Técnicas Avanzadas en Análisis Estadístico

Este curso se enfoca en la aplicación de técnicas estadísticas avanzadas para transformar la información en conocimiento valioso.

Adriana Arango L
adriana.arangol@upb.edu.co

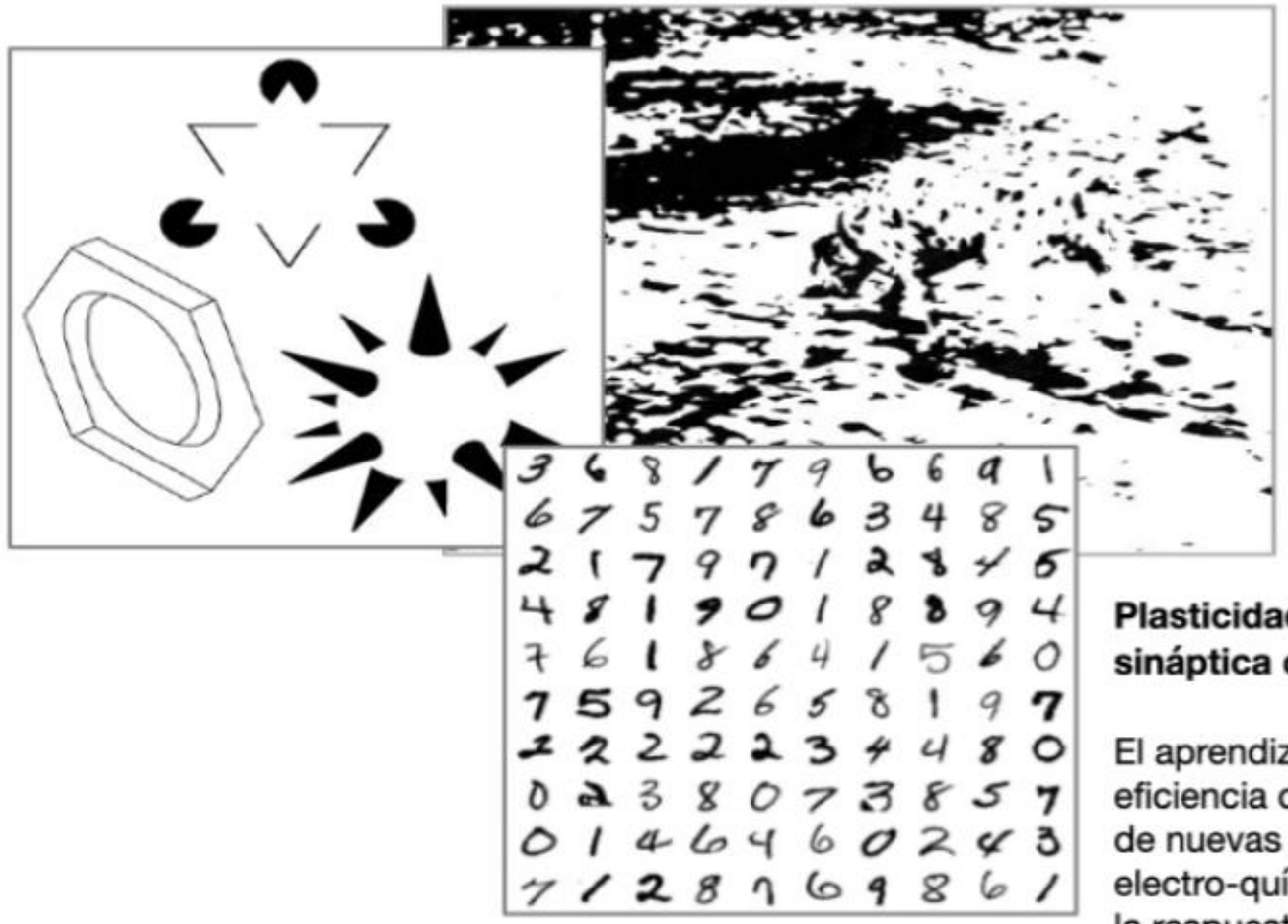
Historia de la Ciencia de Datos

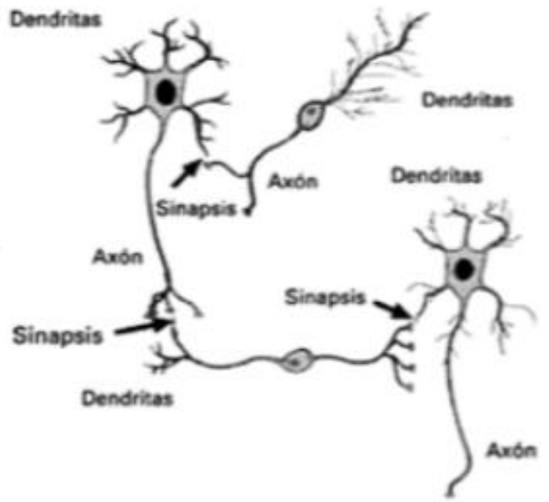
Desde las primeras aplicaciones de la estadística para analizar fenómenos sociales hasta el auge de la informática y el desarrollo de algoritmos de aprendizaje automático, la ciencia de datos ha experimentado una evolución creciente.

Esta disciplina ha sido impulsada por la creciente disponibilidad de datos, el desarrollo de tecnologías de análisis avanzadas y la necesidad de tomar decisiones estratégicas basadas en información sólida.

1894 Surgimiento de la neurociencia

- Interes por entender como “computa el cerebro” y como aprende.
- Teoría de plasticidad neuronal

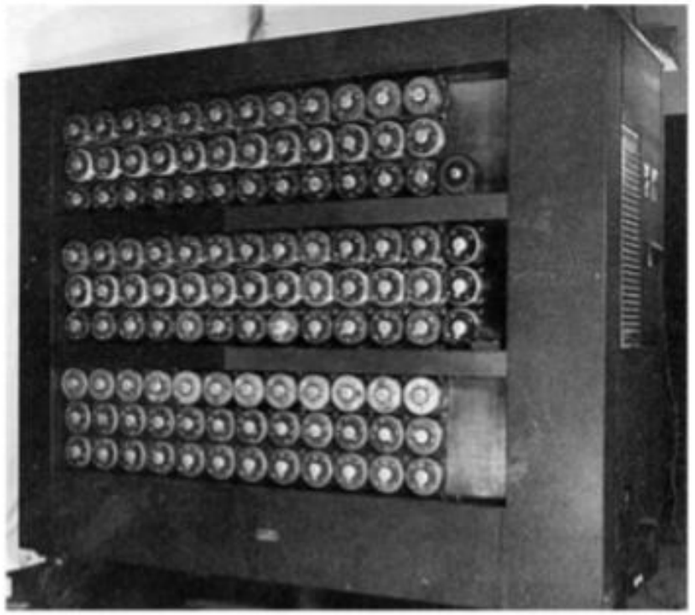




Plasticidad neuronal, neuroplasticidad o plasticidad sináptica o teoría del cerebro plástico

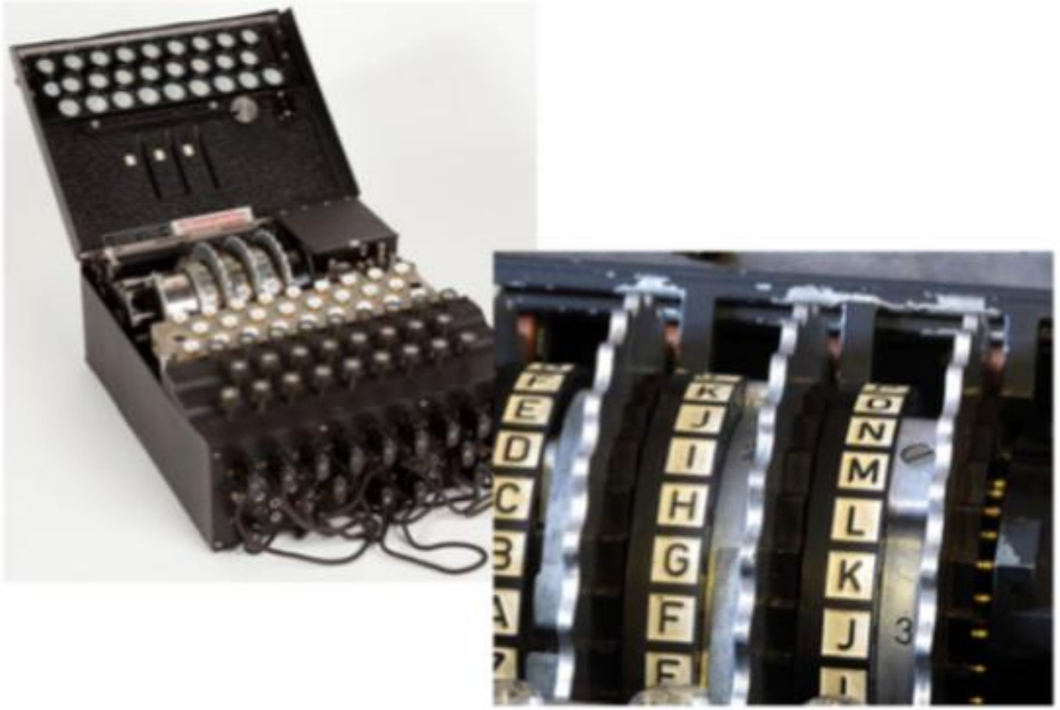
El aprendizaje se da como la modificación de la eficiencia de las conexiones inter-neuronales (y creación de nuevas conexiones) para la transferencia de impulsos electro-químicos modulando tanto la percepción como la respuesta dada ante estímulos del medio.

1940 Bombe Machine

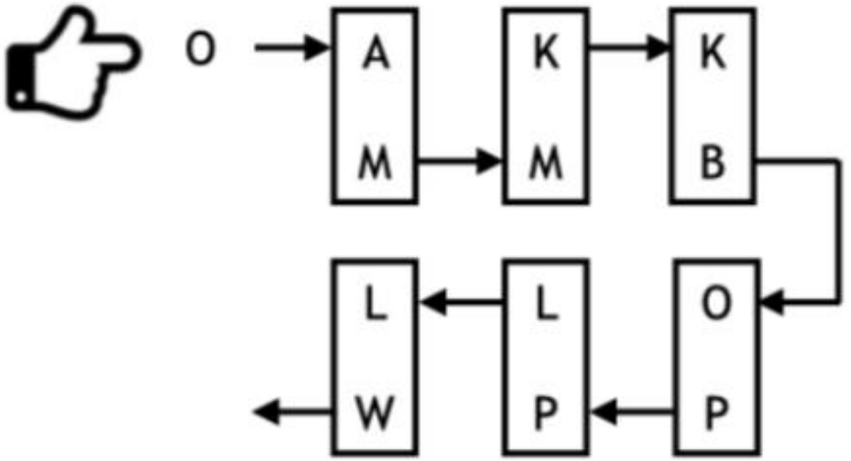


Desarrollada por Alan Turing para decodificar la máquina Enigma.

Máquina Enigma



La máquina enigma usaba tres rotores, de cinco posibles



Por cada pulsación se giraban, al menos, un rotor.

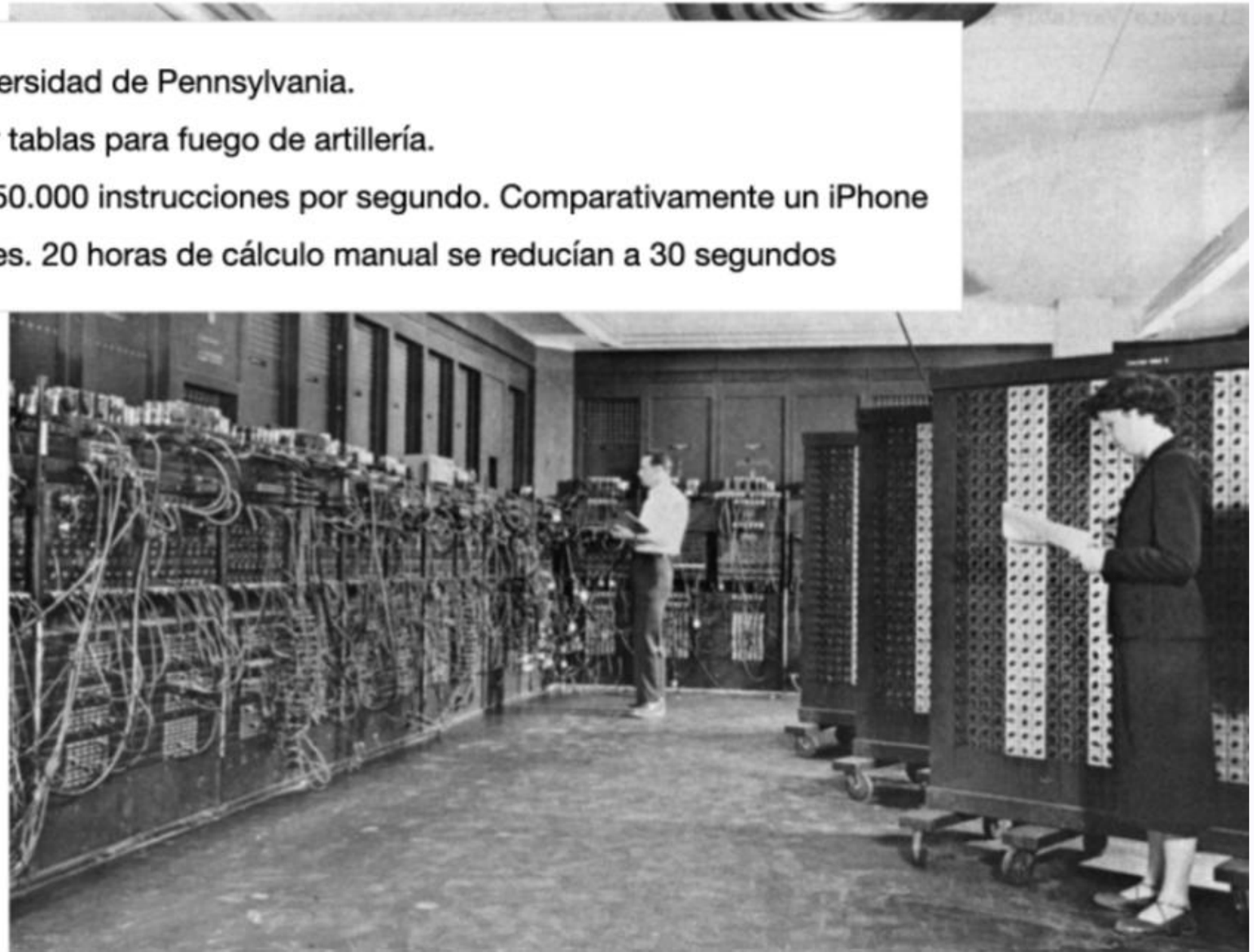
1944 Proyecto Manhattan

Desarrollo de la simulación numérica (método de Monte Carlo) para comprender y pronosticar el comportamiento de reacciones nucleares en cadena usando computadores análogos.

Se basa en la realización de miles de simulaciones de un proceso aleatorio para analizar posteriormente sus resultados.

1946 ENIAC (Electronic Numerical Integrator and Computer)

- Desarrollado en la Universidad de Pennsylvania.
- Su objetivo fue calcular tablas para fuego de artillería.
- Fue capaz de manejar 50.000 instrucciones por segundo. Comparativamente un iPhone puede manejar 5 billones. 20 horas de cálculo manual se reducían a 30 segundos



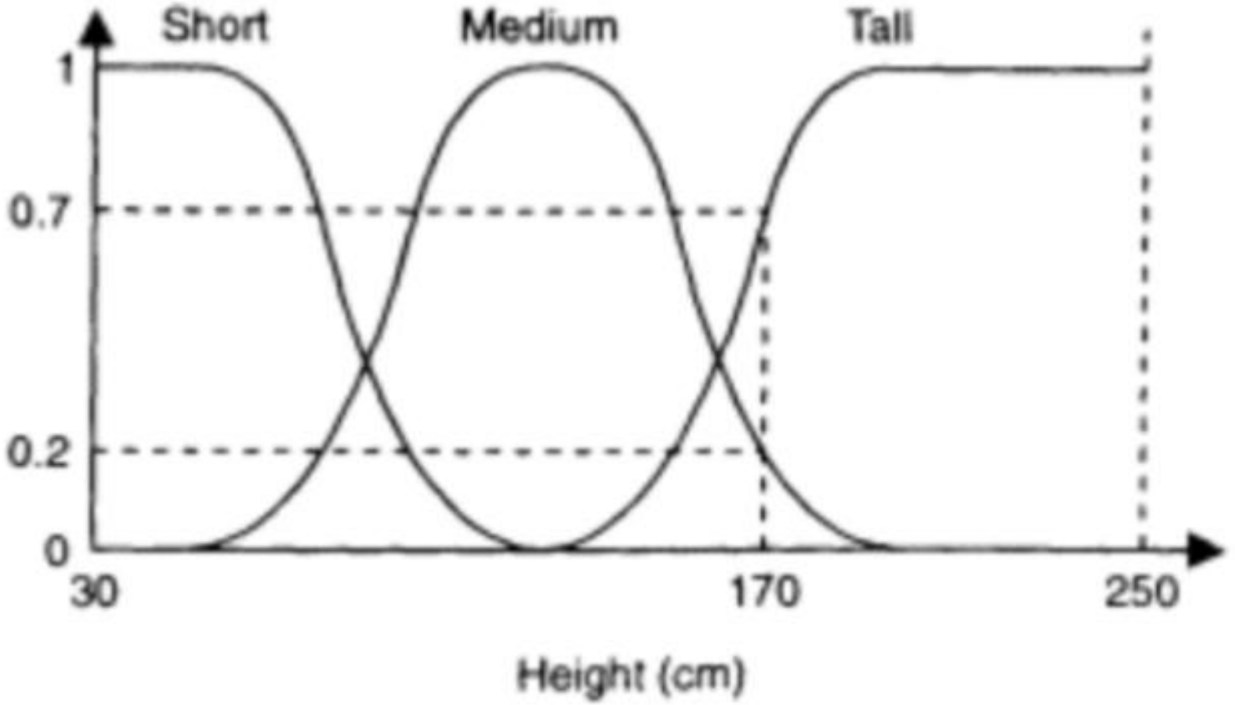
1965 Concepto de Fuzzy Set

Definición tradicional. Función característica:

- $\mu_A(u) = 1$, El elemento u pertence al conjunto A
- $\mu_A(u) = 0$, El elemento u no pertence al conjunto A

Conjunto difuso. Se admite la pertenencia parcial de un elemento a un conjunto.

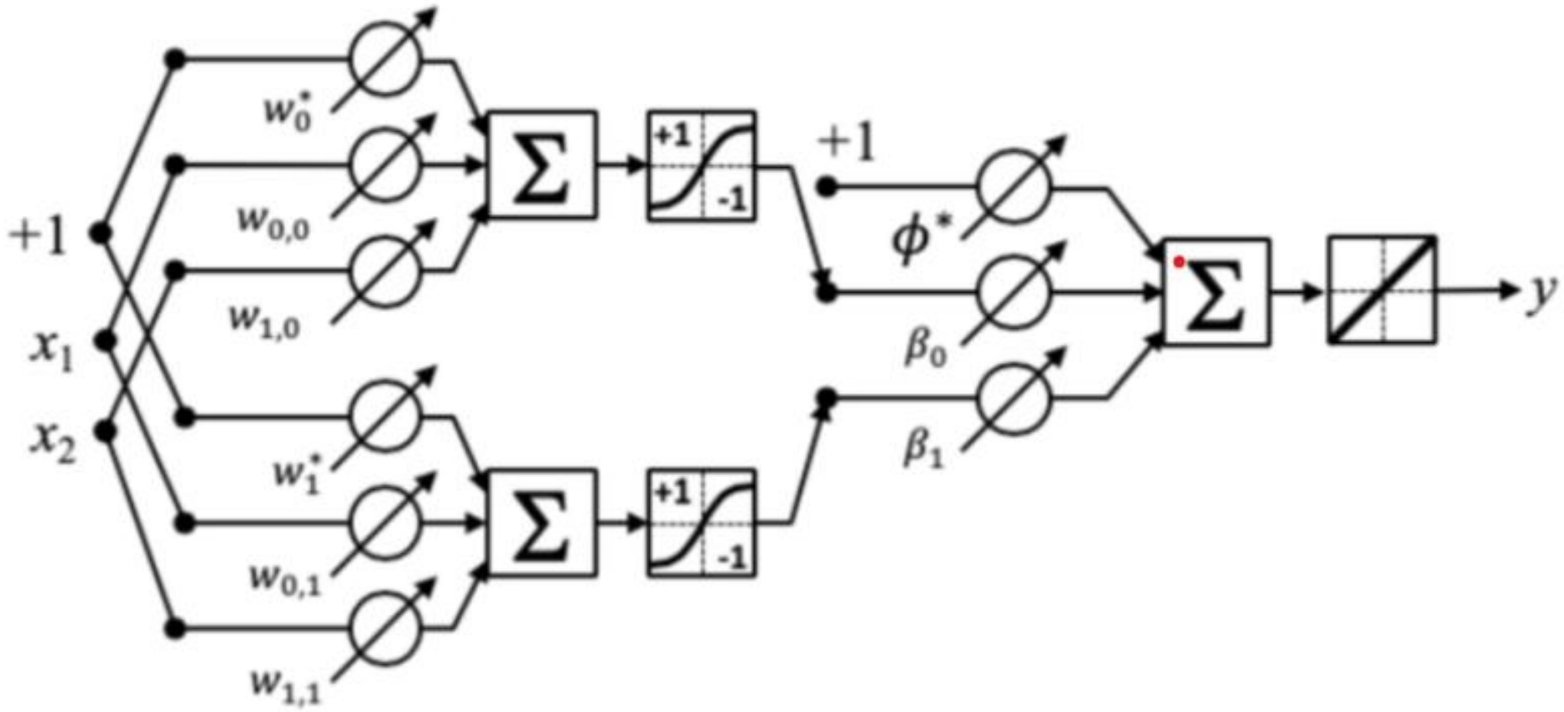
$\mu_A(u) \rightarrow [0,1]$



1974 Backpropagation o Regla Delta generalizada

Tesis doctoral de P. Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*.Harvard, Cambridge, MA, August 1974.

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \mu \frac{\partial}{\partial \mathbf{w}(k)} \left[e^2(k) \right]$$

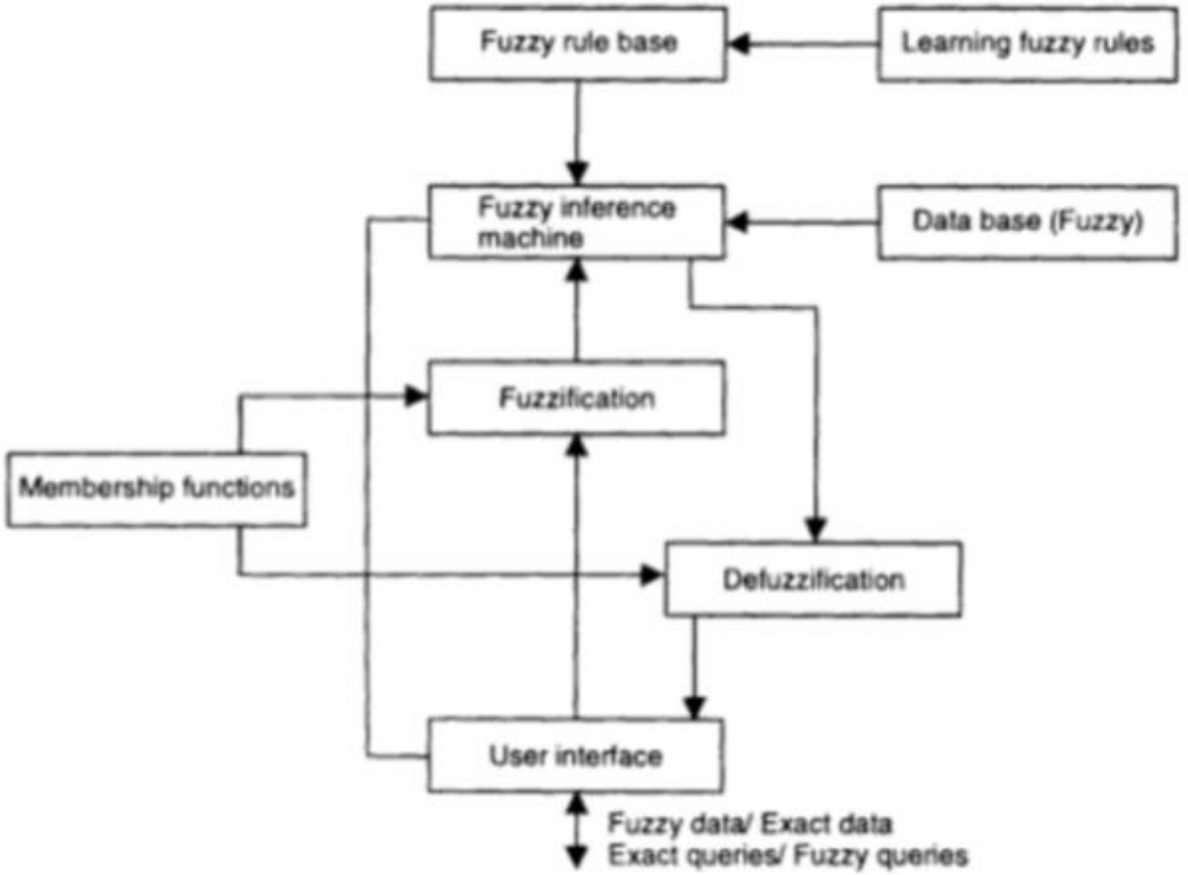


1980 Sistemas difusos de inferencia

Sistemas de representación de datos inexactos y conocimiento heurístico usando conjunto borrosos y reglas difusas en lugar de exactas.

Componentes:

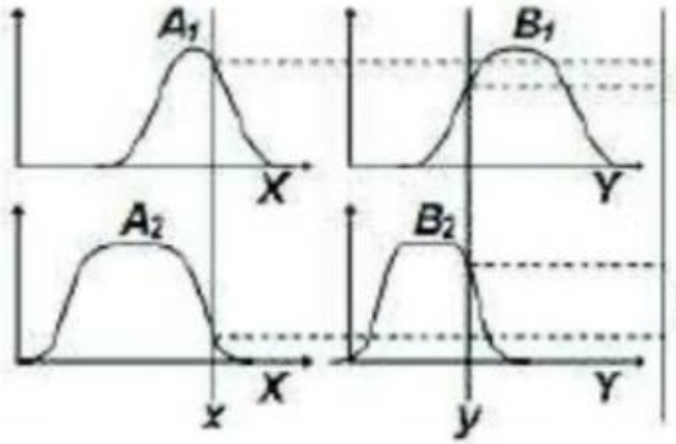
- Variables difusas
- Fuzzificator: convierte entradas precisas en valores de pertenencia
- Reglas difusas: Obtenidas de especialistas o datos numéricos
- Motor de Inferencia Difusa:
- Convierte conjuntos fuzzy en conjuntos fuzzy.
- Determina como las reglas son activadas o combinadas.
- El resultado es un conjunto fuzzy
- Defuzzificator: Convierte el conjunto fuzzy de resultado en un valor preciso



Ventajas:

- Facilidad de implementación
- Mantenimiento barato
- Robustos

1991 Adaptive Neurofuzzy Inference System

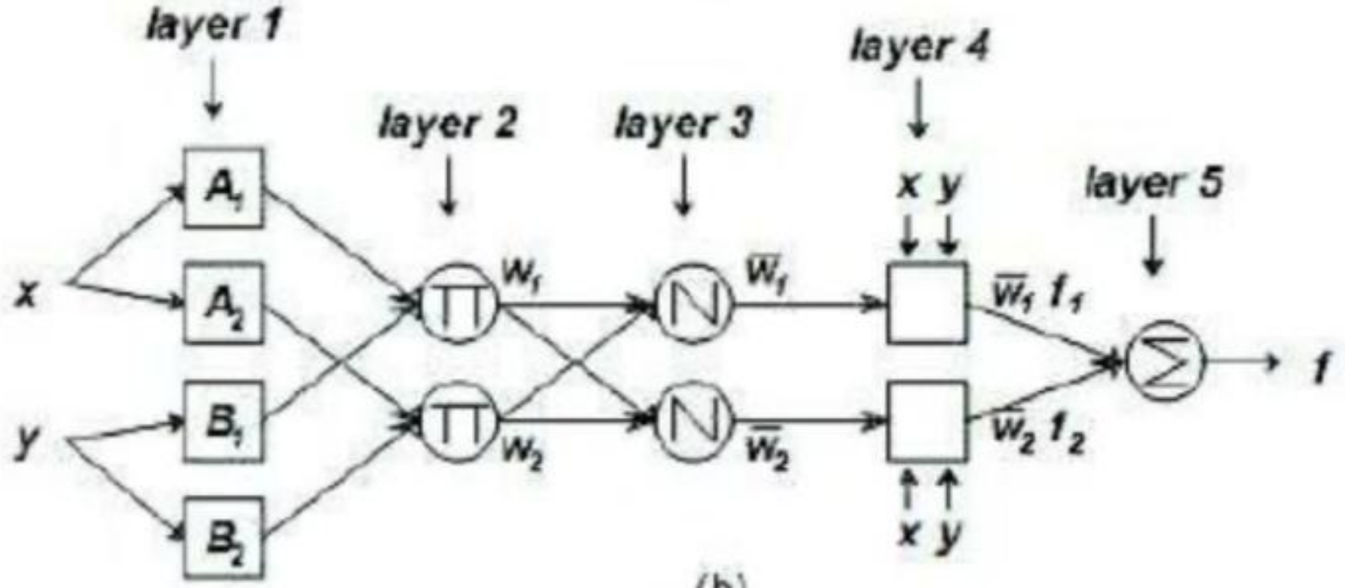


$$f_1 = p_1 x + q_1 y + r_1$$

$$f_2 = p_2 x + q_2 y + r_2$$

$$f = \frac{w_1 f_1 + w_2 f_2}{w_1 + w_2} = \bar{w}_1 f_1 + \bar{w}_2 f_2$$

(a)

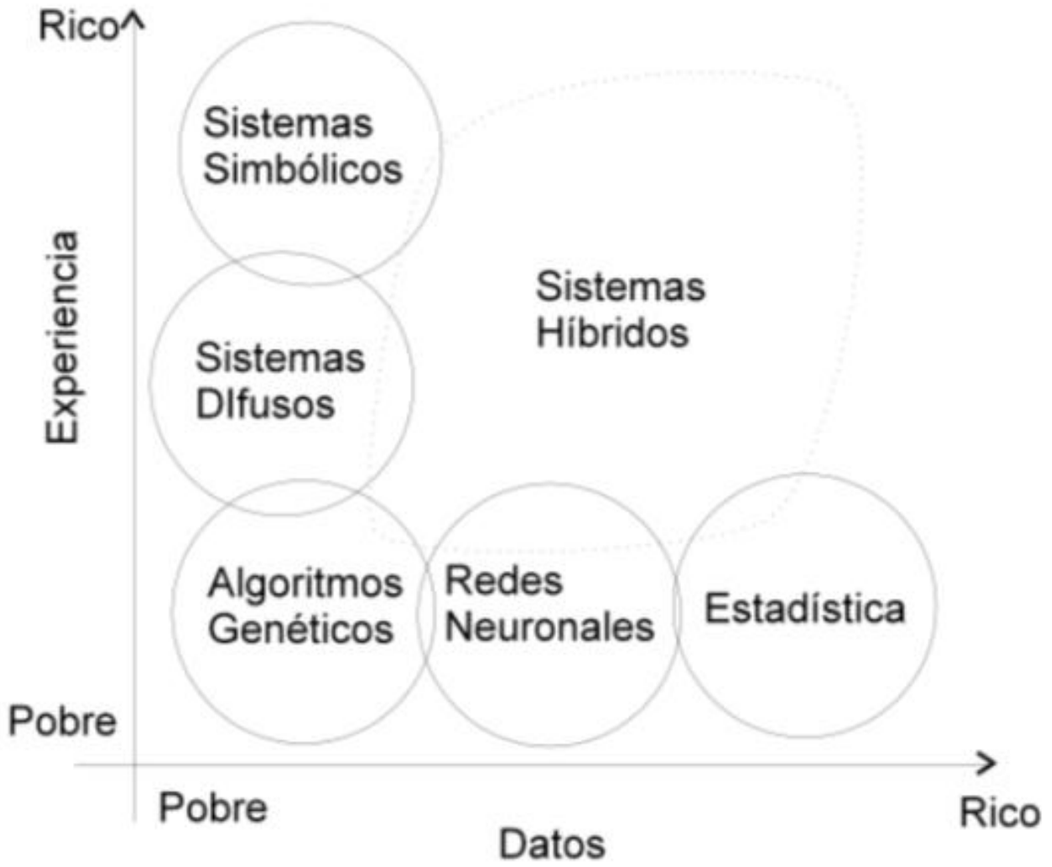


(b)

1995 Ingenieria del Conocimiento

La solución es dependiente del problema y del conocimiento disponible.

- Métodos Estadísticos
- Sistemas simbólicos basados en reglas
- Sistemas difusos
- Redes Neuronales
- Computación evolutiva
- Enjambres de partículas
- …



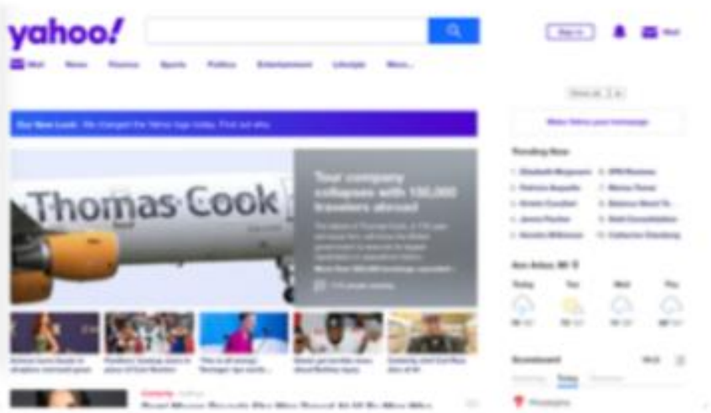
Sistemas que combinan paradigmas

- Sistemas basados en reglas
- Sistemas difusos
- Redes Neuronales
- Otros
  - ✓ Algoritmos Evolutivos
  - ✓ Razonamiento probabilístico
  - ✓ Etc.

1995 Explosión de la Web

- De uso científico y gubernamental a comercial y de consumo
- Creación de Amazon en 1994, y venta de su primer libro en 1995.
- Creación de Yahoo! en 1994.
- Creación de eBay y AltaVista en 1995
- Creación de google en 1996 (algoritmo PageRank)

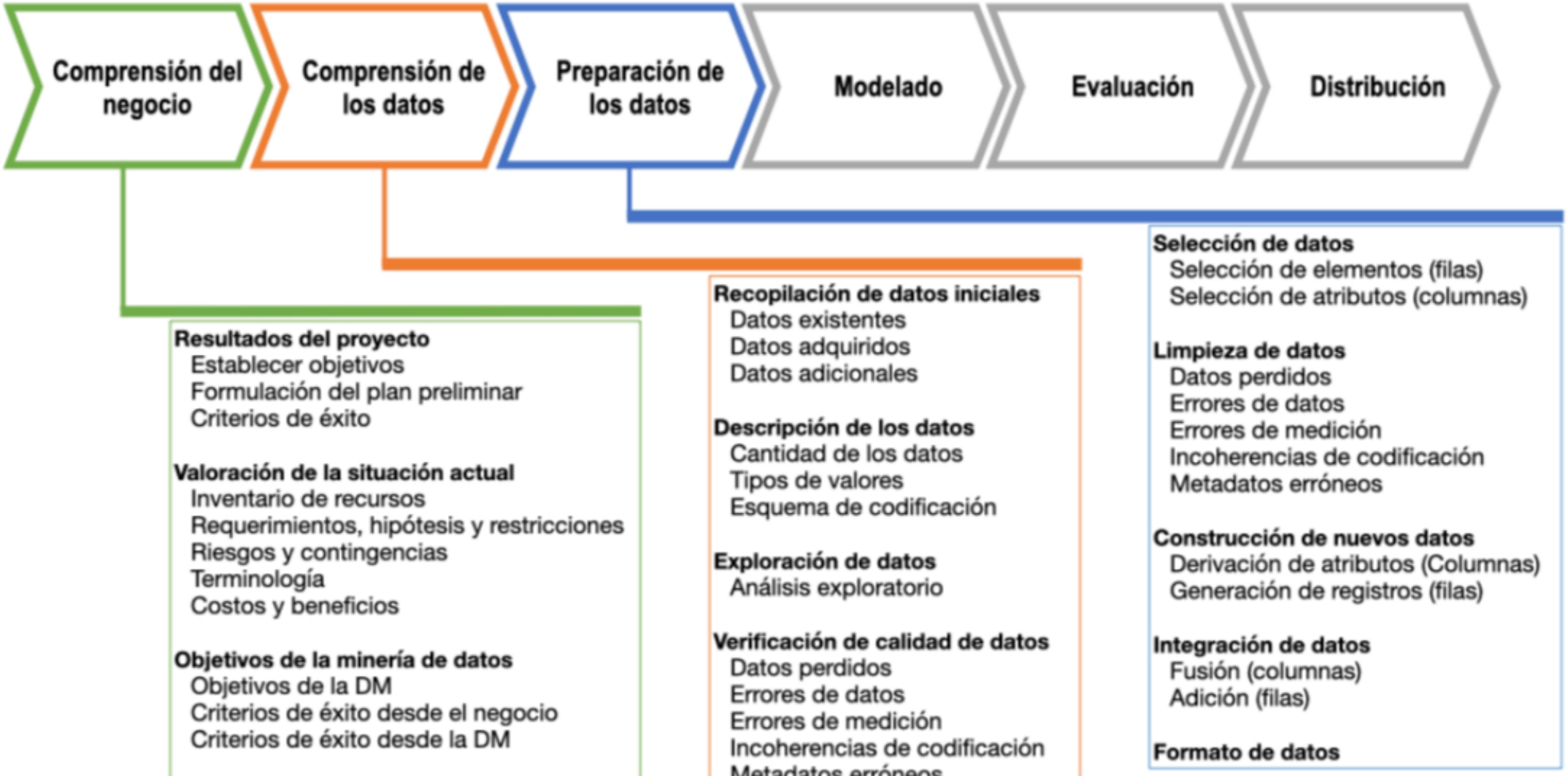
Precursores en la Personalización de la Experiencia online!



1996

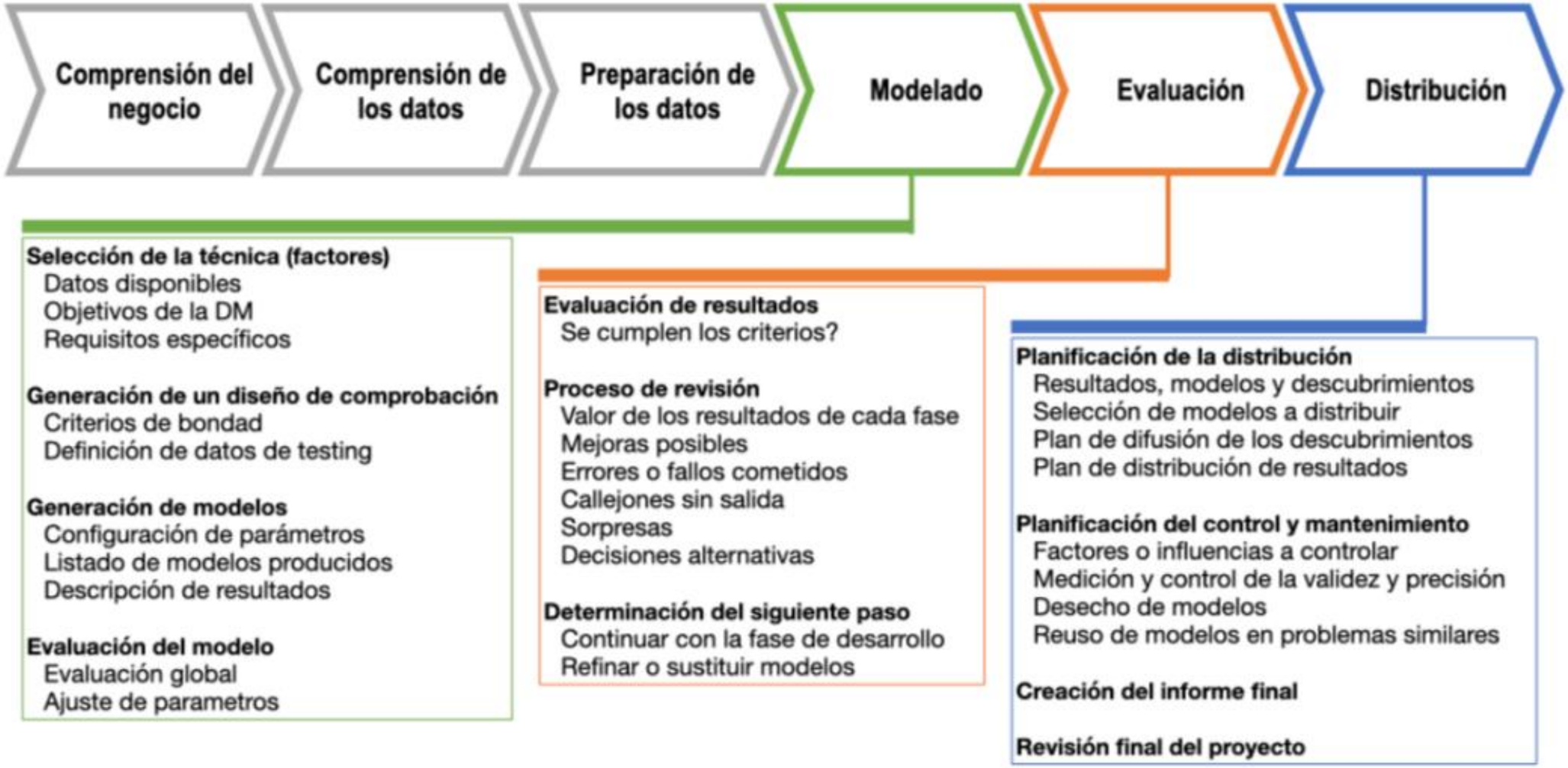
CRISP-DM

Cross-industry standard process for data mining



1996 CRISP-DM

Cross-industry standard process for data mining



1996 Data Science

Programación

Visualización de Datos

Aprendizaje de Máquinas

Inferencia Estadística

Limpieza de Datos

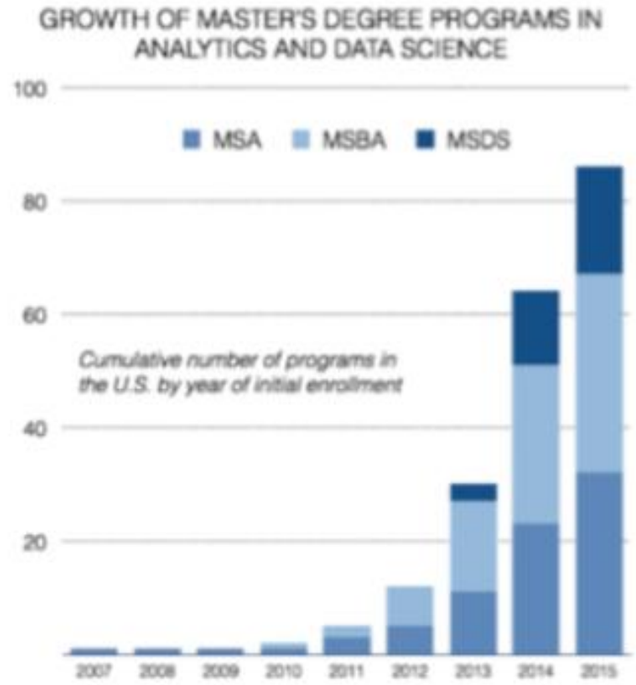
Adquisición de Datos

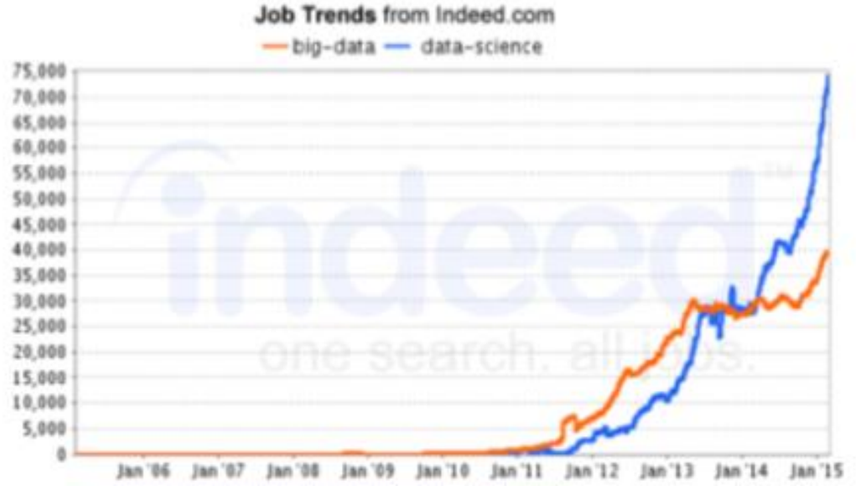
Computación Reproducible

Modelos Estadísticos

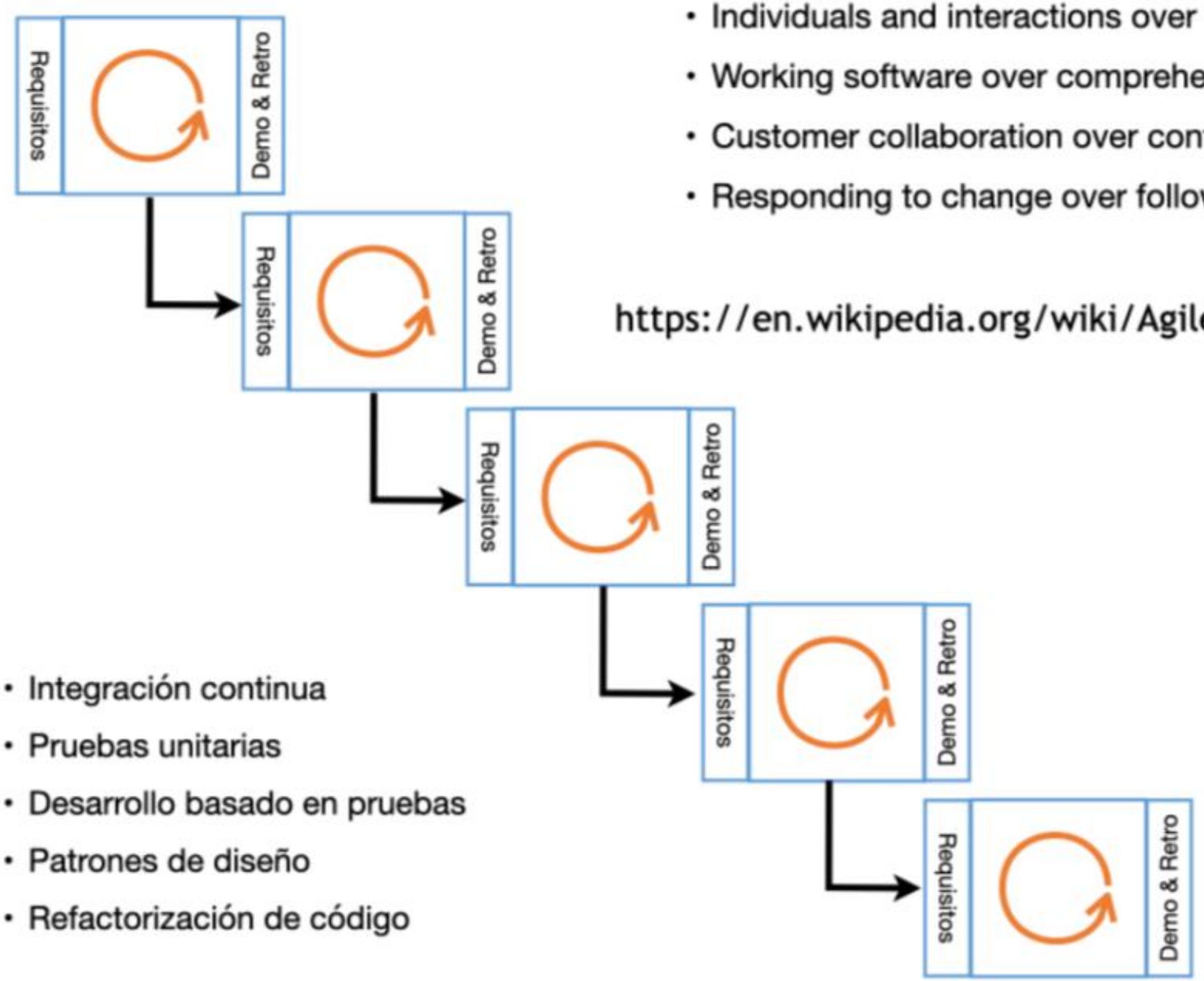
Productos de Datos

Experticia en el problema





2001 Métodos Agiles y manifiesto ágil



- Individuals and interactions over processes and tools
- Working software over comprehensive documentation
- Customer collaboration over contract negotiation
- Responding to change over following a plan

https://en.wikipedia.org/wiki/Agile_software_development

- Integración continua
- Pruebas unitarias
- Desarrollo basado en pruebas
- Patrones de diseño
- Refactorización de código

2005 Traditional Analytics

2006 Analítica Predictiva

Aprendizaje de Máquinas
Aprendizaje (autónomo) a partir de los datos disponibles

Analítica predictiva
Uso de datos históricos para construir modelos matemáticos que capturan tendencias para pronosticar eventos futuros

Metodologías
Regresión lineal
Regresión logística
Agrupamiento
Árboles de decisión
Inteligencia Artificial/Aprendizaje de máquinas
Redes neuronales y aprendizaje profundo
Procesamiento del lenguaje natural

Tipos de modelos
Pronóstico
Clasificación
Outliers
Series de tiempo
Agrupamiento

Aplicaciones
Detección de fraude
Optimización de campañas de marketing
Mejora de operaciones (inventario y manejo de recursos)
Reducción de riesgo
Mantenimiento/abandono/captura de clientes
Análisis del comportamiento
Venta cruzada
Valoración

REPORTE: ¿Qué pasó?
Herramientas de búsqueda, reportaría y consulta.

ANALÍSIS: ¿Por qué pasó?
OLAP y herramientas de visualización

MONITOREO: ¿Que está pasando ahora?
Dashboards, scoreboards

PREDICCIÓN: ¿Qué podría pasar?
Analítica predictiva

Planeamiento de la demanda
Servicio al cliente
Mejora de la calidad
Presupuesto
Gestión de personal
Retail
Lenguaje natural

Técnicas Avanzadas en Análisis Estadístico

1 **Análisis Multivariante**

Explorar relaciones complejas entre múltiples variables.

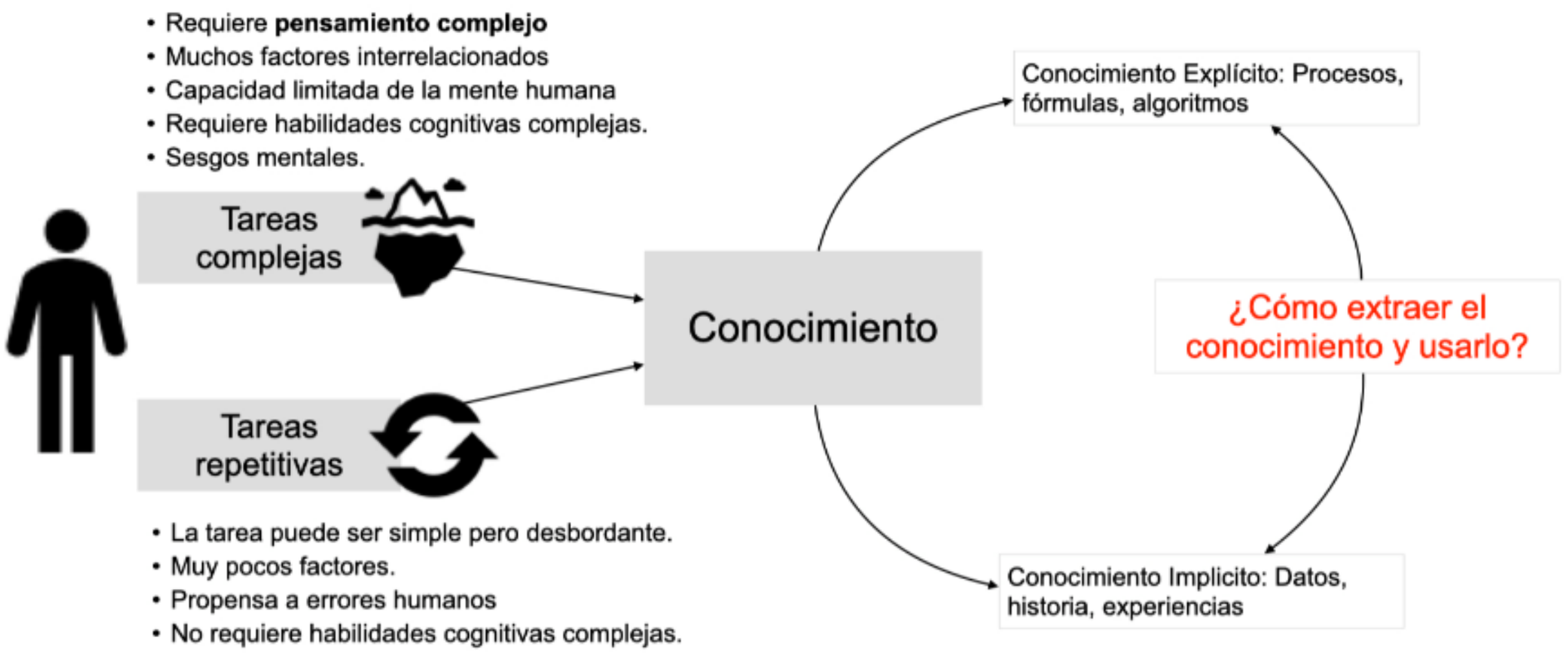
2 **Procesos Estocásticos**

Modelar fenómenos aleatorios que cambian con el tiempo.

3 **Series de Tiempo**

Analizar datos secuenciales para identificar patrones y realizar pronósticos.

¿Por qué se deben tomar decisiones basadas en datos?



La Importancia de la Ciencia de Datos

Toma de Decisiones

Los datos permiten a las empresas tomar decisiones más informadas y estratégicas.

Innovación

Los datos impulsan la innovación y el desarrollo de nuevos productos y servicios.

Eficiencia

Los datos permiten optimizar procesos y mejorar la eficiencia operativa.

Experiencia del Cliente

Los datos permiten personalizar la experiencia del cliente y ofrecer productos y servicios más relevantes.

Análisis Multivariante: Conceptos y Aplicaciones

Análisis de Cluster

Agrupar datos en grupos basados en similitudes.

Análisis de Regresión

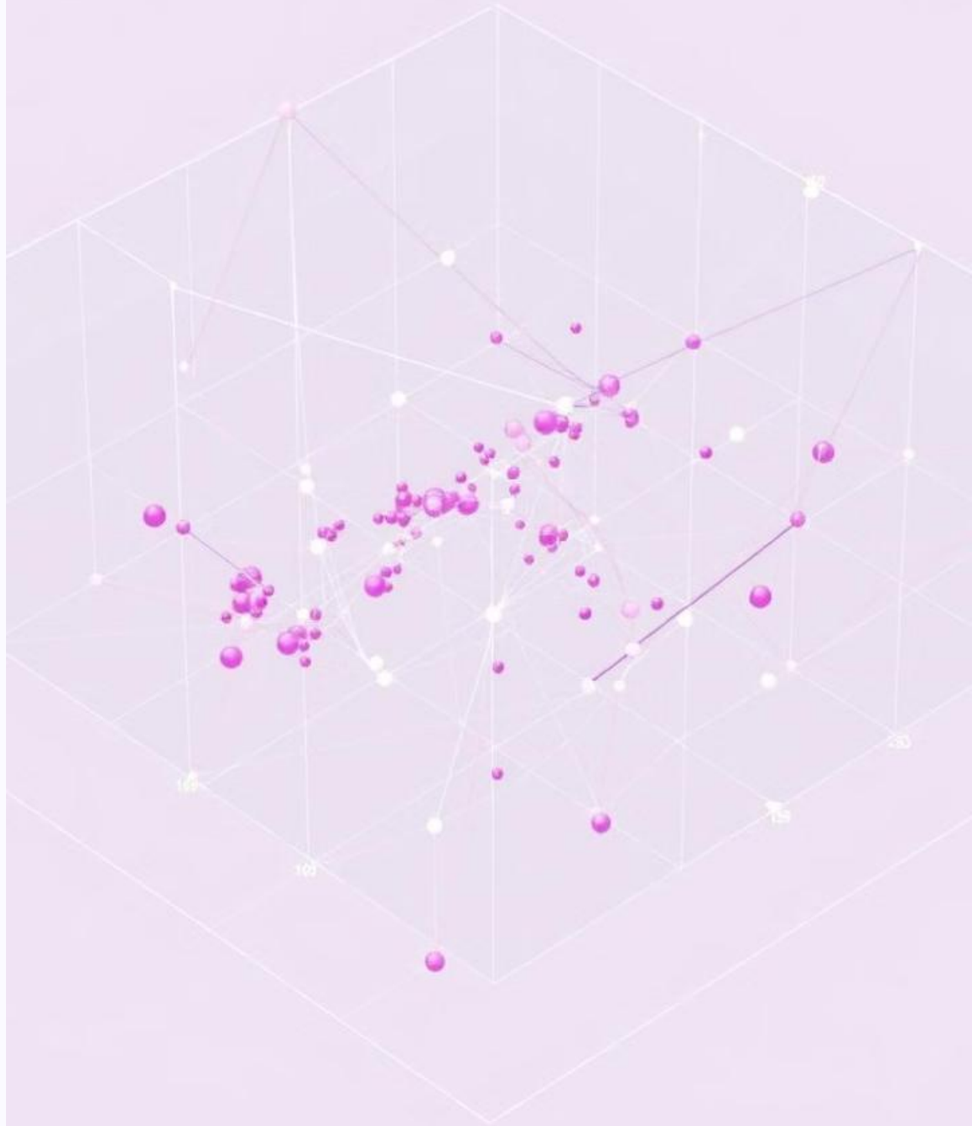
Modelar la relación entre variables para predecir resultados.

Análisis de la Varianza (ANOVA)

Comparar grupos de datos para determinar diferencias significativas.

Análisis de Componentes Principales (PCA)

Reducir la dimensionalidad de los datos para simplificar el análisis.



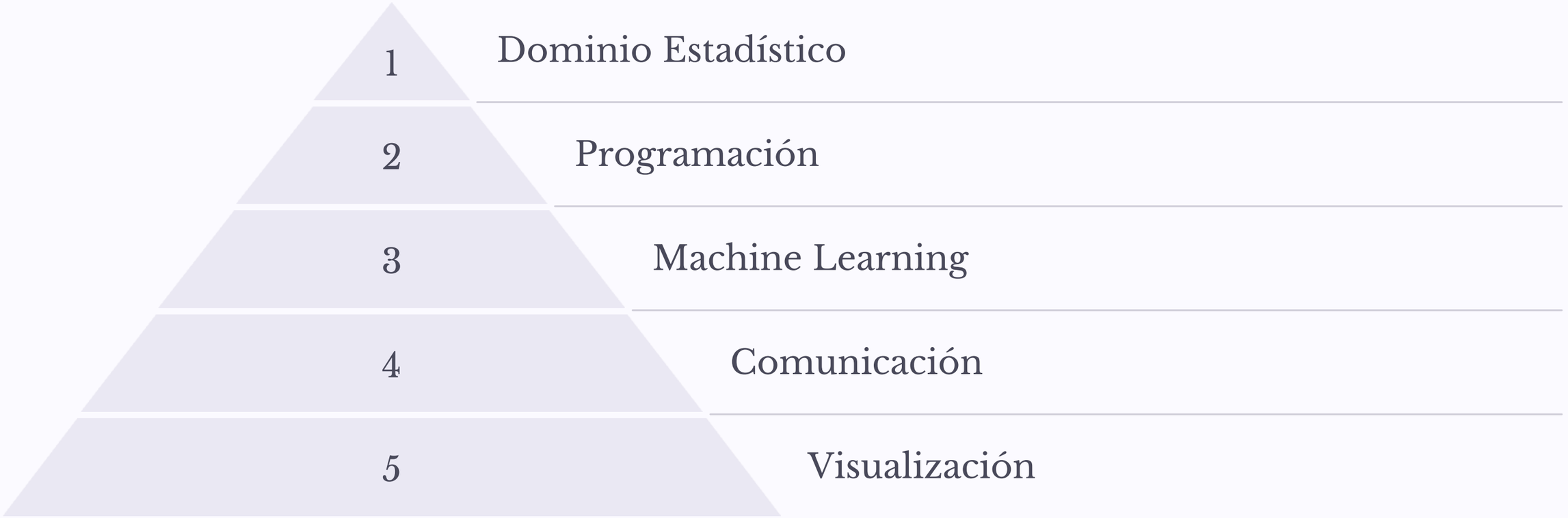
Procesos Estocásticos: Modelación y Predicción

1. Los procesos estocásticos modelan eventos aleatorios que evolucionan con el tiempo.
2. Estos procesos se utilizan en campos como la finanzas, la meteorología y la biología.
3. Las técnicas de modelado y predicción permiten comprender y predecir el comportamiento de estos sistemas.

Series de Tiempo: Análisis y Pronósticos

1. Identificar patrones y tendencias en datos secuenciales.
2. Analizar la estacionalidad, la tendencia y la aleatoriedad.
3. Realizar pronósticos de eventos futuros basados en datos históricos.
4. Optimizar procesos, planificar recursos y tomar decisiones estratégicas.

Habilidades en Ciencia de Datos: Objetivos de Aprendizaje



Casos de Éxito y Ejemplos Prácticos

1

Salud

Análisis de datos para mejorar el diagnóstico y la atención médica.

2

Finanzas

Detección de fraudes, gestión de riesgos y predicción de mercados.

3

Marketing

Segmentación de clientes, personalización de campañas y análisis de la experiencia del cliente.

Conclusiones y Próximos Pasos

1

Desbloquea el potencial de los datos

La ciencia de datos es una herramienta poderosa para resolver problemas y crear valor.

2

Amplía tus habilidades

El dominio de las técnicas estadísticas avanzadas es fundamental para el éxito en la era de los datos.

3

Aplica el conocimiento

Utiliza las habilidades aprendidas para mejorar la toma de decisiones.