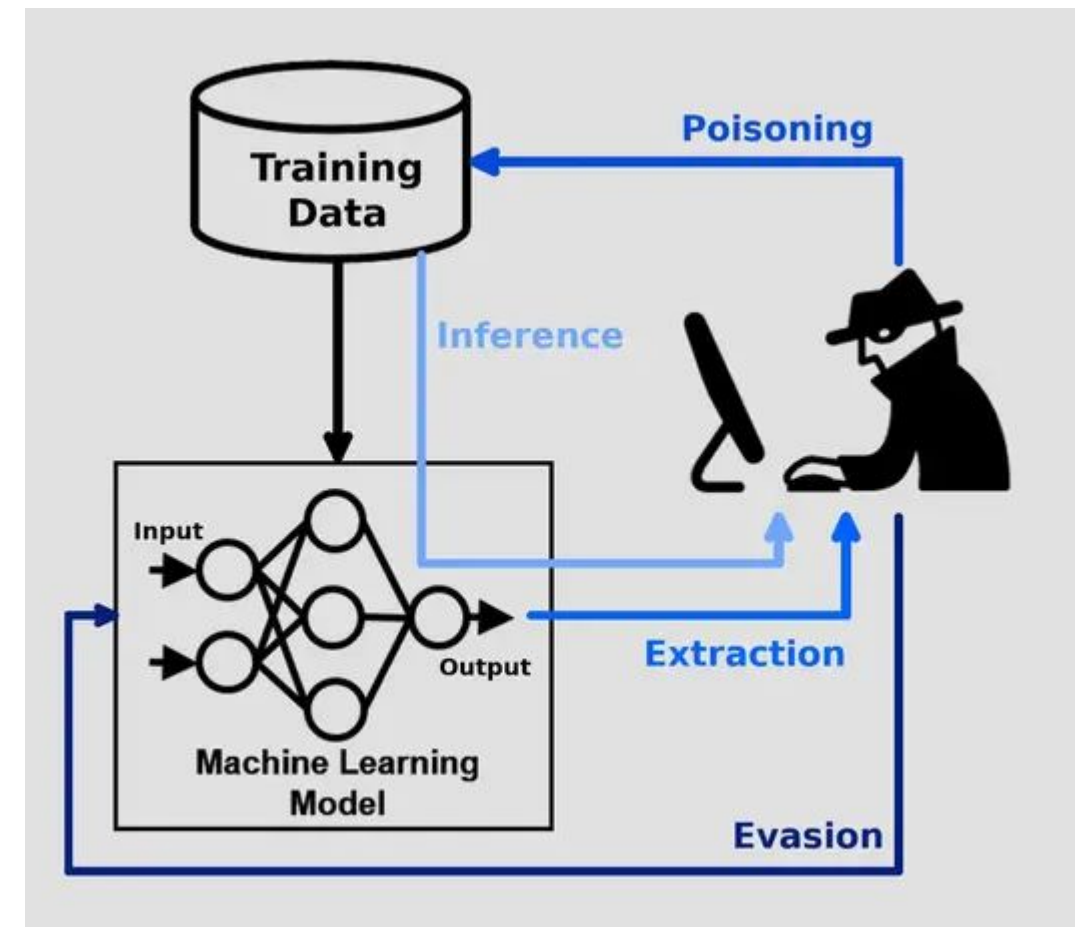


Extremadamente relevante en casos donde los errores pueden causar daño en el mundo real como diagnósticos médicos o vehículos autónomos.

- Precisión y Confiabilidad
- Seguridad contra Ataques Adversariales
- Robustez en Condiciones del Mundo Real

Riesgos de Seguridad

- **Envenenamiento de Datos de Entrenamiento:** Riesgo de manipulación de datos de entrenamiento.
- **Entradas Adversarias:** Riesgo de ataques con datos diseñados para engañar al sistema.
- **Ataques de Inversión y Inferencia Adversarial:** Riesgo de exponer parámetros del modelo o datos de entrenamiento.



Transparencia

Transparencia se refiere a que **usuarios y partes interesadas tengan acceso a la información necesaria para tomar decisiones informadas** sobre el algoritmo, modelo, estudio, etc.

Involucra tanto al modelo en sí como al proceso desde la creación hasta su uso.

Transparencia

Rastreabilidad: Documentar claramente metas, definiciones, decisiones de diseño y suposiciones.

Comunicación: Ser transparente sobre el uso de tecnología de ML y sus limitaciones.

Inteligibilidad: Los interesados deben poder entender y supervisar el comportamiento de los sistemas de ML para alcanzar sus objetivos ([Vaughan], [EGTAI]).

[Vaughan] Wortman; Vaughan; Wallach. [A Human-Centered Agenda for Intelligible Machine Learning](#).

[EGTAI] [Ethics Guidelines for Trustworthy AI](#).

Transparencia

Interpretabilidad: Observar la relación causa-efecto dentro de un sistema.

Explicabilidad: Explicar en términos humanos los mecanismos internos de un sistema de aprendizaje automático o profundo.

Cuando no se tienen, hablamos del problema de la caja negra. Común en enfoques de ML más complejos como las redes neuronales. [Gall]

Responsabilidad

ML se usa en áreas de alto impacto (salud, bienestar, justicia penal), lo que amplifica las consecuencias cuando algo sale mal.

- Actores en la cadena de ML deben asumir la responsabilidad de considerar el impacto y rendir cuentas cuando las cosas salen mal.
- "El algoritmo lo hizo" no es una excusa aceptable ante errores o consecuencias no deseadas ([FATML]).

Responsabilidad

- **La transparencia habilita la responsabilidad** al permitir ver lo que está yendo mal y dónde.
- Requiere:
 - Procesos adecuados para considerar riesgos.
 - Documentación de políticas y procesos.
 - Medios de reparación para afectados
 - Responsabilidad por uso de ML de terceros

Parte 2. Análisis de Caso



Análisis de Caso

1. Dividirse en grupos
2. Cada grupo analizará un caso hipotético identificando potenciales riesgos éticos
3. Cada grupo presenta brevemente los riesgos encontrados y su reflexión

Caso 1: Industria de la Salud

Se desarrolla un algoritmo de inteligencia artificial para apoyar la toma de decisiones. El algoritmo genera un índice de riesgo para cada nuevo paciente que predice si requiere cuidados adicionales y prioridad de atención médica.

Para el entrenamiento del modelo se usan registros históricos de pacientes. Las características consideradas por el modelo incluyen datos personales de cada paciente y un listado de enfermedades subyacentes. Como variable objetivo se utiliza una variable construida a partir del número total de atenciones médicas y el gasto total en atenciones médicas de cada paciente.

Caso 2: Banca y Finanzas

Un equipo de desarrollo en el sector financiero está creando un modelo de inteligencia artificial para evaluar la solvencia crediticia de los solicitantes de préstamos. El modelo se propone automatizar el proceso de toma de decisiones, asignando puntajes de riesgo crediticio a los clientes potenciales.

El modelo utiliza datos financieros como ingresos, historial crediticio, deudas, etc. y datos demográficos como edad, género, ubicación, estado civil, etc.

La variable objetivo a considerar es la solvencia crediticia, es decir un puntaje que refleja la probabilidad de que un solicitante cumpla con sus obligaciones crediticias.

Caso 3: Proceso de Contratación

Un equipo de desarrollo está diseñando un modelo de inteligencia artificial para automatizar el proceso de contratación en una empresa. El objetivo es agilizar la selección de candidatos y mejorar la eficiencia en la toma de decisiones.

El modelo utiliza datos históricos de procesos de contratación realizados en últimos años en la empresa para su entrenamiento. Estos incluyen características extraídas del curriculum como experiencia laboral, habilidades, educación, etc. No se utilizan datos demográficos como género, edad o ubicación para evitar introducir sesgos.

Como variable objetivo se considera si la persona fue contratada.

Caso 4: Educación

Un colegio decide utilizar modelos generativos de lenguaje y de imágenes como chatGPT y Dall.e para generar cuentos sobre diferentes profesiones. El objetivo es proporcionar a los estudiantes historias creativas que inspiren interés en diversas carreras y fomenten la imaginación.

Para la creación de los cuentos los profesores y profesoras utilizan *prompts* como “*Crea un cuento corto y realista para niños sobre un día en un hospital que muestre el trabajo que realizan las personas que trabajan en ese lugar*”

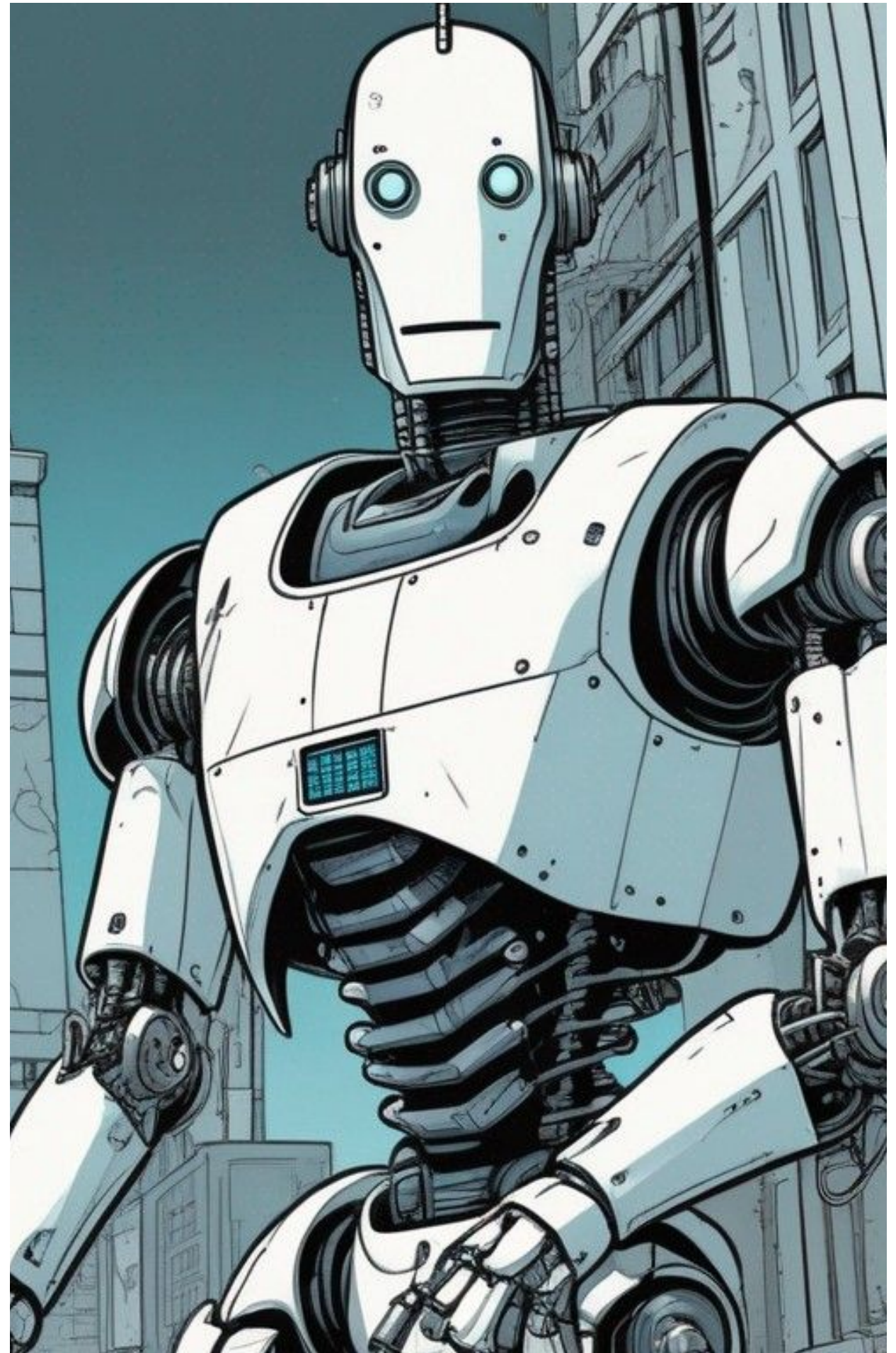
Caso 5: Detección de actitud criminal

Una municipalidad decide implementar tecnología avanzada basada en cámaras en autobuses y calles para detectar actitudes criminales sospechosas. El objetivo es mejorar la seguridad pública mediante la identificación temprana de comportamientos potencialmente delictivos.

Para su desarrollo utilizan cámaras de vigilancia avanzadas instaladas en autobuses y áreas urbanas.

El sistema utiliza algoritmos de reconocimiento de imágenes y comportamiento que analizan vestimentas, gestos, movimientos y comportamientos para identificar actitudes criminales sospechosas.

Parte 3. Medidas de Mitigación



Mitigación

Trabajo en Progreso

- Múltiples enfoques
- Aún no existen estándares de mitigación para todos los riesgos posibles
- (En 2023) No existe regulación legal para el uso de la IA específicamente

Mitigación

Sesgos y Justicia (Bias & Fairness)

- Definición previa de criterios de justicia que se espera garantizar con la solución
- Datos de entrenamiento diversos y representativos

Mitigación

Sesgos y Justicia (Bias & Fairness)

- Mediciones en laboratorio y en entorno real usando herramientas especializadas
 - [IBM 360 degree toolkit](#)
 - [FairLearn](#)
- Auditoría Algorítmica
 - Generalmente realizadas por terceros

Mitigación

Privacidad

- Ajustarse a normativas estrictas de protección de datos. Ej. GDPR
 - Consentimiento de los usuarios
 - Notificar brechas de seguridad
 - Transparencia en el uso de datos
 - Privacidad desde el diseño (ej. solicitar solo lo necesario, no exponer datos en aplicaciones, etc.)
 - Derecho al olvido (eliminación de datos)

Mitigación

Privacidad

- Usar tecnologías de terceros que cumplen normativas estrictas y tienen una política de privacidad clara.
- **Privacidad diferencial** al liberar o compartir datos con terceros.

Mitigación

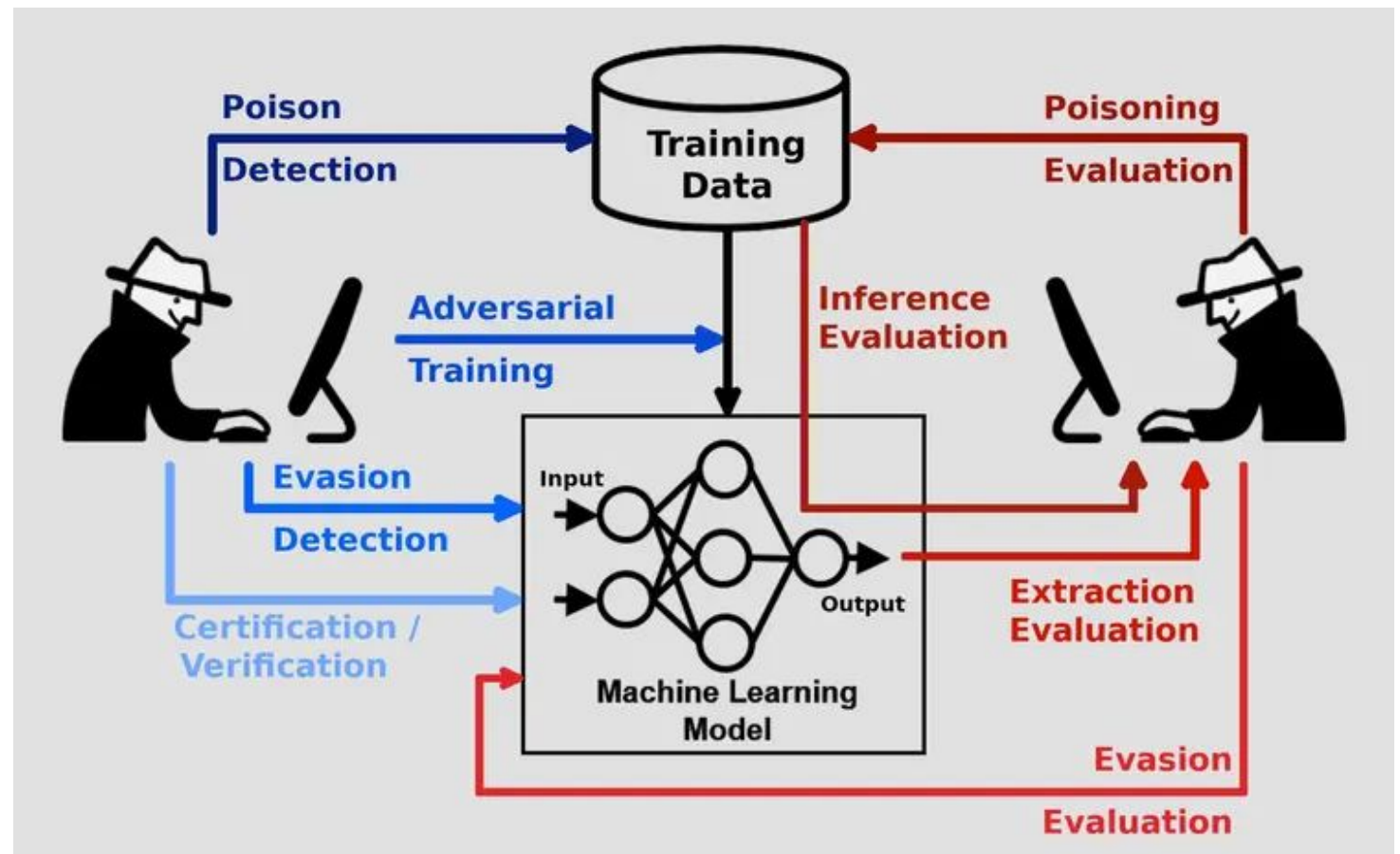
Protección y Seguridad

- Policy Layers
 - Capa lógica que se ubica frente al sistema ML/AI y analiza y filtra el contenido en base a criterios predefinidos
- Supervisión Humana
 - Monitoreo o delegación de la decisión final a una persona que pueda discernir las consecuencias de la acción tomada

Mitigación

Protección y Seguridad

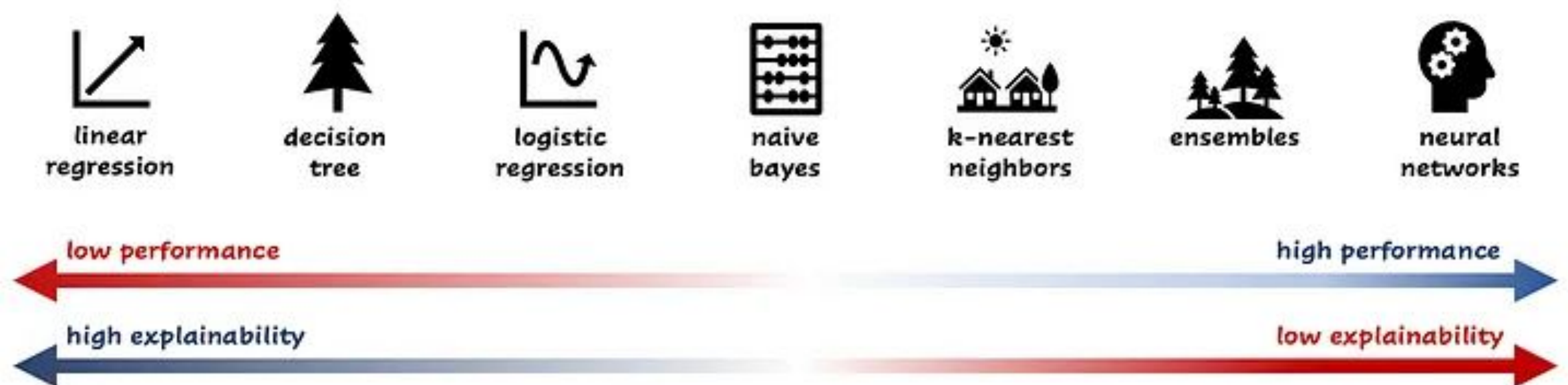
- Entrenamiento Adversarial
 - [IBM Adversarial Robustness Toolbox](#)



Mitigación

Transparencia y Explicabilidad

- Evaluar caso a caso el nivel de transparencia requerido para la solución y seleccionar modelos acorde a lo requerido.



Mitigación

Transparencia y Explicabilidad

- Documentación de los datos utilizados
 - [Datasheet for Datasets](#)
 - [Data Cards & Dataset Cards](#)
- Documentación del modelo creado y sus limitaciones
 - [Model Cards for Model Reporting](#)

Mitigación

Responsabilidad (Accountability)

- Considerar aspectos éticos desde el diseño
- Políticas internas de la empresa, startup, etc.
 - Responsabilidad de crearlas
 - Responsabilidad de conocerlas
- Mantener registro del proceso de diseño y de creación.
- Transparentar limitaciones de nuestras creaciones y las evaluaciones realizadas.

Mitigación

**¿Qué otras
medidas de
mitigación crees
que se podrían
aplicar?**



Recapitulando

- Existen diferentes aspectos Éticos que se deben considerar al desarrollar soluciones basadas en datos y que utilizan machine learning o inteligencia artificial.
- Estos deben ser considerados en todas las etapas del proceso de creación desde el diseño hasta el despliegue y monitoreo.
- Existen medidas de mitigación que se pueden aplicar cuando se detectan riesgos.
- Para evaluar aspectos éticos se trabaja sobre valores que sirven de guía para la toma de decisiones.



dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl

f @ in  / DCCUCHILE