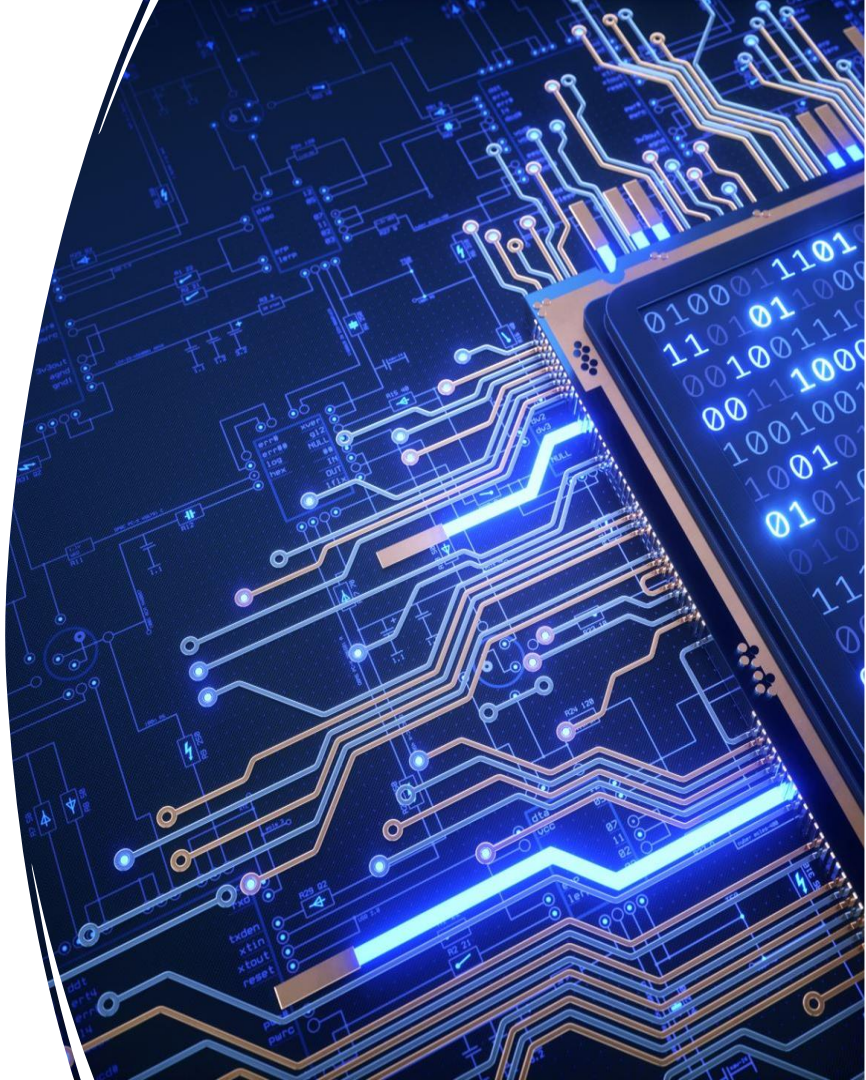
Clase 01: Introducción a la Ciencia de Datos y al Aprendizaje Automático

Walter Gómez
IMA539
Departamento de Ingeniería Matemática y
Minor de Análisis de Datos

Universidad de La Frontera



Contextualización

- Enfrentamos una nueva revolución, un nuevo paradigma, una nueva era… **La era del big data!**

Contextualización: ¿Por qué generamos tantos datos?

Desde los inicios de la civilización queremos conocer el porqué de los fenómenos que ocurren, con el fin de lograr la supervivencia.
Explicar los fenómenos, buscar la verdad sobre Fenómenos extraños que generan incertidumbre.

- **DESARROLLO CIENTÍFICO**
- **FENÓMENOS COMPLEJOS**

Introducción

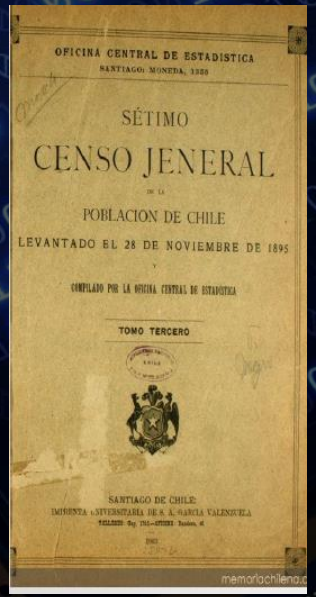
Los modelos matemáticos permiten s**imular el comportamiento de las interacciones y relaciones entre diferentes sistemas y predecir su comportamiento futuro**.

El desarrollo científico nos ha permitido conocer y comprender que vivimos en un mundo complejo, **donde múltiples variables influyen en el comportamiento de los fenómenos**.

Dato

Viene del latín “dtum”, que significa “lo que se da”.

Representación simbólica de información de carácter cuantitativa o cualitativa que es generada por una acción.



Evolución en el registro de la información

	A	B	C	D	E	F	G	H	I	J	K	L
		fecha	zona	evento	tipoevento	unixtime	s_p	duracion	amplitudctas	frecuencia	energia	est
0		2012-11-30 1		2	11301138.9LF	LP	1354275519.4	0.0	23.0	229.0	6.13	35084922264	176
1		2012-12-02 1(		2	12021004.9LF	LP	1354442709.3	0.0	25.0	2684.0	5.05	53463319918	231
2		2012-12-05 1		2	12051440.9V	VT	1354718431.(	0.46	10.0	1211.0	6.18	21663588617	231
3		2012-12-07 0		2	12070920.9V	VT	1354872034.(	1.42	17.0	573.0	7.38	35740891594	176
4		2012-12-07 1		2	12071123.9LF	LP	1354879420.4	0.0	14.0	195.0	5.47	28806465356	176
5		2012-12-07 1		2	12071123.9V	VT	1354879446.5	0.89	26.0	1696.0	8.08	32611598578	176
6		2012-12-07 2		2	12072234.9V	VT	1354919688.7	0.64	15.0	697.0	7.17	18527934831	176
7		2012-12-08 0		2	12080214.9V	VT	1354932903.2	0.83	15.0	500.0	4.77	18527934831	176
8		2012-12-08 0		2	12080855.9V	VT	1354956956.2	1.53	24.0	3020.0	5.2	21551361193	176
9		2012-12-08 0		2	12080951.9LF	LP	1354960316.1	0.0	17.0	246.0	1.11	76384188620	176
10		2012-12-08 1		2	12081727.9LF	LP	1354987642.3	0.0	32.0	218.0	2.11	18609239711	176
11		2012-12-08 1		2	12081734.9LF	LP	1354988080.(	0.0	24.0	492.0	4.37	43499746588	176
12		2012-12-08 1		2	12081812.9V	VT	1354990332.5	1.15	12.0	373.0	30.21	57058686290	176
13		2012-12-09 0		2	12090557.9V	VT	1355032637.5	0.71	5.0	2633.0	5.88	518666566.0	176
14		2012-12-10 0		2	12100730.9V	VT	1355124635.6	1.14	26.0	44814.0	10.52	32611598578	176
15		2012-12-11 0		2	12110521.9LF	LP	1355203293.1	0.0	17.0	584.0	2.52	76384188620	176
16		2012-12-11 1		2	12111107.9V	VT	1355224037.3	0.0	2.0	10170.0	31.24	3304949.0	176
17		2012-12-11 2		2	12112309.9V	VT	1355267355.5	1.84	10.0	1074.0	15.32	21663588617	176
18		2012-12-12 2		2	12122338.9LF	LP	1355355548.7	0.0	19.0	386.0	2.19	13376764949	176
19		2012-12-13 0		2	12130119.9LF	LP	1355361603.(	0.0	22.0	353.0	2.37	28030242349	176
20		2012-12-13 0		2	12130207.9LF	LP	1355364452.(	0.0	16.0	686.0	4.46	56307155159	176
21		2012-12-13 0		2	12130420.9LF	LP	1355372436.1	0.0	20.0	2112.0	5.52	17326174535	176
22		2012-12-13 0		2	12130427.9LF	LP	1355372874.2	0.0	29.0	225.0	1.63	11317559742	176
23		2012-12-16 2(		2	12162048.9V	VT	1355690894.5	0.6	14.0	2292.0	9.85	12884262993	176
24		2012-12-16 2		2	12162309.9V	VT	1355699370.3	0.0	34.0	2933.0	8.67	12969896042	176
25		2012-12-17 0		2	12170346.9V	VT	1355715974.2	0.51	19.0	655.0	7.44	6394159964(	176

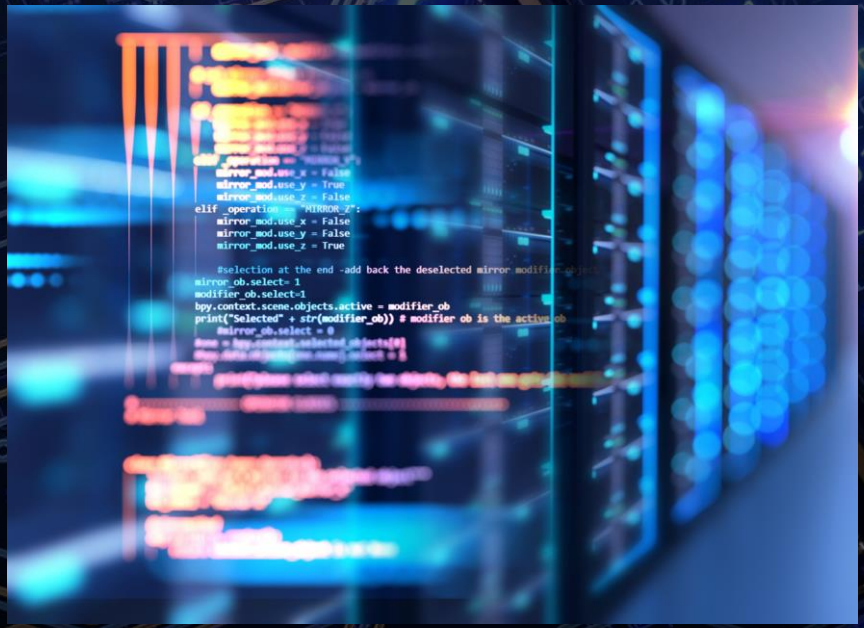
Años 80’ se masifica la digitalización

Años 2000 Registros digitalizados

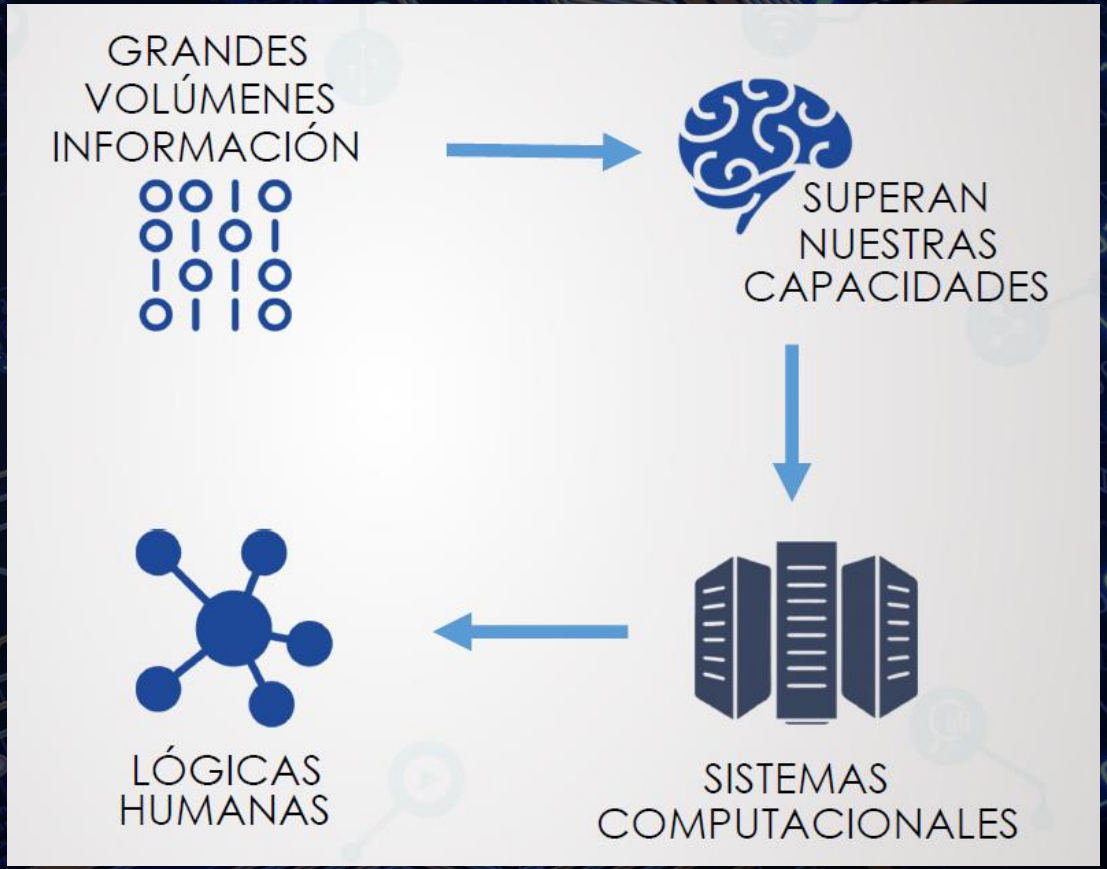
Actualidad: Big Data

Es un fenómeno científico – tecnológico.

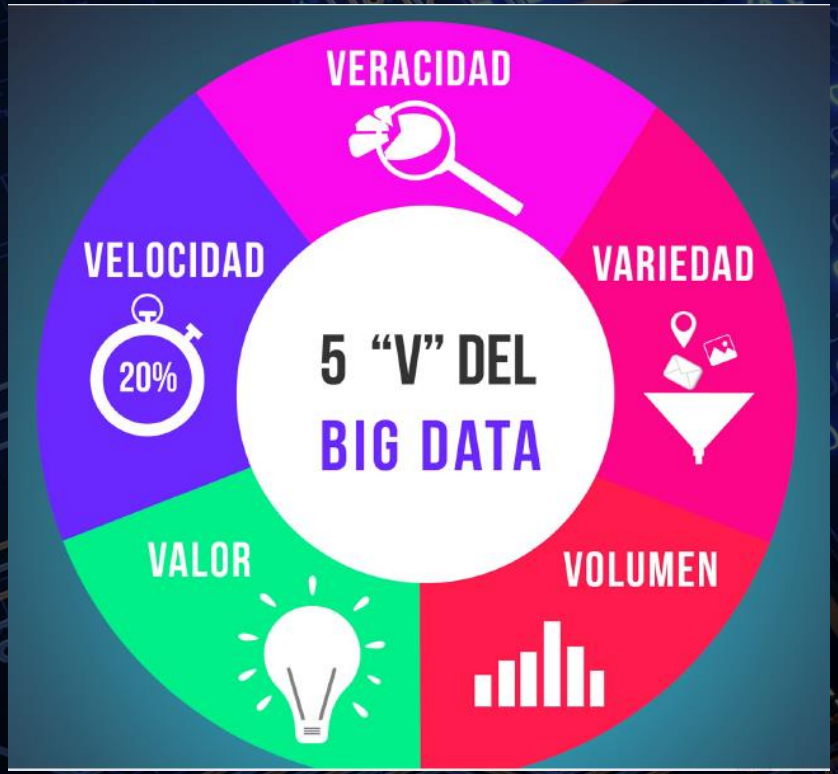
Nos permite transformar la complejidad en simplicidad con ayuda de las tecnologías de la información.



Big Data


GRANDES VOLÚMENES INFORMACIÓN
SUPERAN NUESTRAS CAPACIDADES
SISTEMAS COMPUTACIONALES
LÓGICAS HUMANAS

Las 5 V del Big Data



Big Data

VOLUMEN: La primera característica del big data es que implica grandes volúmenes de datos e información. Actualmente, gracias a los dispositivos tecnológicos, es posible capturar, procesar y analizar miles de datos minuto a minuto, por lo que el primer desafío es desarrollar capacidades técnicas y tecnológicas para el procesamiento y análisis de datos masivos.

VELOCIDAD: El segundo desafío del big data tiene relación con la velocidad de generación y procesamiento. Al trabajar con datos masivos, se hace necesario contar con capacidades robustas que hagan frente a la volatilidad de los datos, ya que muchos de ellos tienen una corta vida útil y es necesario capturarlos y analizarlos en el momento oportuno para que no pierdan valor.

VARIEDAD: Los datos que se recopilan pueden provenir de diferentes fuentes y además podemos encontrarlos en diferentes formatos: datos estructurados y no estructurados. Dado el origen diverso de los datos, la configuración de procesos de análisis considerando la variedad de éstos, es la tercera característica del big data.

Big Data: Variedad

Tipos de datos

Estructurados

Datos que tienen un modelo definido o provienen de un campo determinado en un registro.

- Fichas de clientes
- Transacciones comerciales

Semiestructurados

Datos que no tienen formatos fijos, pero contienen atributos o etiquetas.

- Correos electrónicos
- Fichas con imágenes médicas

No estructurados

Datos que no tienen un modelo predefinido o no están organizados de alguna manera.

- Videos
- Fotografías
- Audios

Big Data

VERACIDAD: El siguiente desafío es resguardar la calidad de los datos. Al trabajar con grandes volúmenes de datos, se pueden presentar problemas como registros incompletos o erróneos, datos faltantes en determinados campos o información que proviene de diferentes fuentes y son discrepantes. La veracidad de los datos es el cuarto desafío en big data.

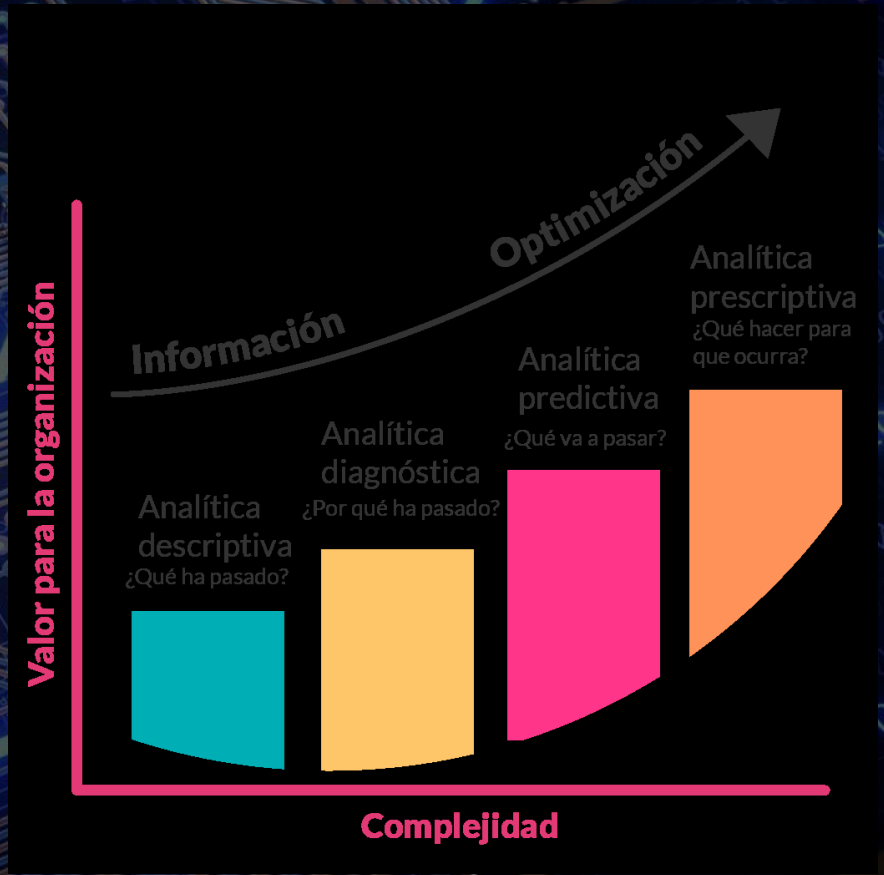
Big Data: Valor

¿Dónde está el valor del big data?

Capacidad analítica

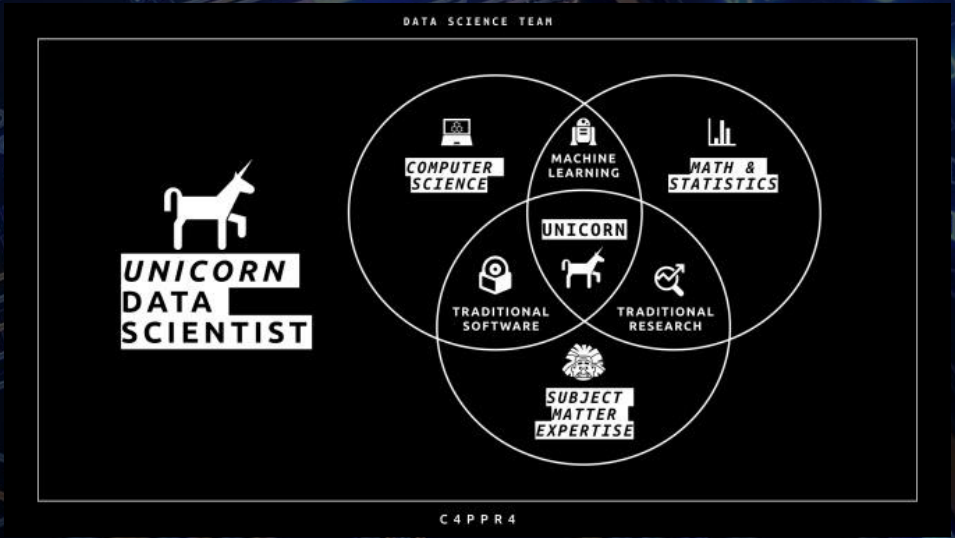
Big Data

La complejidad analítica presenta una relación directa entre el Valor y la dificultad de los procesos de análisis.



¿Qué es un Científico de Datos?

- Un Científico de Datos es un profesional con habilidades en Matemáticas y Estadísticas, de Programación Computacional y uso de Softwares. Además cuenta con habilidades para Presentación y Visualización de resultados (Análisis).
- Sin embargo, independiente del nivel de dominio de esas áreas, es fundamental el conocimiento del negocio o de los fenómenos estudiados (capacidad para entenderlo), para que los datos tengan un sentido y una utilidad para quienes los consumen.



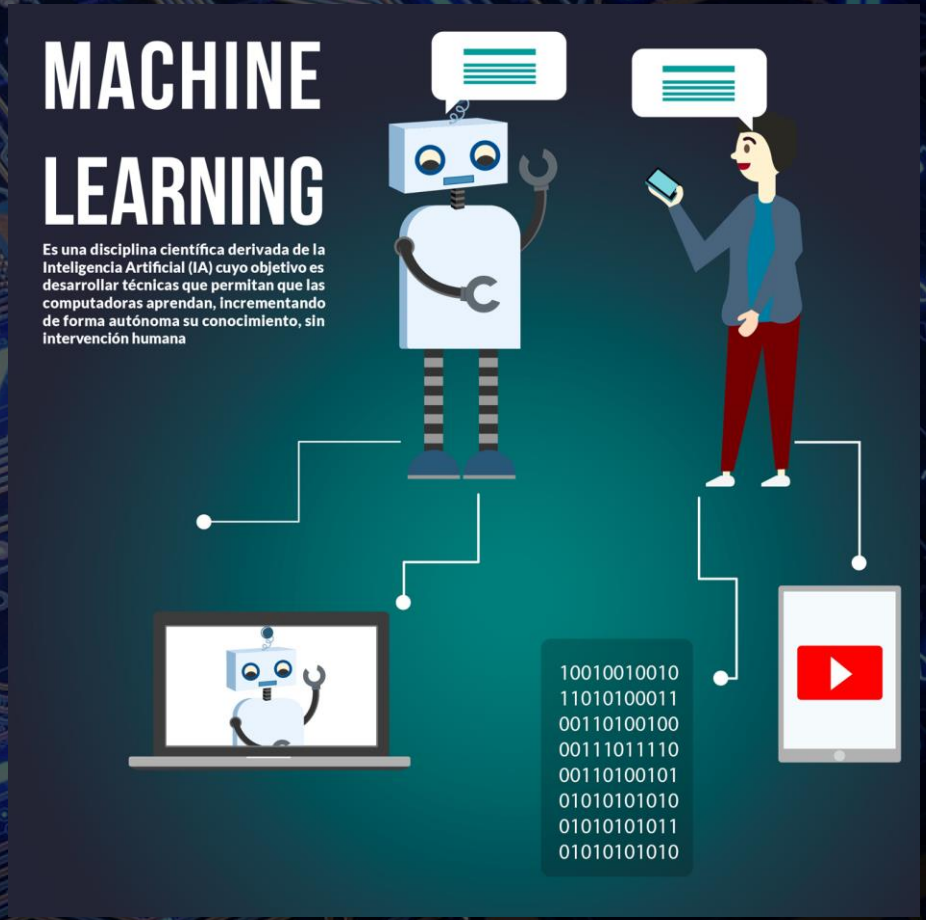
Pilares de un Científico de Datos

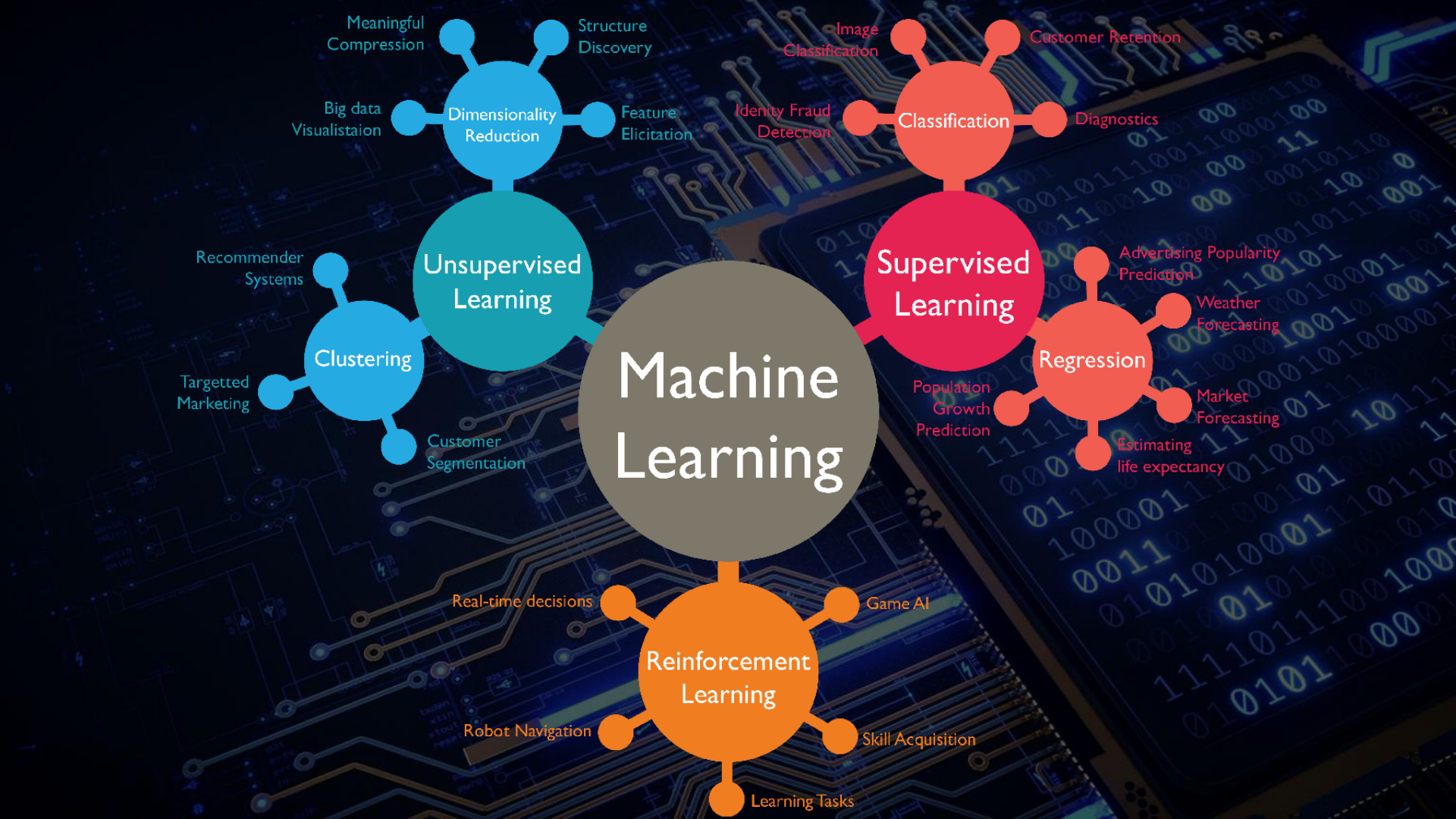
- **Modelamiento Matemático y Estadístico**: Un científico de datos debe contar con conocimientos matemáticos que le permitan diseñar e implementar procesos analíticos en diferentes niveles de complejidad, que le permitan reconocer patrones de comportamiento en los datos, estimar probabilidades de ocurrencia de diferentes fenómenos, realizar pronósticos, simular escenarios futuros, etc.

- **Programación Computacional**:Un data scientist debe manejar lenguaje de programación que lo dote de capacidades para automatizar procesos de análisis. Además, puede ser capaz de diseñar interfaces de aprendizaje automático (Machine Learning) entre humanos y computadoras, estableciendo las bases para soluciones basadas en Inteligencia Artificial.

Machine Learning

El Aprendizaje Automático, más conocido como Machine Learning, es una subdisciplina de la Ciencia de Datos cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan de forma autónoma, sin intervención humana. Sus usos y aplicaciones son considerados un enfoque de Inteligencia Artificial.

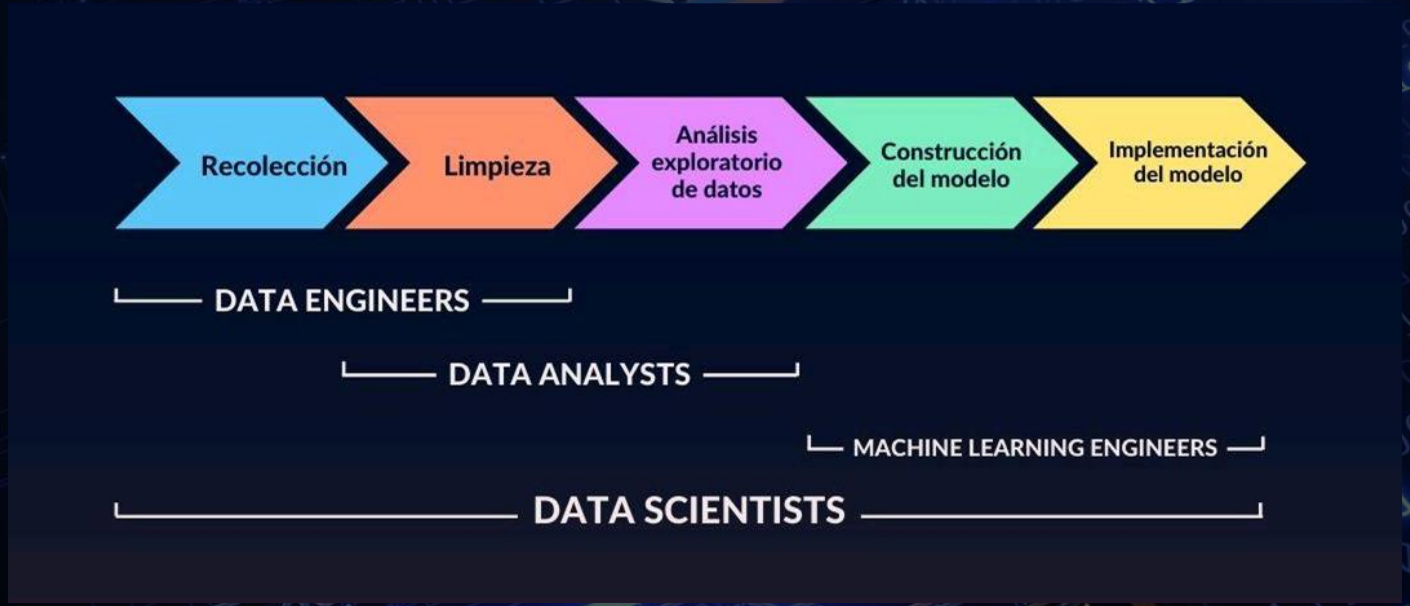



Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Identity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Pilares de un Científico de Datos

Visualización de Datos: Un científico de datos es capaz de contar historias con los datos. Utiliza técnicas narrativas para comunicar los resultados de los procesos de análisis, utilizando criterios para seleccionar las técnicas de visualización y gráficos adecuados para quienes consuman la información interpreten adecuadamente las conclusiones obtenidas.

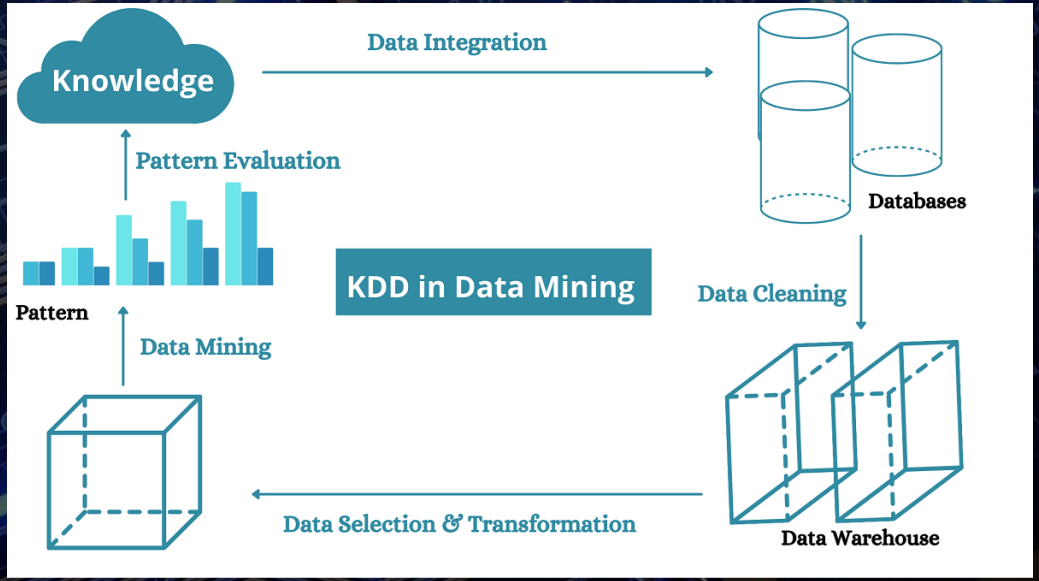
Algunas etapas en la Ciencia de Datos



Metodologías para Proyectos de Ciencias de Datos

KNOWLEDGE DISCOVERY DATABASE (KDD): La minería de datos, también conocida como descubrimiento de conocimientos en bases de datos, se refiere a la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos almacenados en bases de datos. Consta de 6 etapas:

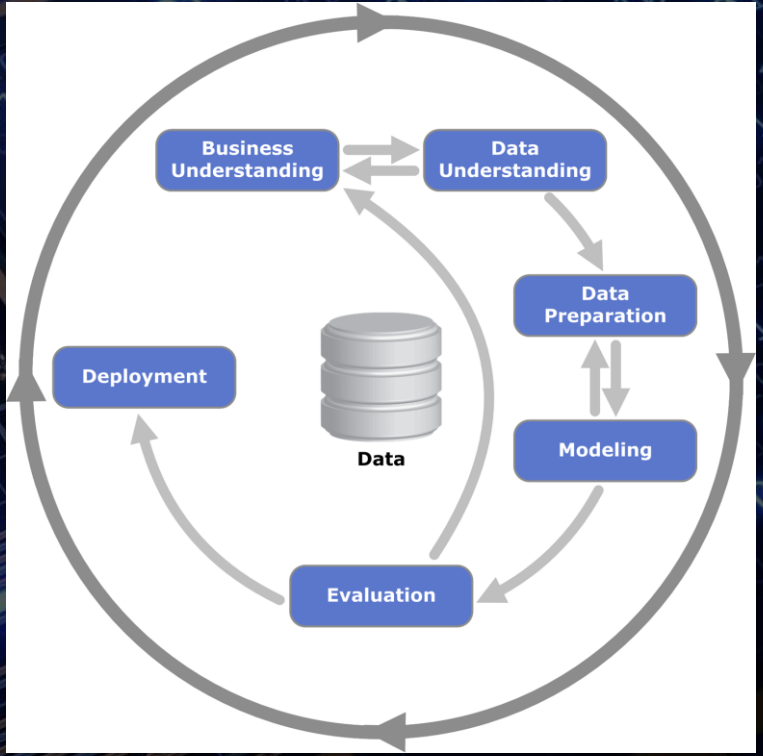
1. Data Integration
2. Data cleaning
3. Data transformation
4. Data mining
5. Pattern Evaluation
6. Knowledge representation



Metodologías para Proyectos de Ciencias de Datos

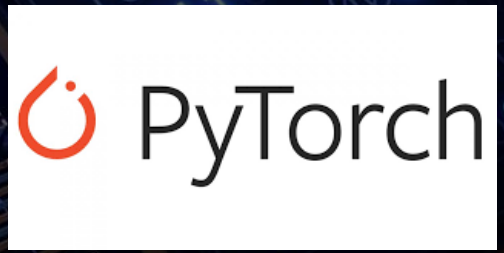
CRISP - DM: Es una metodología desarrollada por la empresa IBM, cuyas siglas significan Cross Industry Standard Process for Data Mining. Proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de ciencia de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software.
Consta de 6 fases:

1. **Business Understanding**. Definición de necesidades del cliente (comprensión del negocio).
2. **Data Understanding**. Estudio y comprensión de los datos.
3. **Data Preparation.** Análisis descriptivo de los datos y selección de las características.
4. **Modeling**. Se seleccionan y aplican técnicas de modelamiento que sean pertinentes al problema.
5. **Evaluation.** En esta etapa del proyecto se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la perspectiva del análisis de datos.
6. **Deployment** (Puesta en producción). Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.



Softwares para Ciencia de Datos

Principales librerías de Python para Ciencia de Datos



Clase 1: Introducción a la Ciencia de Datos y al Aprendizaje Automático

Walter Gómez
IMA539
Departamento de Ingeniería Matemática
y Minor de Análisis de Datos

Universidad de La Frontera