



UNIVERSIDAD
NACIONAL
DE COLOMBIA

APRENDIZAJE DE MÁQUINAS

Adquisición, Procesamiento y Etiquetado de Datos

JOHN W. BRANCH

Profesor Titular

Departamento de Ciencias de la Computación y de la Decisión

Director del Grupo de I+D en Inteligencia Artificial – GIDIA

jwbranch@unal.edu.co

<https://github.com/sroble05/3008422-AprendizajeDeMaquinas>

Datos



En el mundo, la cantidad de datos recolectados cada segundo a través de diferentes dispositivos electrónicos es realmente grande. Con un conjunto de estos, y utilizando técnicas matemáticas y computacionales, somos capaces de crear información.

Existen diferentes métodos y caminos para convertir los datos en información. Algunos buscan entender los datos desde el área estadística, mientras otros buscan predecir ciertos valores (clasificación y regresión).

Datos

En la actualidad, la recolección de datos e información se ha vuelto uno de los objetivos principales en la mayoría de empresas. Con ellos, una compañía puede mejorar sus estrategias de ventas, así como atraer a nuevos clientes dependiendo de sus gustos y/o necesidades.



Datos

Actualmente, la utilización de datos no solo nos sirve para un análisis estadístico. Con los avances computacionales en inteligencia artificial, hemos logrado poder automatizar procesos que, años atrás, nos era poco rentable.

Utilizando diferentes técnicas, se han remplazado procesos muy costosos en tiempo y dinero por sistemas automáticos con el mismo o mejor desempeño obtenido por un grupo de profesionales.



Adquisición de Datos

Los datos pueden ser clasificados en 4 dominios, dependiendo de su origen:

TEXTO

CUANDO **EL** OJO **VE** UN **COLOR** SE **EXCITA** INMEDIATAMENTE, Y ÉSTA **ES** SU **NATURALEZA**, ESPONTÁNEA Y DE **NECESIDAD**, PRODUCIR **OTRA** EN LA QUE **EL** COLOR **ORIGINAL** COMPRENDE LA ESCALA **CROMÁTICA** ENTERA. **UN** ÚNICO **COLOR** EXCITA, **MEDIANTE** UNA **SENSACIÓN** ESPECÍFICA, LA TENDENCIA A LA **UNIVERSALIDAD**. EN **ESTO** RESIDE LA LEY **FUNDAMENTAL** DE **TODA** ARMONÍA **DE** LOS **COLORES...**

AUDIO



VIDEO



IMAGEN














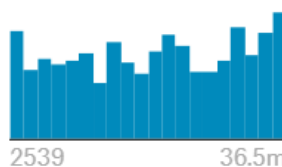
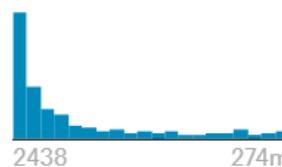
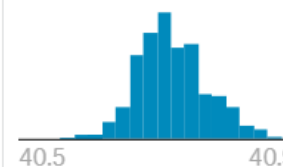
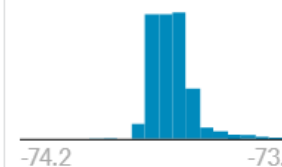
Adquisición de Datos

Con ayuda de estos datos podemos predecir comportamientos y patrones. Esto con el fin de generar información vital para la toma de decisiones.



Datos – Texto (Estructurado)

New York City Airbnb Open Data - Kaggle

AB_NYC_2019.csv (6.75 MB)								
16 of 16 columns ▾ Views   								
	 id ▾ listing ID	 name ▾ name of the listing	 host_id ▾ host ID	 host_name ▾ name of the host	 neighbourhood_gro ▾ location	 neighbourhood ▾ area	 latitude ▾ latitude coordinates	 longitude ▾ longitude coordinates
	 2539 36.5m	47905 unique values	 2438 274m	11452 unique values	Manhattan 44% Brooklyn 41% Other (3) 15%	Williamsburg 8% Bedford-Stuyves... 8% Other (219) 84%	 40.5 40.9	 -74.2 -73.7
1	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237
2	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377
3	3647	THE VILLAGE OF HARLEM...NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419
4	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976
5	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399
6	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975
7	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596
8	5178	Large Furnished Room Near B'way	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493

Datos – Texto (No estructurado)

Cortas frases de textos con el fin de encontrar la similitud de diferentes documentos.

Text Similarity - Kaggle



Datos - Audio

Un conjunto de datos de audios de gatos y perros con el fin de clasificar a que animal pertenece cada sonido.



Data (59 MB)

Data Sources

▼ cats

- cat_110.wav
- cat_112.wav
- cat_115.wav
- cat_126.wav
- cat_129.wav
- cat_130.wav
- cat_133.wav
- cat_135.wav
- cat_137.wav
- cat_14.wav
- ... 10 more

▼ test

- dog_barking_112.wav
- dog_barking_12.wav
- dog_barking_15.wav
- dog_barking_19.wav
- dog_barking_24.wav
- dog_barking_3.wav
- dog_barking_34.wav
- dog_barking_43.wav

Audio Cats and Dogs - Kaggle
Classify raw sound events

Datos - Video

Un conjunto de datos de videos provenientes de la plataforma de Google YouTube con el fin de hacer seguimiento a la tendencia de videos dentro de la plataforma.

Trending YouTube Video Statistics - Kaggle



Datos - Imagen

Natural Images - Kaggle

Clasificación de 8 diferentes clases de un conjunto de imágenes “naturales”.



Preprocesamiento de Datos

Los datos a utilizar deben pasar por un proceso de preprocesamiento. Esto para seguir un estándar en los datos y lograr un mayor desempeño y exactitud a la hora de resolver el problema.

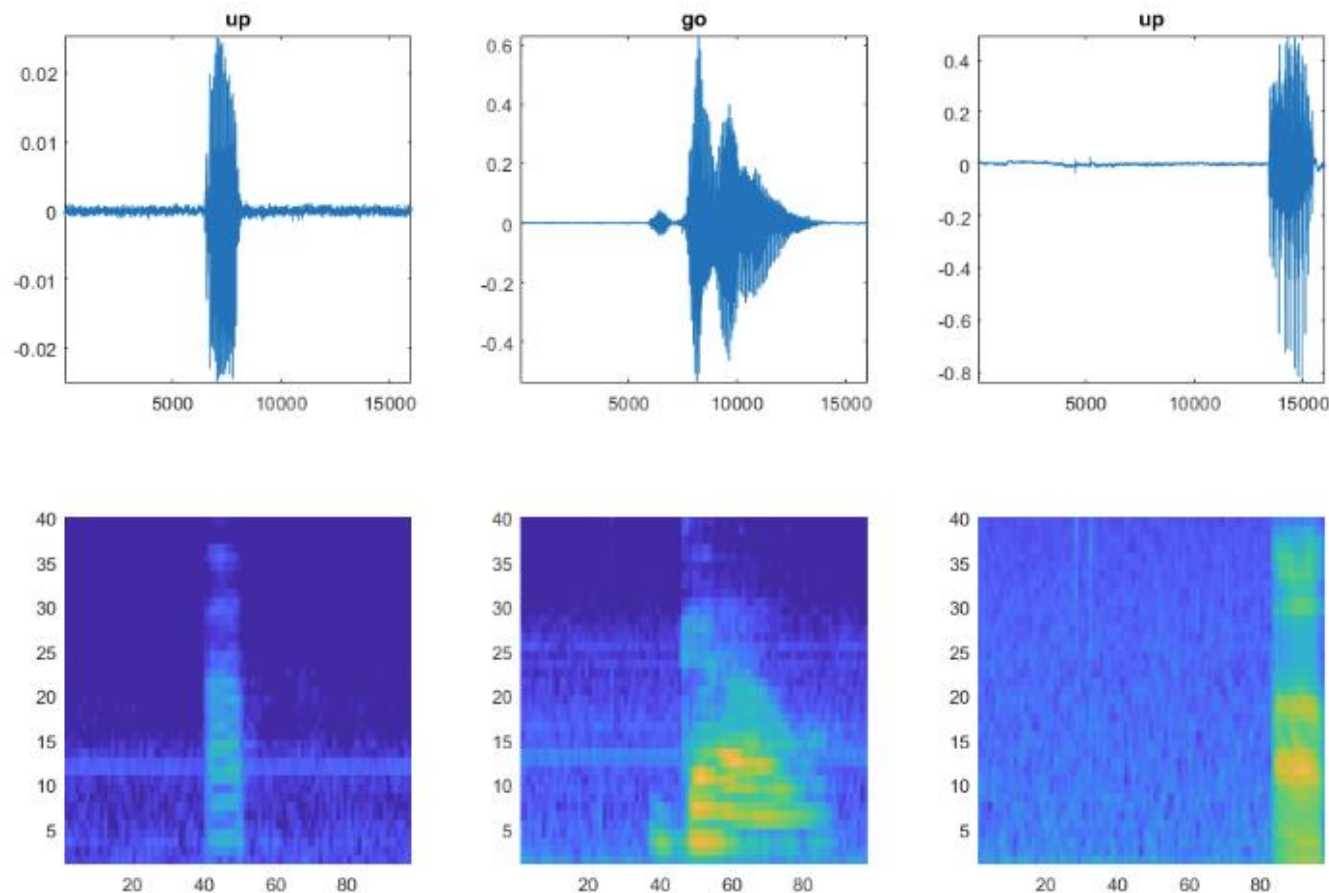
Si los datos no pasan por este proceso, los resultados en las futuras etapas no podrán alcanzar los valores reales de precisión posibles.



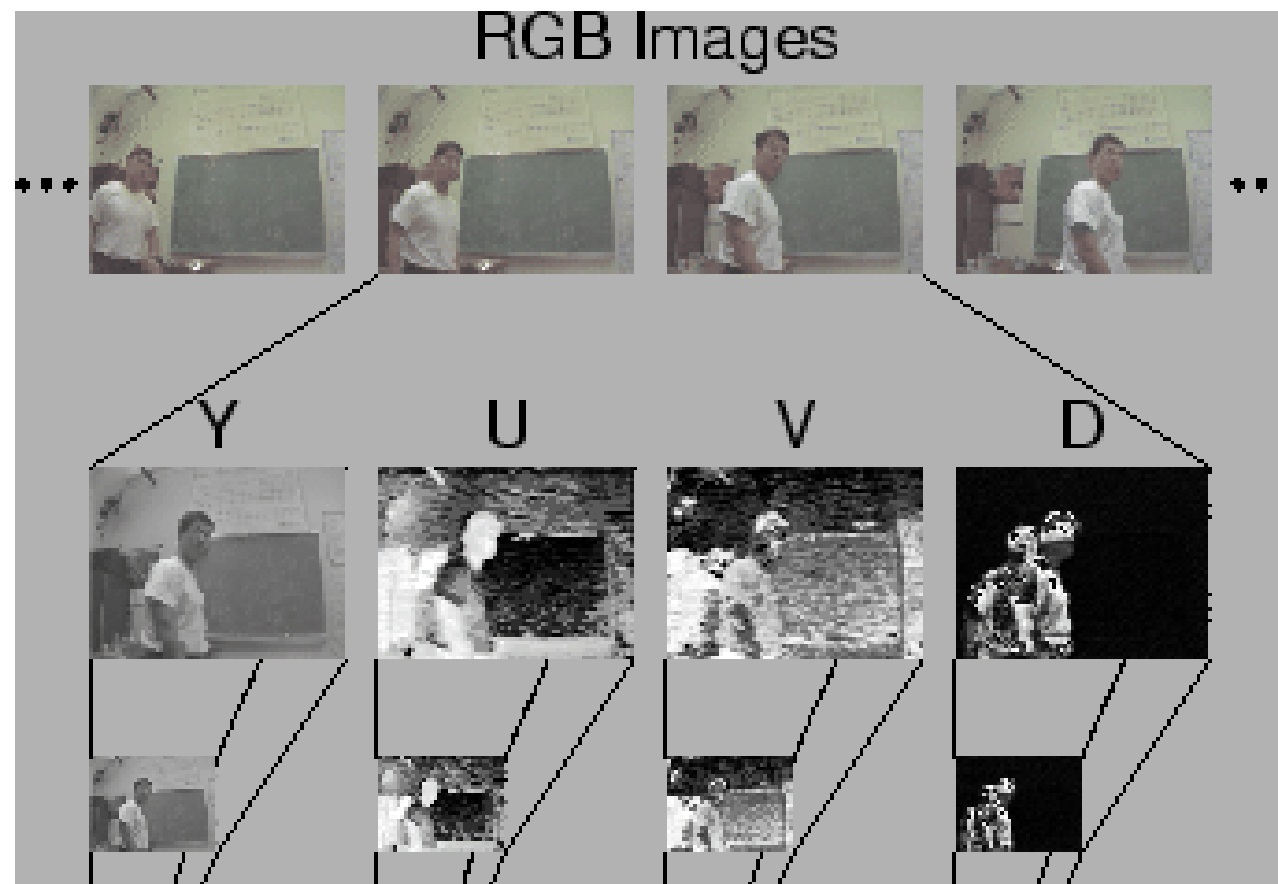
Preprocesamiento de Datos - Texto

Raw	Lowercased
Canada CanadA CANADA	canada
TOMCAT Tomcat toMcat	tomcat

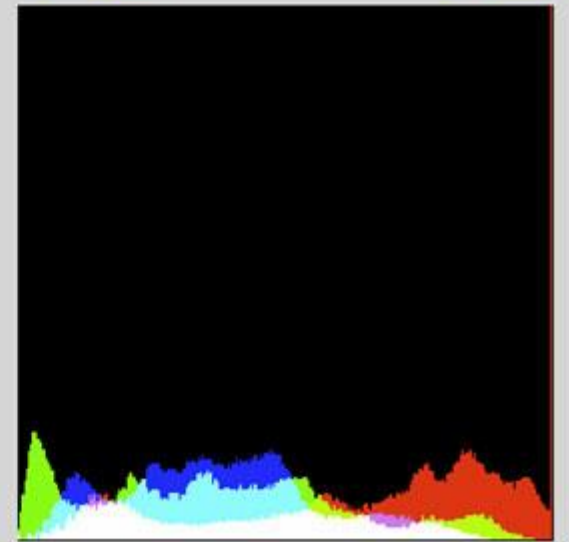
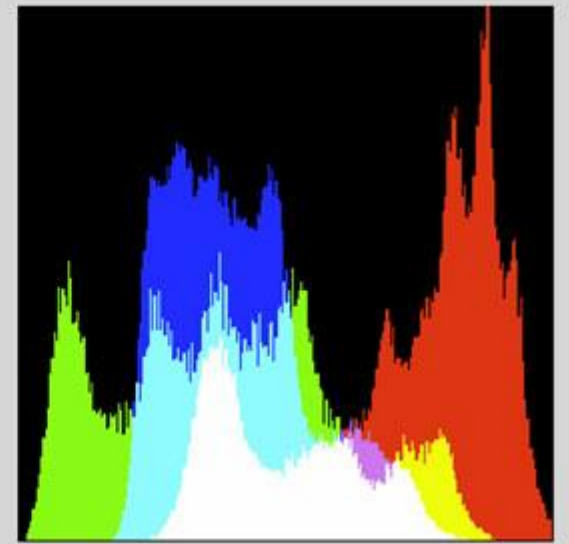
Preprocesamiento de Datos - Audio



Preprocesamiento de Datos - Video



Preprocesamiento de Datos - Imagen



Etiquetado

El etiquetado de datos es una de las tareas más importantes a la hora de extraer información de estos.

Al tener imágenes etiquetadas correctamente, el computador es capaz de aprender a diferenciar entre diferentes clases. Por ejemplo, entre gatos y perros

Machine Learning:

Sample



Label



dog



cat



horse

Human Learning:

We learn through



Examples

Long Ear Black nose



Diagrams

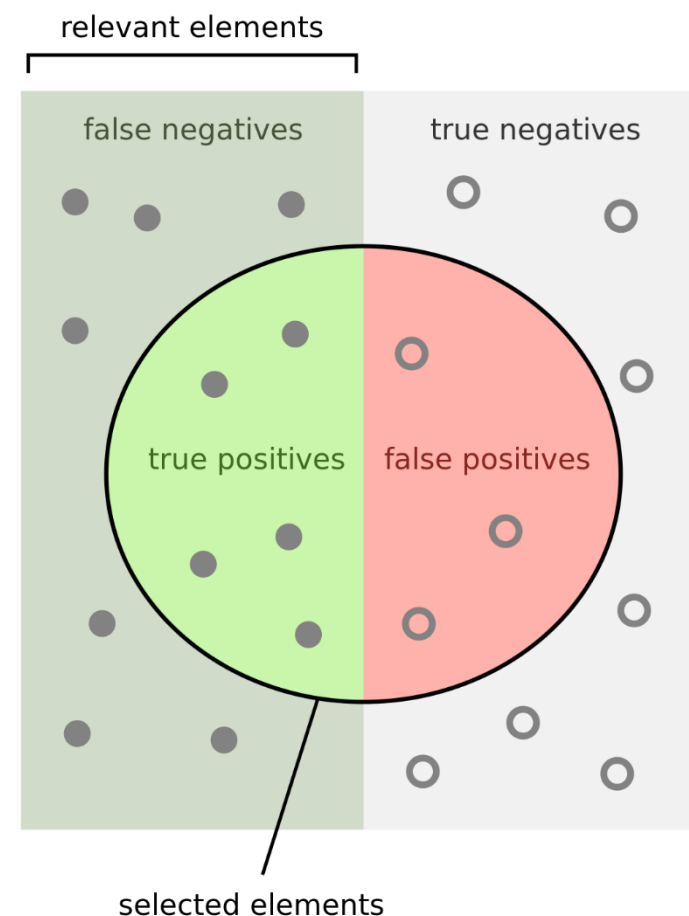


Comparisons

Etiquetado

En imágenes, los datos pre-procesados y útiles (para el problema de clasificación) deben estar compuestos por un ejemplo (la imagen de un tamaño y rango de color específico) y una etiqueta (la clase a la que pertenece tal ejemplo).

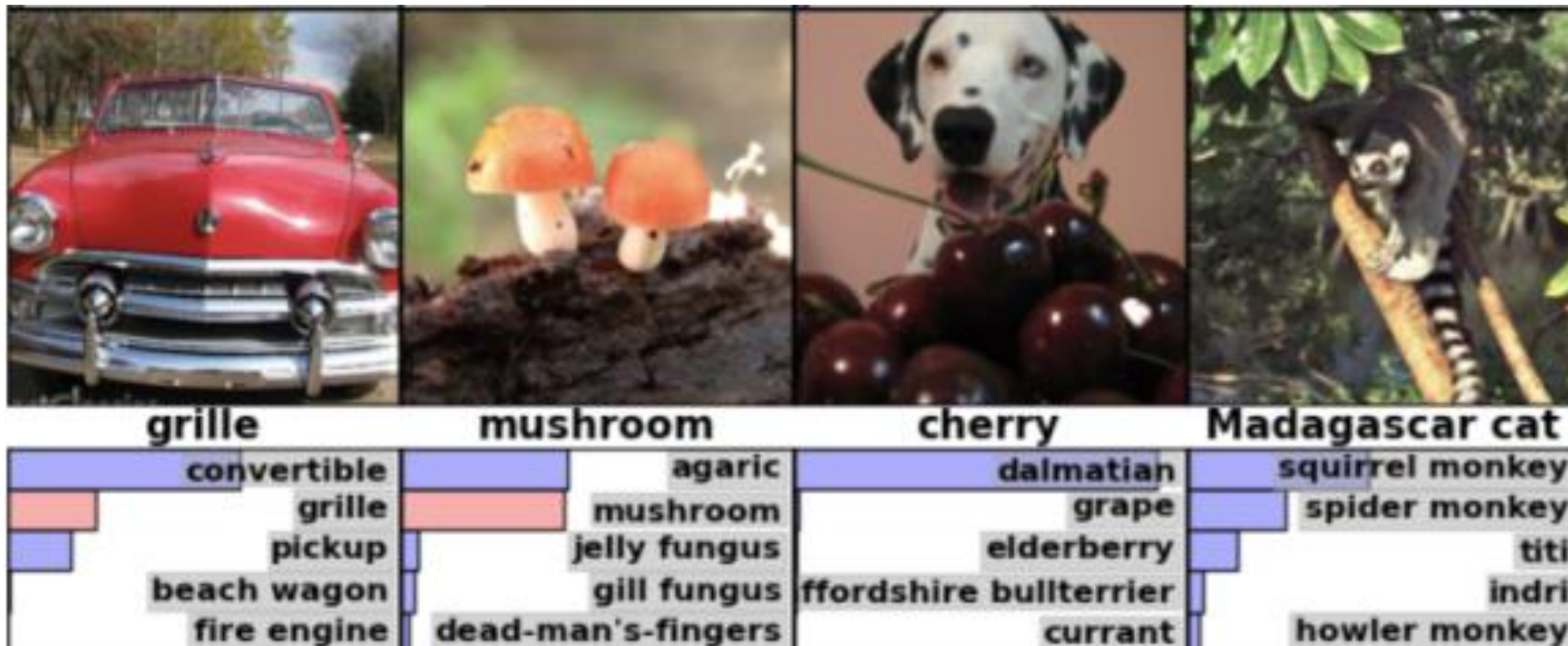
Una etiqueta errada de los ejemplos nos llevaría a tener dato que impediría a nuestro modelo alcanzar su máximo potencial.



Etiquetado



Etiquetado



Etiquetado

En el proceso de clasificación de imágenes es importante que nuestros datos etiquetados cumplan una distribución igual o parecida entre ellos.

Es decir, que nuestras clases tengan el mismo número de ejemplos con la misma desviación estándar en la información de las imágenes.

Image classification

Easiest classes

red fox (100) hen-of-the-woods (100) ibex (100) goldfinch (100) flat-coated retriever (100)



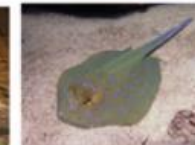
tiger (100)

hamster (100)

porcupine (100)

stingray (100)

Blenheim spaniel (100)



Hardest classes

muzzle (71) hatchet (68) water bottle (68) velvet (68) loupe (66)



hook (66)

spotlight (66)

ladle (65)

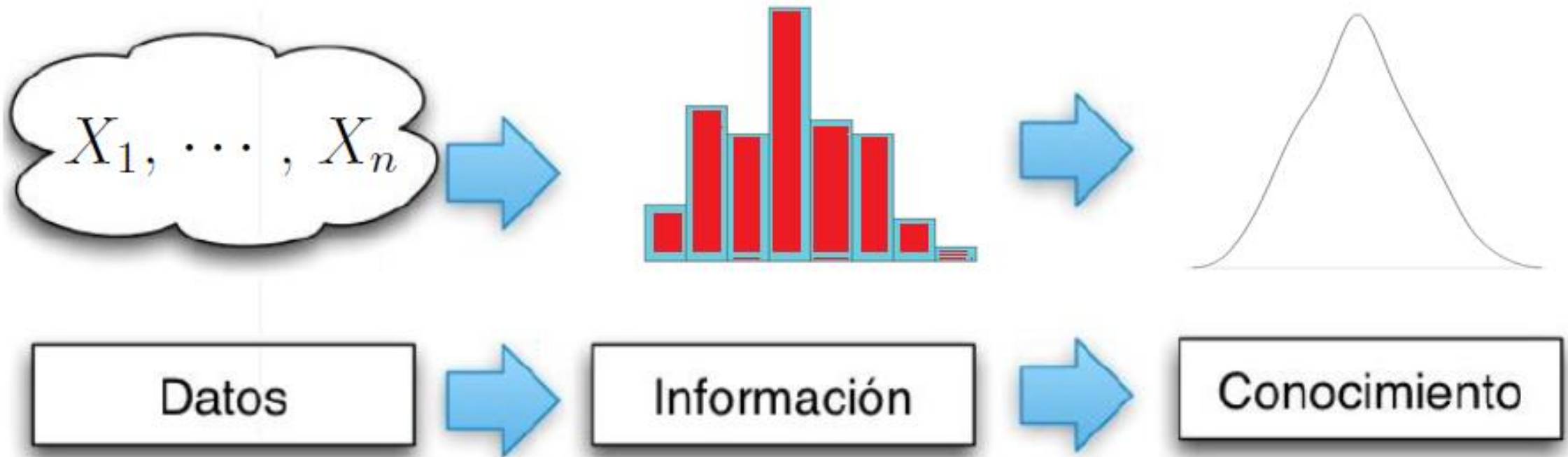
restaurant (64) letter opener (59)



Etiquetado

Si el proceso de etiquetado no se realizó correctamente, este puede traer grandes consecuencias en el resultado final del modelo diseñado.





Datos: Son elementos aislados y en bruto, obtenidos mediante algún proceso de medición, observación o registro, susceptibles de ser transformados para producir información.

Dato Estadístico: Es aquel que se obtiene a través de técnicas, métodos o procedimientos estadísticos. También a través de representaciones numéricas o codificaciones de hechos, cualidades o características.

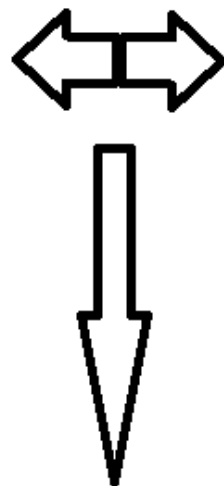
DATO ESTADÍSTICO

MICRODATOS

DATOS ELEMENTALES SOBRE OBJETOS INDIVIDUALES (EMPRESAS, FAMILIAS, PERSONAS, ETC.) OBTENIDOS DIRECTAMENTE DE DONDE SE PRODUCEN. RECOPIRADOS A TRAVÉS DE ENCUESTAS, CENSOS Ó REGISTROS ADMINISTRATIVOS.
SEXO, EDAD, PROFESIÓN, INGRESO MENSUAL

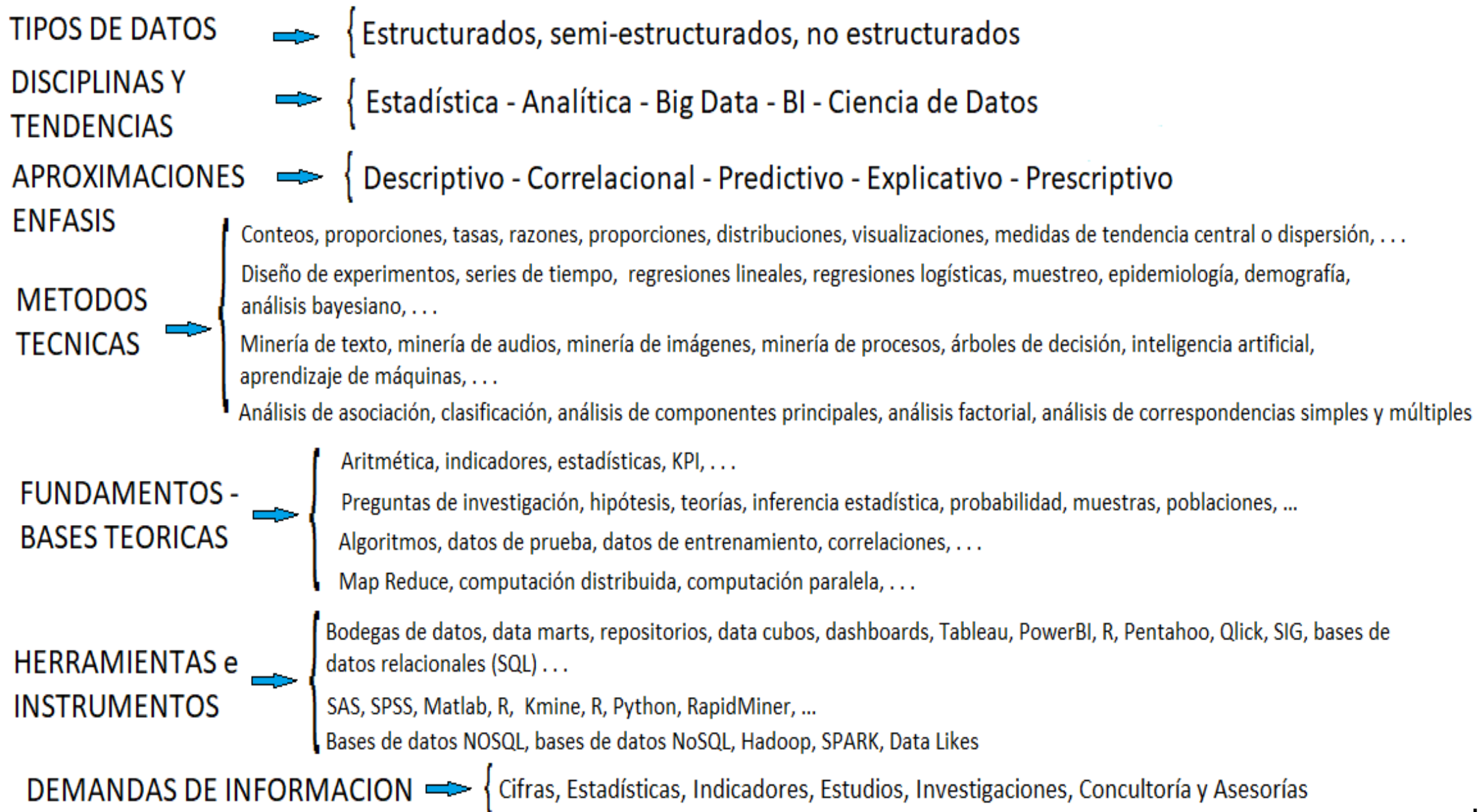
MACRODATOS

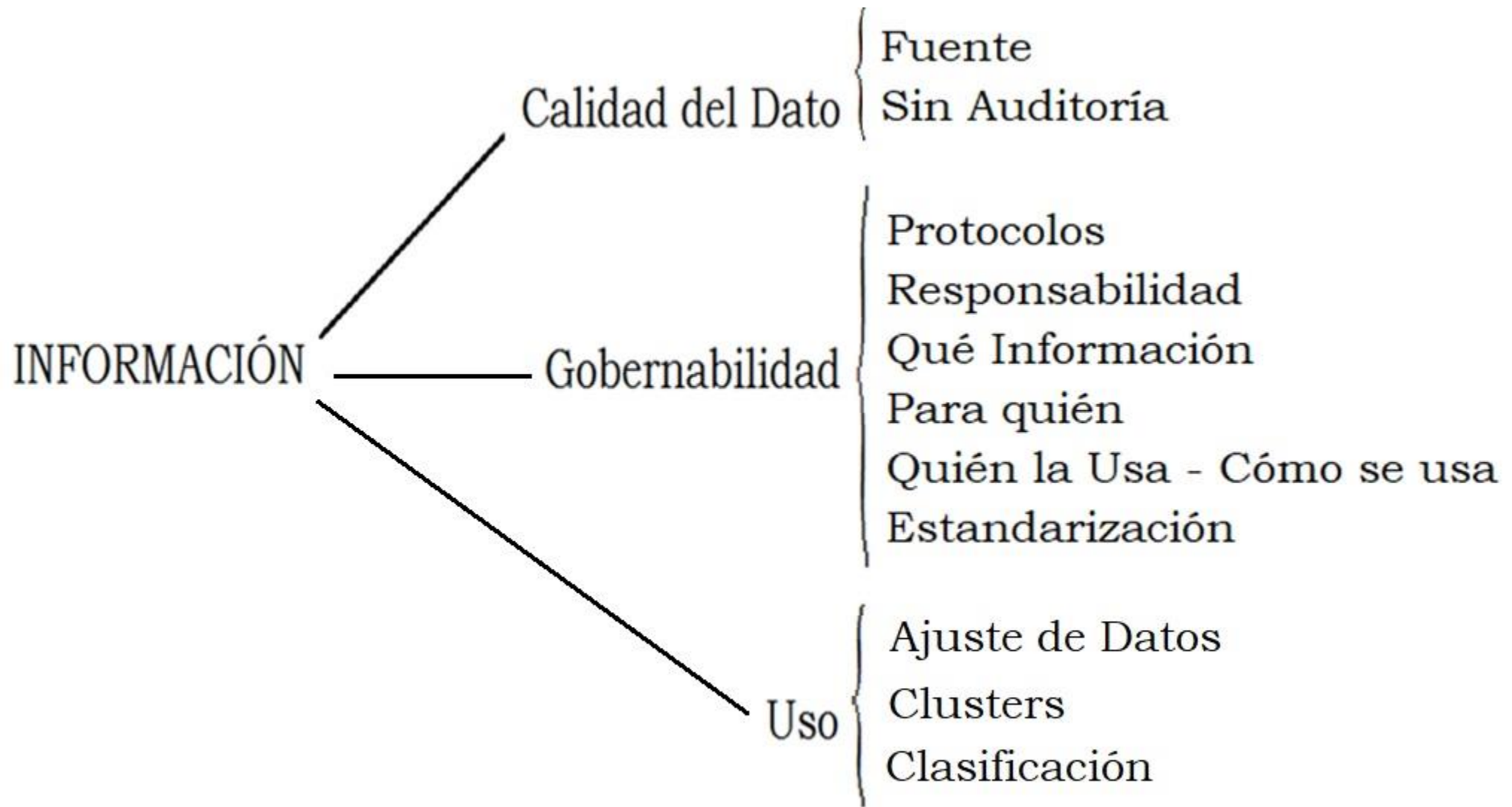
SON LOS DATOS QUE SE OBTIENEN A PARTIR DE LOS MICRODATOS MEDIANTE UN PROCESO DE AGREGACIÓN. DATOS ORIGINADOS POR APLICACIÓN DE MÉTODOS ESTADÍSTICOS TALES COMO: TOTALES, PROMEDIOS, FRECUENCIAS, SOBRE GRUPOS O AGREGACIONES DE



CALIDAD DEL DATO

Datos





Estándares jerárquicos de calidad para los datos:

Disponibilidad.

Accesibilidad: los datos pueden hacerse fácilmente públicos o fáciles de adquirir.

Oportunidad: los datos llegan a tiempo. Los datos se actualizan regularmente.

Usabilidad o Credibilidad

Confiabilidad

Exactitud: Los datos proporcionados son precisos.

Consistencia: Todos los datos son consistentes o verificables.

Integridad: El formato de los datos es claro y cumple los criterios

Compleitud: Una deficiencia de un componente afectará la precisión y la integridad de los datos

Pertinencia.

Conveniencia: Los datos recogidos exponen completamente el tema de interés o parte de él.

Calidad de presentación:

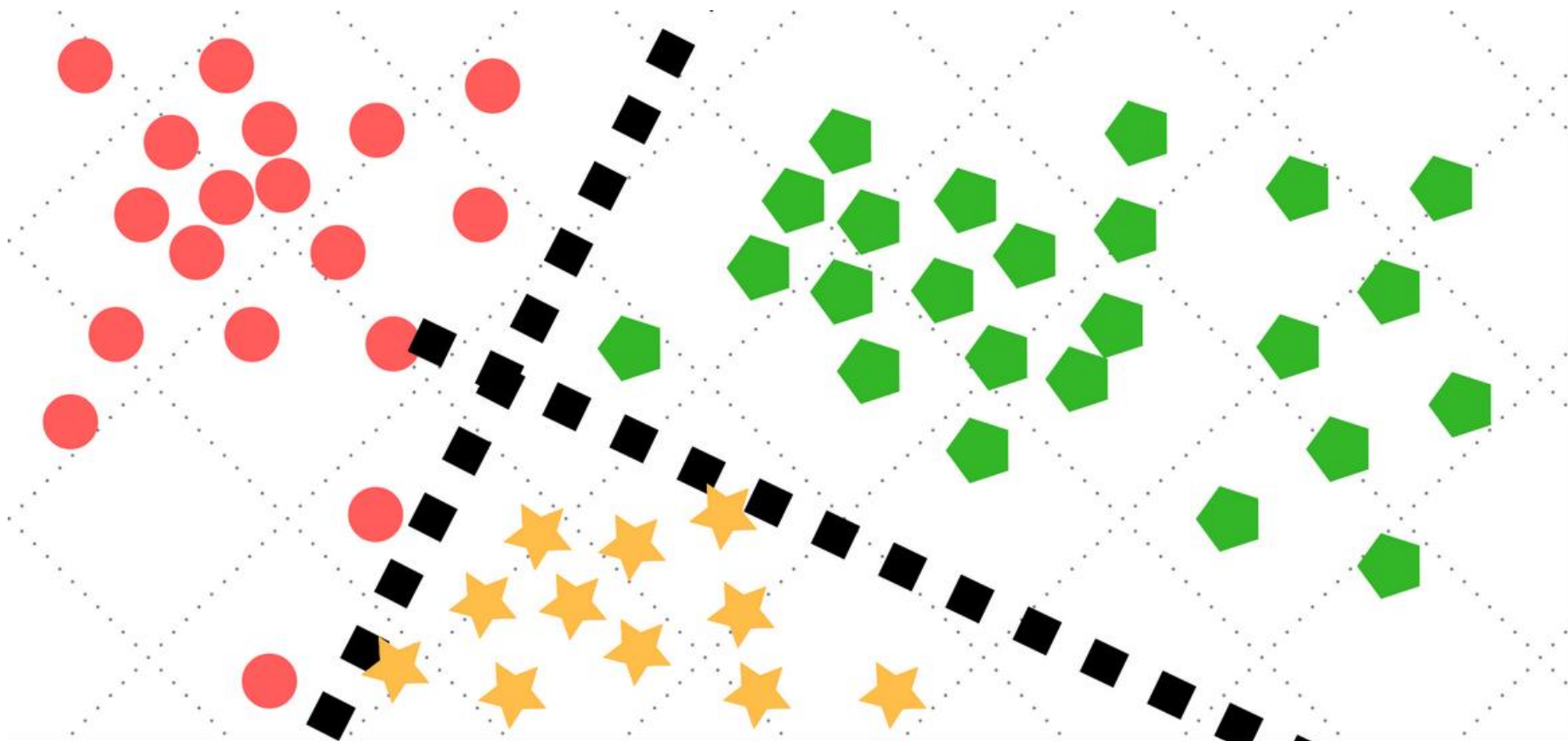
Legibilidad:

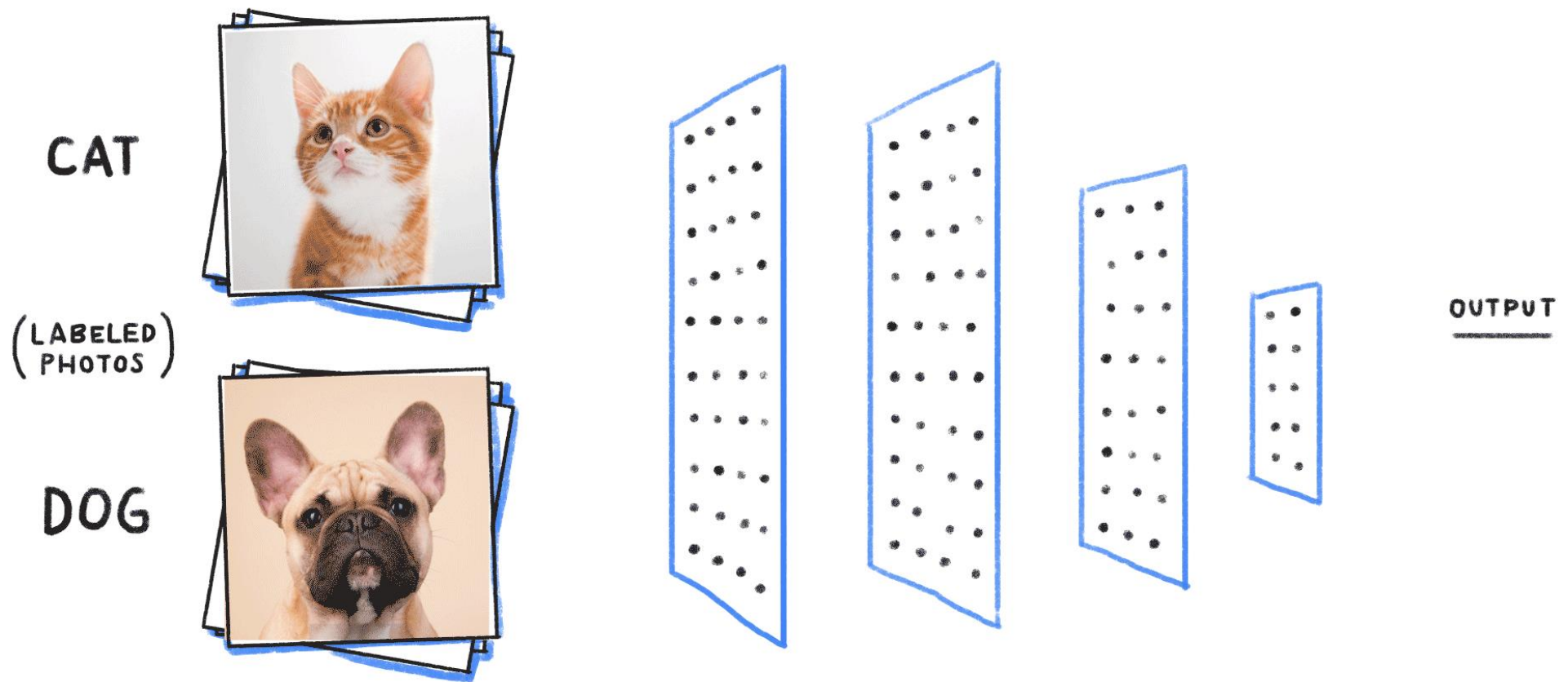
Los datos (contenido, formato, etc.) son claros y comprensibles

La descripción de los datos, la clasificación y el contenido de codificación satisfacen la especificación y son fáciles de entender

Etiquetado

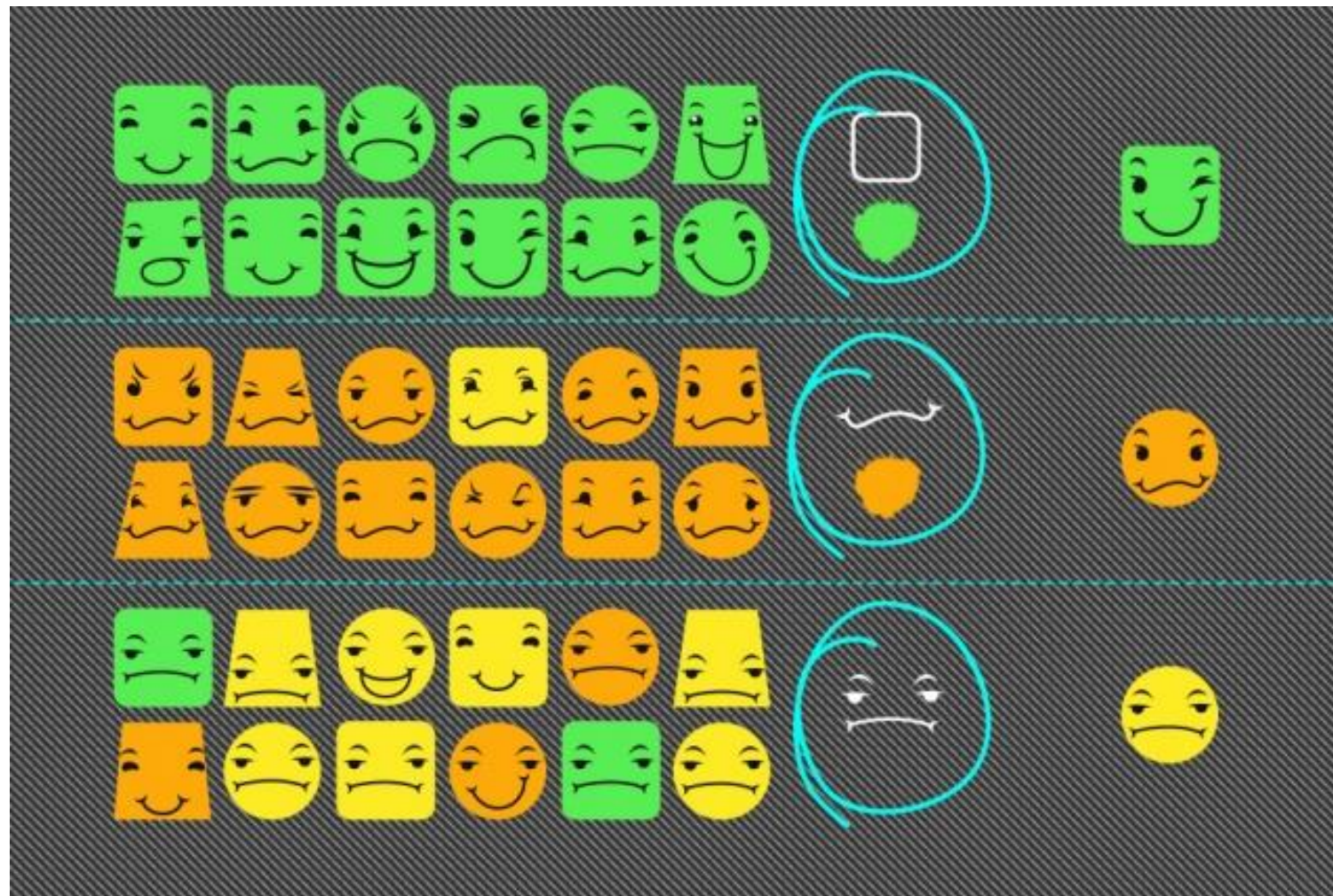
Todo esto, con el fin de realizar tareas como:



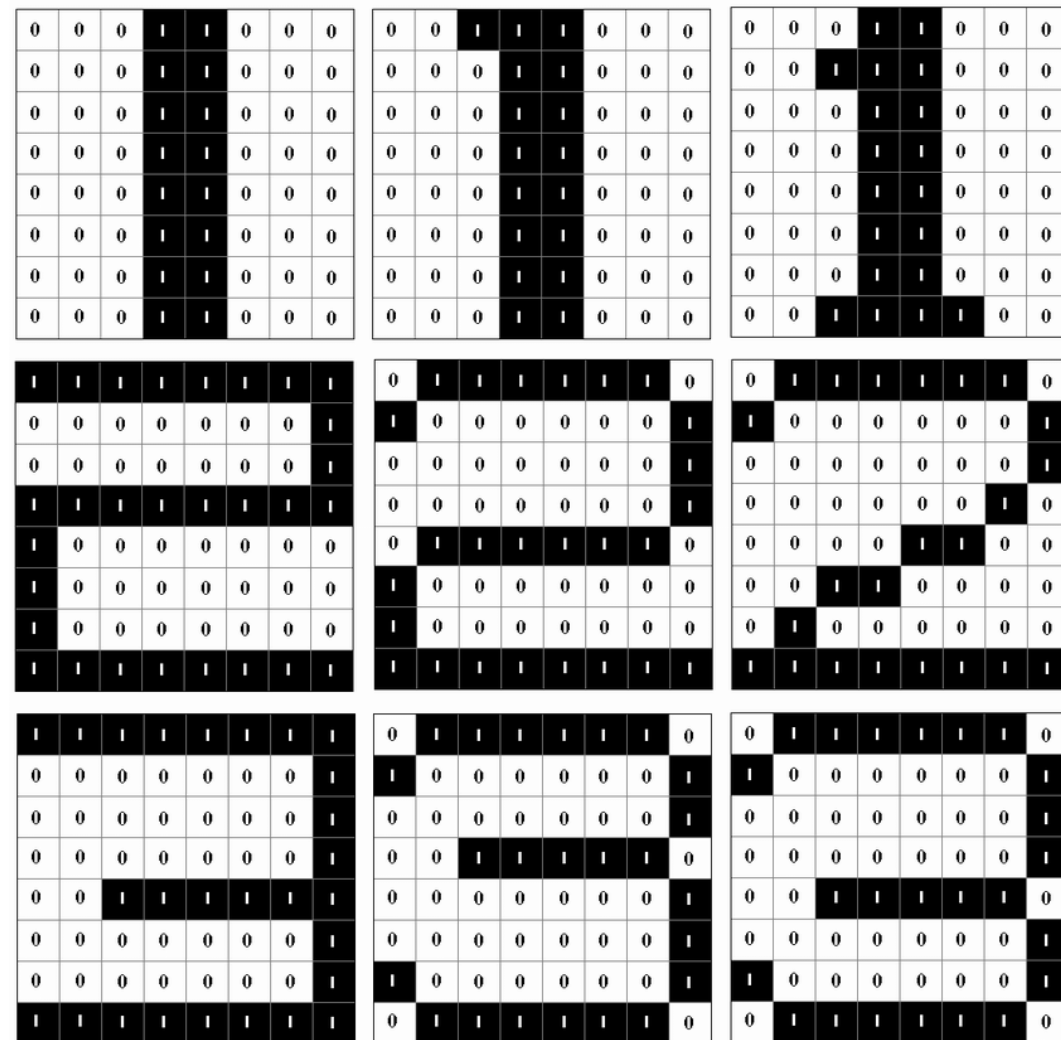
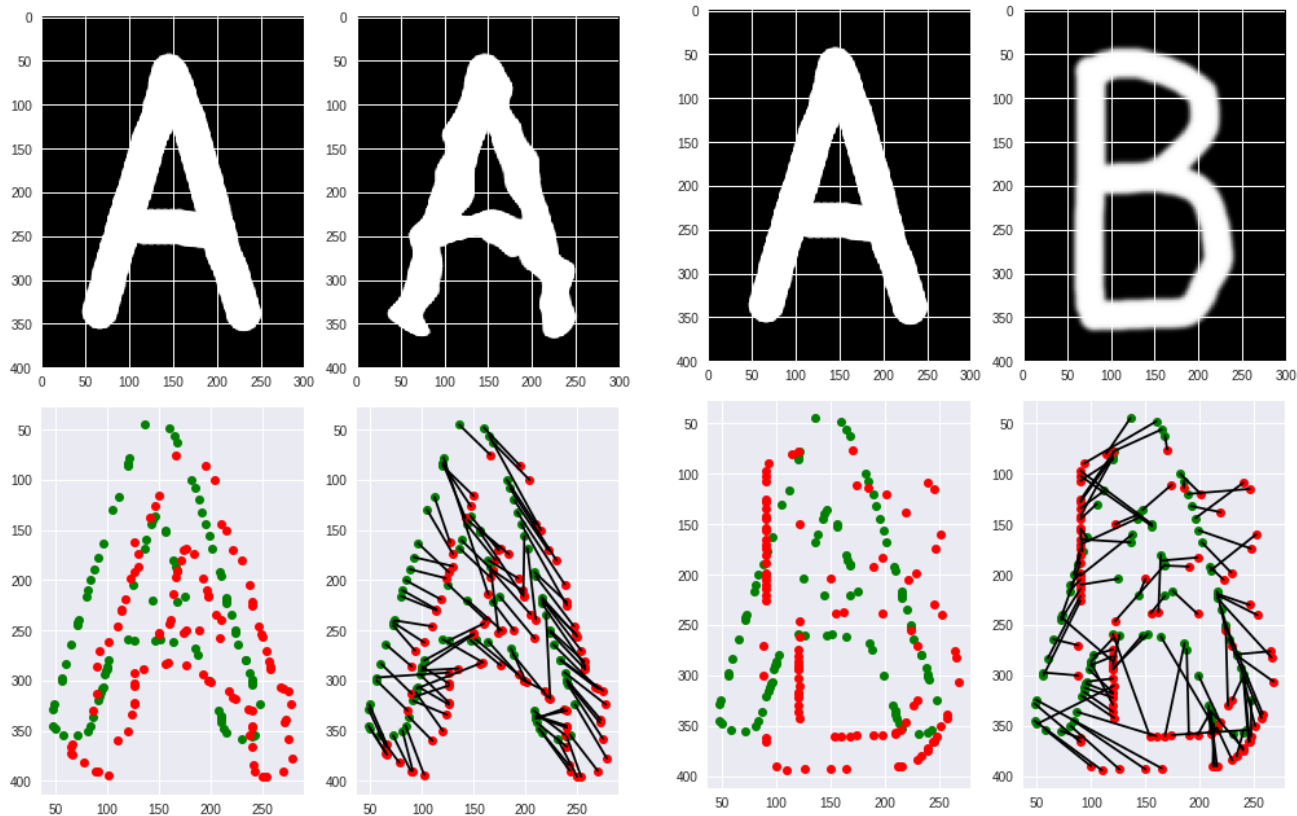


Ejemplos

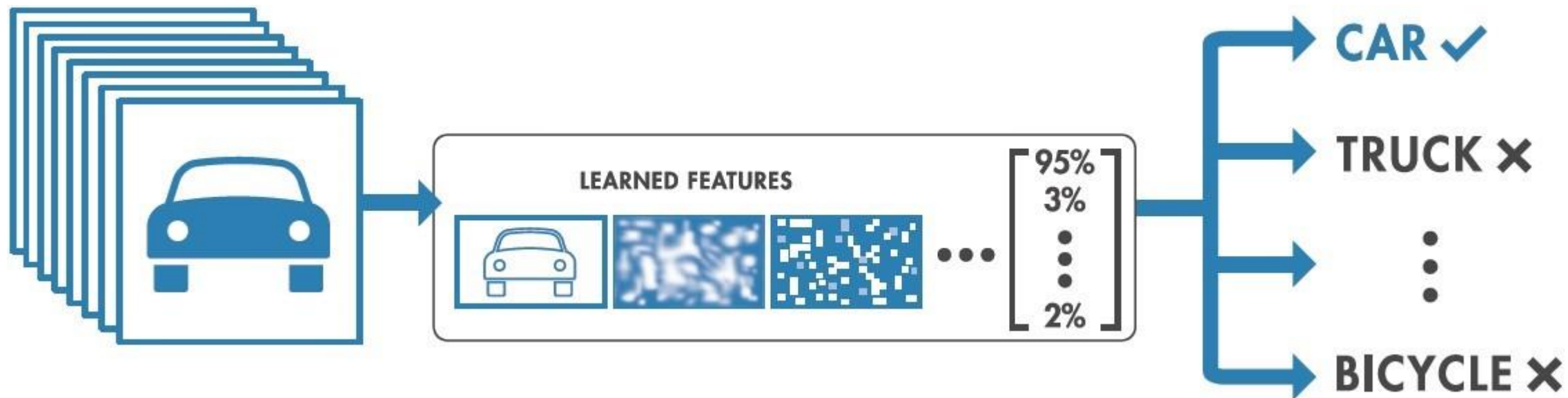
Clasificación y Reconocimiento
de Patrones



Ejemplos



Ejemplos



IMPORTAR

EXPORTAR

DATOS



