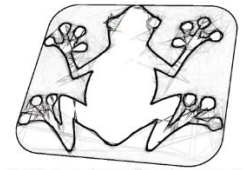


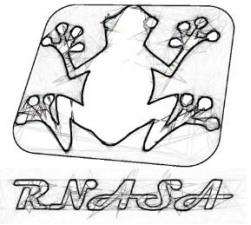
Fundamentos del Big Data

Conceptos Básicos

Grupo de Investigación RNASA-IMEDIR

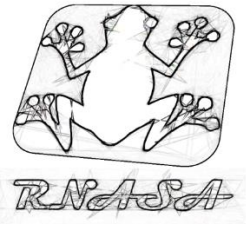
¿Qué es Big Data?

Hechos (en un minuto)



- Email users send more than 204 million messages
- Mobile Web receives 217 new users
- Google receives over 2 million search queries
- YouTube users upload 48 hours of new video
- Facebook users share 684,000 bits of content
- Twitter users send more than 100,000 tweets
- Apple receives around 47,000 application downloads
- Brands receive more than 34,000 Facebook 'likes'
- Instagram users share 3,600 new photos

¿Qué es big data?

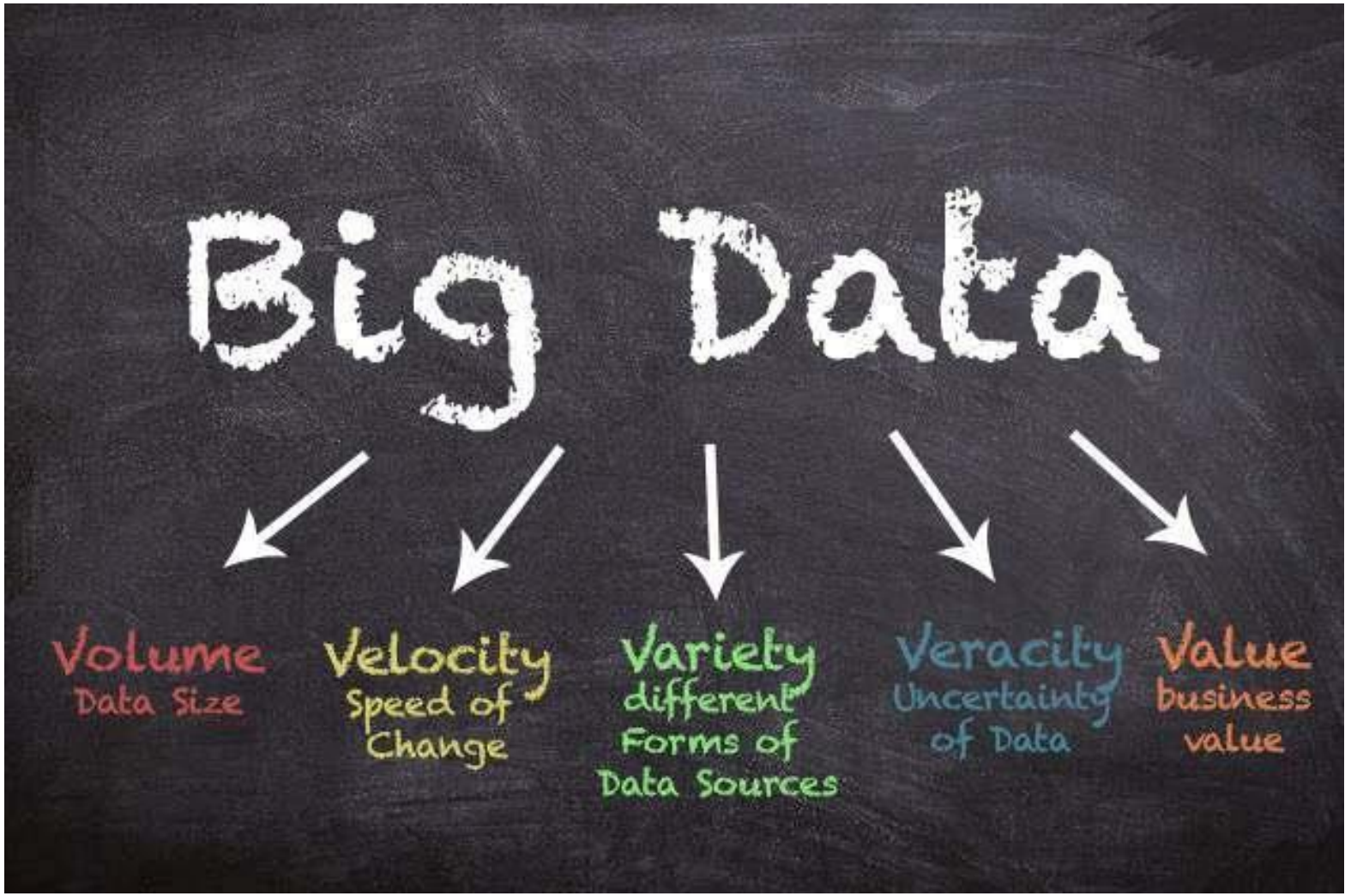


- Volúmenes de datos masivos que no pueden ser tratados mediante técnicas convencionales:

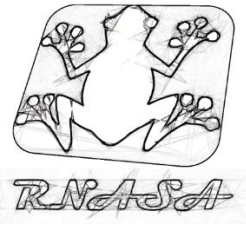
  - Captura
  - Transferencia
  - Almacenamiento
  - Gestión / Mantenimiento / Consulta
  - Análisis (Extracción de conocimiento)
  - Visualización

¿Qué es Big Data?

Características

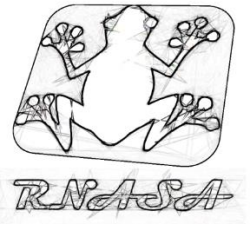


¿Qué es Big Data?
Características



- Las V’s del Big Data:
  - Volumen → Gran cantidad de datos
  - Variedad → Diferente naturaleza
  - Velocidad → Captura / Análisis / Crecimiento
  - Veracidad → Calidad de los datos
  - Valor → Análisis / Extracción de conocimiento

¿Qué es Big Data?
Características



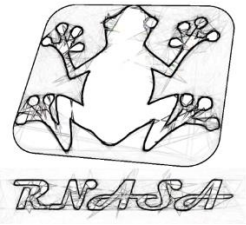
- Volumen

  - Se refiere al gran volumen de información que se maneja.

  - Los datos se acumulan con un crecimiento exponencial, requiriendo ampliar continuamente el almacenamiento de datos.

  - Cuando se habla de bases de datos masivas se refiere a magnitudes del orden de petabytes o exabytes

¿Qué es Big Data?
Características



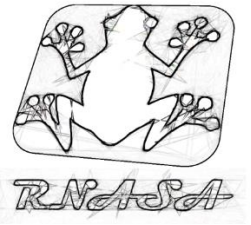
- Variedad

  - Necesidad de agregar información procedente de una amplia variedad de fuentes de información independientes: redes sociales, sensores, máquinas o personas individuales.

  - En general son datos desestructurados, así como gráficos, texto, sonido o imágenes.

  - Estos datos no pueden gestionarse fácilmente con bases de datos relacionales y las herramientas de inteligencia de negocio tradicionales.

¿Qué es Big Data?
Características

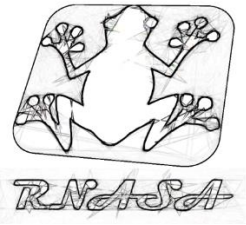


- Velocidad

  - Es la enorme velocidad en la generación, recogida y proceso de la información.

  - Hay que ser capaz de almacenar y procesar en tiempo real millones de datos generados por segundo por fuentes de información tales como sensores, cámaras de videos, redes sociales, blogs, páginas webs,...

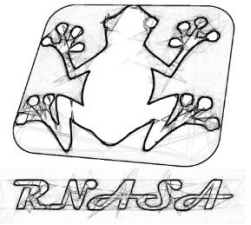
¿Qué es big data?
Características



- Veracidad

  - Se debe analizar inteligentemente un gran volumen de datos con la finalidad de obtener una información verídica y útil que nos permita mejorar nuestra toma de decisiones.

¿Qué es Big Data?
Características



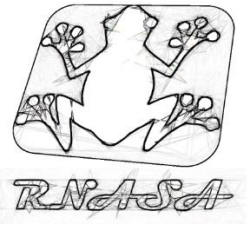
- Valor

  - Es la creación de una ventaja competitiva al identificar y procesar los datos claves, permitiendo así, por ejemplo:

    - Monetizar los datos.
    - Obtener nuevos clientes.
    - Generar fidelidad.
    - Reducir costes.
    - Mejorar la imagen de marca

DIFERENCIAR BIG DATA DE...

BUSSINESS INTELLIGENCE

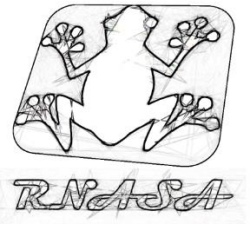


- Big Data se puede entender como una evolución del concepto Business Intelligence.

- En Business Intelligence:
  - Se captura información de una organización y tras un análisis, se obtienen resultados que tienen la finalidad de ayudar a la toma de decisiones estratégicas en la empresa.
  - La información proviene de fuentes de datos estructuradas.
  - Los datos se agrupan en un servidor central y se analizan de forma offline, estructurándose en una base de datos relacional.

DIFERENCIAR BIG DATA DE...

BUSSINESS INTELLIGENCE

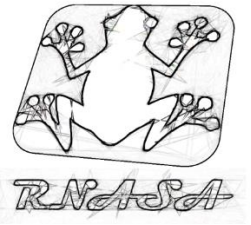


- En Big Data:

  - Se procesa información no estructurada: lenguaje natural, información de redes sociales, información sensores, dispositivos (wearables), etc
  - Los datos se almacenan de forma distribuida que permiten manejar cantidades más grandes de información de forma más ágil.
  - Se emplea procesamiento paralelo masivo de datos, mejorando la velocidad del análisis.

DIFERENCIAR BIG DATA DE...

DATA WAREHOUSE

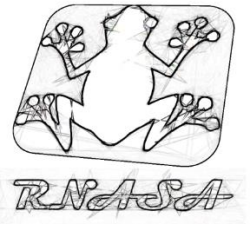


•Un **Data Warehouse** es un almacén de datos, creado con el fin de permitir la toma de decisiones en la organización.

• Una **solución Big Data** es al mismo tiempo almacén de datos y una tecnología de proceso, análisis y visualización de datos.

DIFERENCIAR BIG DATA DE...

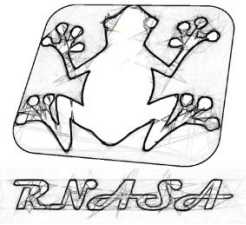
DATA MINING



• Big Data y minería de datos se relacionan por el uso de grandes conjuntos de datos para su procesamiento y análisis, sin embargo, divergen en su operativa.

• La minería de datos busca información concreta accediendo a partes pequeñas y específicas de los datos dentro de esos grandes conjuntos.

Análisis en el Big Data

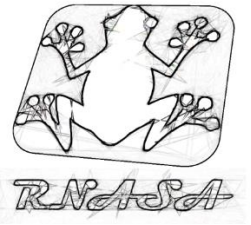


• La cuestión clave no es tener la capacidad para recolectar y almacenar una gran cantidad de datos.

• Con la acumulación de los datos no sé alcanzan ventajas competitivas: es necesario saber organizarlos, refinarlos, y convertirlos en información relevante que permita ganar posiciones en el mercado.

• Los datos tienen sólo valor potencial, es su análisis y sistematización el que permite incrementar la capacidad de innovar y obtener ventajas en las organizaciones.

Análisis en el big data



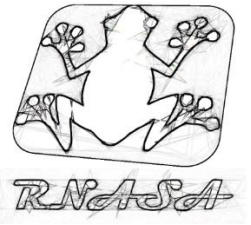
• Estas afirmaciones, nos lleva a hacer preguntas y encontrar respuestas, para la empresa y la sociedad.

• La gestión correcta de los datos genera una conciencia en las administraciones, empresas y ciudadanos, que los datos y su análisis son un activo de las sociedades modernas.

• Es evidente que se tienen que desarrollar nuevos perfiles para cuidar y sacar el máximo de esos activos.

PERFILES PROFESIONALES

LOS 7 PERFILES PROFESIONALES DEL BIG DATA

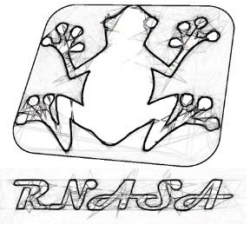


• **Chief Data Officer** (CDO): es el responsable de asegurar que la organización es data driven. Lidera la gestión de datos y analítica asociada por el negocio y, por tanto, es responsable de los diferentes equipos especialidades en datos.

• **Data Scientists** (científico de los datos): son los miembros clave del equipo de ciencia de datos. Permiten extraer conocimiento e información valiosa de los datos. Tienen visión general del proceso de extremo a extremo y pueden resolver problemas de ciencias datos, la construcción de modelos analíticos y algoritmos.

PERFILES PROFESIONALES

LOS 7 PERFILES PROFESIONALES DEL BIG DATA

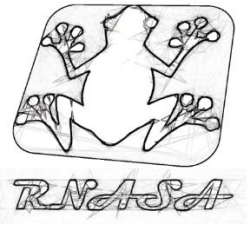


• **Citizen Data Scientist**: puede extraer valor, a través de su experiencia, explorando los datos, desde las unidades de negocio. Pueden ejecutar una serie simple de tareas analíticas utilizando herramientas de descubrimiento de datos.

• **Data Engineer** (ingeniero de datos): Se encarga de proporcionar los datos de una manera accesible y apropiada a los usuarios y Data scientists. Es un perfil especializado en infraestructura big data. Desarrolla y explota técnicas, procesos, herramientas y métodos que deben servir para el desarrollo de aplicaciones big data.

PERFILES PROFESIONALES

LOS 7 PERFILES PROFESIONALES DEL BIG DATA



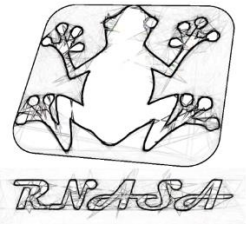
• **Data Steward** (administrador de datos): es responsable de mantener la calidad, disponibilidad y seguridad de los datos.

• **Business Data Analyst** (analista de datos): participa en las iniciativas y proyectos de análisis de datos. Es la persona que recoge las necesidades de los usuarios de negocio para los Data Scientist y presenta resultados obtenidos.

• **Data Artist:** son los responsables de crear los gráficos, infografías y otras herramientas visuales para ayudar a las diferentes personas a comprender datos complejos.

NECESIDADES

NUEVA INFRAESTRUCTURA TECNOLÓGICA



• La capacidad de aportar escalamiento de procesamiento masivo permite la identificación continua de información útil sepultada dentro de Big data.

• Integrar metodologías y tecnología para el descubrimiento y entendimiento de información basado en fuentes altamente escalables. Por ejemplo, Open Data, Linked Data, Social Data, Sentiment Analysis, Online Stream Analysis, Web Intelligence.

NECESIDADES

NUEVA INFRAESTRUCTURA TECNOLOGICA

Para ello es necesario que:

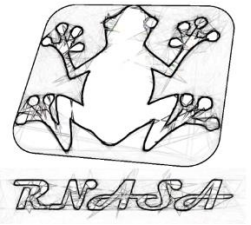
- Sea escalable de forma masiva a petabytes de datos (en la actualidad).

- Soporte y acceso a datos de baja latencia y toma de decisiones

- Tenga análisis integrado para acelerar el modelado de análisis avanzado y de los procesos.

NECESIDADES

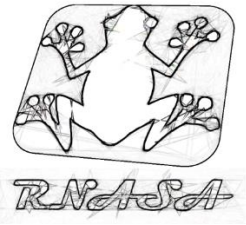
NUEVA INFRAESTRUCTURA TECNOLÓGICA



• Identificar las oportunidades de transformación y generación de valor basadas en el análisis de los datos, proveniente tanto de fuentes internas como externas a la organización.

• Desarrollar soluciones que permitan generar valor añadido y diferenciación a partir de los procesos de análisis de información sobre Big Data.

Riesgos en el Análisis de datos



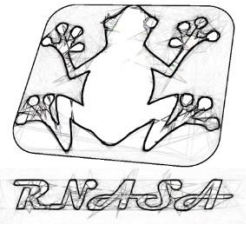
• Uno de los riesgos que presenta la búsqueda de información en el Big Data, es el descubrimiento de patrones no significativos.

• Estos patrones no relevantes se conocen en la estadística como principios de Bonferroni.

• Una gran cantidad datos como los que se analizan en los entornos Big Data, pueden “validar” cualquier patrón.

DATA MINING

ANÁLISIS DE LA INFORMACIÓN

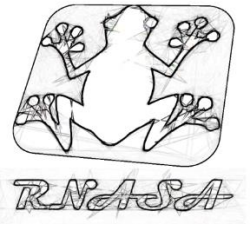


- Técnicas interdisciplinares (Inteligencia Artificial / Machine Learning) en entornos de procesamiento distribuido:

  - Redes de Neuronas Artificiales (ANN / DL)
  - Árboles de Decisión (DT)
  - Regresión (Simple / Multiple)
  - Redes Bayesianas
  - Support Vector Machines (SVM)

Tipos de Análisis

Modelos predictivos



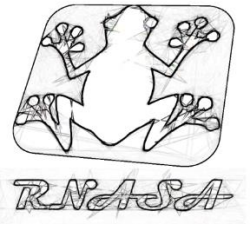
- **Evalúan qué probabilidad tiene un individuo de mostrar un comportamiento específico en el futuro.**

- **Buscan patrones discriminadores** en los datos para responder comportamiento.

- **Realizan cálculos en tiempo real** para evaluar un determinado riesgo u oportunidad, a fin de orientar una decisión adecuada.

TIPOS DE ANÁLISIS

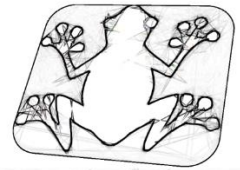
MODELOS PREDICTIVOS



• Describen las relaciones entre los datos para poder clasificar a los individuos en grupos.

• Identifican diferentes relaciones entre individuos que pueden ser utilizadas para predecir también acciones futuras.

• Describen la relación entre todos los elementos de una decisión, la decisión a tomar y las variables y valores que determinan la propia decisión, con la finalidad de predecir los resultados mediante el análisis de muchas variables.

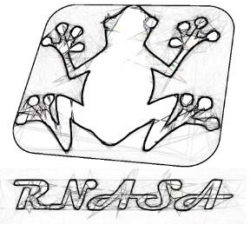


El dato como activo de valor

Fundamentos de Big Data

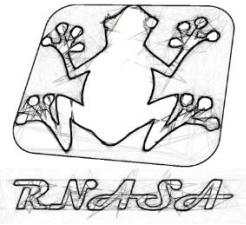
Grupo de Investigación RNASA-IMEDIR

EL DATO COMO ACTIVO DE VALOR



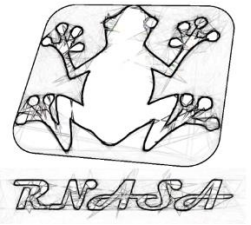
• Vivimos una época de profunda transformación de las organizaciones, fundamentada en tecnología (como cloud computing, impresión 3D, etc.), en la que la gran mayoría de sus procesos de negocio y su cadena de valor están siendo revisados e interpretados de nuevo.

El dato como activo de valor



• Esta transformación tiene distintos efectos. Entre los mismos, destaca la generación de datos como subproducto o como intención principal de la transformación. Cuando, de forma consciente, la organización gestiona el dato, es capaz de tomar decisiones más eficaces y eficientes, y de competir en el mercado de una forma distinta.

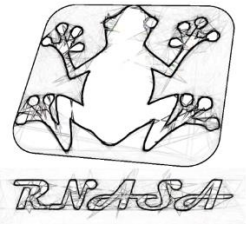
El dato como activo de valor



• Es destacable, por ejemplo, el caso de Netflix, que usa los datos de cliente (su comportamiento y preferencias) combinados con algoritmos para evitar el abandono de clientes. De hecho, su conocimiento y la eficiencia de sus algoritmos se incrementan a medida que la compañía tiene más clientes, lo que genera un efecto de red.

• Hasta tal punto es importante esta iniciativa que, según Netflix, su impacto en el negocio es de un billón de dólares.

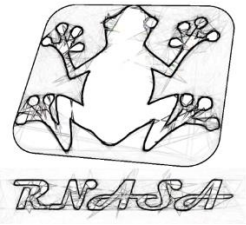
EL DATO COMO ACTIVO DE VALOR



• Estas organizaciones han cambiado su percepción respecto al dato. Han pasado de considerar el dato como un activo tóxico, cuyo uso es necesario controlar y limitar, a un activo de valor, cuyo uso hay que gestionar y maximizar.

• Es decir, aquellas empresas que se están transformando en organizaciones orientadas al dato contemplan el dato como un recurso que genera **ventajas competitivas**.

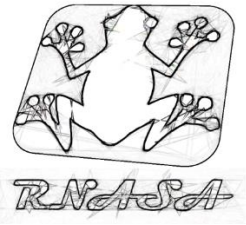
EL DATO COMO ACTIVO DE VALOR



• El foco del dato no solo está en la eficiencia, como en el caso de Netflix. Ian Davis, analista de McKinsey y gurú del Big Data, postula que las organizaciones deben ser ágiles para adaptarse a los cambios del mercado. Y defiende el rol del dato como fuente de la agilidad y el aprendizaje continuo. Para que una empresa sea flexible ante el mercado, ha de poder adaptar sus decisiones de forma continua y, por lo tanto, será tan flexible como lo sean sus datos y sus capacidades de análisis.

El dato como activo de valor

Características de los datos

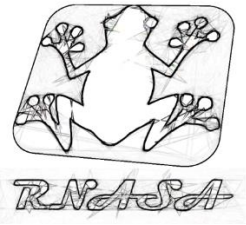


Desde la perspectiva de un activo, el dato tiene sus peculiaridades si lo comparamos con otro tipo de activos:

- Los datos frecuentemente tienen una naturaleza transitoria, es decir, están vinculados a un tiempo y un momento. Por ejemplo, los valores bursátiles, como activo, solo tienen sentido si son capturados y analizados de forma continua. Esto es muy diferente, por ejemplo, en el caso de un edificio, cuyo valor presenta una duración más amplia.

EL DATO COMO ACTIVO DE VALOR

CARACTERÍSTICAS DE LOS DATOS

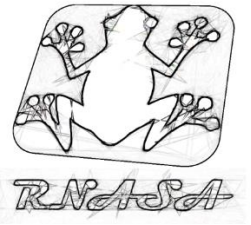


• El dato requiere un mantenimiento continuo, es decir, es necesario manipularlo para mantener o incrementar su valor.

• El dato puede reusarse, es decir, a diferencia de otros activos que con el uso se desgastan, como por ejemplo una pelota, el dato tiene la posibilidad de ser usado en diferentes escenarios, a un coste marginal.

• El dato posee un gran potencial como activo, puesto que tiene una baja o nula trasparencia, transferencia y replicación, y una larga duración.

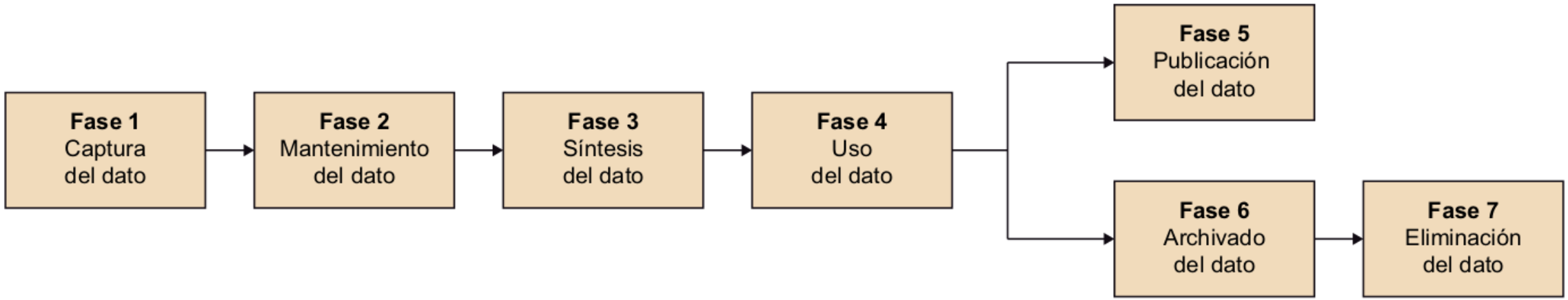
Ciclo de vida del dato



• En el momento en el que una organización identifica el dato como un activo de valor, el siguiente paso es la gestión de dicho activo de forma precisa. Para ello, es necesario poder conocer y asociar a cada dato lo que definimos como el ciclo de vida de un activo:

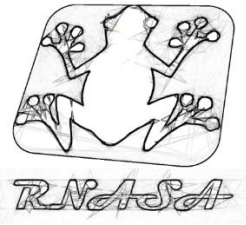
• **Entendemos como ciclo de vida de un activo las diferentes etapas por las que pasa un activo desde su nacimiento hasta el fin.**

Ciclo de vida del dato



Ciclo de vida del dato

Fases



• **Fase 1. Captura del dato**. Esta fase puede considerarse como el acto de crear datos que no existen aún en la organización y que nunca han existido en la misma. Encontramos distintos métodos para la captura de datos, entre los que destacan:

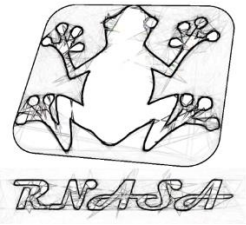
1) Adquisición de datos: la ingesta de datos que han sido creados y existen fuera de la organización.

2) Introducción de datos: la creación de nuevos datos en la organización por personal humano, o generados mediante dispositivos dentro de la propia organización.

3) Recepción de señales: la captura de datos creados por dispositivos, normalmente en sistemas de control, pero cada vez más importantes para los sistemas de información como el Internet de las cosas.

Ciclo de vida del dato

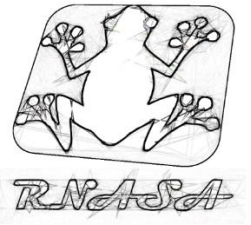
Fases



• **Fase 2. Mantenimiento del dato**. El objetivo de esta fase es procesar el dato, pero sin que se genere aún un valor claro para la organización. Este procesamiento incluye tareas de integración, limpieza, enriquecimiento, así como procesos de extracción, transformación y carga del dato (conocidos en inglés como ETL - extract, transform, and load). Debido a la diversidad de actividades en esta fase, existen numerosos retos asociados, como, por ejemplo, cómo mejorar el proceso de envío del dato al destino final para su síntesis y uso, previniendo que se genere un elevado número de movimientos de datos durante todo el procesamiento, de inicio a fin.

Ciclo de vida del dato

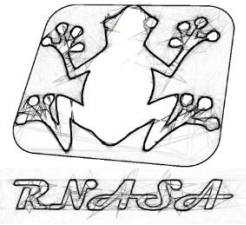
Fases



• **Fase 3. Síntesis del dato.** Esta fase consiste en la creación de datos de valor aplicando un procesamiento o lógica inductiva determinada, usando otros datos como fuente. Esta es el área del procesamiento analítico donde se usa el modelado de datos, como, por ejemplo, el modelo de riesgos de una organización. La lógica inductiva requiere algún tipo de experiencia o conocimiento como parte de la lógica de negocio, como por ejemplo la forma en la que se crean los informes de créditos bancarios.

Ciclo de vida del dato

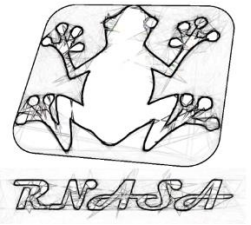
Fases



• **Fase 4. Uso del dato**. Una vez que el dato ha sido capturado y transformado dentro de la organización y se ha usado como fuente en la fase anterior, el dato se usa para beneficio de la propia organización, en tareas que la organización utiliza y gestiona. Aunque normalmente se trata de tareas que no pertenecen al ciclo de la vida del dato, el dato es cada vez más considerado como un activo fundamental en los modelos de negocios de muchas organizaciones. En esta fase, también existen retos importantes, como por ejemplo el uso permitido del dato, o lo que es lo mismo, si es legal o no el uso del dato de la manera en la que los usuarios de negocio pretenden usarlo.

Ciclo de vida del dato

Fases



• **Fase 5. Publicación del dato**. Esta fase puede ser definida como el envío del dato. Este envío puede ser interno (a una intranet) o externo (a un lugar fuera de la organización, gestionado por terceros). Un ejemplo es una agencia de inversión que envía informes mensuales a sus clientes. Una vez que el dato ha sido enviado fuera de la organización, es imposible recuperarlo para una posterior modificación. El gobierno del dato ayuda a decidir cómo actuar con los datos incorrectos o incompletos que han sido enviados fuera de la organización. Los accesos no autorizados a los datos también se incluirían en esta fase.

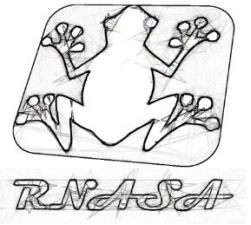
Ciclo de vida del dato

Fases

• **Fase 6. Archivado de datos.** Esta fase consiste en copiar los datos en un entorno donde son almacenados, en caso de que se vuelvan a necesitar en el futuro en un entorno activo de producción, y la completa eliminación de estos datos en todos los entornos activos. Un archivo de datos es simplemente un almacenamiento de datos, pero no de mantenimiento, uso o publicación de datos. En caso necesario, los datos pueden ser recuperados en un entorno donde es posible llevar a cabo cualquiera de estas actividades.

Ciclo de vida del dato

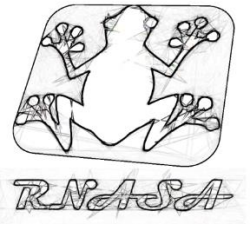
Fases



• **Fase 7. Eliminación del dato**. Esta es la fase final del ciclo de vida del dato, y consiste en la destrucción de cualquier copia del dato que exista dentro de la organización. Idealmente, este proceso será llevado a cabo mediante un archivado de datos. Un reto, en esta fase, sería verificar que la eliminación se ha llevado a cabo de manera satisfactoria.

CICLO DE VIDA DEL DATO

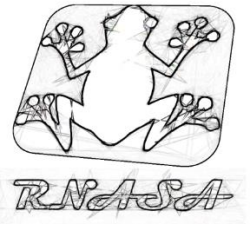
CAMBIOS EN LOS DATOS



• El dato no es un activo estático durante su ciclo de vida. Dentro de las organizaciones, se crean nuevas fuentes de datos continuamente, y es necesario mantener un registro del dato a la vez que se mueve a través de distintos sistemas dentro de la organización. Para ello, necesitamos establecer lo que se conoce como trazabilidad del dato o linaje del dato.

CICLO DE VIDA DEL DATO

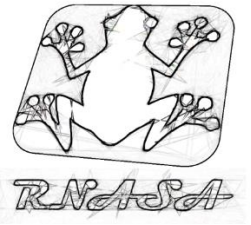
CAMBIOS EN LOS DATOS



• Se entiende como **linaje del dato** la capacidad de conocer todo el ciclo de vida de un dato, desde la fecha y hora exacta en la que fue extraído, el momento en que se produjo su transformación, y hasta el instante en que tuvo lugar su carga desde un entorno fuente (servidor, fichero, tabla, campo, etc.) a otro de destino. En inglés, data lineage.

CICLO DE VIDA DEL DATO

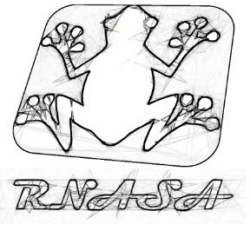
CAMBIOS EN LOS DATOS



• Mediante el linaje del dato, las organizaciones pueden obtener una mejor comprensión de lo que sucede con el dato, lo que posibilita el rastreo e identificación de errores y, así, aplicar protocolos más rigurosos de gobierno del dato.

CICLO DE VIDA DEL DATO

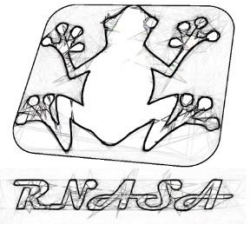
CAMBIOS EN LOS DATOS



• El linaje del dato provee a los profesionales del dato una representación visual que permite visualizar el flujo del dato. De este modo, se puede identificar cómo y cuándo el dato es modificado en la organización. Por ejemplo, identificar los cambios efectuados en el dato por los diferentes procesos de extracción, transformación o carga definidos en la arquitectura de datos de la organización. Cabe también destacar que, gracias a la capacidad de monitorizar el dato de una forma continuada, los errores relativos al dato pueden ser detectados antes de que aparezcan y, por tanto, corregidos.

CICLO DE VIDA DEL DATO

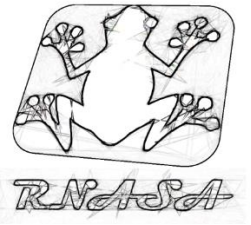
CAMBIOS EN LOS DATOS



• Un escenario de uso común del linaje del dato se da en el área de la inteligencia de negocio (business intelligence). Por ejemplo, esta capacidad muestra cómo se ha obtenido cierta información relativa al negocio y qué papel puede desempeñar en los distintos métodos disponibles de integración de datos en la organización.

CICLO DE VIDA DEL DATO

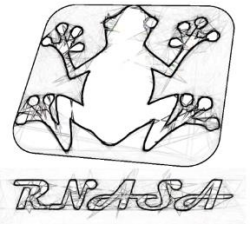
CAMBIOS EN LOS DATOS



• Otro escenario habitual tiene que ver con la reducción de riesgos y la protección de datos. Los profesionales del dato pueden usar el linaje del dato para gestionar de una manera óptima el dato y, en todo momento, controlar dónde se encuentran los datos sensibles, y evitar así su exposición o minimizar los efectos de una posible violación en la seguridad en la organización.

CICLO DE VIDA DEL DATO

CAMBIOS EN LOS DATOS



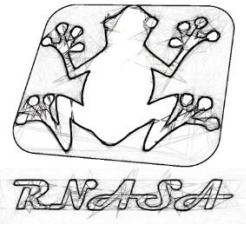
- Otros casos de uso incluyen:

  - **Resolución de errores o conflictos**: en la creación de informes, la trazabilidad del dato permite conocer cómo se han construido las métricas que se incluyen, qué transformaciones se han hecho y de dónde provienen.

  - **Análisis del impacto**: en el desarrollo y evolución de sistemas de información, el linaje del dato habilita comprender de antemano qué sucederá con el dato y qué medidas se deben tener en cuenta.

CICLO DE VIDA DEL DATO

CAMBIOS EN LOS DATOS



• **Informes de conformidad**: en ciertos sectores, se han impuesto regulaciones y normas para una mayor transparencia. Data lineage habilita reducir los errores humanos y las brechas en la conformidad respecto a la normativa vigente.

En esencia, el linaje del dato es un paso necesario, pero no suficiente, hacia la gestión eficiente del dato.

Convertir datos en conocimiento

Tipos de aprendizajes

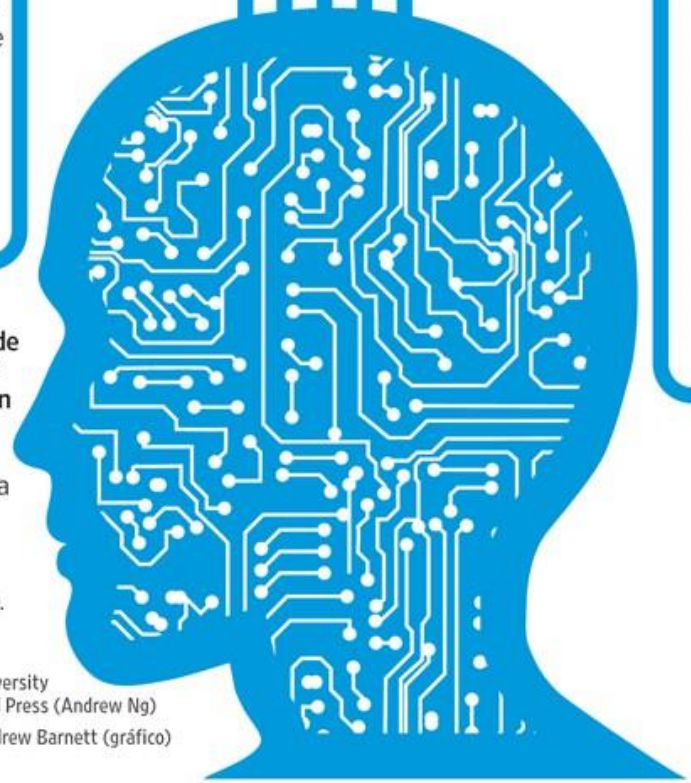
Grupo de Investigación RNASA-IMEDIR

Enseñándole a las máquinas

Enfocado fuertemente en inteligencia artificial (IA), Sillicon Valley atrae y financia a profesores de renombre en el campo.

Demis Hassabis
Vicepresidente de ingeniería de DeepMind Technologies en Google. Graduado de la Universidad de Cambridge. Después de comprar DeepMind, Google contrató a varios expertos en IA de la Universidad de Oxford y ofreció financiación a sus departamentos de ingeniería y ciencias informáticas.

Yann LeCun
Director de investigación de IA en Facebook y profesor de ciencias informáticas en la Universidad de Nueva York. Desarrolló el reconocimiento de escritura a mano. Pionero del aprendizaje automático, la visión computarizada y el procesamiento de lenguaje.



Geoff Hinton
Investigador en Google y profesor de ciencias informáticas en la Universidad de Toronto.
Se considera el padrino del aprendizaje automático.

Carlos Guestrin
Brasileño. Profesor de aprendizaje automático en la Universidad de Washington.
Su cátedra y la de su esposa son financiadas por Amazon. Reconocido por esfuerzos en el acceso al aprendizaje automático a través de herramientas de desarrolladores.

Andrew Ng
Director de IA en Baidu y profesor asistente en la Universidad de Stanford.
Fundó el proyecto Brain en Google antes de pasar a Baidu. Líder en la investigación de IA.

Fotos: Google (Demis Hassabis); Facebook (Yann LeCun); University of Toronto (Geoff Hinton); Dato (Carlos Guestrin); Associated Press (Andrew Ng)

Fuentes: reportaje de WSJ; Amir Mizroch (investigación); Andrew Barnett (gráfico)

THE WALL STREET JOURNAL.

¿Qué es aprender?

- La capacidad de aprender es uno de los atributos distintivos del ser humano

- El aprendizaje humano es diverso e incluye:
  - Adquisición de conocimiento
  - Desarrollo de habilidades a través de instrucción y practica
  - Organización de conocimiento
  - Descubrimiento de hechos
  - …

¿Qué es aprender?

De la misma forma,

MACHINE LEARNING estudia y modela computacionalmente los procesos de aprendizaje en sus diversas manifestaciones

¿Y EN EL CASO DE LAS MÁQUINAS?

- Aprendizaje: Es el campo de estudio que le da a las computadoras la habilidad de aprender sin ser programadas explícitamente [Samuel, 59]

- Aprendizaje: Cambios adaptivos en el sistema para hacer la misma tarea(s) de la misma población de una manera más eficiente y efectiva la próxima vez [Simon, 83]

- Aprendizaje: Un programa de computadora se dice que aprende de experiencia E con respecto a una clase de tareas T y medida de desempeño D, si su desempeño en las tareas en T , medidas con D, mejora con experiencia E [Mitchell, 97].

Aprendizaje

- Desde un punto de vista mas tradicional (hablando de representaciones simbólicas/reglas,...), podemos decir que una buena parte de ML esta dedicada a inferir reglas a partir de ejemplos.
- Descripciones generales de clases de objetos, obtenidas a partir de un conjunto de ejemplos, pueden ser usadas para clasificar o predecir.
- En general, el interes no esta en aprender conceptos de la forma en que lo hacen los humanos, sino aprender representaciones simbólicas de ellos.

Aprendizaje Computacional

Otros

- **Aprendizaje genético:** Aplica algoritmos inspirados en la teoría de la evolución para encontrar descripciones generales a conjuntos de ejemplos.
- **Aprendizaje conexionista:** Busca descripciones generales mediante el uso de la capacidad de adaptación de redes de neuronas artificiales
- **Aprendizaje por analogía:** intenta emular la capacidad humana de recordar la solución de problemas previos ante la aparición de problemas parecidos
- **Aprendizaje multiestrategia:** o combinación de diferentes tipos de estrategias y/o diferentes tipos de aprendizaje.

OTRAS CLASIFICACIONES

Otro criterio puede ser basar la clasificación en el objetivo o propósito principal del proceso de aprendizaje:

- ***Aprendizaje sintético:*** el objetivo es adquirir nuevo conocimiento e ir más allá del conocimiento poseído (**inducción y analogía**)

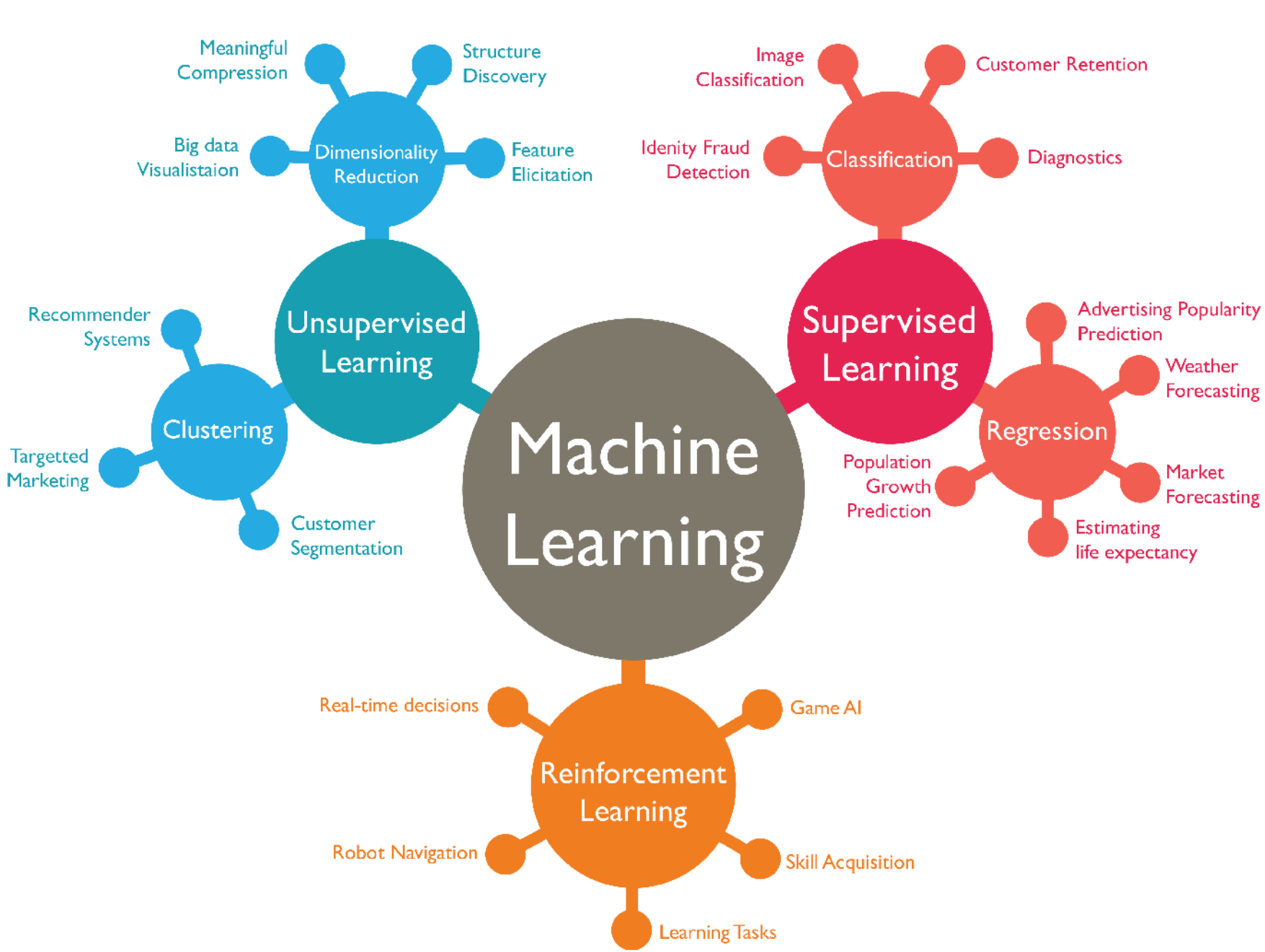
- ***Aprendizaje analítico:*** en el que poseemos un conocimiento general y lo particularizamos para hacerlo más efectivo (**deducción**)

Tipos de Aprendizaje

- ***Aprendizaje supervisado***, donde se va dirigiendo al sistema en el proceso de entrenamiento.

- ***Aprendizaje no supervisado***, donde no se corrige al sistema en su proceso de entrenamiento.

- ***Aprendizaje por refuerzo***, en el que no se le dice la salida, sólo si ha clasificado bien o no.


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

Tipos de Aprendizaje

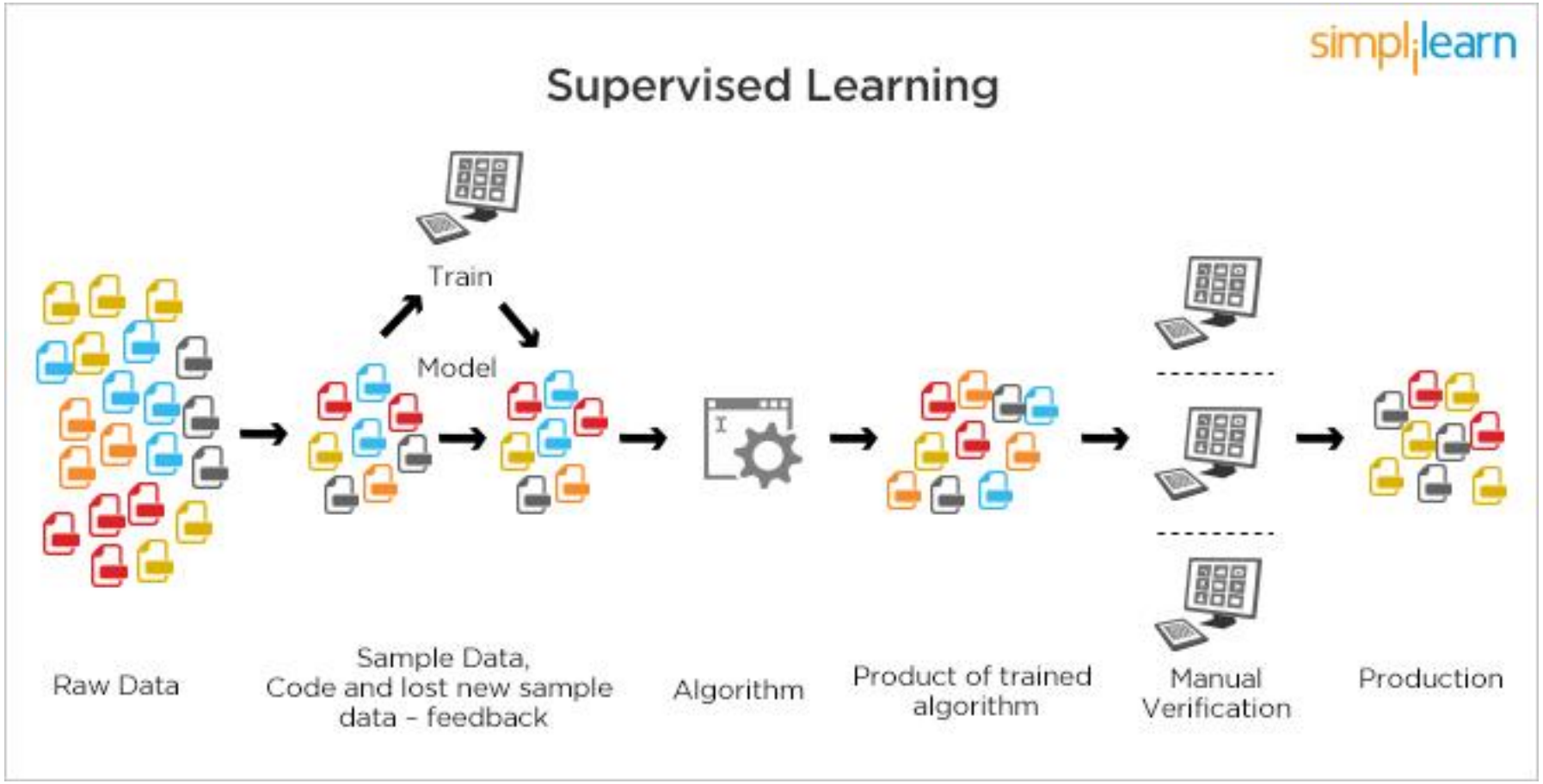
- Supervisado
  - Predicción: clasificación
  - Regresión: las clases son continuas.

- No Supervisado
  - Clustering

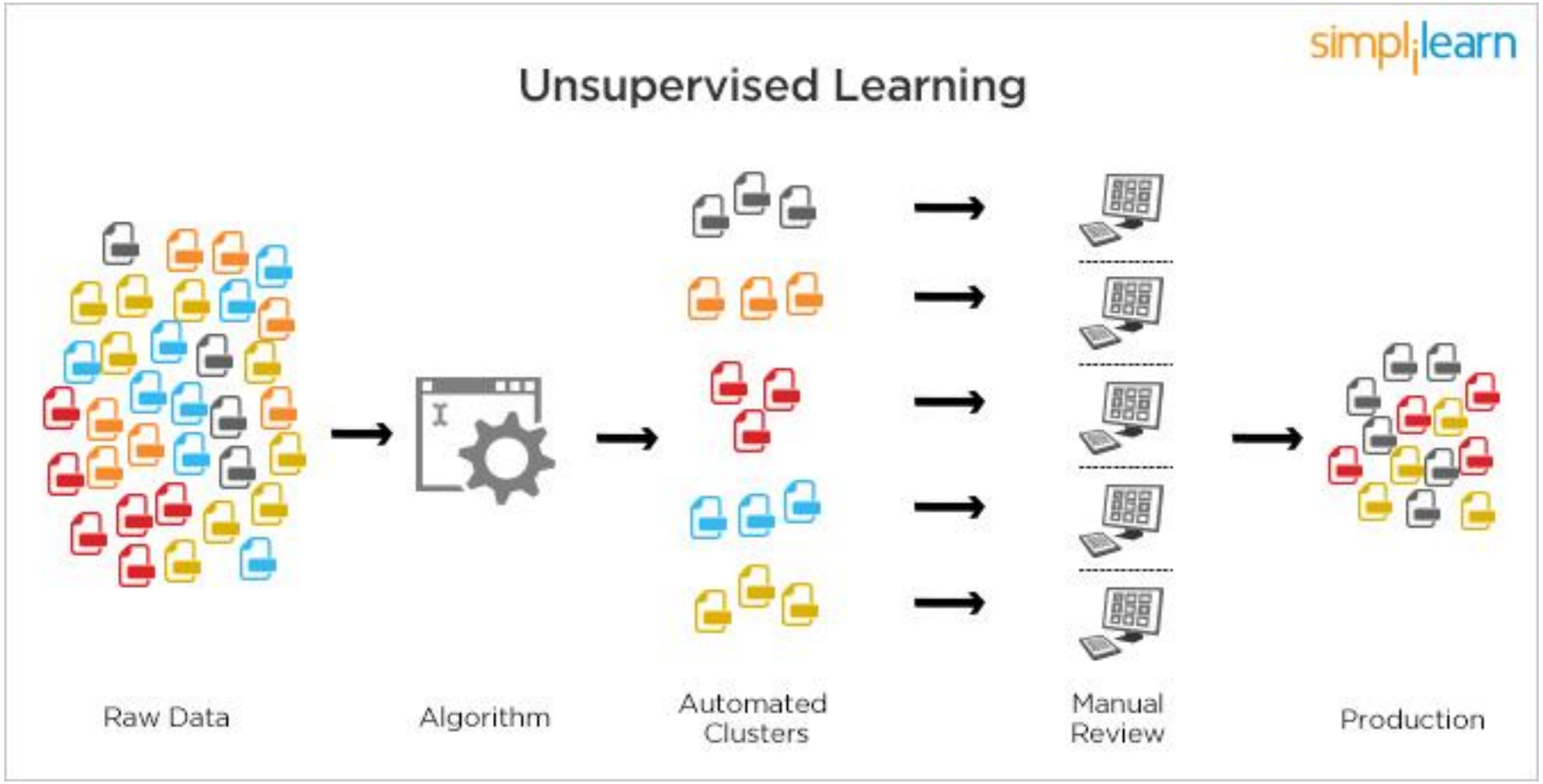
La meta es inducir un modelo para poder predecir el valor de la clase dados los valores de los atributos.

Se usan por ejemplo, arboles de regresión, regresión lineal, redes neuronales, LWR, etc.

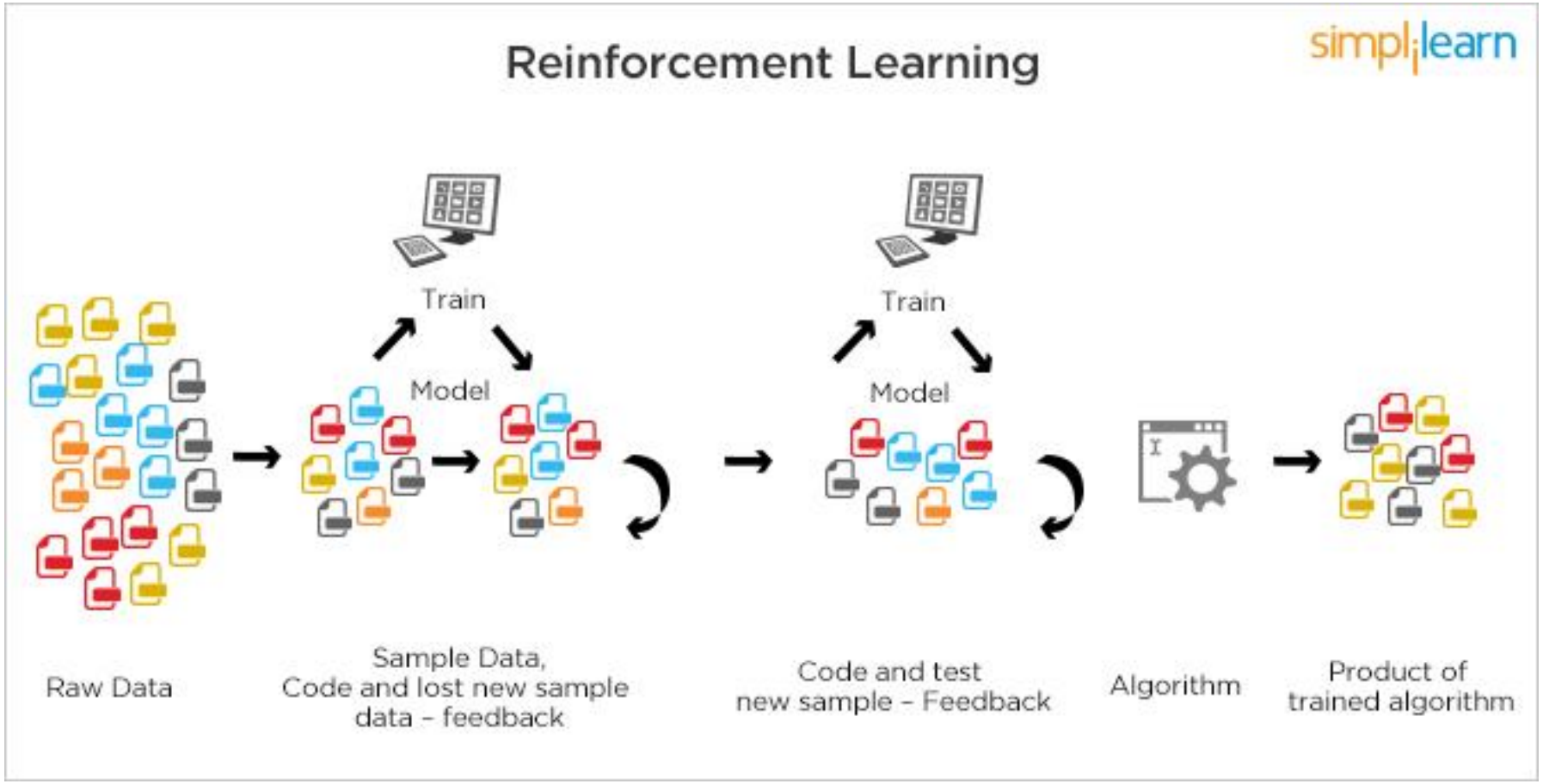
Aprendizaje Supervisado



Aprendizaje No Supervisado



Aprendizaje por Refuerzo



Aprendizaje Inductivo

- El aprendizaje inductivo puede verse como el proceso de aprender una función.
- Por ejemplo, en aprendizaje supervisado, al elemento de aprendizaje se le da un valor correcto (o aproximadamente correcto) de una función a aprender para entradas particulares y cambia la representación de la función que esta infiriendo, para tratar de aparear la información dada por la retroalimentación que ofrecen los ejemplos.

INDUCTIVO SUPERVISADO

Los métodos más utilizados en aplicaciones provienen del aprendizaje inductivo supervisado:

- **Inducción:** Pasamos de lo específico a lo general
- **Supervisión:** Conocemos el concepto al que pertenece cada ejemplo

A partir de un conjunto de ejemplos etiquetados obtenemos un modelo:

- El modelo generaliza los ejemplos, representando los conceptos que definen las etiquetas
- Obtenemos lo que es común entre los ejemplos de un concepto que les diferencia de los otros

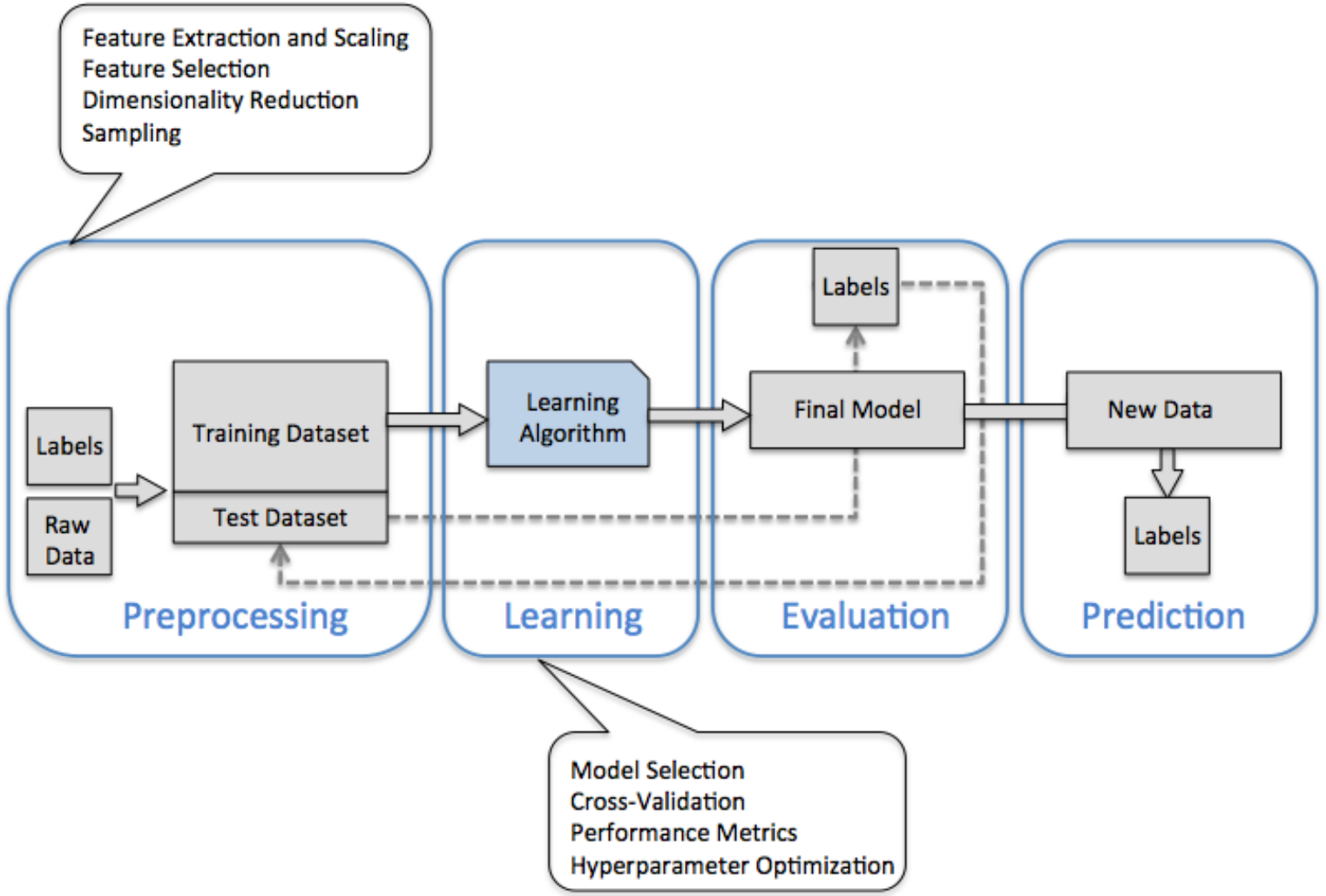
TIPOS DE MÉTODOS DE APRENDIZAJE INDUCTIVO SUPERVISADO

- Modelos caja blanca (podemos inspeccionar el modelo)
  - Árboles de decisión/reglas de inducción
  - Modelos probabilísticos

- Modelos caja negra
  - Redes de neuronas artificiales
  - Máquinas de soporte vectorial

- Podemos plantear el problema como:
  - Clasificación: un conjunto finito de conceptos
  - Regresión: una función continua

Workflow Básico



Workflow "Científico"