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

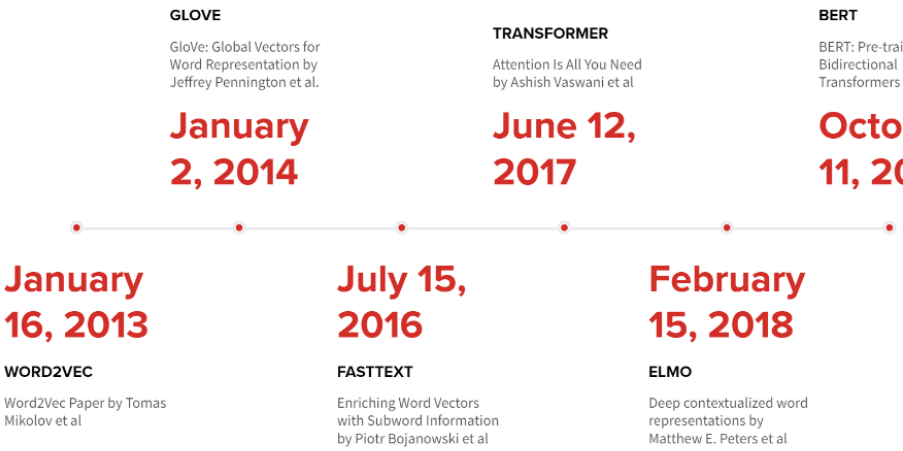
Roberto González

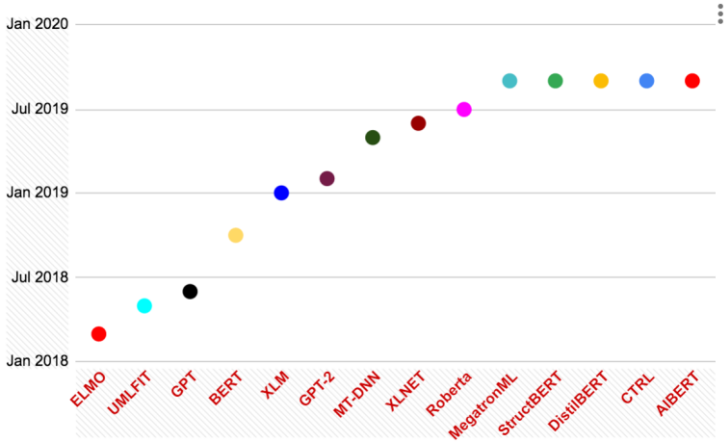
regonzar@uc.cl

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





Abril 2020
General Language Understanding - Human parity
Conversational Question Answering System CoQA

Natural Language Processing

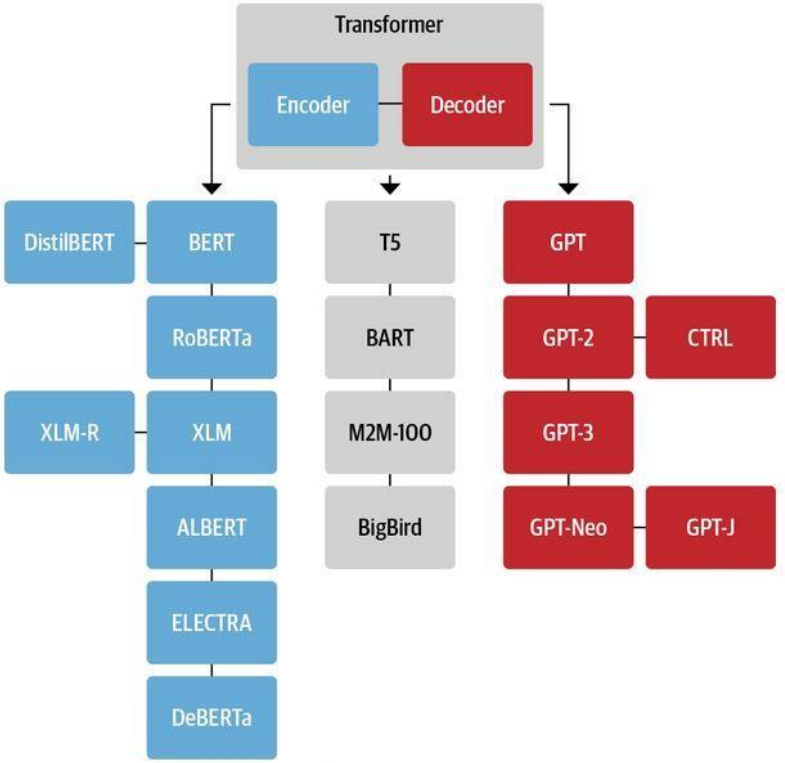
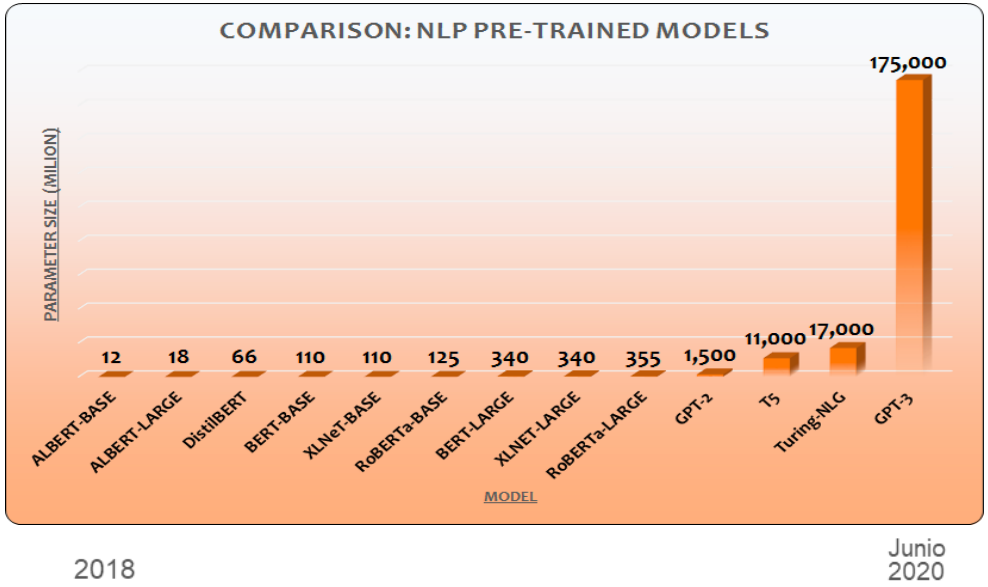


Figure 3-8. An overview of some of the most prominent transformer architectures

45 TB
texto
Train Set

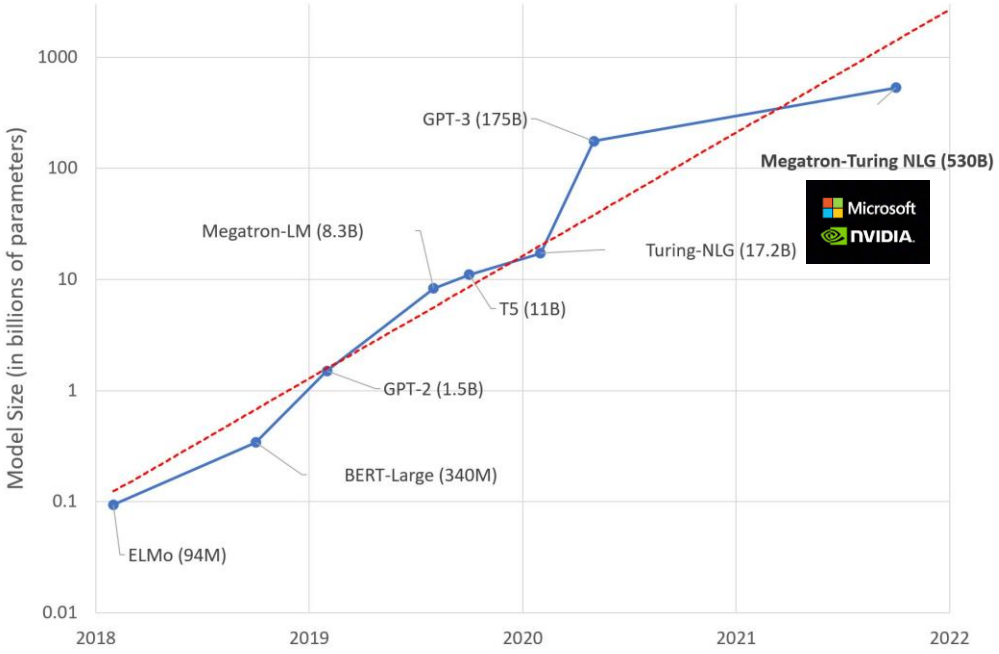


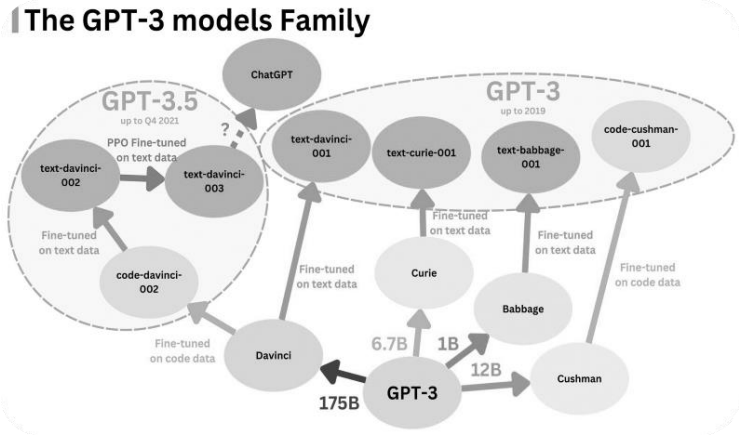
GPT-3 y LLMs

https://openai.com/

Artificial General Intelligence(AGI) para todos

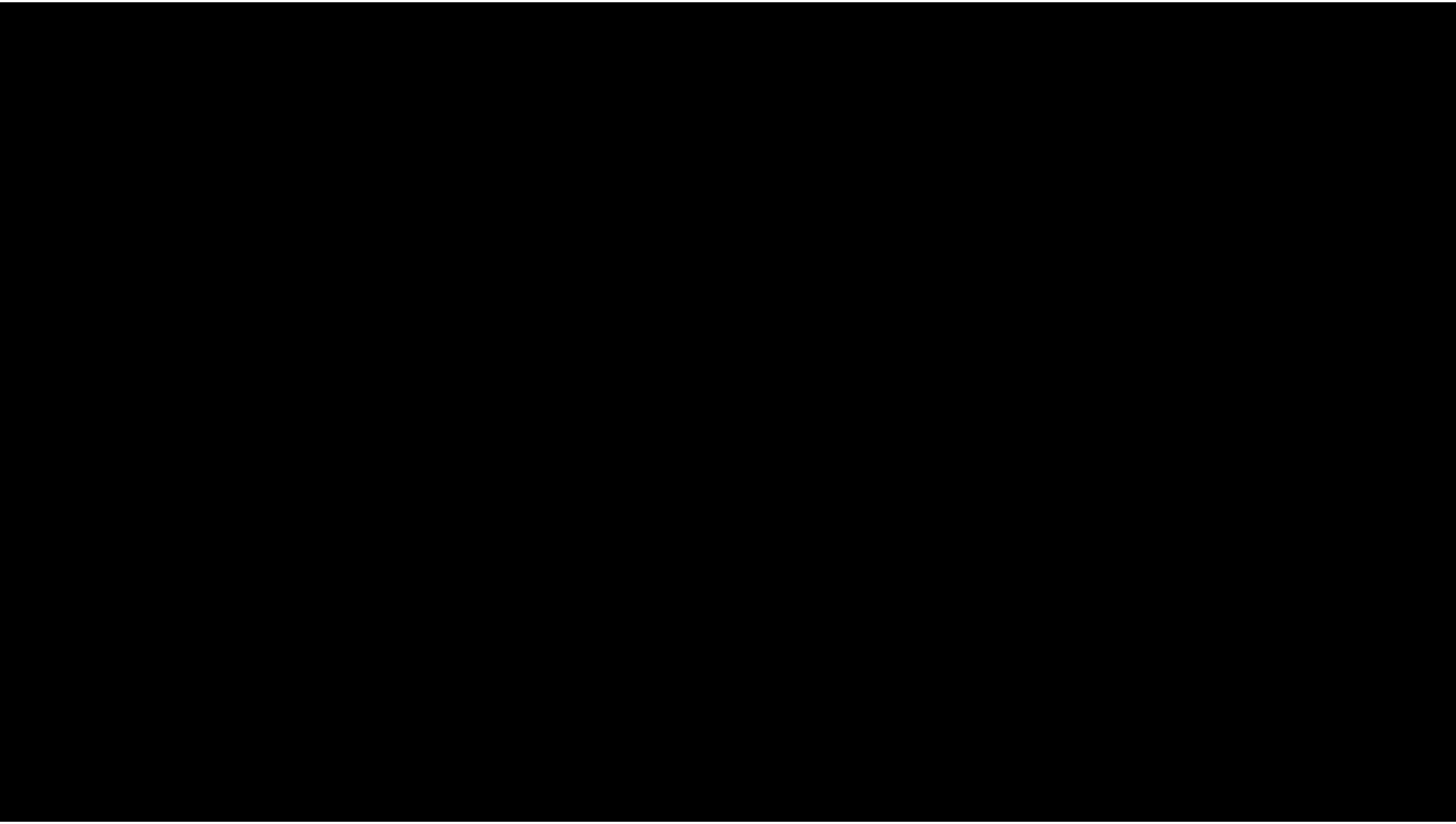
Modelo de NLP desarrollado por OpenAI(US$4.6M training)



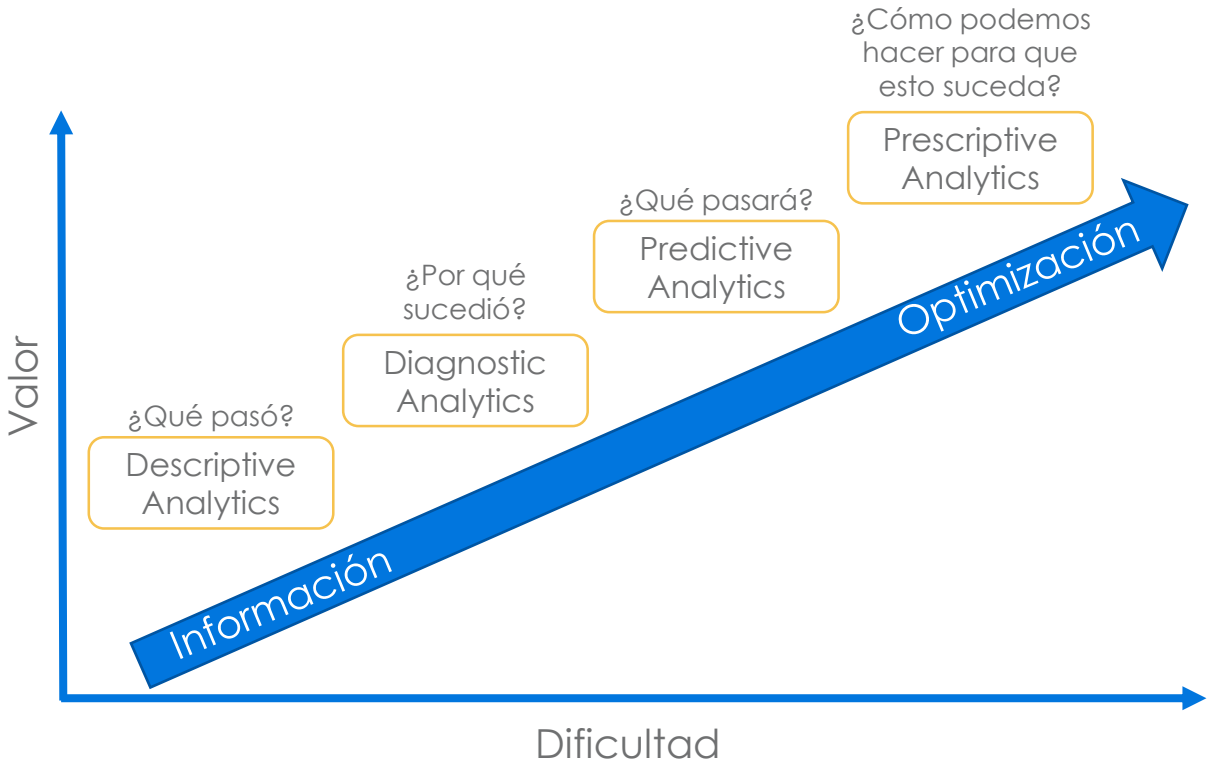


Parámetros se duplican cada 10 meses - Nueva ley de moore?

Dall-E 2



La evolución de los temas analíticos



Fuente: Gartner Business Intelligence & Analytics Summit 2013

https://youtu.be/UFUwNBCkmYs

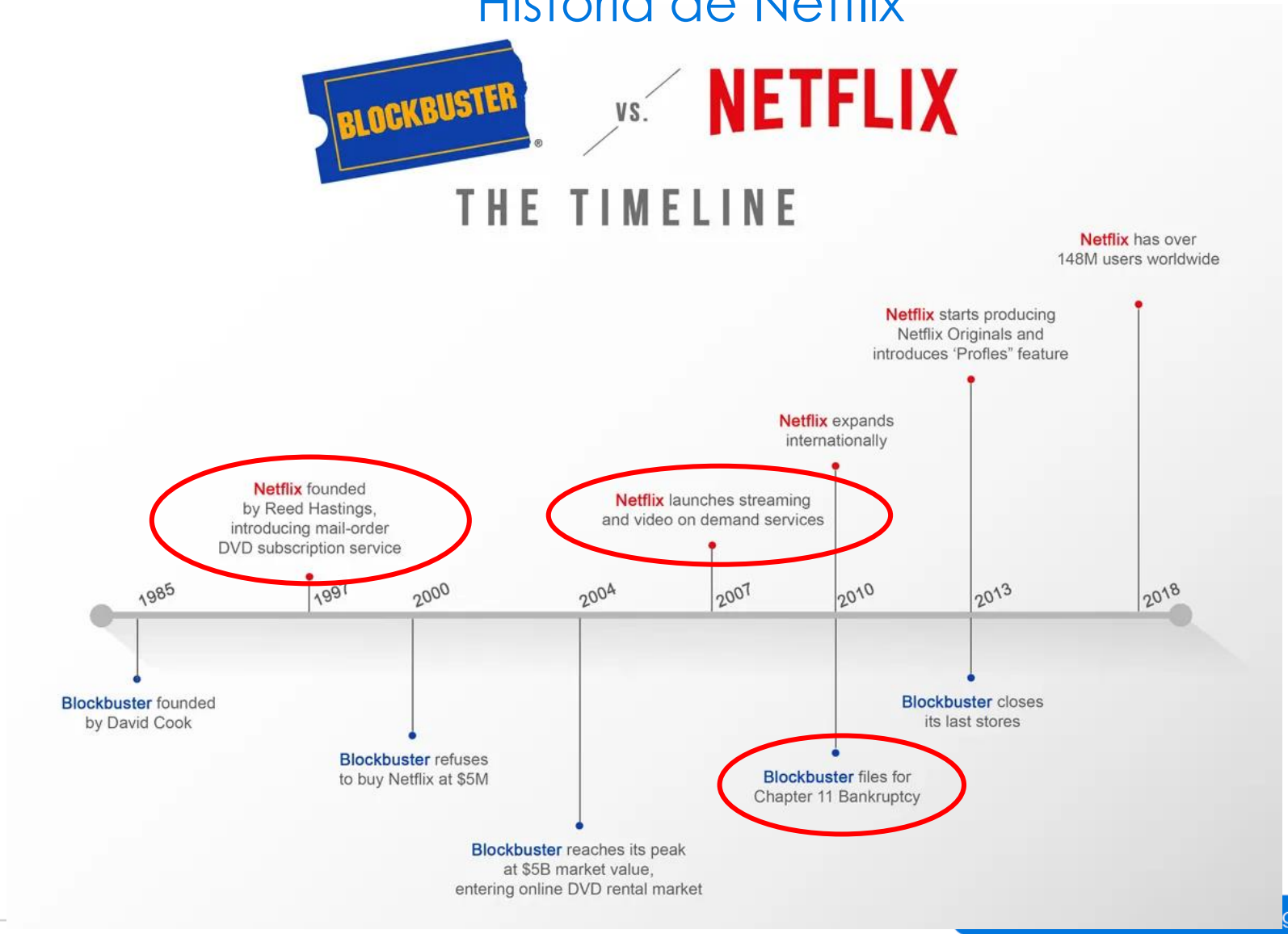
Blockbuster vs Netflix

	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

Netflix touts $900k AI jobs amid Hollywood strikes

By Max Matza
BBC News

28 July 2023

Netflix has triggered an angry response from striking Hollywood actors and writers after posting a job advert for an artificial intelligence (AI) expert.

Clase 02: Metodologías de Análisis de Datos

METODOLOGÍAS

Profesor: **Roberto González**

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.

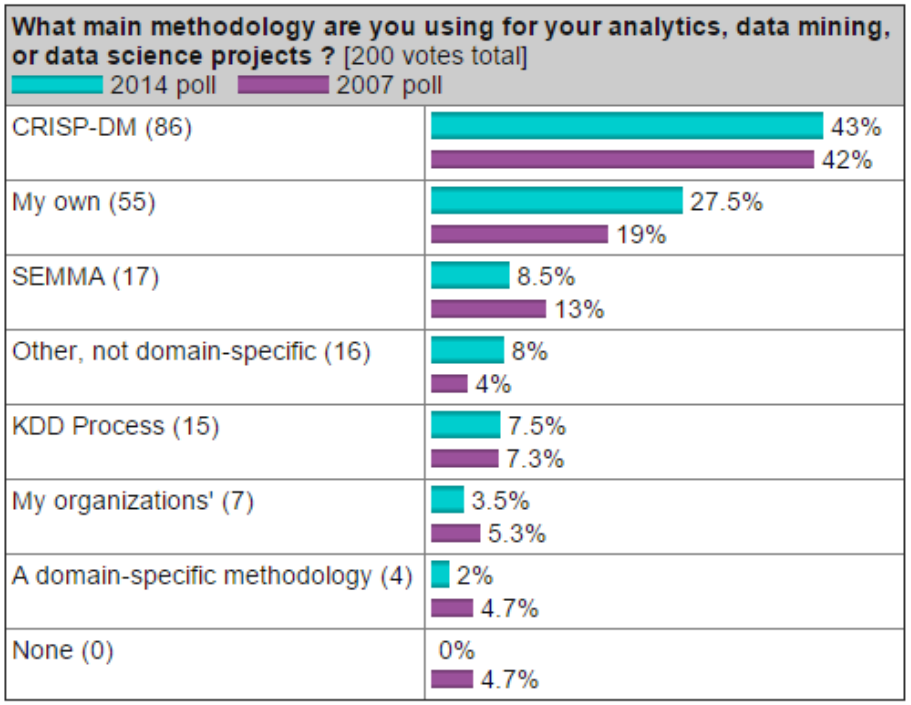
¿Por qué se utilizan?

- 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

Clase 02: Metodologías de Análisis de Datos

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

Profesor: **Roberto González**

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

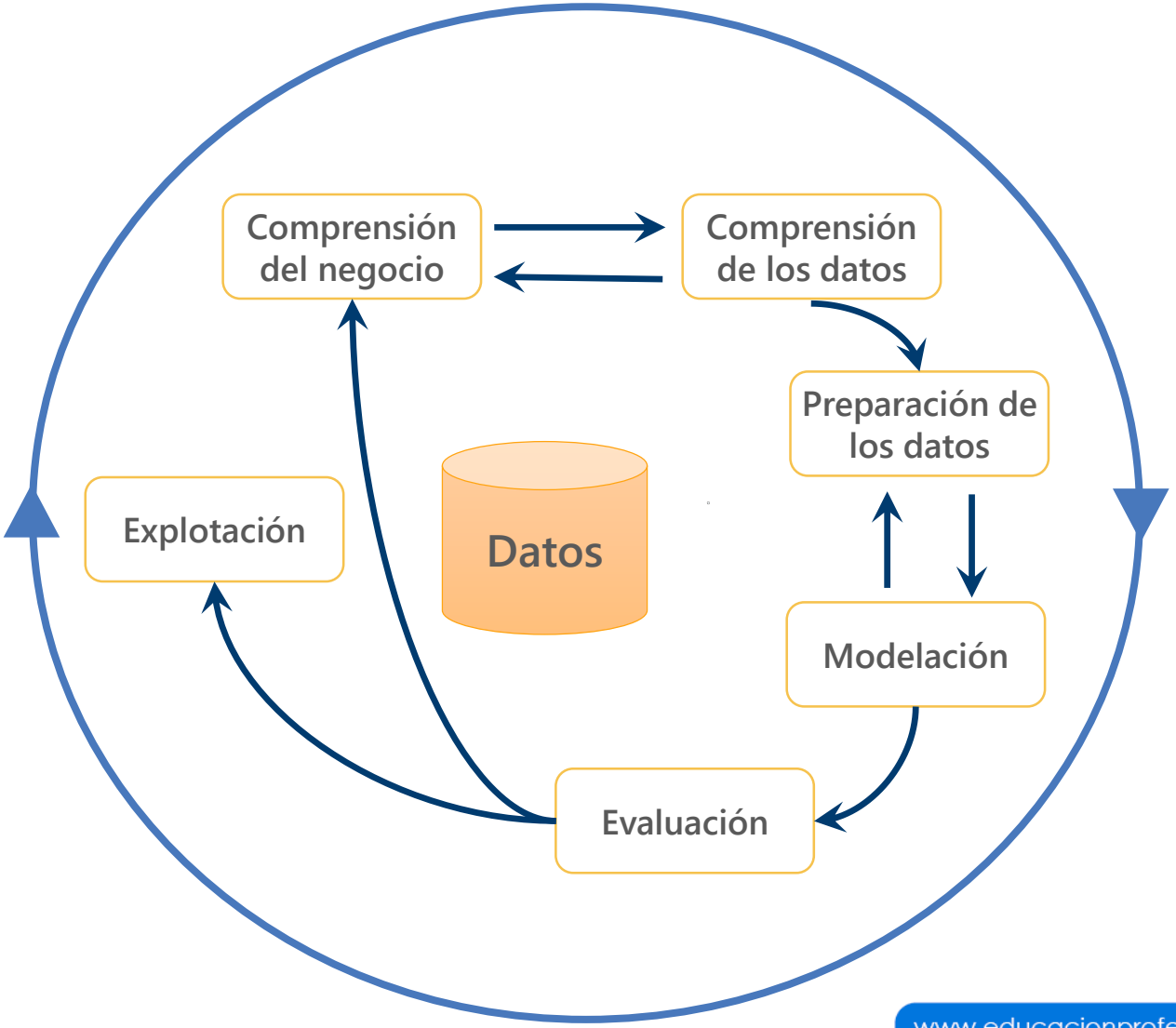
como podemos extraer valor a los datos?

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.

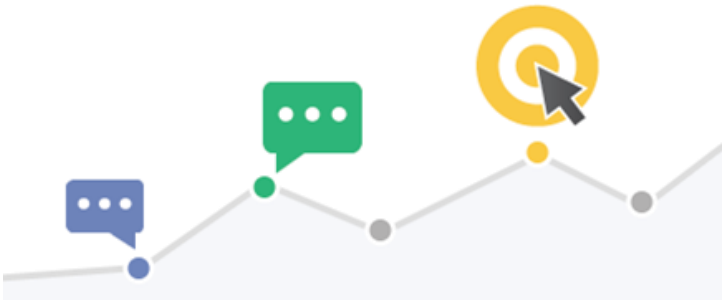
En resumen, para extraer valor de los datos es necesario seguir un proceso riguroso que implica recolectar, limpiar, analizar e interpretar los datos para tomar decisiones informadas.

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.

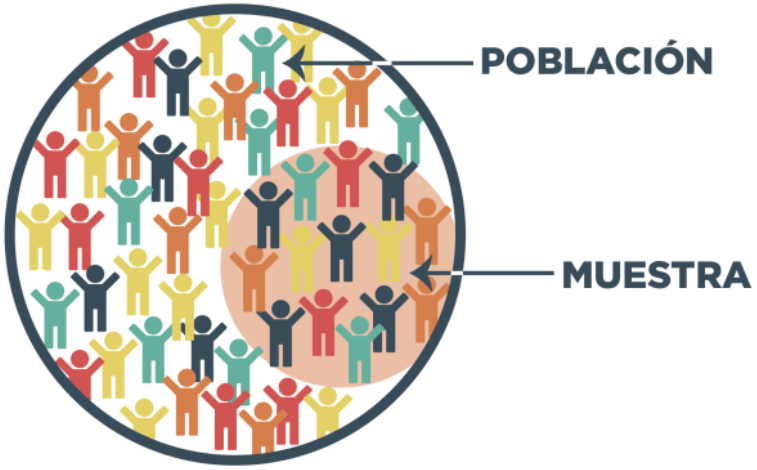
- ¿Cómo?
  - Uso de herramientas de visualización y técnicas de estadísticas descriptivas.

- 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

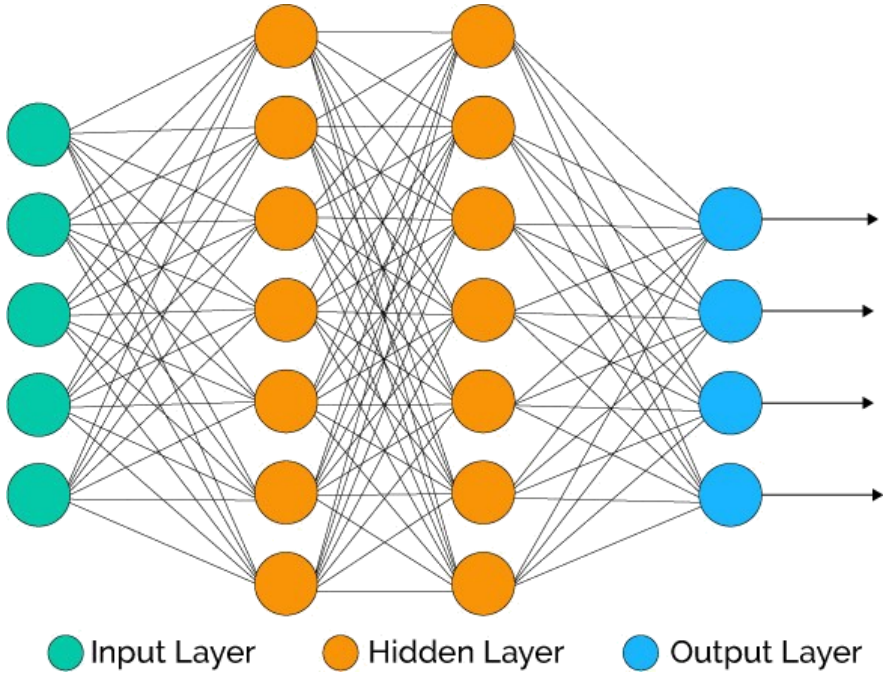
- Normalización de datos
- Construcción de nuevas variables que faciliten el proceso de minería de datos.
- Reducción de Dimensionalidad
  - Variables Correlacionadas
- Discretización de variables continuas

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

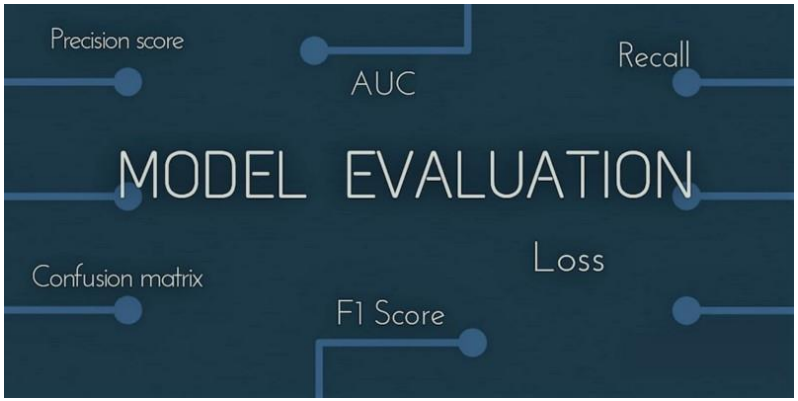
- Las técnicas más utilizadas:
  - Métodos de clustering
  - Análisis de regresión
  - Redes neuronales
  - Árboles de decisión
  - Reglas de asociación
  - Etc.

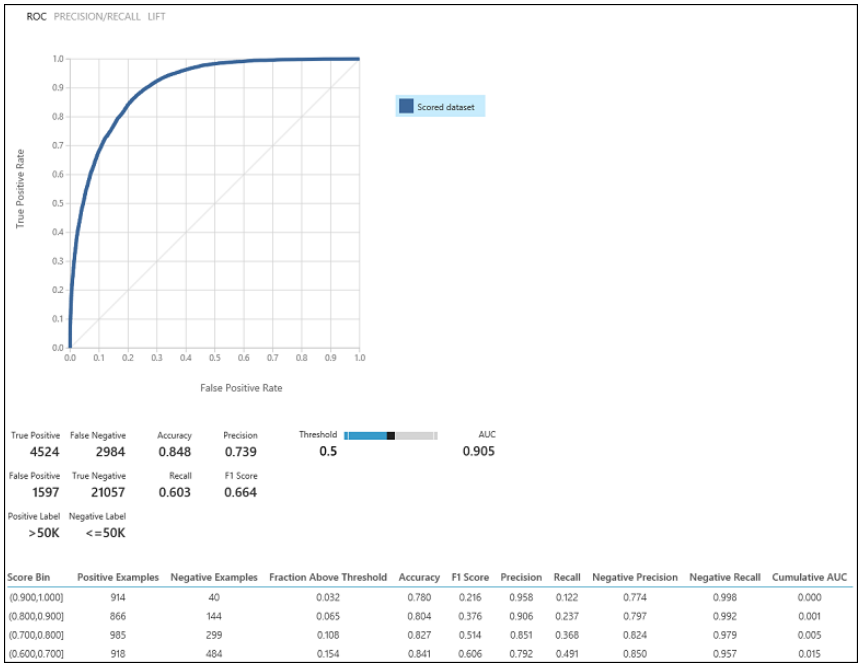


Fase 5: Evaluación

- Valora los resultados mediante el análisis de bondad del modelo.

- Contrasta con otros métodos estadísticos o con nuevas muestras.





Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	914	40	0.032	0.780	0.216	0.958	0.122	0.774	0.998	0.000
(0.800,0.900]	866	144	0.065	0.804	0.376	0.906	0.237	0.797	0.992	0.001
(0.700,0.800]	985	299	0.108	0.827	0.514	0.851	0.368	0.824	0.979	0.005
(0.600,0.700]	918	484	0.154	0.841	0.606	0.792	0.491	0.850	0.957	0.015

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

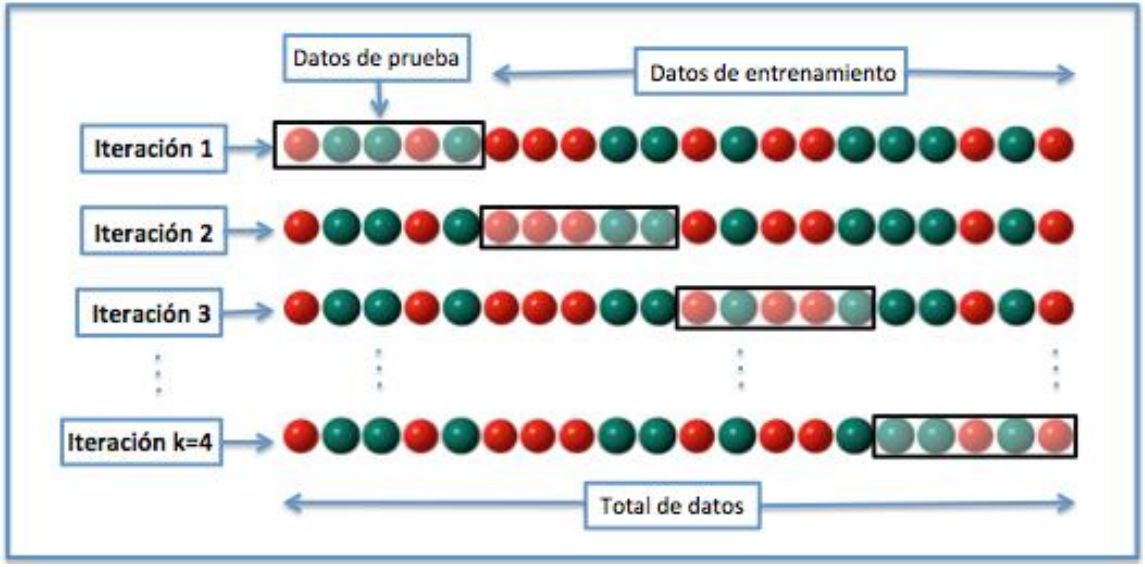
Fase 5: Evaluación
Validación Simple

- 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.

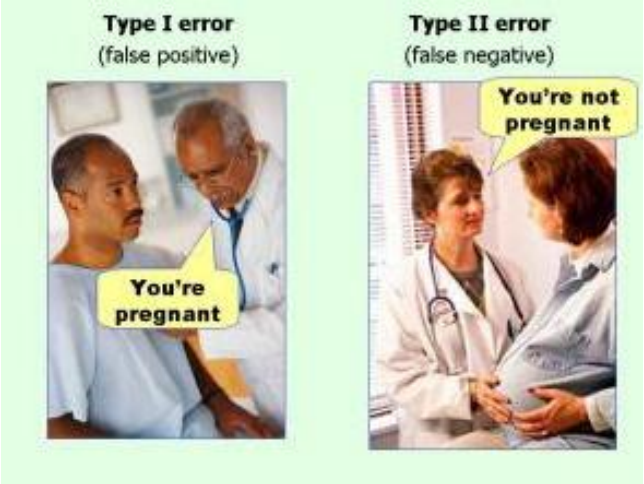
Fase 5: Evaluación
k-fold Cross-Validation

- 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

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

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

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

Fase 5: Evaluación
Precisión

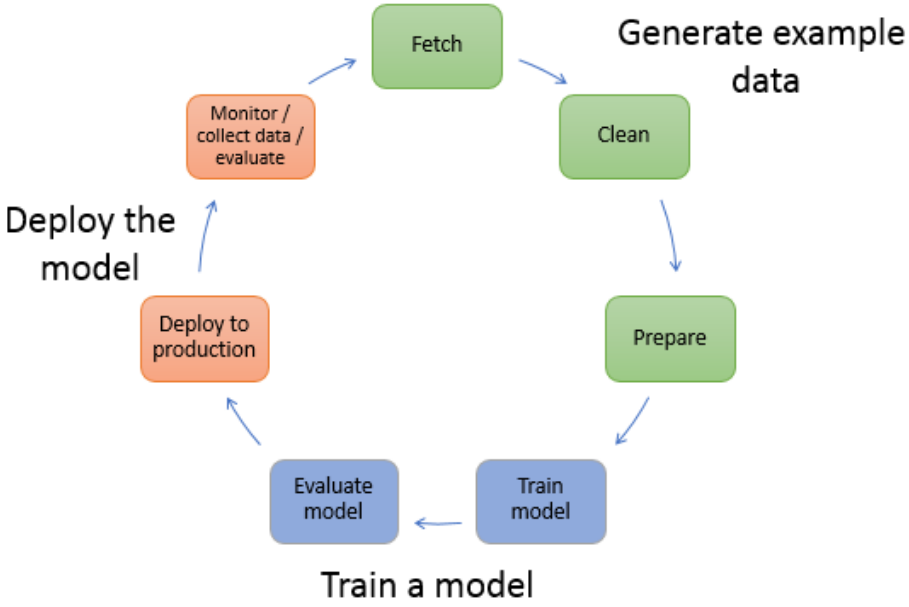
- Limitaciones de la precisión:
  - Supongamos un problema con 2 posibilidades:
    - 4.990 clientes que se mantienen leales.
    - 10 clientes que fugan de la compañía.
- Si el modelo nos indica que todos los clientes se mantendrán en la compañía, su precisión es:
  - 4.990/5.000 = 99,8%
- … Pero a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



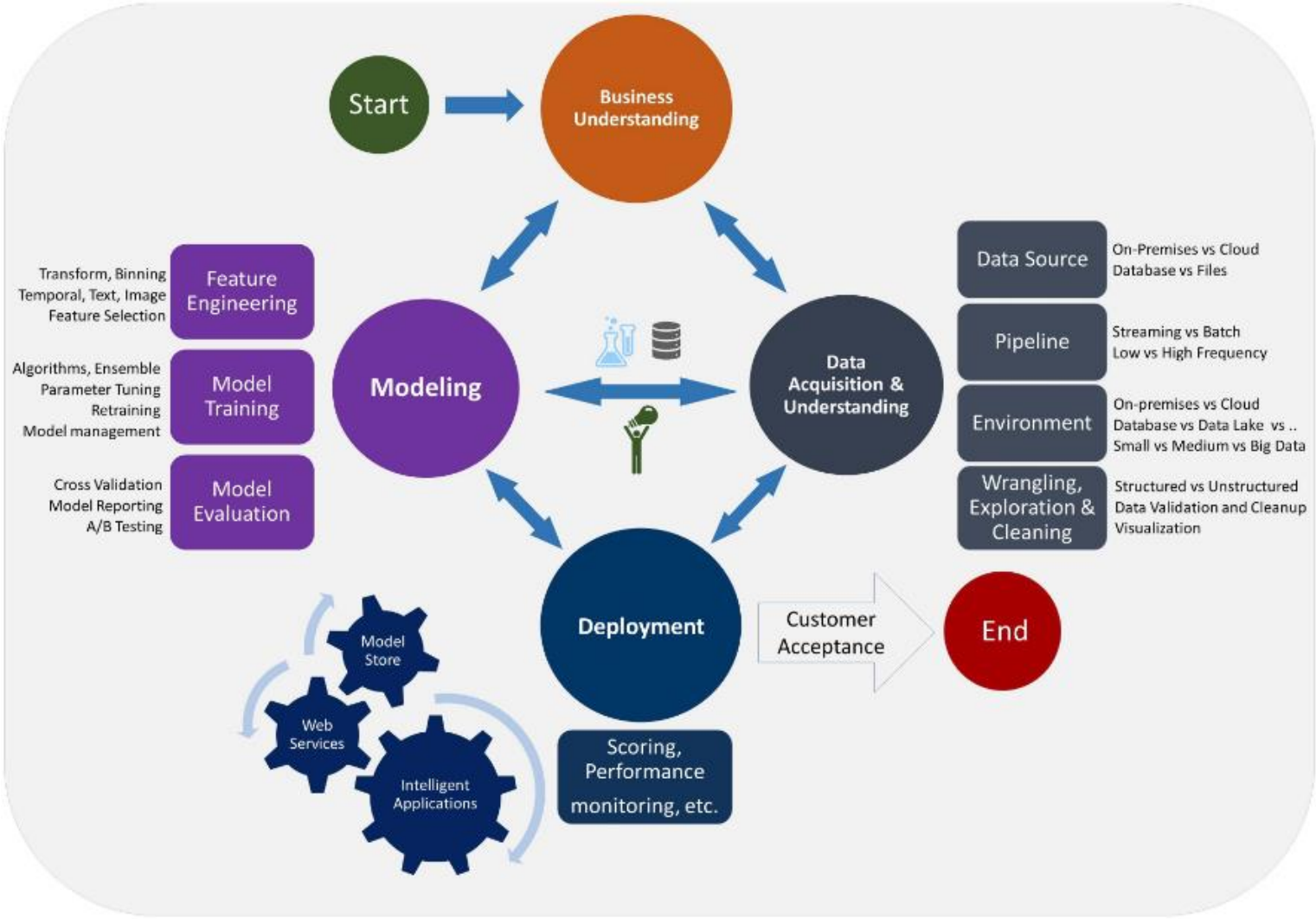
Clase 02: Metodologías de Análisis de Datos

TDSP: FASES

TDSP

- Team Data Science Process
- Metodología de Data Science agil 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 satisfice necesidades del negocio

Clase 02: Metodologías de Análisis de Datos

ANALYTICS Y DATA SCIENCE EN CHILE

Profesor: **Roberto González**

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

Noticia ›

Ministerio de Ciencia, Tecnología e Innovación de Chile, con apoyo de UNESCO, inicia actualización de política de Inteligencia Artificial

El trabajo de renovación comenzó con el taller denominado "El futuro de la Inteligencia Artificial en el Estado", actividad en la que participaron más de 30 expertos y expertas en esta materia.

15 de Junio de 2023

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

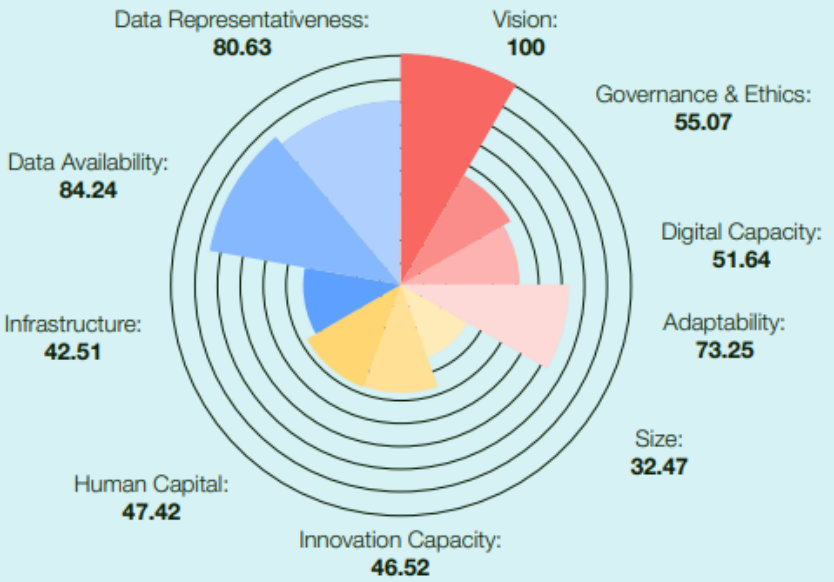
Government AI Readiness Index 2021

Chile

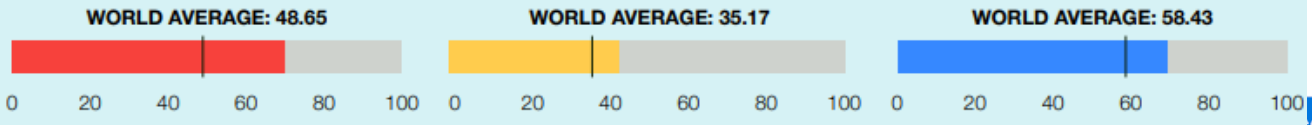
Index Score
60.42/100

Rank
41/160

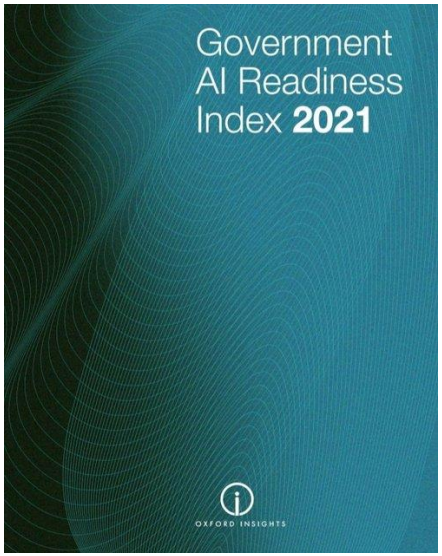
Regional Rank
2/26



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



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

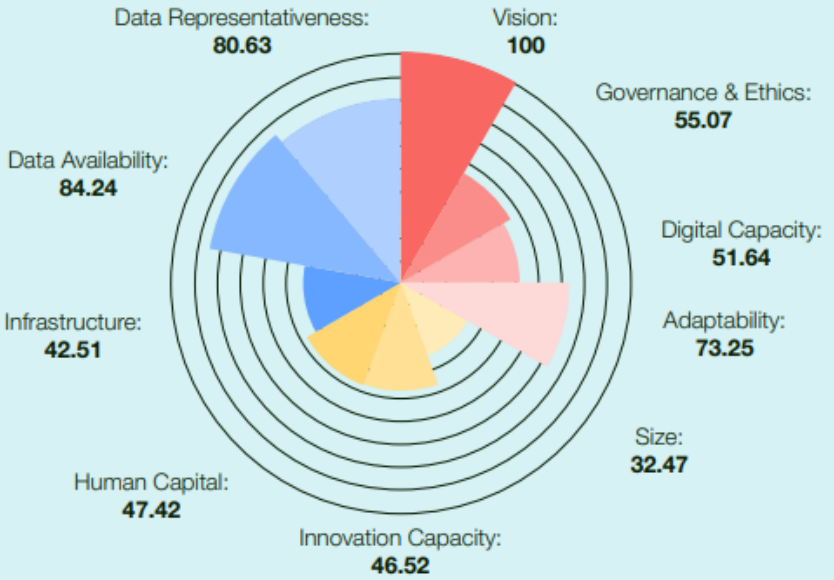


https://www.oxfordinsights.com/government-ai-readiness-index2021

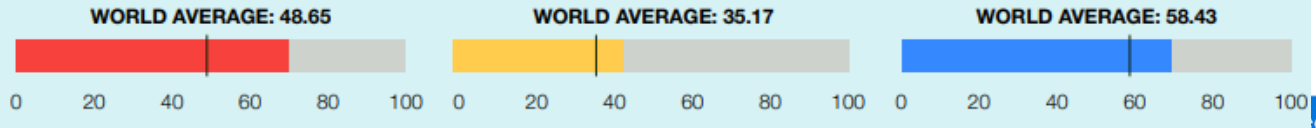
Government AI Readiness Index 2022

Chile

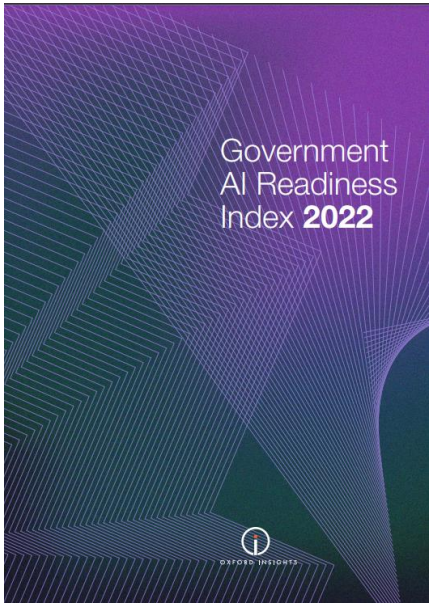
2021	2022
Index Score	
60.42/100	**62.52**
Rank	
41/160	**35/181**
Regional Rank	
2/26	**1/31**



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



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



Otro índice de IA en Latam

Índice **latinoamericano** de inteligencia artificial

Lanzamiento 11 Ago 2023

https://indicelatam.cl/

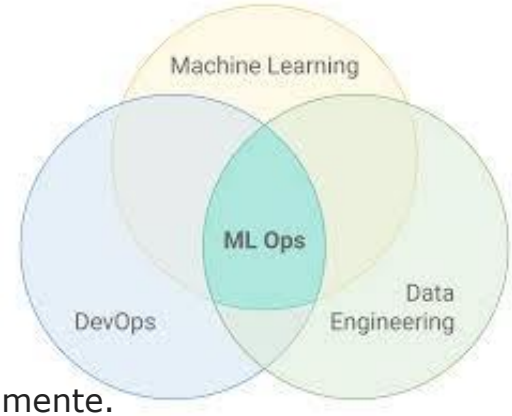
https://www.youtube.com/watch?v=dBcHYgCq8jM&t=8110s

Clase 02: Metodologías de Análisis de Datos

IMPLEMENTACIÓN EN PRODUCCIÓN

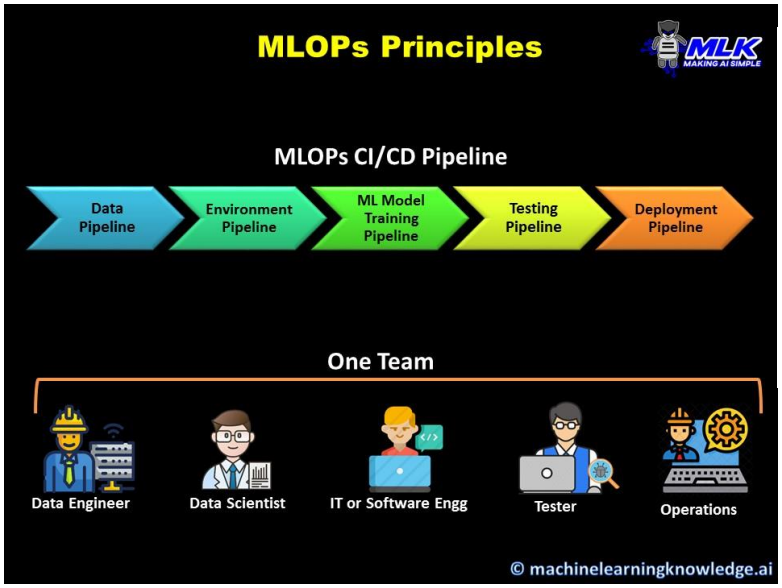
Profesor: **Roberto González**

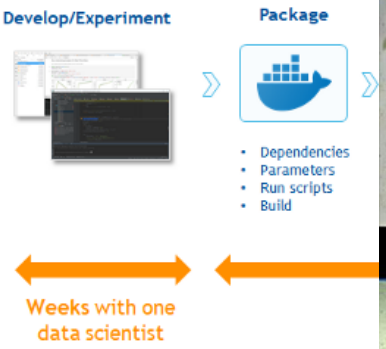
Tendencias



MLOPS – Machine Learning Ops.

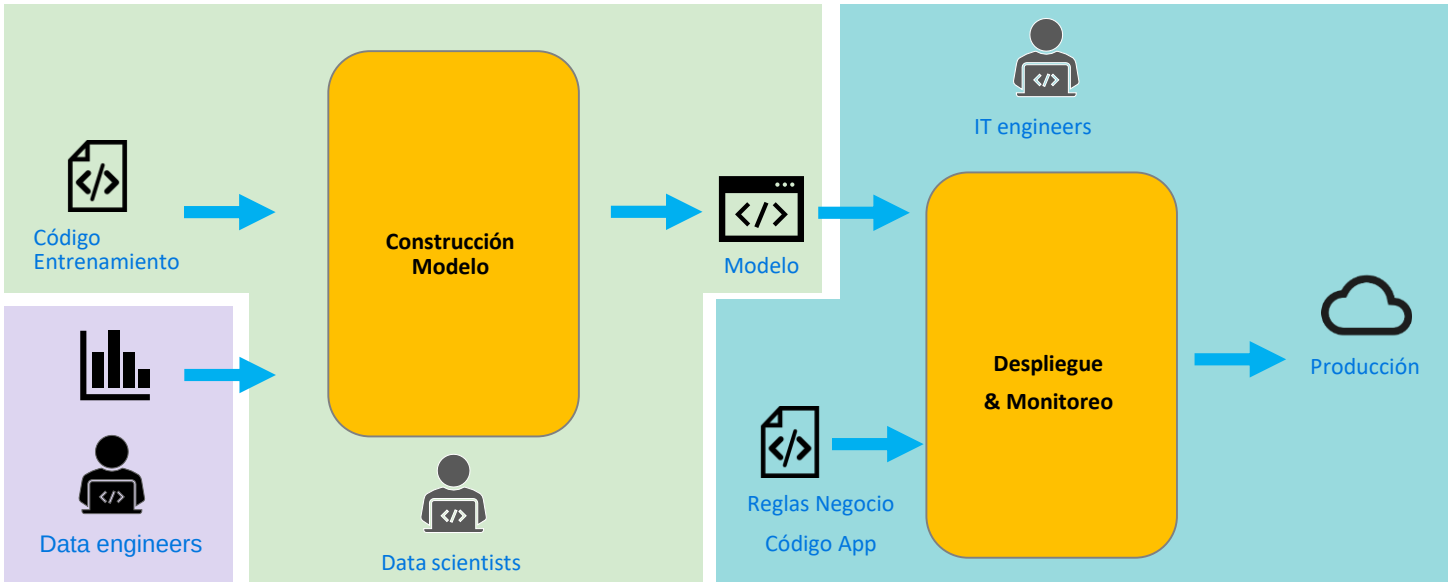
- Solo **22%** de los proyectos de ML son implementados en producción existosamente.
- **39%** de los data scientist enfrenta grandes desafíos para poder administrar dependencias y ambientes en la etapa de implementación productiva.
- **38%** de data scientist admiten falta de conocimientos para implementar una solución productiva.
- **43%** de encuestados dicen encontrar dificultad para escalar su modelo y satisfacer las necesidades productivas de su organización. (Encuesta Anaconda 2020)
- Continouos Integration/Delivery (CI/CD)





Desafío organizacional: divisiones aisladas

En muchas ocasiones las organizaciones están formadas por grupos funcionales aislados(equipos TI, SW Eng, DS, etc) creando barreras y debilitando la habilidad de automatizar procesos end-to-end relativos ML.



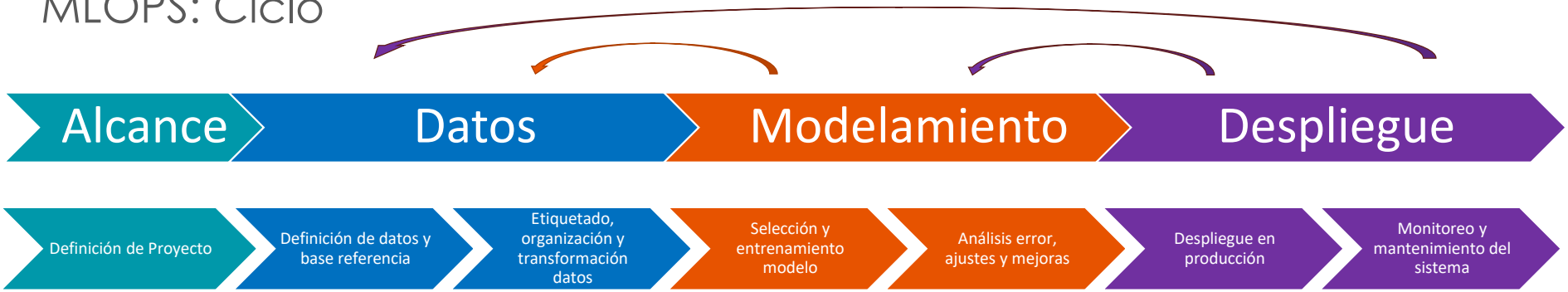
Tendencias

MLOPS: Centrado en el Modelo → Centrado en los Datos

Enfoque Centrado en el Modelo	Enfoque Centrado en los Datos
• Datos estáticos, mejorar iterativamente el código/modelo con métricas de rendimiento • Asociado en general a la academia, donde se busca perfeccionar modelos para un problema fijo(i.e. kaggle data challenges) • Maximizar rendimiento, Human-level performance(HLP)	• Código/Modelo fijo y mejorar iterativamente el proceso y datos • Mejorar consistencia y calidad de datos(Labeling, features): Big Data → Good Data • Datos cambian constantemente en ambientes productivos(Data Drift) • Asociado en general a casos aplicados en la industria(i.e. sistema recomendador productos que cambian constantemente) • Resultados (iteraciones rápidas y continuas), rendimiento alineado a métricas del negocio • Datos libres de sesgos, es justo?

Tendencias

MLOPS: Ciclo



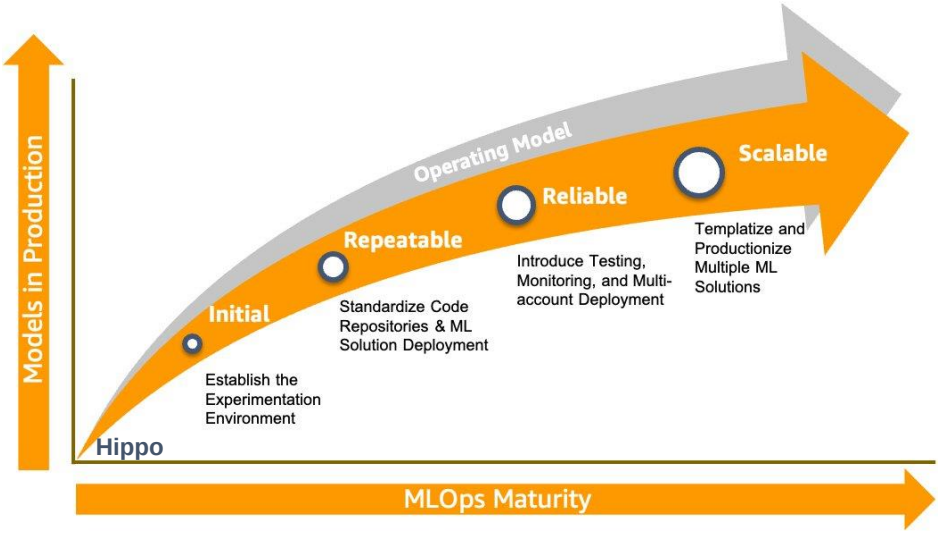
- Big data >> Good data
- Raw data >> Features
- Ingesta de datos entrenamiento e inferencia
- Consistencia y sesgos
- Transformaciones
- Esquema datos y validación
- Bases de referencia para entrenar, validar, servir
- Versionamiento de datos y esquemas

- Modelamiento centrado en datos
- Versionamiento de modelos
- Evaluación
- Métricas y rendimiento
- Re-entrenamientos
- Sesgos modelo, Fairness
- Ajuste parámetros
- Alcance modelo-datos, bias & variance
- Consideraciones escalabilidad

- Pipelines y orquestación
- Monitoreo modelo producción
- Auditoria rendimiento y uso
- Trazabilidad de métricas y rendimiento
- Data drift / concept drift / evolución de sesgos
- Software/env/workflows
- Metadata /Data provenance/Data Lineage
- Web services
- Escalabilidad costo-efectiva

Tendencias

¿Cuándo introducir MLOps?



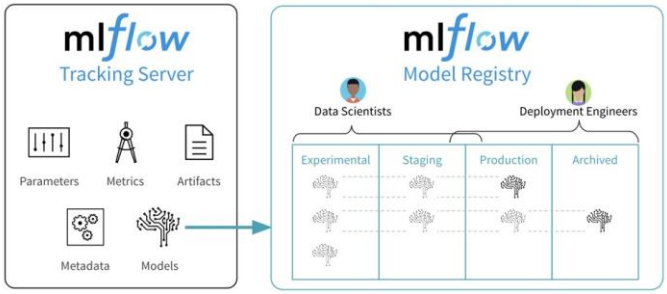
Algunas razones que pueden obligar a implementar MLOps en la industria:

- Servir modelos hacia usuarios
- Crecimiento de células de Analítica
- Mantenimiento y mejoras de múltiples o complejos modelos

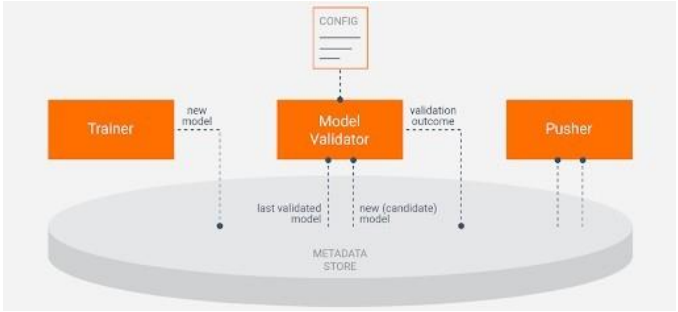
"With the quantity and quality of data available today, it is just poor business for organizations to ignore data in favor of making decisions solely based on what the HiPPO wants done."

Bernard Marr Forbes

MLOps: Algunas Plataformas para implementarlo



TensorFlow Extended



Kubeflow

Azure Machine Learning

Apache Airflow

Clase 2: Metodologías

ENSAYO 1

Profesor: **Roberto González**

Contenido Ensayo

Los ensayos deben ser entregados por email al ayudante (regonzar@uc.cl) a mas tardar **el miércoles 6 Septiembre a las 18:30hrs**, se penalizará la entrega atrasada.

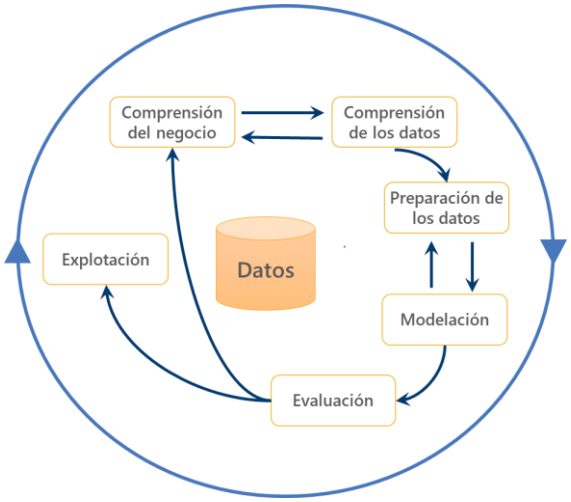
Los ensayos deben ser breves y precisos con un mínimo de 2 planas de texto (2000 o más caracteres) y un máximo de 4 planas de texto (no mas de 4000 caracteres en total). Se penalizaran documentos fuera del rango de caracteres.

Pueden ser grupos de 2 personas, se permiten imágenes o diagramas que no serán considerados en la cuenta de caracteres

Contenido Ensayo

1. Escoja un caso especifico de analítica de datos asociado al trabajo actual de alguno de los miembros del grupo (*o al rubro/sector especifico asociado donde tenga conocimientos del negocio*). Si no esta trabajando, escoja un caso ficticio asociado al rubro en el cual desea trabajar a futuro. Describa brevemente el caso, los datos disponibles y el objetivo especifico que busca resolver(*puede ser ficticio pero consistente*).

2. Ahora suponga que el caso se debe resolver utilizando la metodología **CRISP-DM**, y describa brevemente que se debe realizar en especifico y que desafíos podrían presentarse **para cada una de las etapas** descritas en clases (que hará en cada etapa *específicamente, no repetir definiciones genéricas vistas en clases*)



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

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

regonzar@uc.cl
jcarceo@uc.cl

www.educacionprofesional.ing.uc.cl