



ESCUELA DE INGENIERÍA  
FACULTAD DE INGENIERÍA

EDUCACIÓN  
PROFESIONAL

# Diplomado en Big Data y Ciencia de Datos

## Curso: *Ciencia de Datos y sus Aplicaciones*

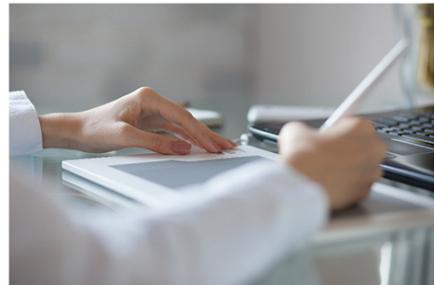
Educación Profesional  
Escuela de Ingeniería UC

 regonzar@uc.cl

 rmunoz@uc.cl

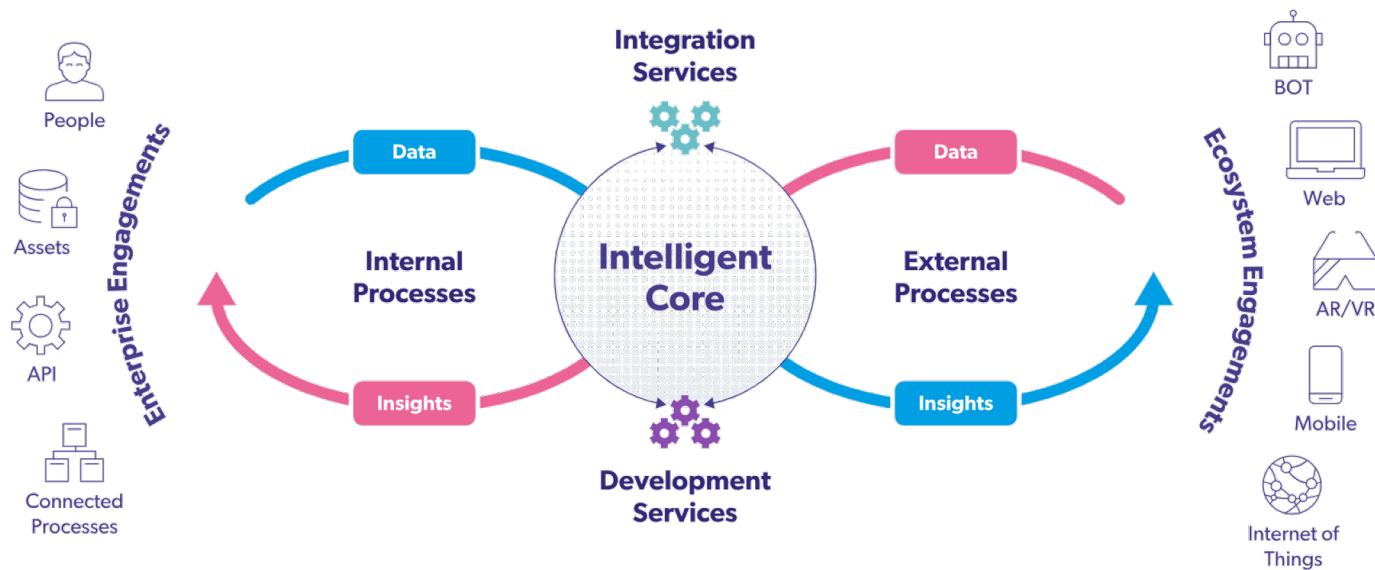
 jcaiceo@uc.cl

Roberto González, Roberto Muñoz, Jaime Caiceo



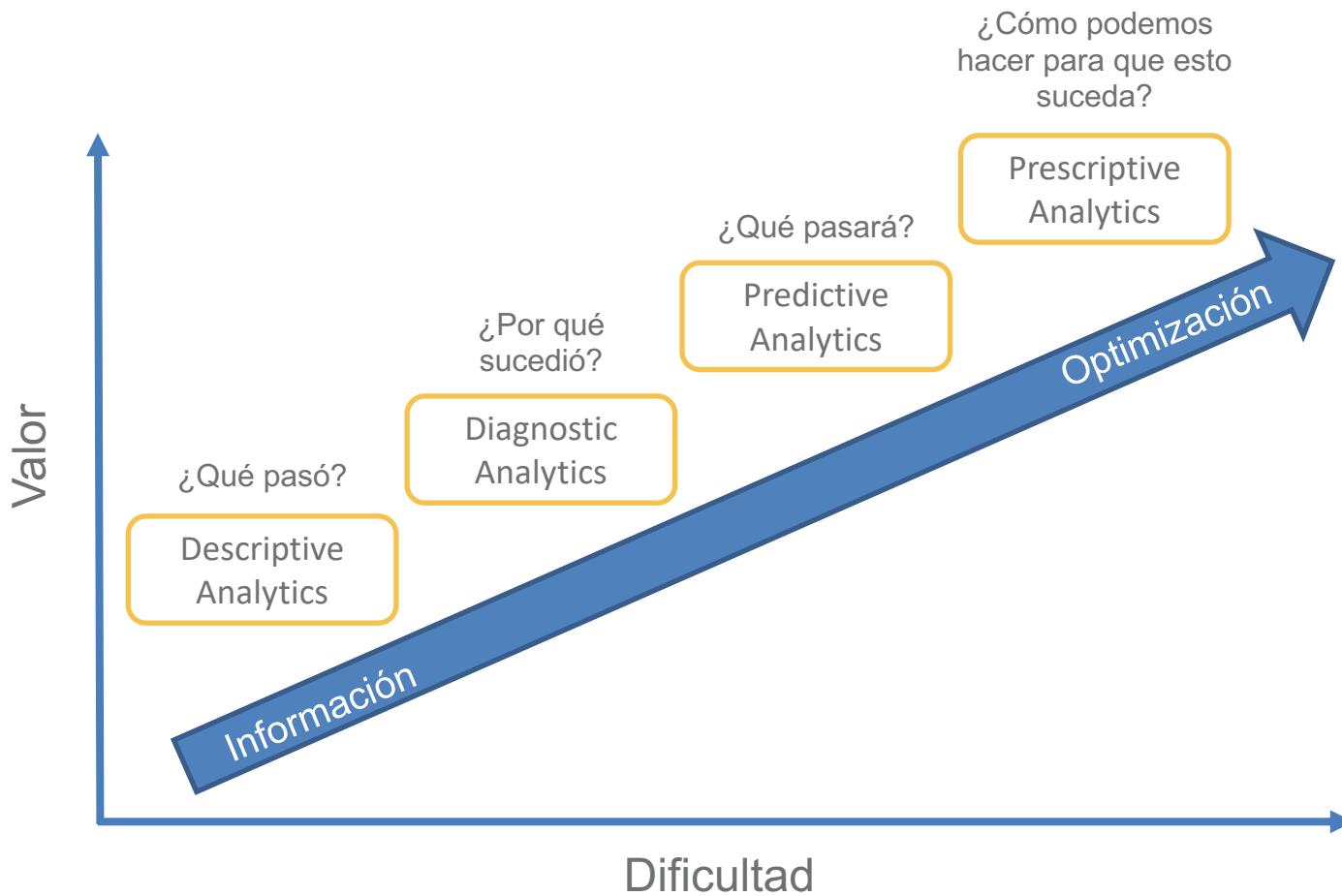
# Transformación basada en inteligencia

- Las organizaciones cuentan con múltiples fuentes de datos
- Interacción con usuarios a través de múltiples canales



Fuente: IDC, Noviembre 2017

# La evolución de los temas analíticos



Fuente: Gartner Business Intelligence & Analytics Summit 2013

<https://youtu.be/UFUwNBCkmYs>



ESCUELA DE INGENIERÍA  
FACULTAD DE INGENIERÍA

EDUCACIÓN  
PROFESIONAL

# Netflix vs Blockbuster

Escriba tres palabras que resuman las razones de  
porqué Netflix perduró en el tiempo y le robó el  
mercado a Blockbuster  
(Use el espacio para separarlas)

<https://pollev.com/robertomunoz211>

**Escriba tres palabras que resuman las razones de porqué Netflix perduró en el tiempo y le robó el mercado a Blockbuster**

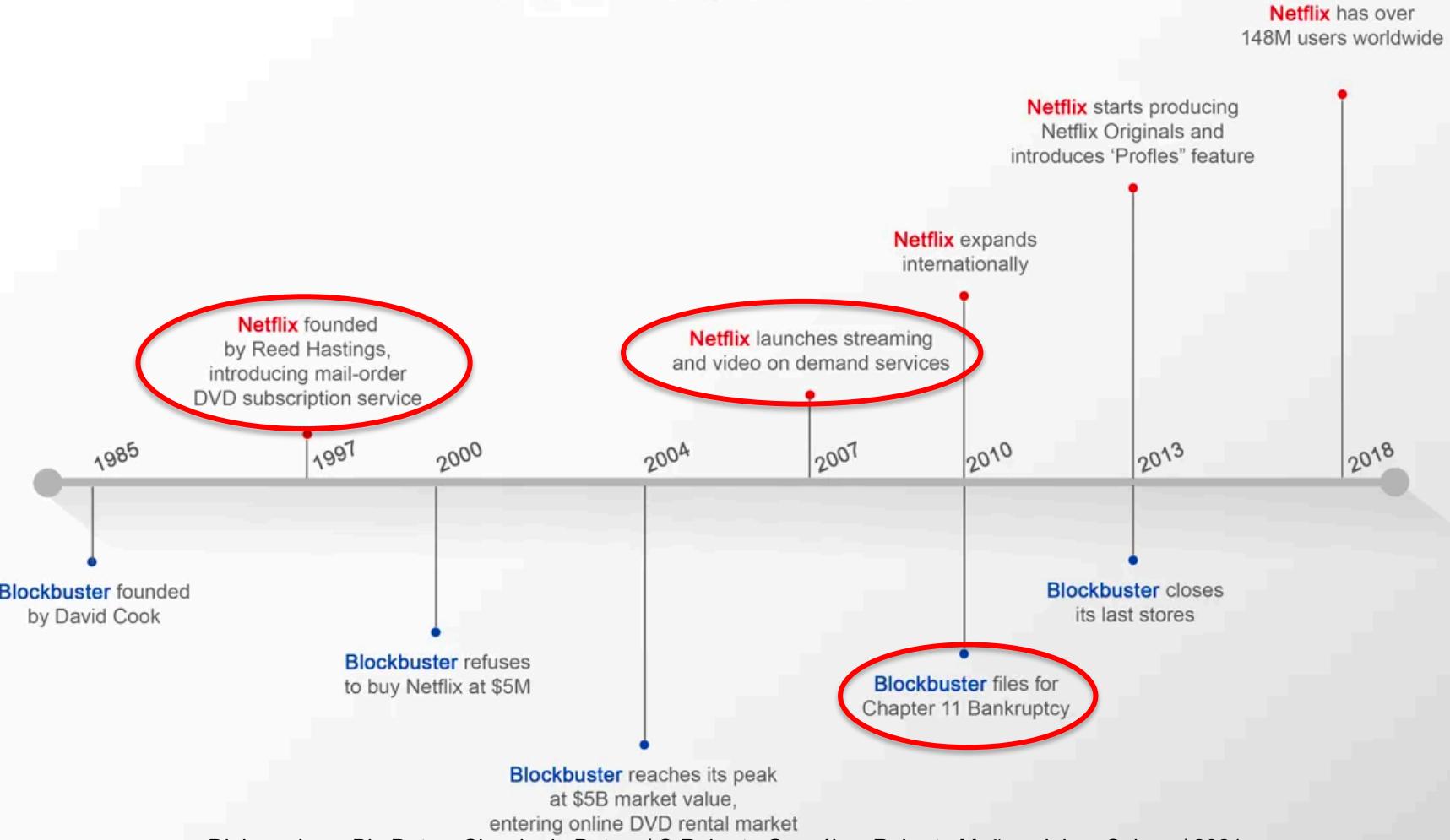
# Blockbuster vs Netflix



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



## THE TIMELINE



# 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?

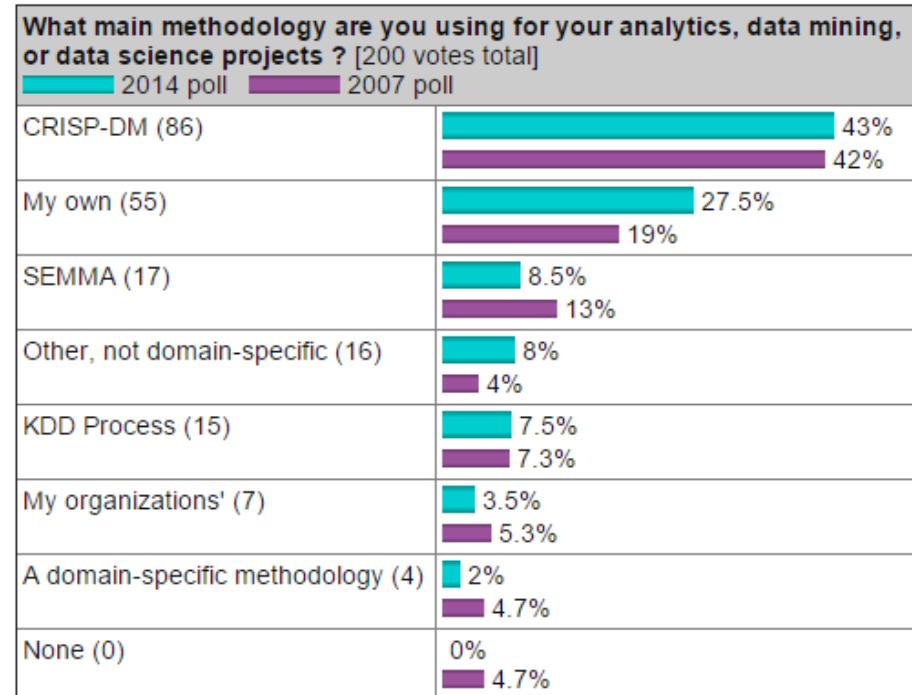
- Las metodologías aseguran que un enfoque consistente se aplicará a todos los proyectos.
- Reducen el riesgo asociado a errores y “atajos”.



# 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>

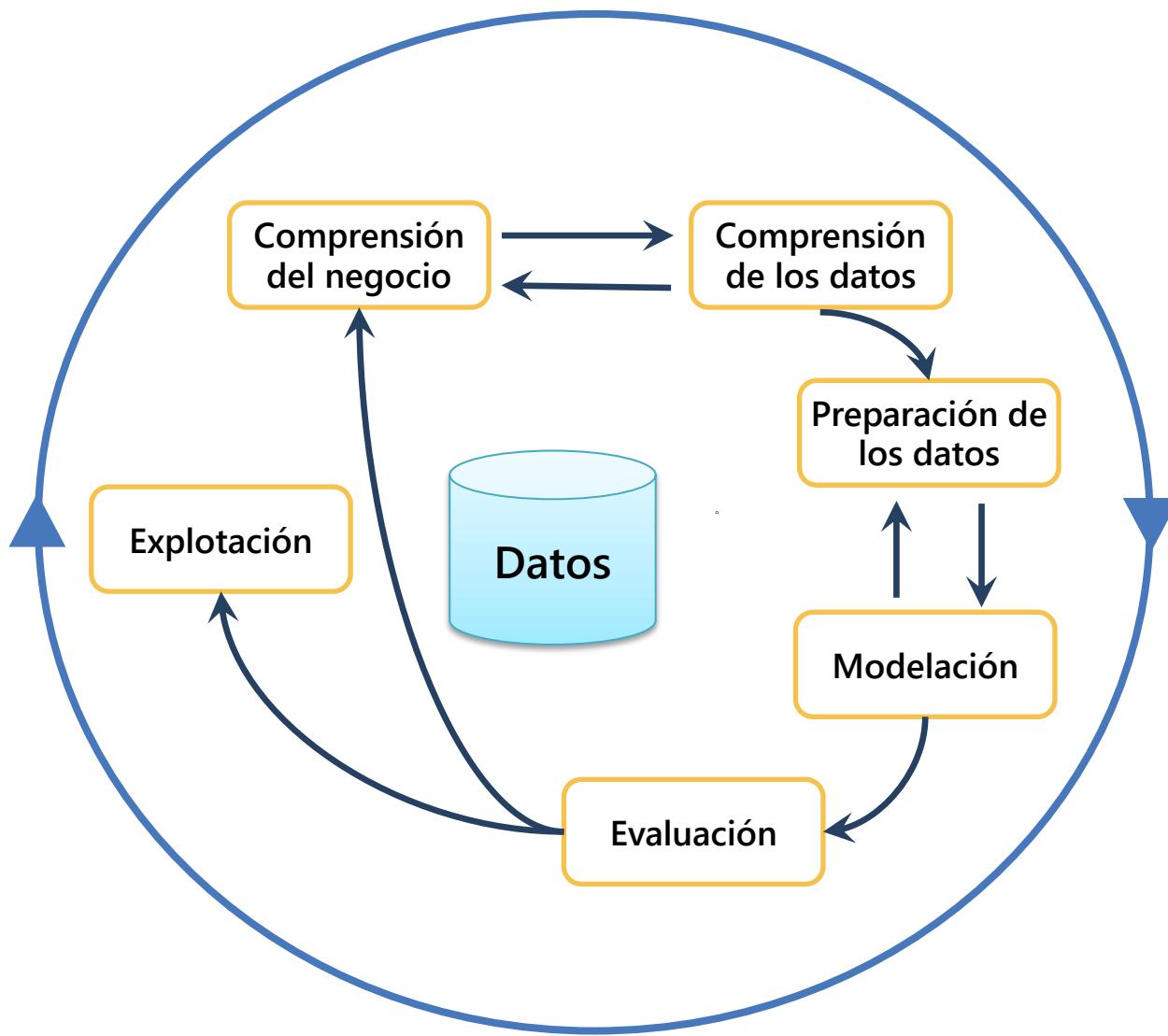
Clase 02: Metodologías de Análisis de Datos

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

# CRISP - DM

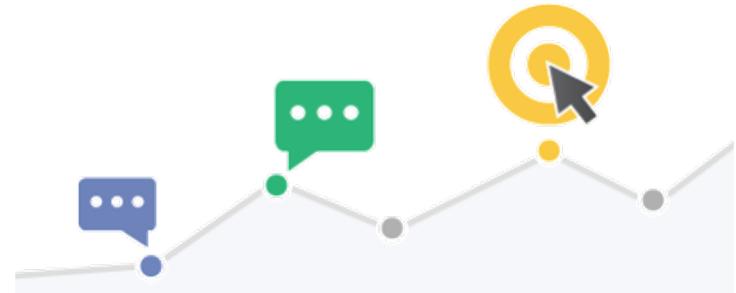
- CRoss-Industry Standard Process for Data Mining.
- Metodología para el proceso de Minería de Datos
  - Valida el proceso, ayuda a planear y administrar proyectos.
- Desarrollado el año 2000 por algunas compañías: SPSS/ISL, NCR, OHRA.
- Está enfocado en el negocio y al análisis técnico.

# Visión General



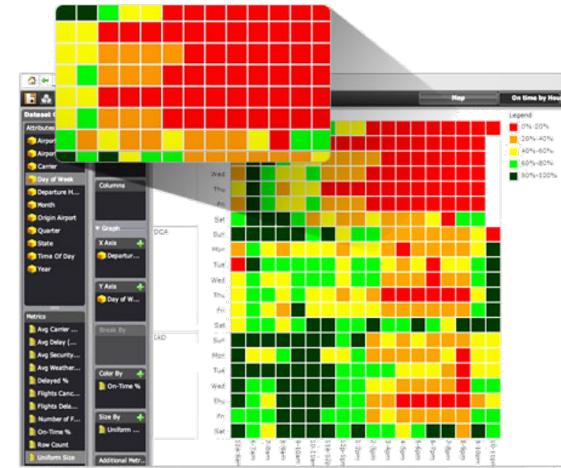
# Fase 1: Comprensión del Negocio

- 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

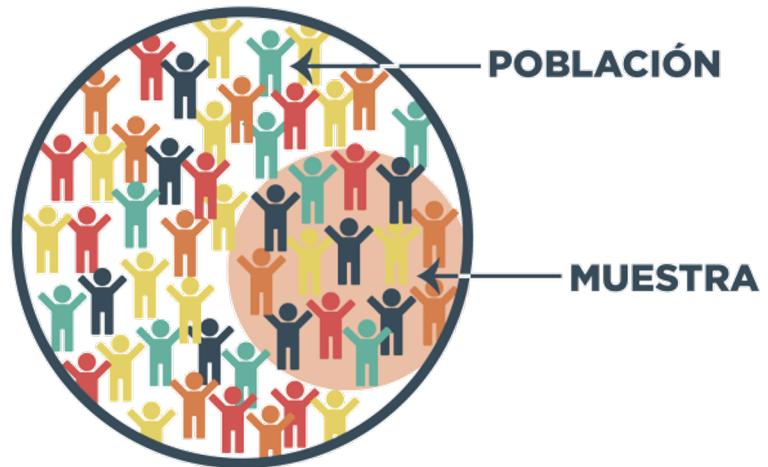
- Objetivo:
  - Simplificar el problema y optimizar la eficiencia del modelo.
- ¿Cómo?
  - Uso de herramientas de visualización y técnicas de estadísticas descriptivas.
- Es relevante también determinar la calidad de los datos.



# 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

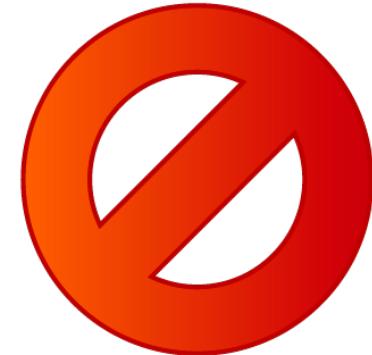
Limpieza: ¿En qué centrarse?

- 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

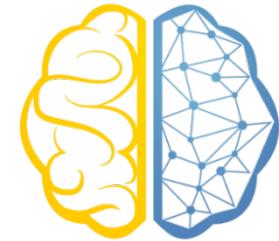
## Transformación de Datos

- 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

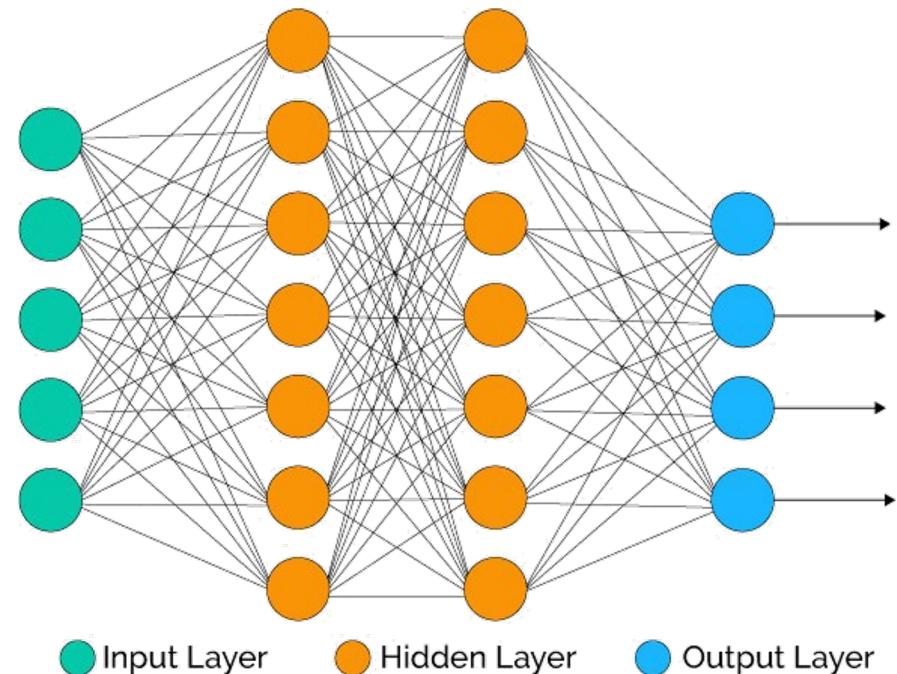
- El objetivo:
  - Satisfacer las metas del planteadas en los primeros pasos, a través de un método particular de Minería de Datos.
  - Por tanto es crucial:
    - Seleccionar el algoritmo correcto a partir del problema que tenemos que abordar y las metas esperadas.



# Fase 4: Modelación

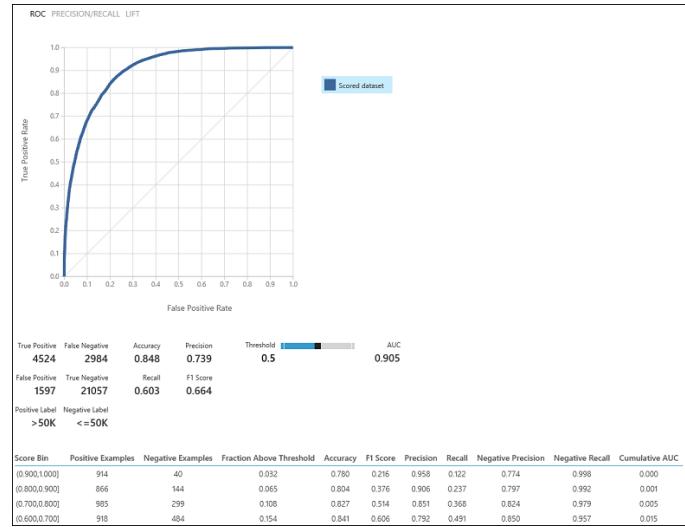
## Técnicas

- 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.
- Contrastá con otros métodos estadísticos o con nuevas muestras.



# Fase 5: Evaluación

- Precisión
  - Porcentaje de casos bien clasificados.
- Eficiencia
  - Tiempo necesario para construir/uso del 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

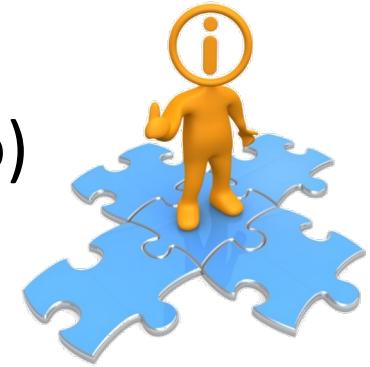
- Técnicas de evaluación generales:
  - Validación simple, validación cruzada
- Clasificación supervisada:
  - Porcentaje de bien clasificados
  - Matriz de confusión



# 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

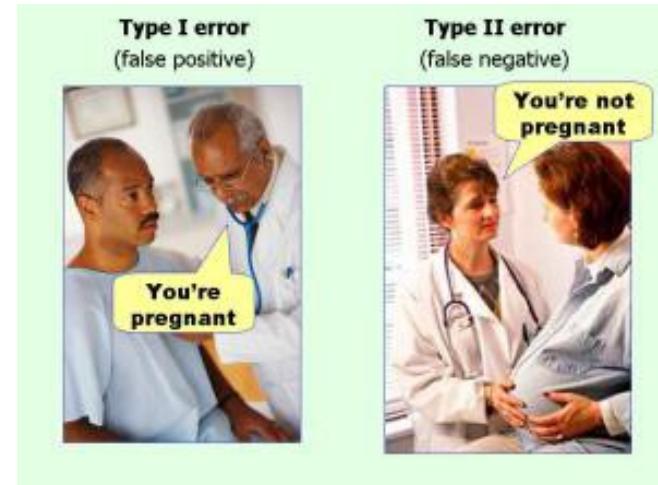
## k-fold Cross-Validation

- 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	
		$C_P$	$C_N$
<b>Valor Real</b>	$C_P$	<b>VP:</b> Verdadero Positivo	<b>FN:</b> Falso Negativo
	$C_N$	<b>FP:</b> Falso Positivo	<b>VN:</b> Verdadero Negativo



# 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



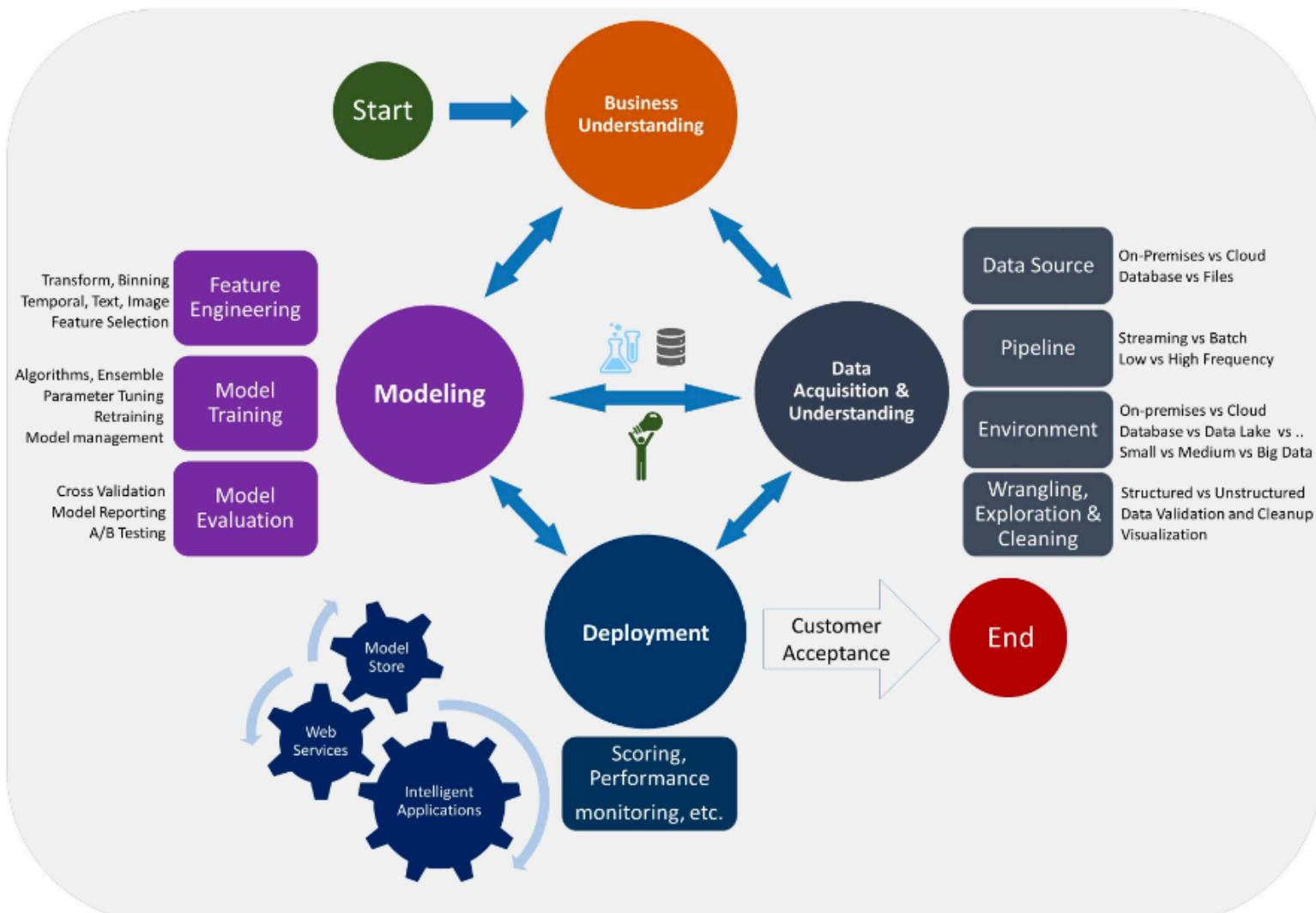
Clase 02: Metodologías de Análisis de Datos

# **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 satisface necesidades del negocio

Clase 08: Cierre

# **ANALYTICS Y DATA SCIENCE EN CHILE**

# Índice Madurez Digital

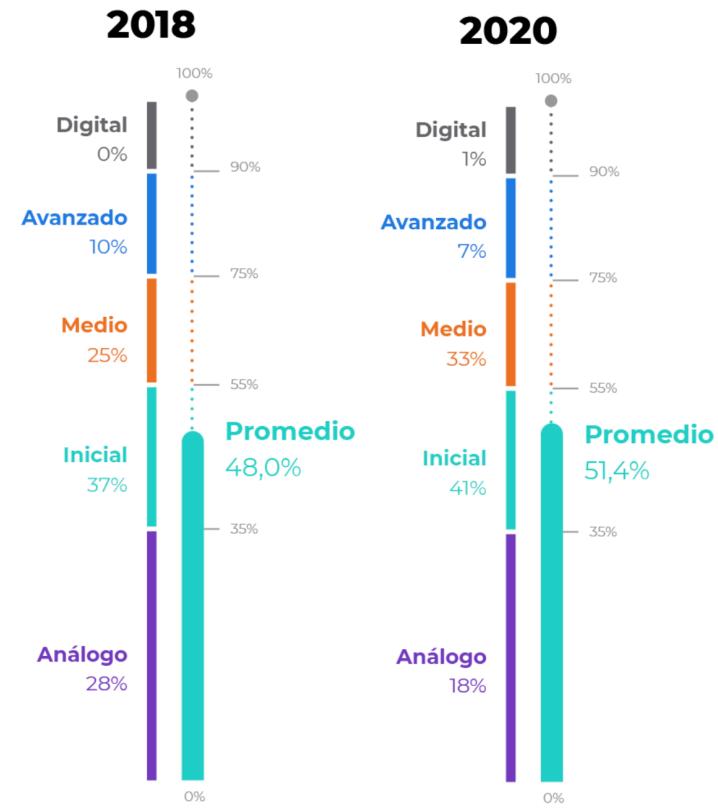
<https://www.somosvirtus.com/imdv-chile-2020>



## Incial

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.



# Potencial de los datos

## ¿Estamos aprovechando todo el potencial de la data?

Con más y mejores datos, los insights generados pueden “desatar” 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%**

Grandes empresas

**73%**

Pymes

**73%**

Sartups

**98%**

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

**69%**

Instituciones públicas

**86%**

Centros de conocimiento

**60%**

Gremios

**50%**

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

**59%**

Grandes empresas

**36%**

Pymes

**39%**

Sartups

**23%**

Instituciones públicas

**54%**

Centros de conocimiento

**20%**

Gremios

**33%**

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

**37%**

Grandes empresas

**28%**

Pymes

**42%**

Sartups

**23%**

Instituciones públicas

**32%**

Centros de conocimiento

**7%**

Gremios

**32%**

señala que su organización realiza Data Analytics avanzado

**37%**

Grandes empresas

**21%**

Pymes

**42%**

Sartups

**31%**

Instituciones públicas

**38%**

Centros de conocimiento

**7%**

Gremios



# “2021 el año de la IA”

18  
2021 EL AÑO DE LA INTELIGENCIA ARTIFICIAL

# LOS PASOS QUE VIENEN PARA LA IA EN CHILE

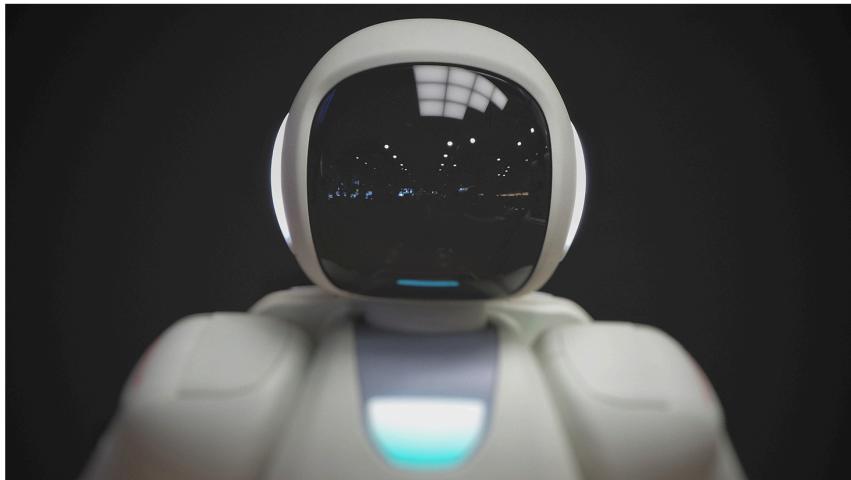
Aunque hay consenso sobre los beneficios económicos y sociales que la Inteligencia Artificial traerá para el país, una reciente consulta pública mostró que prevalecen dudas éticas y regulatorias sobre ella. Mientras, la Política Nacional que fijará sus lineamientos estratégicos ya está en su última etapa.

POR AIRAM FERNÁNDEZ

**71%**  
DE LOS EJECUTIVOS CHILENOS ADOPTARÁ IA EN LOS PRÓXIMOS MESES, SEGÚN ACCENTURE.

# 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**



# Zippedi: Robótica e Inteligencia Artificial



# The Not Company: Alimentos basados en plantas





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.