



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

Diplomado en Big Data y Ciencia de Datos
Ciencia de Datos y sus Aplicaciones

Clase 02: Metodologías y Actualidad

Roberto González



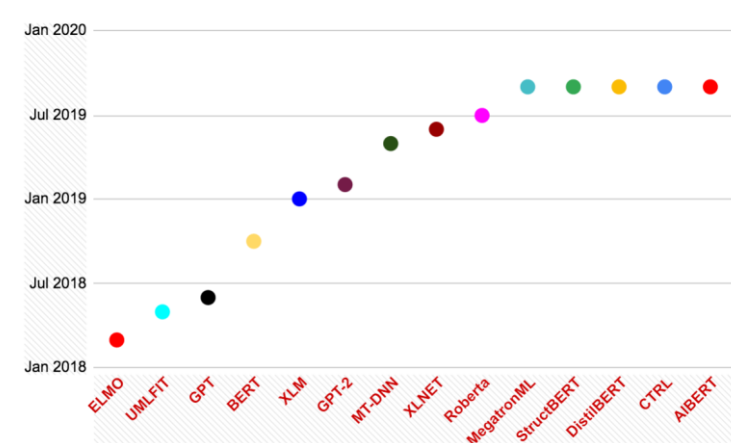
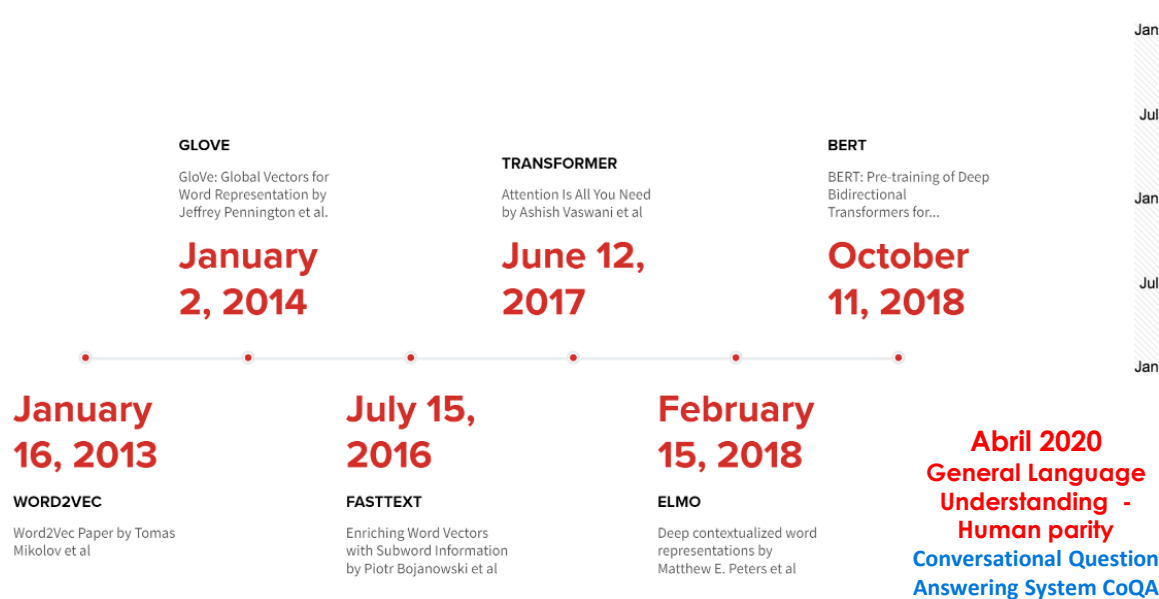
regonzar@uc.cl



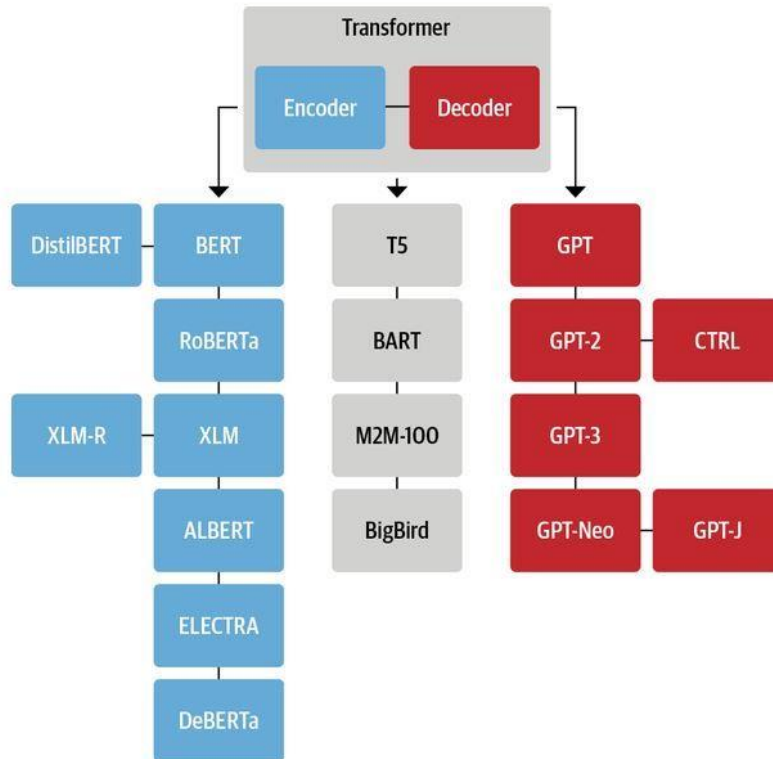
Natural Language Processing

Procesamiento de lenguaje natural

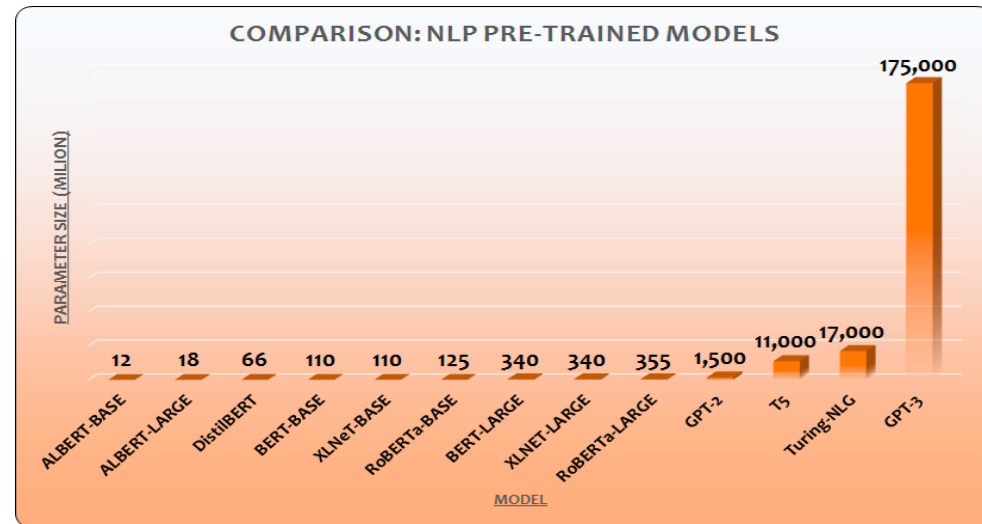
- 1) Gran cantidad de información almacenada en archivos de texto, chats y logs
- 2) Nuevos modelos de Deep Learning han permitido avanzar en análisis semántico de texto



Natural Language Processing



45 TB
texto
Train Set



2018

Junio
2020

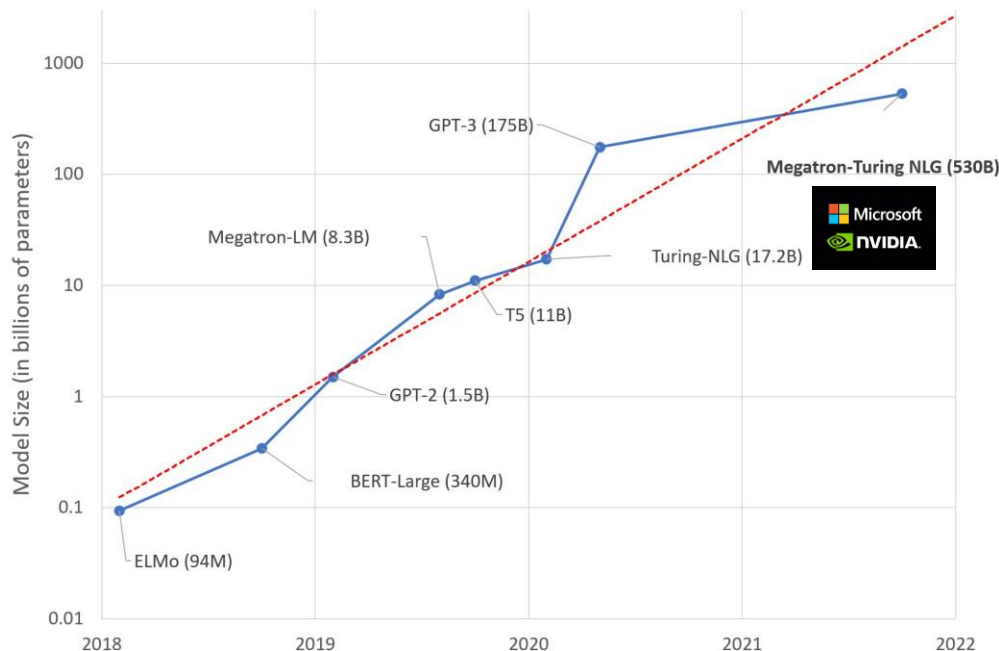
Figure 3-8. An overview of some of the most prominent transformer architectures

GPT-3 y LLMs

<https://openai.com/>

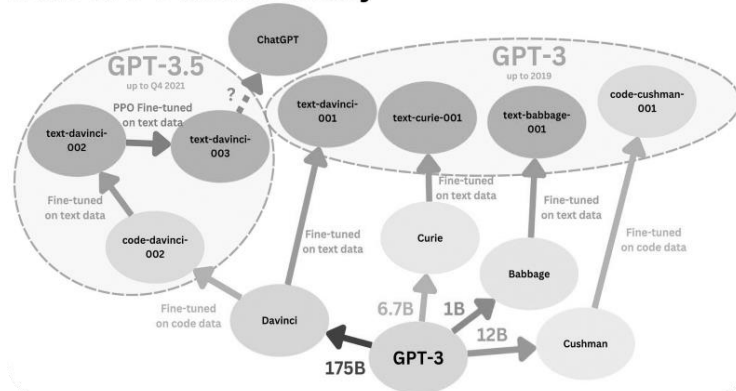


Artificial General Intelligence (AGI) para todos
Modelo de NLP desarrollado por OpenAI (US\$4.6M training)



Parámetros se duplican cada 10 meses - Nueva ley de moore?

The GPT-3 models Family

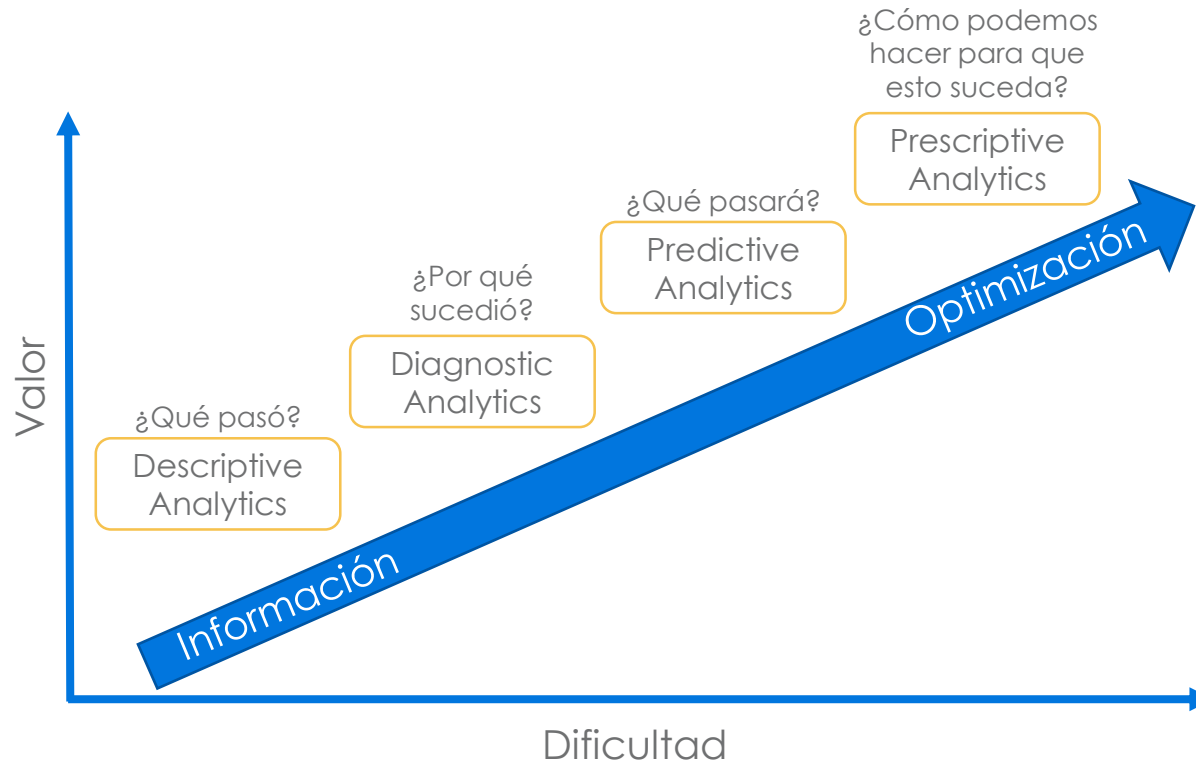




Dall-E 2



La evolución de los temas analíticos



Fuente: Gartner Business Intelligence & Analytics Summit 2013



<https://youtu.be/UFUwNBCkmYs>

www.educacionprofesional.ing.uc.cl

Blockbuster vs Netflix



	Blockbuster	Netflix
Año Fundación	1985	1997
Ventas 2004	\$ 6 mil millones	\$ 500 millones
Ventas 2018	\$ 0	\$ 16 mil millones

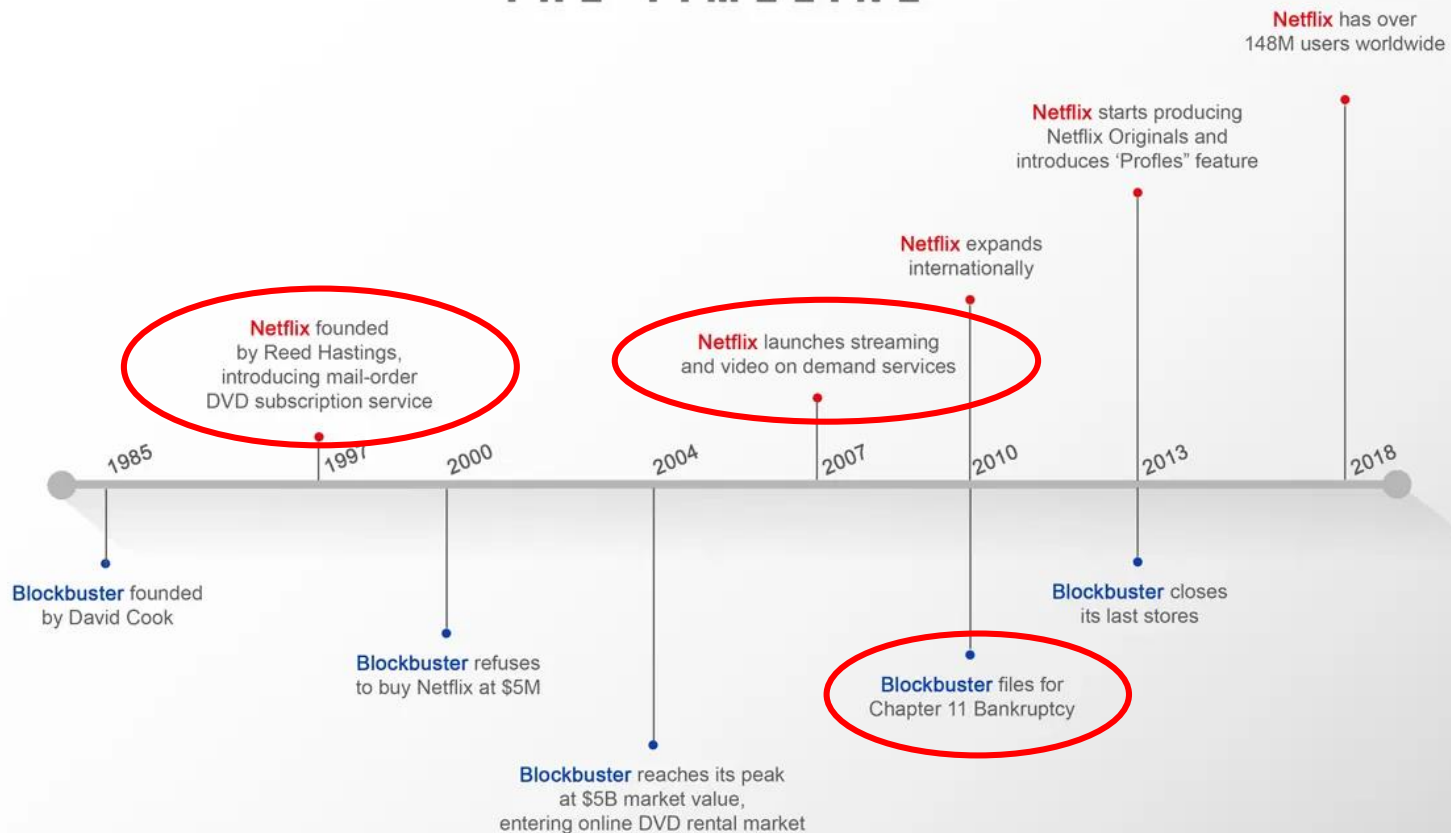




Netflix vs Blockbuster

¿Porqué Netflix perduró en el tiempo y le robó el mercado a Blockbuster?

Historia de Netflix





Propuesta Netflix

- Machine learning directamente ligado al modelo de negocios de Netflix
- Apoyar a áreas de marketing y publicidad
- Proponer nuevos títulos e identificar nuevos tipos de clientes
- ML aplicado en un producto real
- Modelos ML de impacto a nivel mundial
- Equipo diverso, múltiples habilidades



Clase 02: Metodologías de Análisis de Datos

METODOLOGÍAS

Una metodología

- Es un proceso preciso y formal.
- Una metodología incluye:
 - Actividades paso a paso para cada fase.
 - Roles individuales para cada actividad.
 - Productos y niveles de calidad para cada actividad.
 - Herramientas y técnicas que se usarán para cada actividad.



¿Por qué se utilizan?

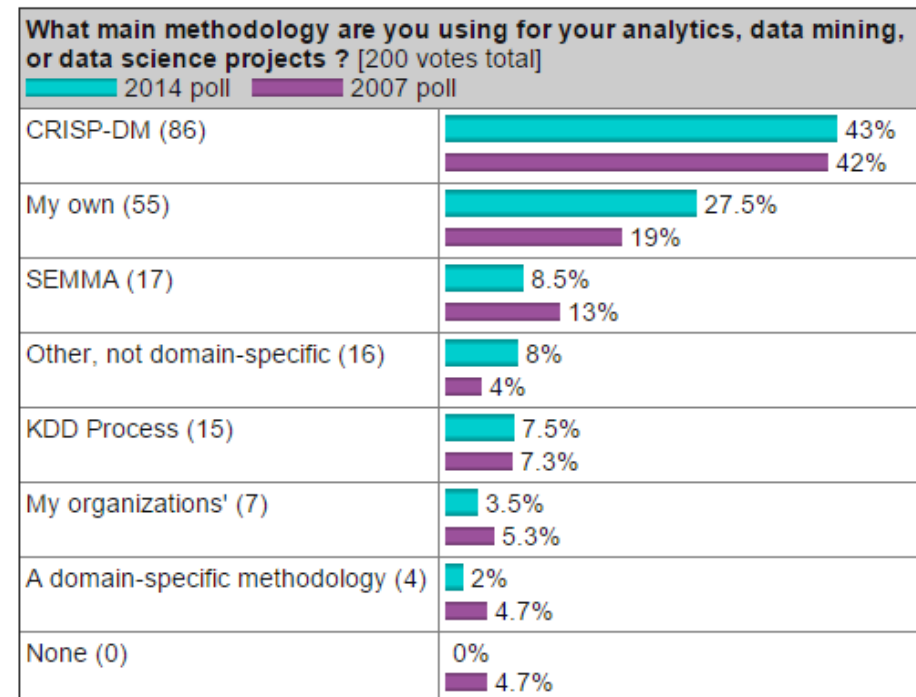
- Las metodologías aseguran que un enfoque consistente se aplicará a todos los proyectos.
- Reducen el riesgo asociado a errores y “atajos”.



Metodologías más utilizadas para Análisis de Datos

- Distribución regional de los votantes

- US/Canada 45.5%
- Europe 28.5%
- Asia 14.0%
- Latin America 9.5%
- Other 2.5%



Fuente: KDNuggets Poll, Octubre 2014
<http://www.kdnuggets.com>

www.educacionprofesional.ing.uc.cl



Clase 02: Metodologías de Análisis de Datos

CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)



CRISP - DM

- CRoss-Industry Standard Process for Data Mining.
- Metodología para el proceso de Minería de Datos
 - Valida el proceso, ayuda a planear y administrar proyectos.
- Desarrollado el año 2000 por algunas compañías: SPSS/ISL, NCR, OHRA.
- Está enfocado en el negocio y al análisis técnico.

CRISP - DM



como podemos extraer valor a los datos?

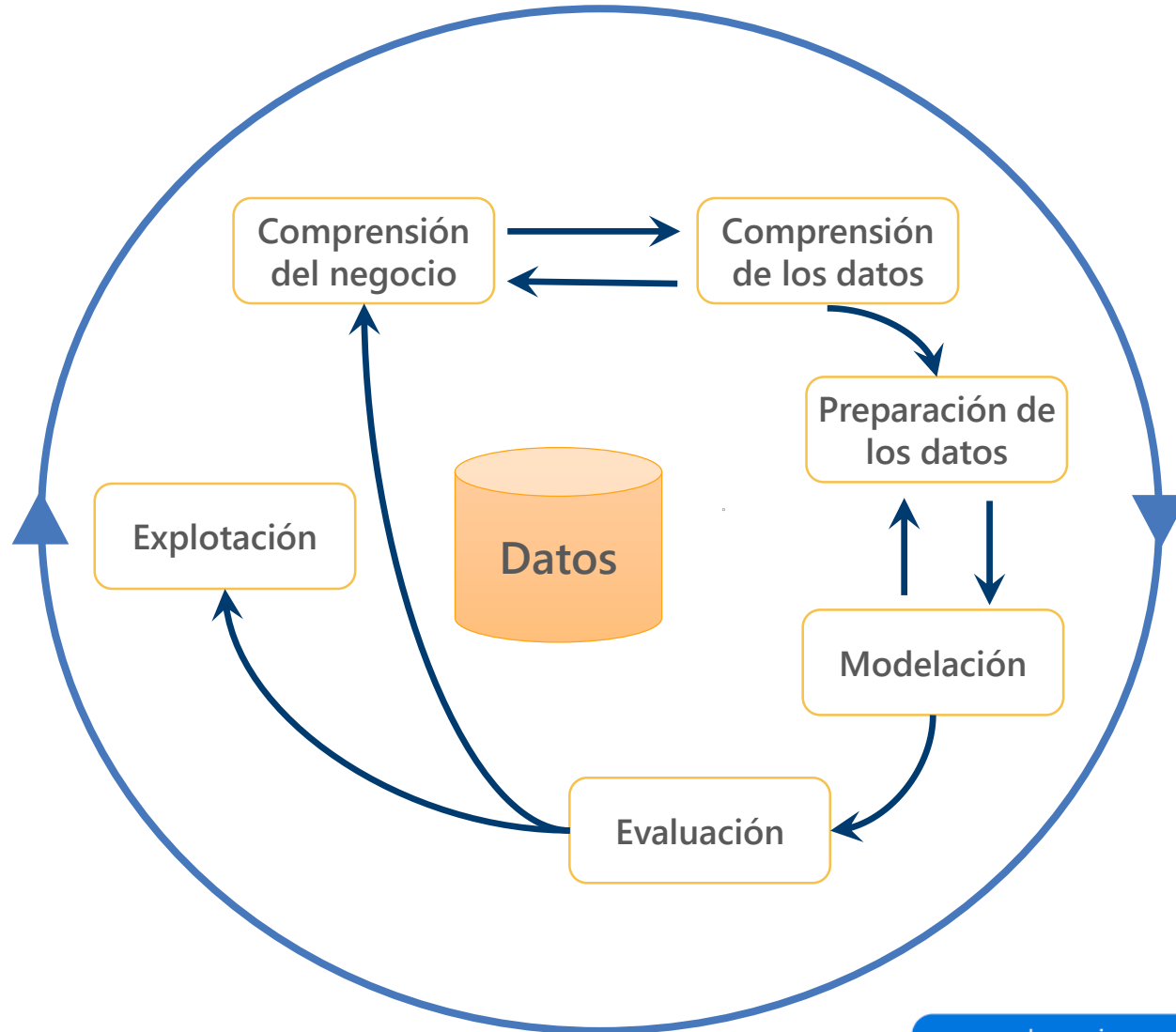


Para extraer valor de los datos, es necesario seguir un proceso que implica varias etapas. A continuación, se presentan algunas de las etapas más importantes para extraer valor a partir de los datos:

1. Definir el problema: En primer lugar, es necesario definir claramente el problema que se quiere resolver o la pregunta que se quiere responder. Esto permitirá enfocar la recolección y el análisis de los datos.
2. Recolectar los datos: Luego, es necesario recolectar los datos necesarios para abordar el problema definido. Esto puede implicar la recolección de datos de diferentes fuentes, incluyendo bases de datos, encuestas, sensores y redes sociales, entre otras.
3. Limpiar y preparar los datos: Una vez que se han recolectado los datos, es necesario limpiarlos y prepararlos para su análisis. Esto implica eliminar valores atípicos, corregir errores y formatear los datos de manera que sean fáciles de analizar.
4. Analizar los datos: Después de preparar los datos, es necesario analizarlos para extraer información y patrones relevantes. Esto puede implicar el uso de técnicas de minería de datos, aprendizaje automático, estadísticas y visualización de datos.
5. Interpretar los resultados: Una vez que se han analizado los datos, es necesario interpretar los resultados y extraer conclusiones. Esto implica entender lo que los datos dicen sobre el problema definido y cómo se pueden utilizar para resolverlo.
6. Tomar decisiones: Finalmente, se deben tomar decisiones basadas en los resultados obtenidos. Esto puede implicar hacer cambios en una organización, tomar decisiones de inversión o planificar futuras acciones.

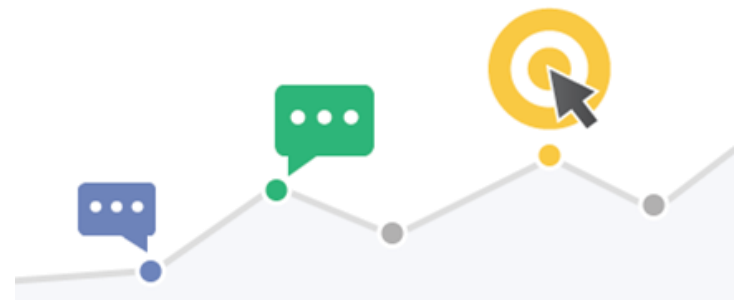
En resumen, para extraer valor de los datos es necesario seguir un proceso riguroso que implica recolectar, limpiar, analizar e interpretar los datos para tomar decisiones informadas.

Visión General



Fase 1: Comprensión del Negocio

- Determinar los objetivos de negocio
 - Dentro de este contexto es importante definir los criterios de éxito del negocio
- Levantamiento de requerimientos, riesgos, supuestos y beneficios
- Definir los objetivos del proyecto
 - Dentro de este contexto es importante definir los criterios de éxito del proyecto
- Generar planificación inicial



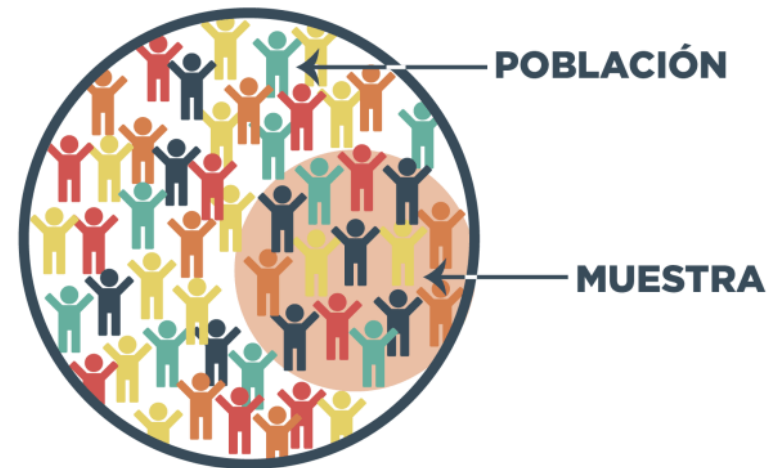
Fase 2: Comprensión de los Datos

- Objetivo:
 - Simplificar el problema y optimizar la eficiencia del modelo.
- ¿Cómo?
 - Uso de herramientas de visualización y técnicas de estadísticas descriptivas.
- Es relevante también determinar la calidad de los datos.



Fase 3: Preparación de los Datos Selección

- Seleccionar el conjunto de datos o las variables o muestras sobre los cuales el proceso de análisis va a ser ejecutado.
- Selección de muestras.



Fase 3: Preparación de los Datos Limpieza de Datos

- La calidad del conocimiento a descubrir depende (además de otros factores) de la calidad de los datos analizados.
- Nuestro Objetivo:
 - Mejorar la calidad de los datos.





Fase 3: Preparación de los Datos

Limpieza: ¿En qué centrarse?

- Datos necesarios que no están a disposición
 - Estrategias para obtener datos
- Presencia de datos faltantes (missing values)
 - Estrategias para tratamiento de datos faltantes.
- Presencia de datos que no se ajustan al comportamiento general de los datos (outliers)

Fase 3: Preparación de los Datos

Missing values

- Es posible que los métodos que utilizaremos en fases posteriores no traten bien los campos con missing values.
- Hay que detectarlos y tratarlos.
- Posibles estrategias:
 - Ignorarlos
 - Eliminar variable
 - Filtrar registro
 - Reemplazar el valor
 - Etc.





Fase 3: Preparación de los Datos

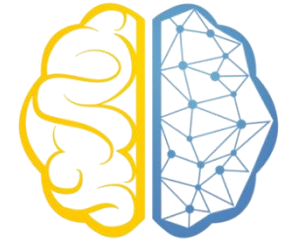
Transformación de Datos

- Normalización de datos
- Construcción de nuevas variables que faciliten el proceso de minería de datos.
- Reducción de Dimensionalidad
 - Variables Correlacionadas
- Discretización de variables continuas





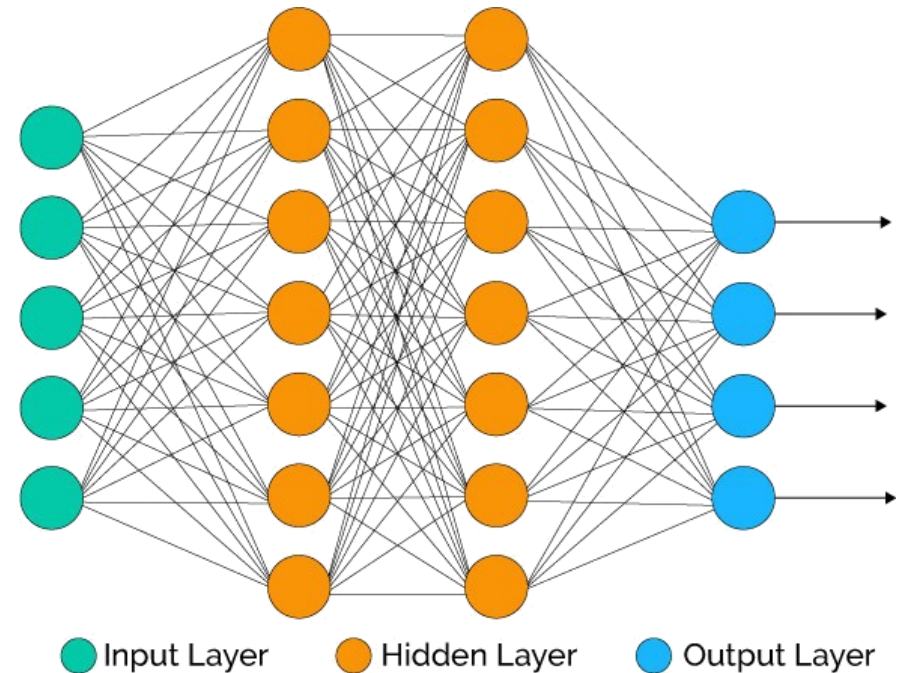
Fase 4: Modelación



- El objetivo:
 - Satisfacer las metas del planteadas en los primeros pasos, a través de un método particular de Minería de Datos.
- Por tanto es crucial:
 - Seleccionar el algoritmo correcto a partir del problema que tenemos que abordar y las metas esperadas.

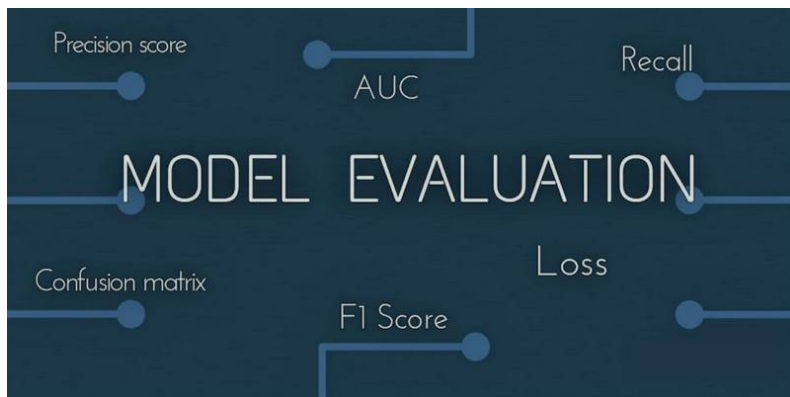
Fase 4: Modelación Técnicas

- Las técnicas más utilizadas:
 - Métodos de clustering
 - Análisis de regresión
 - Redes neuronales
 - Árboles de decisión
 - Reglas de asociación
 - Etc.



Fase 5: Evaluación

- Valora los resultados mediante el análisis de bondad del modelo.
- Contrasta con otros métodos estadísticos o con nuevas muestras.



Fase 5: Evaluación

- Precisión
 - Porcentaje de casos bien clasificados.
- Eficiencia
 - Tiempo necesario para construir/usar el modelo.
- Robustez
 - Frente a ruido y valores nulos.
- Interpretabilidad y Complejidad
 - Economía del pensamiento
 - En igualdad de condiciones la solución más sencilla es probablemente la correcta.



Fase 5: Evaluación Algunas Técnicas

- Técnicas de evaluación generales:
 - Validación simple, validación cruzada
- Clasificación supervisada:
 - Porcentaje de bien clasificados
 - Matriz de confusión



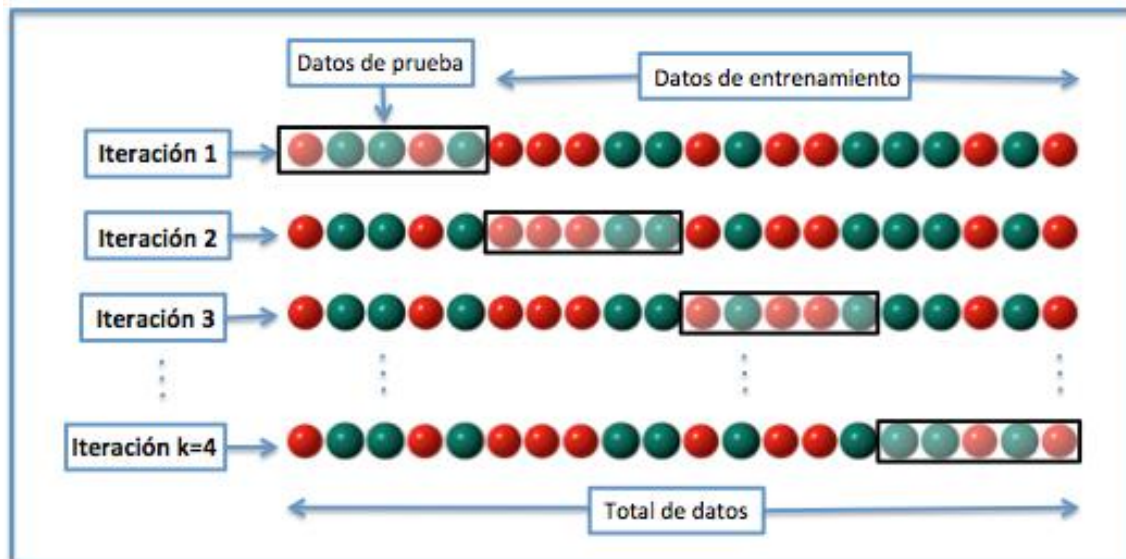
Fase 5: Evaluación Validación Simple

- Separar los datos disponibles en dos subconjuntos de datos:
 - Entrenamiento (para generar un modelo)
 - Test (el resto de los datos)
- Sobre el set de datos de test se estima el error del modelo obtenido con el set de entrenamiento.



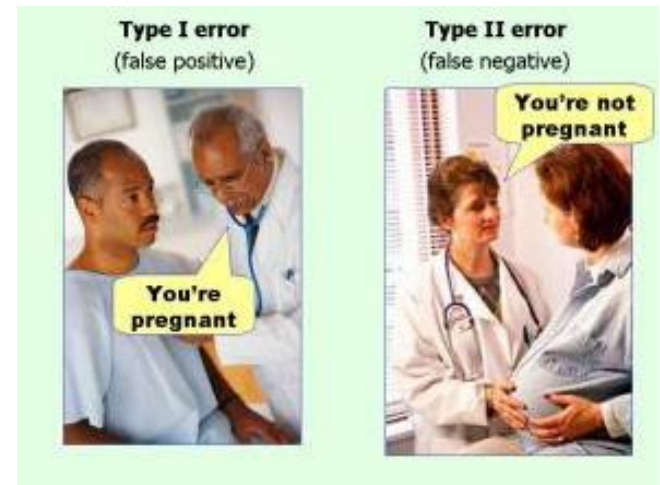
Fase 5: Evaluación k-fold Cross-Validation

- Se divide aleatoriamente el conjunto de datos en k subconjuntos de intersección vacía (más o menos del mismo tamaño). Por lo general se usan 10 partes, “10 fold cross-validation”.
- En la iteración i , se usa el subconjunto i como conjunto de prueba y los $k-1$ restantes como conjunto de entrenamiento.
- Como medida de evaluación del método de clasificación se toma la media aritmética de las k iteraciones realizadas.



Fase 5: Evaluación Matriz de Confusión

		Predicción	
		C_P	C_N
Valor Real	C_P	VP: Verdadero Positivo	FN: Falso Negativo
	C_N	FP: Falso Positivo	VN: Verdadero Negativo



Recall, Precision, Accuracy

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

- **Entrenamiento:** Ajustar los parámetros del algoritmo de forma tal de que se minimicen la cantidad de predicciones que no correspondan a la etiqueta original.
- **Recall:** Porcentaje de clasificados correctamente como positivos sobre todos los que realmente eran positivos.
- **Precision:** Porcentaje de clasificados correctamente como positivos sobre todos los clasificados como positivos.
- **Accuracy:** Porcentaje de clasificados correctamente.

Recall, Precision, Accuracy



Recall sobre Precisión

Detección posible fallo de una maquina minera muy costosa que no debe parar. Importa mas capturar el máximo de posibles casos reales de fallo aunque existan muchas falsas alarmas.

Muchos fp, pocos fn

$$\text{Recall} = \frac{tp}{tp + fn}$$



Precisión sobre Recall

Detección de evasión / infracción. Importa mas no tener falsas alarmas(y acusar injustamente), a que se me pasen algunos casos de evasión.

Muchos fn, pocos fp

$$\text{Precision} = \frac{tp}{tp + fp}$$

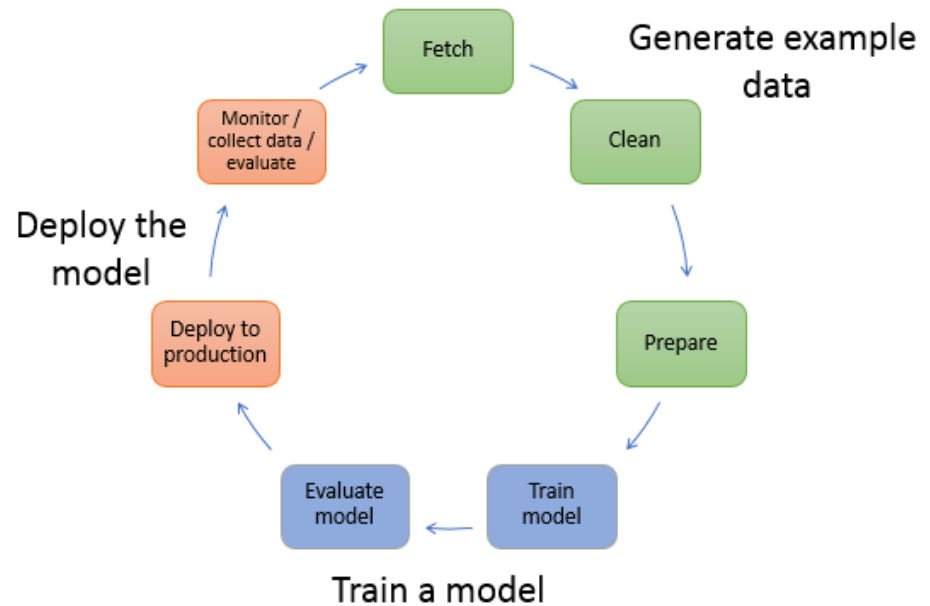


Fase 5: Evaluación Precisión

- Limitaciones de la precisión:
 - Supongamos un problema con 2 posibilidades:
 - 4.990 clientes que se mantienen leales.
 - 10 clientes que fugan de la compañía.
- Si el modelo nos indica que todos los clientes se mantendrán en la compañía, su precisión es:
 - $4.990/5.000 = 99,8\%$
- ... Pero a pesar de esa precisión, el modelo tiene un problema, ya que nunca detectaremos clientes fugados.

Fase 6: Explotación

- Es necesario distribuir, comunicar a los posibles usuarios, integrar lo descubierto al know-how de la organización.
- Medir la evolución del modelo a lo largo del tiempo (los patrones pueden cambiar)
- Modelo debe cada cierto tiempo ser:
 - Reevaluado
 - Reentrenado
 - Reconstruido





Clase 02: Metodologías de Análisis de Datos

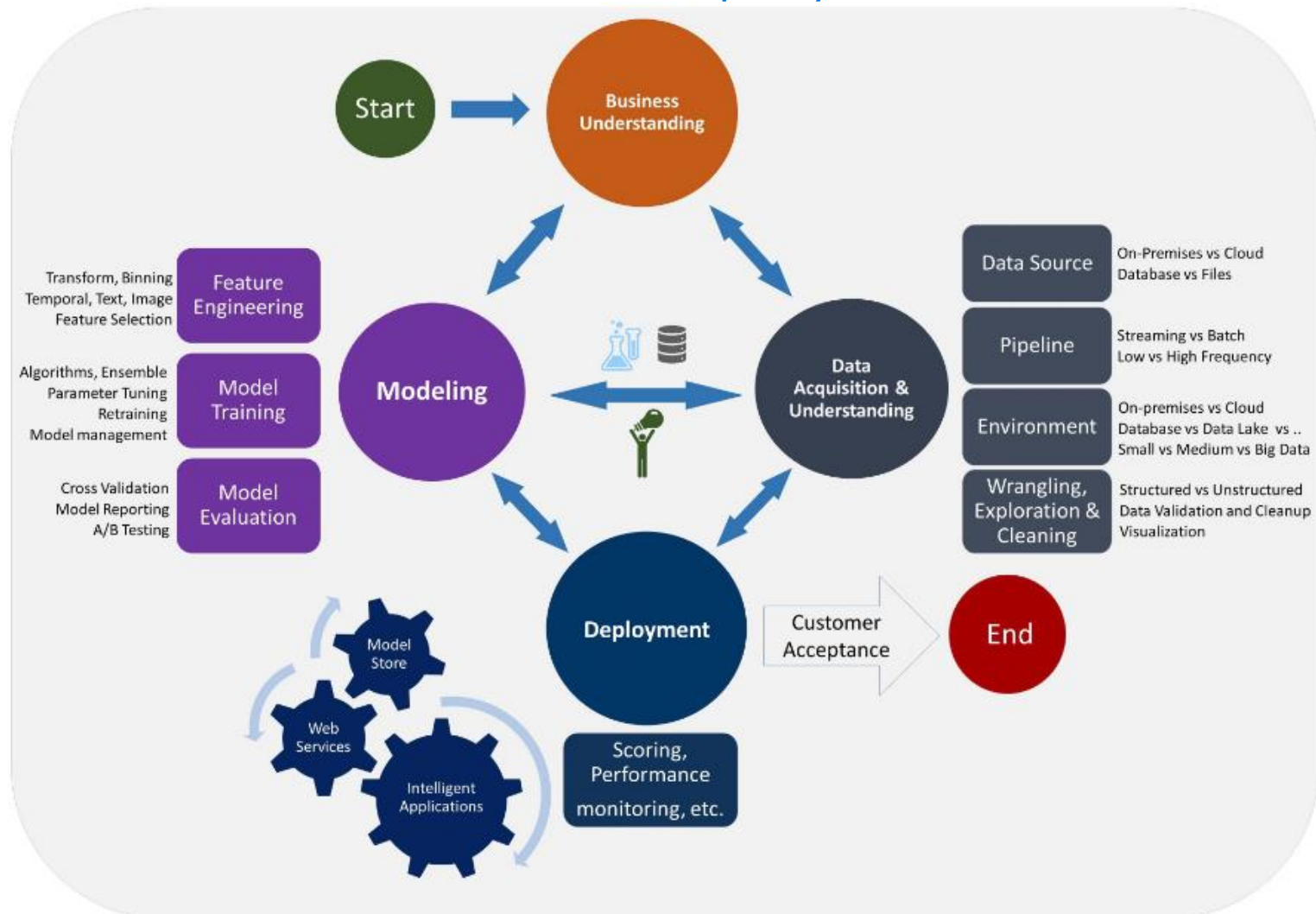
TDSP: FASES



TDSP

- Team Data Science Process
- Metodología de Data Science ágil e iterativa
 - Entregar soluciones analíticas y aplicaciones inteligentes de manera eficiente
- Desarrollado el año 2016 por Microsoft.
- Una mezcla de Scrum y CRISP-DM

Ciclo de vida proyecto





Fases

1. **Entendimiento del negocio:** Definir objetivos e identificar fuentes de datos
2. **Captura y entendimiento de datos:** Ingresar datos y determinar si se pueden responder las preguntas del levantamiento
3. **Modelamiento:** Ingeniería de features y entrenamiento de modelos
4. **Deployment:** Llevar a producción los algoritmos y modelos. Ambiente de producción.
5. **Aceptación del cliente:** Validar con el cliente si el sistema satisface necesidades del negocio



Clase 02: Metodologías de Análisis de Datos

ANALYTICS Y DATA SCIENCE EN CHILE Y EL MUNDO



Chile presenta la primera Política Nacional de Inteligencia Artificial

El trabajo pionero del Ministerio de Ciencia, Tecnología, Conocimiento e Innovación contempla el desarrollo de factores habilitantes, el uso y desarrollo de esta tecnología, y aspectos de ética y seguridad. Junto a esta estrategia nacional, el ministro Andrés Couve presentó un plan de acción que reúne 70 acciones prioritarias y 185 iniciativas desde distintos servicios públicos centradas en aspectos sociales, económicos, y en la formación de talentos con un horizonte de 10 años.

Comparte:  

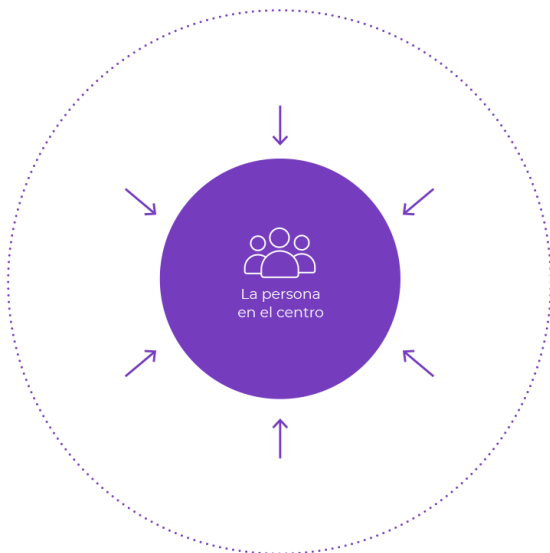
28 Octubre 2021



<https://minciencia.gob.cl/areas-de-trabajo/inteligencia-artificial/politica-nacional-de-inteligencia-artificial/>

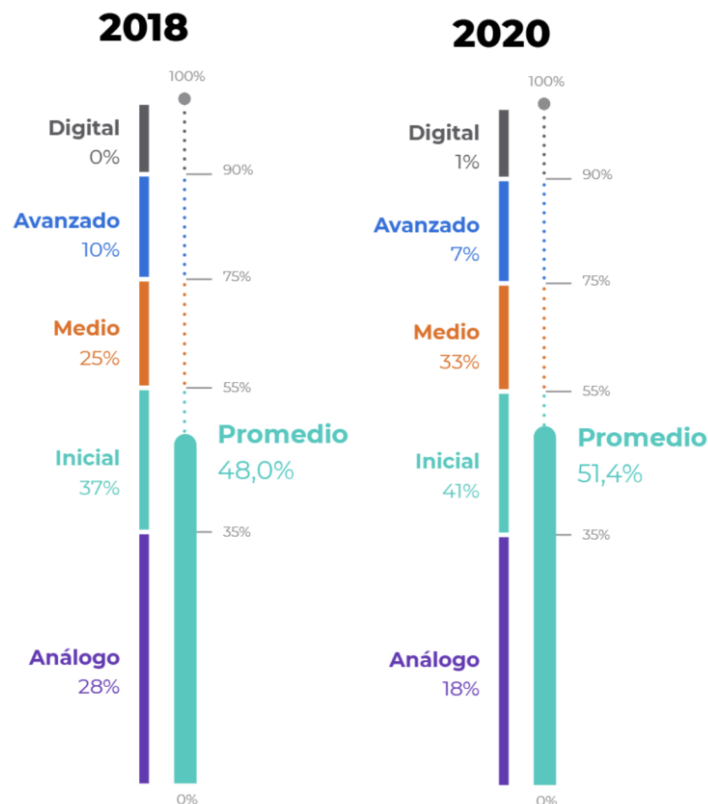
Indice Madurez Digital

<https://www.somosvirtus.com/contenidos/imdv-chile>



Inicial
35% - 55%

Han comenzado un proceso de madurez en la mayoría de las dimensiones. Mientras que en temas estratégicos se encontrarían entrando a un nivel Medio de madurez, en la dimensión de "Tecnología, procesos y operaciones" recién estarían comenzando a madurar. En dimensiones como "Experiencia del cliente", "Innovación y nuevos modelos de negocio", "Data & Analytics", y "Cultura y gestión del cambio", ya habrían dado un primer paso hacia la digitalización, encontrándose en un nivel Inicial.



57.4% 2021 ↑

www.educacionprofesional.ing.uc.cl

Potencial de los datos

¿Estamos aprovechando todo el potencial de la data?

Con más y mejores datos, los insights generados pueden “destrabar” valor sustancial para las organizaciones, tanto de cara al cliente como para la operación interna. Sin embargo, su administración y análisis también se complejiza, por lo que se requiere incorporar nuevas capacidades en un alto nivel de la organización, que permitan darles el mejor uso y obtener el mayor provecho.

➤ Las organizaciones están haciendo un gran esfuerzo por recopilar y analizar datos, pero pocas los usan para mejorar la experiencia de sus clientes y/o para la toma de decisiones de negocios

78%

afirma que su organización está realizando esfuerzos relevantes por contar con más y mejores datos

81%	73%	73%
Grandes empresas	Pymes	Sartups
69%	86%	60%
Instituciones públicas	Centros de conocimiento	Gremios

98%

de los ejecutivos de las organizaciones más avanzadas señalan que están haciendo esfuerzos relevantes por contar con más y mejores datos

50%

afirma que su organización cuenta con un equipo interno de inteligencia de negocios o inteligencia de clientes

59%	36%	39%
Grandes empresas	Pymes	Sartups
23%	54%	20%
Instituciones públicas	Centros de conocimiento	Gremios

33%

sostiene que su organización utiliza los datos para gestionar el negocio de manera efectiva

37%	28%	42%
Grandes empresas	Pymes	Sartups
23%	32%	7%
Instituciones públicas	Centros de conocimiento	Gremios

32%

señala que su organización realiza Data Analytics avanzado

37%	21%	42%
Grandes empresas	Pymes	Sartups
31%	38%	7%
Instituciones públicas	Centros de conocimiento	Gremios

Government AI Readiness Index 2021

Chile



Index Score

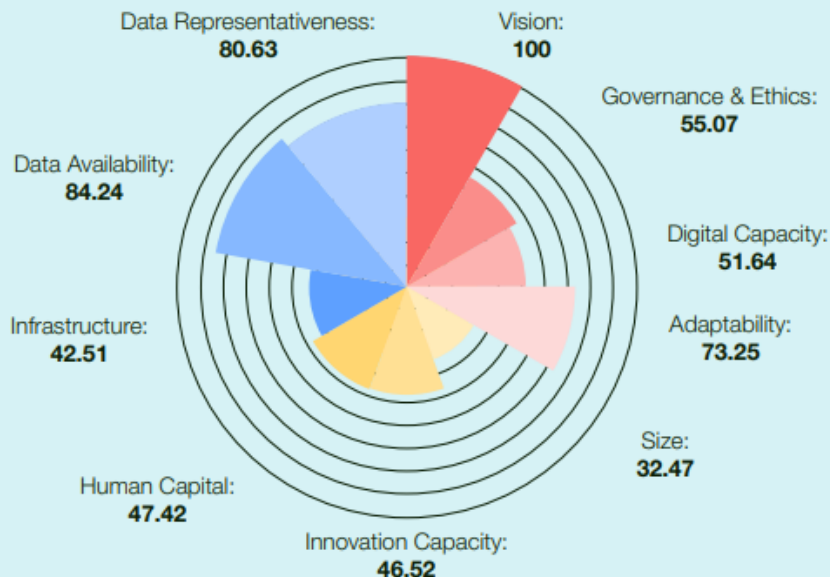
60.42/100

Rank

41/160

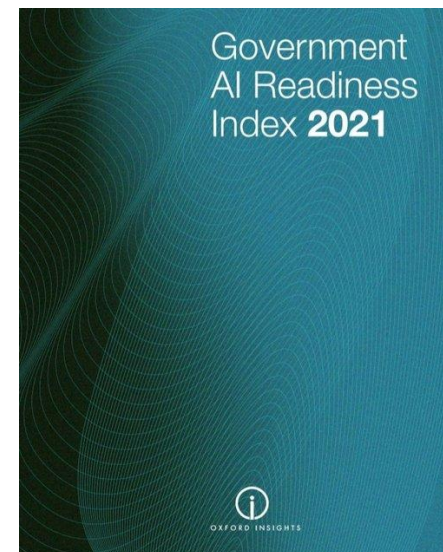
Regional Rank

2/26



The 2021 AI Readiness Index, ranks 160 countries by how prepared their governments are to use AI in public services.

Researchers found that:

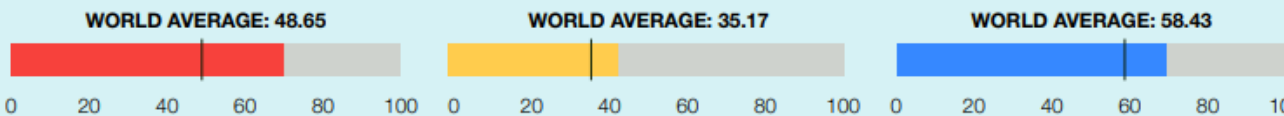


Government

Technology Sector

Data and Infrastructure

Score	Rank	Region Rank	Score	Rank	Region Rank	Score	Rank	Region Rank
69.99/100	33/160	2/26	42.14/100	42/160	2/26	69.13/100	52/160	3/26



Government AI Readiness Index 2022

Chile



2021
Index Score
60.42/100

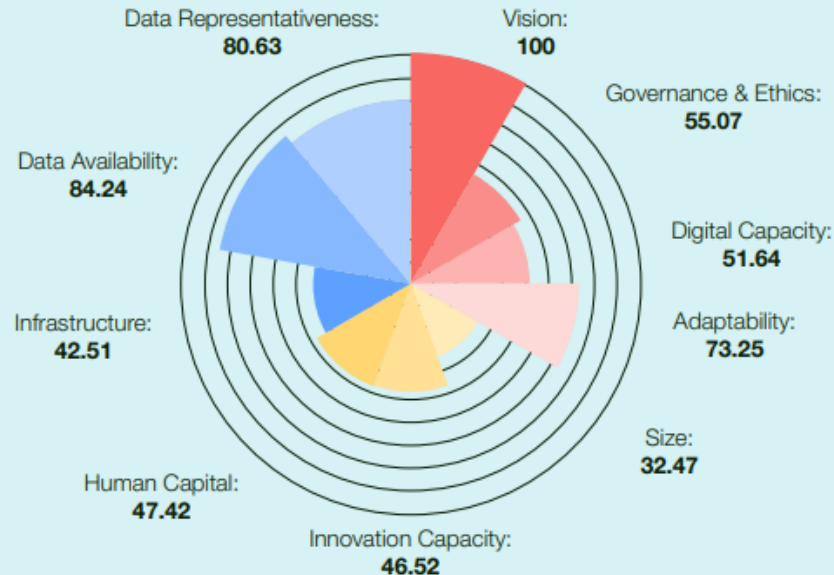
2022
↑ **62.52**

Rank
41/160

↑ **35/181**

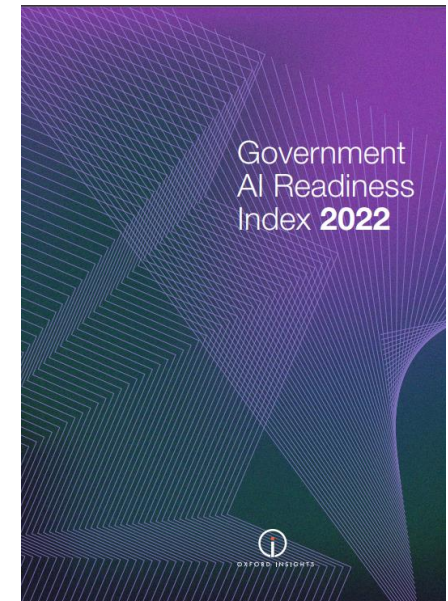
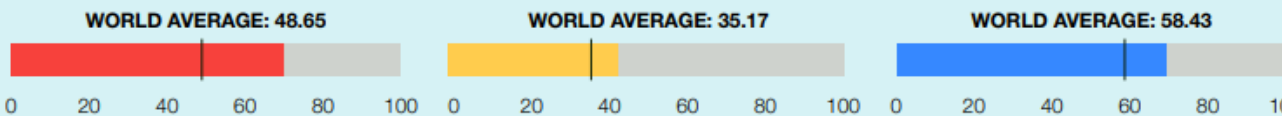
Regional Rank
2/26

↑ **1/31**



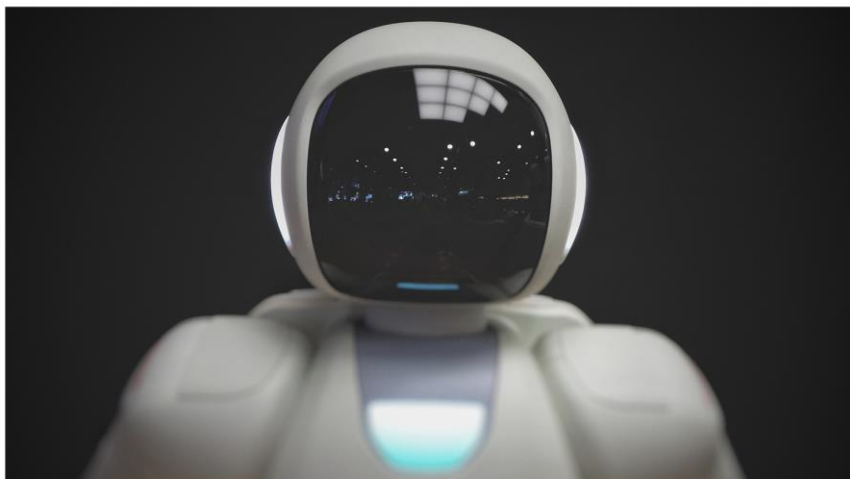
The 2022 AI Readiness Index, ranks 181 countries by how prepared their governments are to use AI in public services. Researchers found that:

Government			Technology Sector			Data and Infrastructure		
Score	Rank	Region Rank	Score	Rank	Region Rank	Score	Rank	Region Rank
69.99/100	33/160	2/26	42.14/100	42/160	2/26	69.13/100	52/160	3/26



Data science y Covid-19

<https://uddventures.udd.cl/blog/ciencia-de-datos-e-inteligencia-artificial-para-contener-la-pandemia>



EMPRENDEDORES

Ciencia de datos e inteligencia artificial para contener la pandemia.

18 de agosto, 2020 / por Jaime Caiceo



N personas: 39 N Sin Riesgo: 29 N Riesgo Bajo: 1 N Riesgo Alto: 9
Análisis de distanciamiento social COVID-19

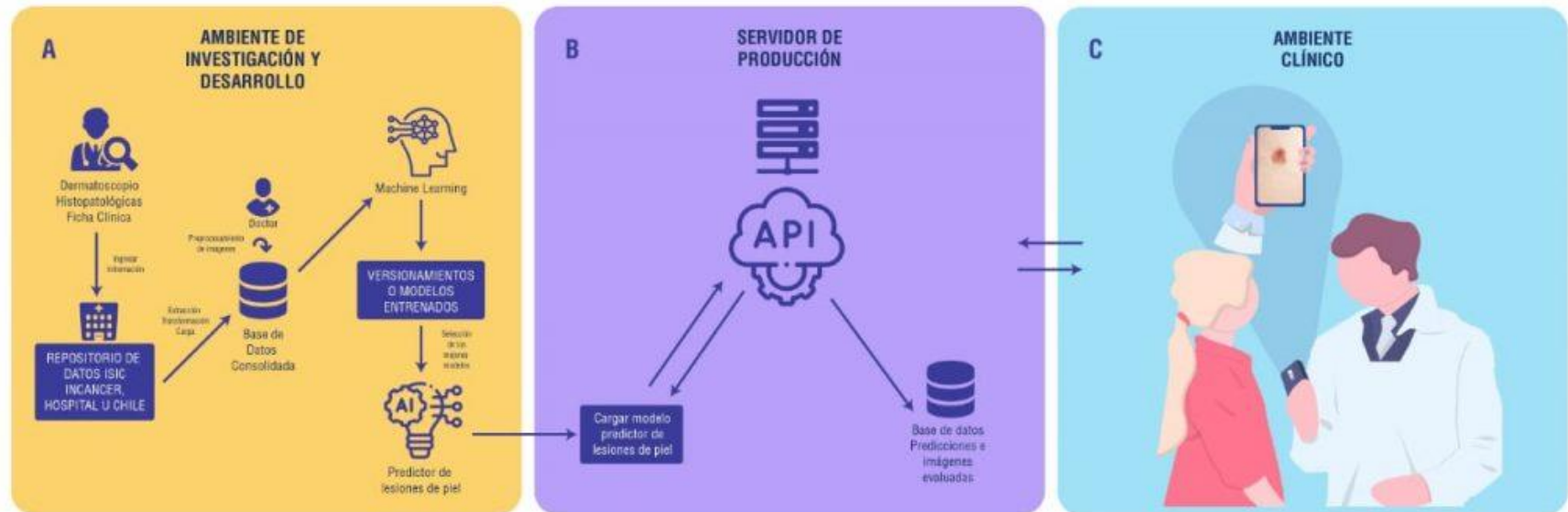
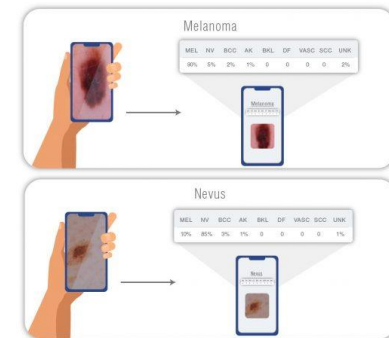
Líneas muestran proximidad entre los personas.		Rectángulos muestran el nivel de riesgo de la persona.	
---	Amarillo: Cercano	---	Bajo Riesgo Alto
---	Verde: Sin Riesgo	---	Alto Riesgo Bajo
---	Verde: Sin Riesgo	---	Verde: Sin Riesgo

www.educacionprofesional.ing.uc.cl

Detección de Melanomas

Miércoles 27 abril de 2022 | 12:59

Investigadores chilenos desarrollan inteligencia artificial para detectar melanomas con una imagen



<https://www.tourinnovacion.cl/tecnologia-ciencia/inteligencia-artificial-desarrollada-en-chile-detecta-melanoma/>

www.educacionprofesional.ing.uc.cl



Zippedi: Robótica e Inteligencia Artificial



The Not Company: Alimentos basados en plantas



En seis años la empresa recaudó cientos de millones de dólares en inversiones, salió de Chile y está presente en más de una decena de países; captó la atención e inversión de una de las personas más ricas del mundo, el fundador de Amazon, Jeff Bezos, y alcanzó una valoración superior a los US\$ 1.000 millones.



Desde la combinación perfecta entre el horario de estudiantes, profesores y disponibilidad de salas, pasando por la medición del aprendizaje del alumno, hasta la capacidad de analizar si un estudiante podría abandonar una carrera, son parte de las áreas de estudio y propuestas de solución que generan los softwares desarrollados por U-Planner, basados en Inteligencia Artificial.

AI – Defensa



Múltiples aplicaciones de AI en aspectos de Defensa, desde modelos de optimización a técnicas de visión computacional.





ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

Diplomado en Big Data y Ciencia de Datos *Ciencia de Datos y sus Aplicaciones*



www.educacionprofesional.ing.uc.cl