Universidad de Buenos Aires
Facultad de Ciencias Económicas

Carrera de Actuario.
Cátedra de Análisis Numérico.
Prof. Titular: Javier García Fronti.

Curso modalidad Virtual

Prof. Adjunto: R. Darío Bacchini

Melisa Elfenbaum

Introducción al aprendizaje automático y minería de datos

El impacto de los datos masivos (Big Data)

Últimos años de la era digital:

- Aumento exponencial de datos generados directa o indirectamente a través de nuevas fuentes (clics en la web, redes sociales, celulares, IoT)
- Mayor poder y velocidad de procesamiento
- Costos de almacenamiento más bajos
- Big Data: datos grandes, diversos y complejos generados por instrumentos, sensores, transacciones en internet, e-mails, videos y todos las fuentes digitales disponibles en la actualidad y en el futuro

El impacto de los datos masivos (Big Data)

Definición de las 3Vs: se alimentan mutuamente y engloban los principales atributos del Big Data:

- Volumen
- Velocidad (tiempo real)
- Variedad (estructurados y no estructurados)

Definición de las 5Vs:

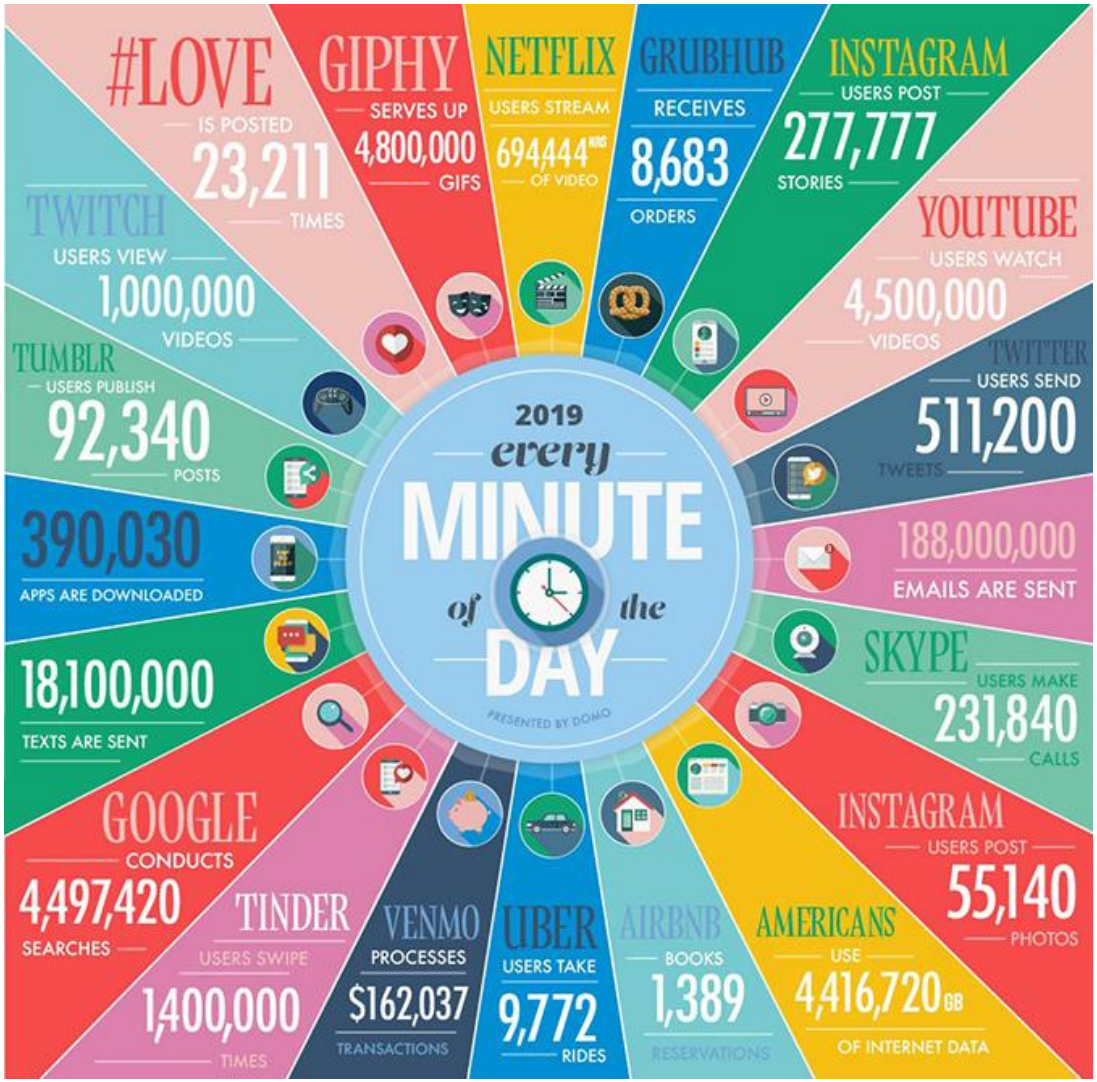
- Veracidad: la necesidad de gestionar la incertidumbre proveniente de los datos
- Valor: cuando los datos se convierten en información, representa el aspecto más relevante del Big Data

El impacto de los datos masivos (Big Data)

Impacto del Big Data

- Tecnología: asociado con el almacenamiento, la obtención y el análisis de los volúmenes cada vez mayores de datos (nuevos tipos de base de datos y almacenamiento en la nube)
- Impacto social: problemas de privacidad y seguridad de datos, la falta de regulación para el uso comercial, la propiedad intelectual y la responsabilidad
- Creación de valor: el valor que se puede obtener a partir de la toma de decisiones basadas en todos los datos disponibles, por la optimización de procesos y descubrimiento de nuevas oportunidades

El impacto de los datos masivos (Big Data)



Procesamiento de datos

- Dato: conjunto de valores que por sí solos son irrelevantes.
- Información: conjunto de datos procesados e interrelacionados, que tienen un significado
- Conocimiento: información integrada con la experiencia, acumulada luego de un proceso de análisis para la toma de decisiones
- Ejemplo:
  - Dato: notas de los parciales
  - Información: cantidad de aprobados
  - Conocimiento: aprobados de otros cuatrimestres/temas con más dificultades
  - Acción: reforzar la explicación sobre ciertos temas

Procesamiento de datos

Procesamiento de datos: Recolección

- **Datos públicos:** se pueden descargar en formato csv, json o xml.

  Ejemplos:

  - Datos de organismos públicos: https://datos.gob.ar/dataset
  - Datos financieros: https://finance.yahoo.com/quote/AAPL/history?p=AAPL

Procesamiento de datos: Recolección

- **API (Application Programming Interface)**: interfaz de programación de aplicaciones, es un intermediario que permite que dos aplicaciones se comuniquen entre sí.
  - Se pueden utilizar APIs para acceder a datos del clima, financieros, noticias, redes sociales, etc.
  - Algunas se pueden acceder de forma gratuita.
  - En general los sitios que proporcionan las APIs tienen una sección en donde explican como acceder. Ejemplo: https://developer.twitter.com/

Procesamiento de datos: Recolección

- **Web Scraping**: mediante un lenguaje de programación se recorre una página web y extraen datos específicos (urls, imágenes, subtítulos, etc).
  - No todas las páginas, ni todas las secciones de una página son accesibles legalmente.
  - La información a la que se puede acceder viene dada por un archivo llamado “robots.txt” que por convención se encuentra en la raíz de la página.
    Ejemplo: https://www.amazon.com/robots.txt

Procesamiento de datos: Almacenamiento

- **Archivos** xml, csv, txt, entre otros.
- **Bases de datos:** Colección de datos interrelacionados almacenados en conjunto sin redundancias. Ventajas: mantener la integridad, consistencia y seguridad de los datos, reducir redundancia, optimizar el rendimiento.

  Tipos de bases de datos

  - BD relacional o SQL
  - BD no relacional o NoSQL (Not only SQL)

Procesamiento de datos: Almacenamiento

- BD relacional o SQL: Almacenan datos con un modelo relacional (grupo de tablas representan los datos y las relaciones entre ellos).
- BD no relacional o NoSQL: Tienen esquemas flexibles para crear aplicaciones complejas cuando las BD relacionales generan problemas de escalabilidad y rendimiento. Crecieron por el surgimiento del Big Data.

Procesamiento de datos: Análisis

- El objetivo de descubrir patrones en los datos que pueden usarse para especificar la estrategia del negocio o para identificar comportamientos fuera de lo común.
- Data mining (minería de datos): se obtienen los patrones sobre datos estructurados. Incluyen análisis analíticos, histogramas, análisis de series de tiempo, cuadros de mando, comparaciones con períodos anteriores, entre otros.

Procesamiento de datos: Análisis

- Text mining (minería de textos): busca extraer información útil de datos textuales no estructurados a través de la identificación y exploración de patrones. Incluye análisis de sentimiento y minería de opiniones.
- Machine learning (aprendizaje automático): técnicas avanzadas de programación y estadísticas con el objetivo de realizar automáticamente análisis de grandes volúmenes de datos y detección de patrones sin intervención humana.

El aprendizaje automático

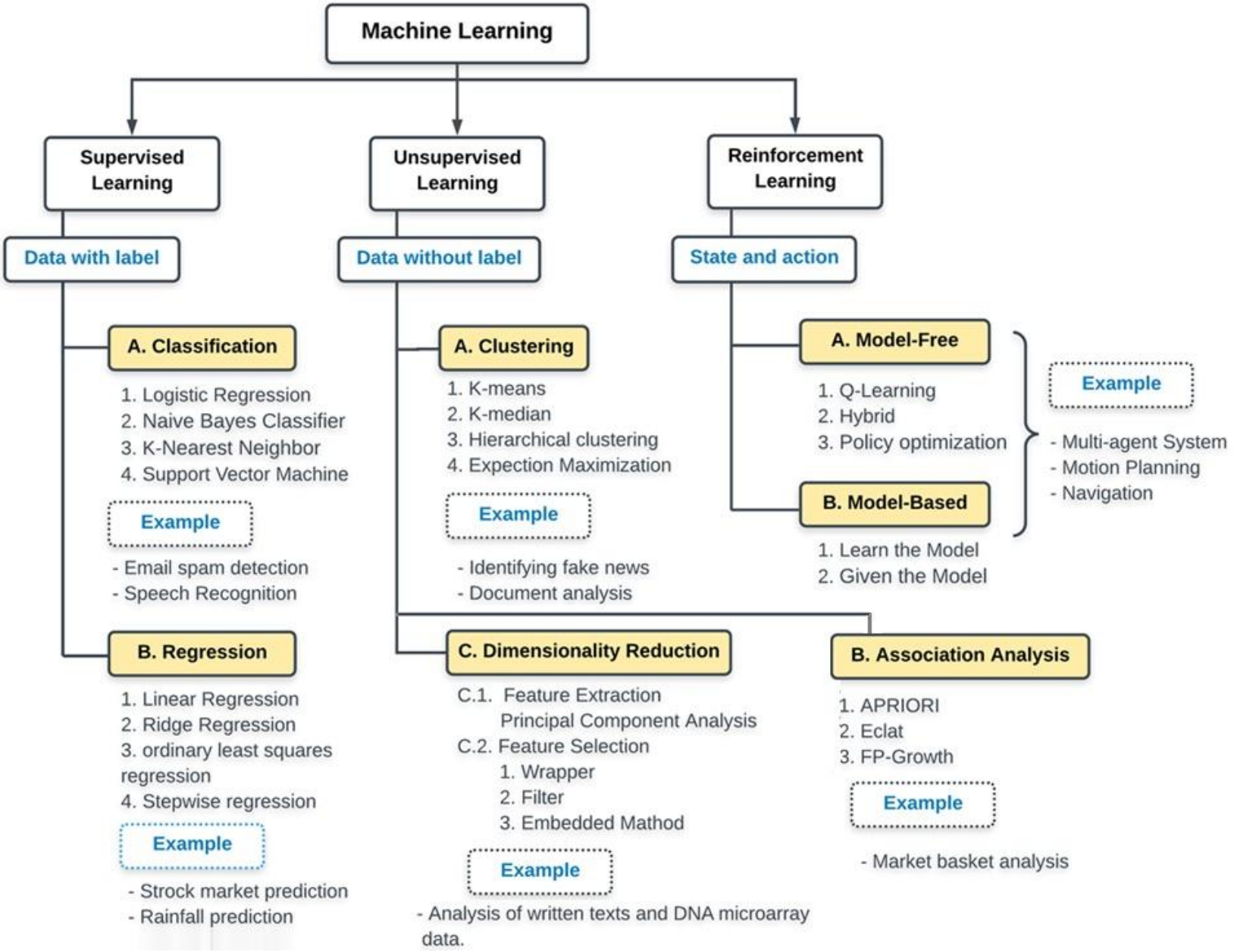
- Conjunto de métodos que permiten detectar automáticamente patrones en función de la experiencia pasada.
- A partir del Big Data, las organizaciones utilizan métodos de aprendizaje automático para aprovechar todos los datos disponibles de las crecientes fuentes externas (redes sociales, datos del mercado, meteorológicos, de los clientes, etc).

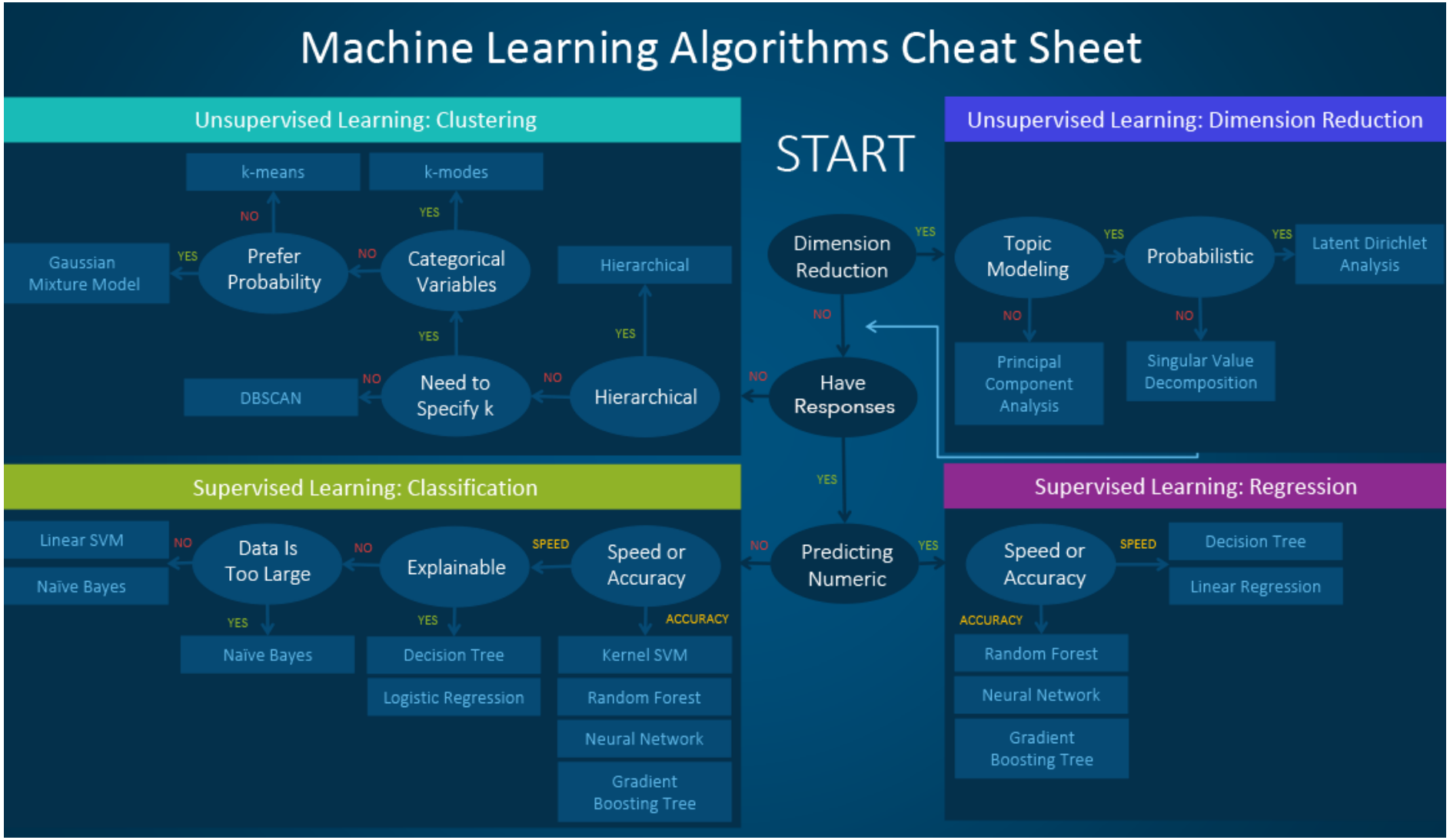
El aprendizaje automático

- Aplicaciones
  - Autos que se manejan solos
  - Filtrado de correo electrónico no deseado
  - Segmentación de clientes
  - Clasificación de páginas web
  - Reconocimiento de voz y autenticación de personas
  - Detección de objetos y clasificación de imágenes
  - Traducción de documentos
  - Detección de valores atípicos
  - Sistemas de recomendación

El aprendizaje automático

- Aprendizaje supervisado: a cada uno de los datos se les asigna una salida o etiqueta
- Aprendizaje no supervisado: datos sin etiquetas asociadas.
- Aprendizaje reforzado: explorar diferentes opciones, evaluando cada resultado para determinar cuál es el óptimocon el fin de maximizar alguna noción de "recompensa". Debido a que requiere datos y entorno simulados, es más difícil de aplicar a situaciones comerciales prácticas.

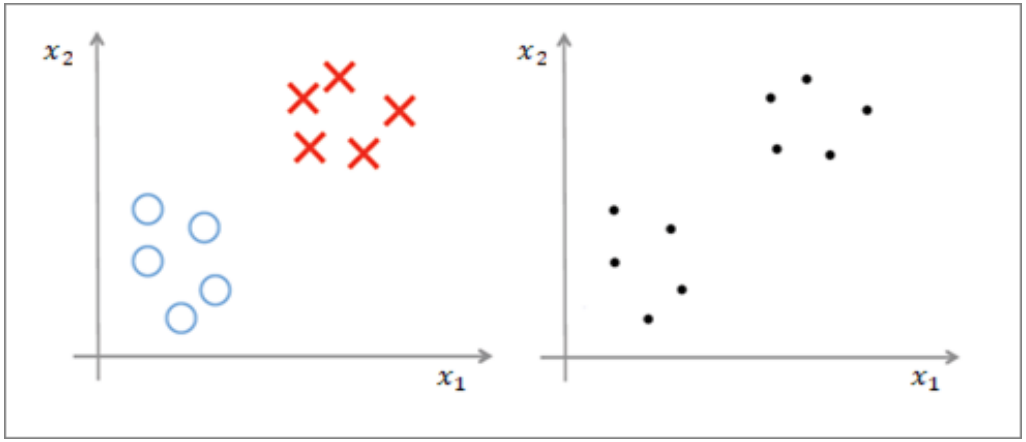

Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Data with label
Data without label
State and action
A. Classification
1. Logistic Regression
2. Naive Bayes Classifier
3. K-Nearest Neighbor
4. Support Vector Machine
Example
- Email spam detection
- Speech Recognition
B. Regression
1. Linear Regression
2. Ridge Regression
3. ordinary least squares regression
4. Stepwise regression
Example
- Strock market prediction
- Rainfall prediction
A. Clustering
1. K-means
2. K-median
3. Hierarchical clustering
4. Expection Maximization
Example
- Identifying fake news
- Document analysis
C. Dimensionality Reduction
C.1. Feature Extraction
Principal Component Analysis
C.2. Feature Selection
1. Wrapper
2. Filter
3. Embedded Mathod
Example
- Analysis of written texts and DNA microarray data.
A. Model-Free
1. Q-Learning
2. Hybrid
3. Policy optimization
B. Model-Based
1. Learn the Model
2. Given the Model
Example
- Multi-agent System
- Motion Planning
- Navigation
B. Association Analysis
1. APRIORI
2. Eclat
3. FP-Growth
Example
- Market basket analysis


Machine Learning Algorithms Cheat Sheet
START
Unsupervised Learning: Clustering
k-means
k-modes
NO
YES
Gaussian Mixture Model
YES
Prefer Probability
NO
Categorical Variables
Hierarchical
YES
YES
DBSCAN
NO
Need to Specify k
NO
Hierarchical
Unsupervised Learning: Dimension Reduction
Dimension Reduction
YES
Topic Modeling
YES
Probabilistic
YES
Latent Dirichlet Analysis
NO
NO
NO
Principal Component Analysis
Singular Value Decomposition
NO
Have Responses
YES
Supervised Learning: Classification
Linear SVM
NO
Data Is Too Large
NO
Explainable
SPEED
Speed or Accuracy
Naïve Bayes
YES
YES
ACCURACY
Naïve Bayes
Decision Tree
Kernel SVM
Logistic Regression
Random Forest
Neural Network
Gradient Boosting Tree
NO
Predicting Numeric
YES
Supervised Learning: Regression
Speed or Accuracy
SPEED
Decision Tree
Linear Regression
ACCURACY
Random Forest
Neural Network
Gradient Boosting Tree

El aprendizaje automático

Ejemplo: Diagnóstico médico

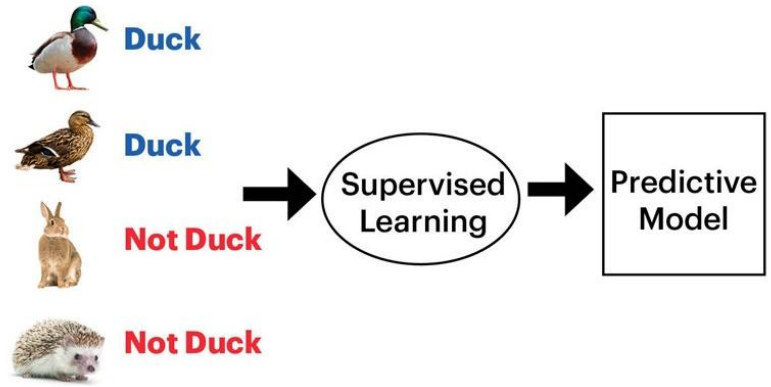
- Datos: grado de sedentarismo (x1) y colesterol (x2).
- Si se sabe que un subconjunto de esos pacientes tiene diabetes -cruces rojas- y otro no -círculos azules- se puede aplicar un método supervisado (clasificación), sino uno no supervisado (agrupamiento)

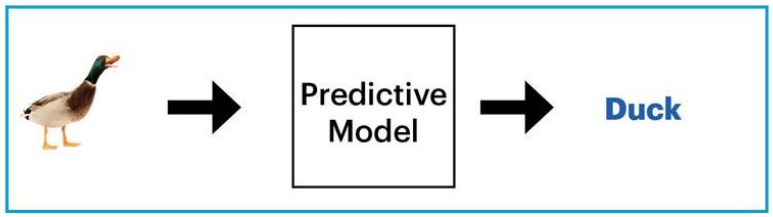


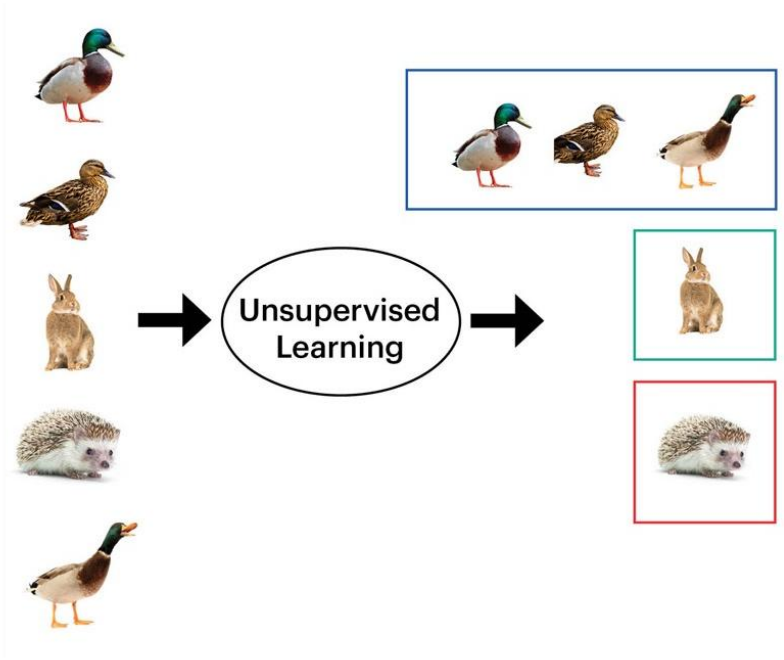
El aprendizaje automático

Ejemplo: Clasificación de imágenes

- Datos: imágenes de animales, que pueden estar etiquetadas o no







El aprendizaje automático

Ejemplos de APIs

- Cloud Vision: detección de imágenes
  cloud.google.com/vision
- Cloud Speech: transcripción de voz a texto
  cloud.google.com/speech-to-text
- Cloud Natural Language: comprensión de textos
  cloud.google.com/natural-language

Data Mining sobre datos públicos

- **Datos:**
  - Descargar csv: https://datos.gob.ar/dataset/salud-covid-19-casos-registrados-republica-argentina
- **Librerías en R:**
  - sqldf
  - ggplot2

Data Mining sobre datos públicos

- **Paquetes**

```
#Consultas sql sobre los dataframe
library(sqldf)
options(sqldf.driver = "SQLite")
#Graficos
library(ggplot2)
```

- **Lectura del csv**

```
covid_arg <- read.csv("C:/Covid19Casos.csv", encoding =
"UTF-8")
```

Data Mining sobre datos públicos

- **Cantidad de confirmados por semana**

```
sqldf()

#Con la libreria sqldf se pueden filtrar y agrupar los datos con consultas sql
casos_por_semana <- sqldf("SELECT sepi_apertura as Semana, count(1) as Cantidad
FROM covid_arg
where clasificacion_resumen = 'Confirmado'
group by sepi_apertura")

#Grafico de barras

graf_semanas<-ggplot(data=casos_por_semana, aes(x=Semana, y=Cantidad)) +
geom_bar(stat="identity",fill="#f70388")

graf_semanas
```

Data Mining sobre datos públicos

- **Cantidad de confirmados y fallecidos por grupo etario**

```
sqldf()
casos_por_grupo <- sqldf( "SELECT CASE WHEN
edad_años_meses = 'Meses' OR (edad_años_meses = 'Años' AND
EDAD<10) THEN '0 A 9' WHEN EDAD<20 THEN '10 A 19' WHEN
EDAD<30 THEN '20 A 29' WHEN EDAD<40 THEN '30 A 39' WHEN
EDAD<50 THEN '40 A 49' WHEN EDAD<60 THEN '50 A 59'  WHEN
EDAD<70 THEN '60 A 69' WHEN EDAD<80 THEN '70 A 79'  WHEN
EDAD<90 THEN '80 A 89' WHEN EDAD>=90 THEN '90+' ELSE 'SIN
DATOS' END Grupo, COUNT(1) as Cantidad, SUM(CASE WHEN
FALLECIDO = 'SI' THEN 1 ELSE 0 END) Fallecidos
FROM covid_arg
WHERE clasificacion_resumen = 'Confirmado'
and edad is not null
group by Grupo order by 1")
```

Data Mining sobre datos públicos

- **Cantidad de confirmados y fallecidos por grupo etario**

```
graf_cant_edad<-ggplot(data=casos_por_grupo, aes(x=Grupo,
y=Cantidad,fill=Grupo, group =1)) +
geom_bar(stat="identity") +
geom_line(aes(x=Grupo, y=Fallecidos),stat="identity")+
geom_text(aes(label=Cantidad), vjust=1.6, color="white",
size=3.5)

graf_cant_edad
graf_fallecidos<-ggplot(data=casos_por_grupo,
aes(x=Grupo, y=Fallecidos,fill=Grupo)) +
geom_bar(stat="identity") +
geom_text(aes(label=Fallecidos), vjust=1.6,
color="black", size=3.5)
graf_fallecidos
```

Data Mining sobre datos públicos

- **Cantidad de confirmados y fallecidos por provincia (de las 15 provincias con más casos)**

```
sqldf()
casos_por_provincia <- sqldf( "SELECT
carga_provincia_nombre as Provincia, COUNT(1) as
Cantidad, SUM(CASE WHEN FALLECIDO = 'SI' THEN 1 ELSE 0
END) Fallecidos, carga_provincia_nombre || ' - ' ||
COUNT(1) as Confirmados
FROM covid_arg
WHERE clasificacion_resumen = 'Confirmado'
group by Provincia
order by Cantidad desc limit 15")
```

Data Mining sobre datos públicos

- **Cantidad de confirmados y fallecidos por provincia (de las 15 provincias con más casos)**

```
graf_prov<-ggplot(data=casos_por_provincia,
aes(x=reorder(Provincia, -Cantidad), y=Cantidad,
fill=Confirmados)) + geom_bar(stat="identity") +
geom_text(aes(label=Cantidad), vjust=1.8,
color="black", size=3.5)
graf_prov

graf_prov2<-ggplot(data=casos_por_provincia,
aes(x=reorder(Provincia, -Fallecidos), y=Fallecidos,
fill=Provincia)) + geom_bar(stat="identity") +
geom_text(aes(label=Fallecidos), vjust=1.8,
color="black", size=3.5)
graf_prov2
```