

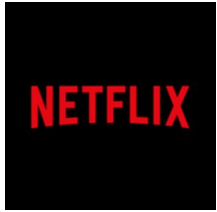


Aprendizaje Automático Datos

Viviana Cotik
1er cuatrimestre 2019



Aprendizaje automático: Algunas aplicaciones



- **Reconocimiento del habla:** Siri, Cortana, Google Now, Alexa
- **Predicción de tiempo de viaje, camino óptimo:** Google Maps, Waze, Uber, Despegar
- **Detección de fraude:** bancos, PayPal, Mercado Libre
- **Publicidad online:** Google Ads

¿Qué cambió?

- mayor disponibilidad de **datos**
- mejor **capacidad de cómputo**
- mejores **algoritmos**

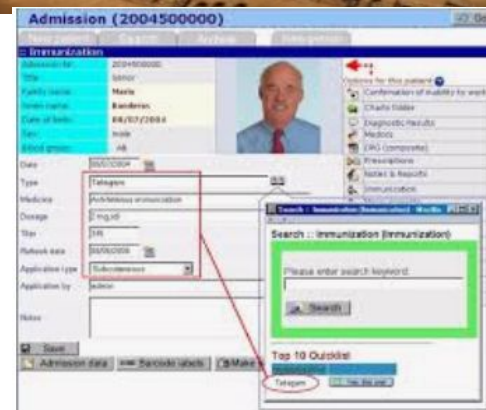
Índice

- Qué cambió
- **Disponibilidad de datos**
- Calidad de datos
- Algunos atributos de calidad. Sesgos y datos abiertos
- Aspectos éticos
- Algunas definiciones

Disponibilidad de Datos

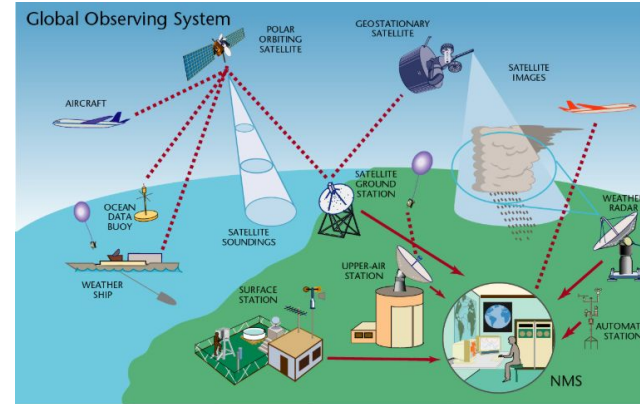


Disponibilidad de datos



Proyecto genoma humano

Disponibilidad de datos



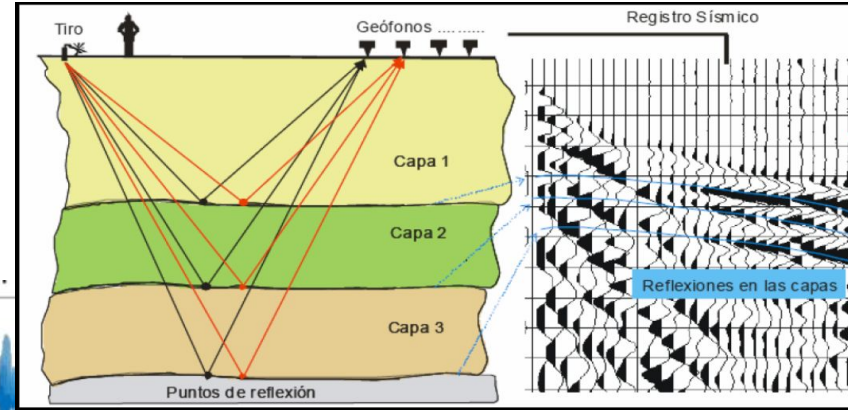
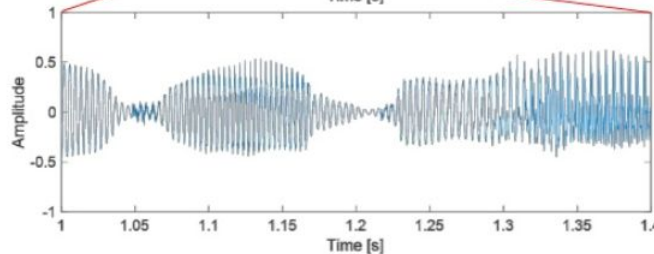
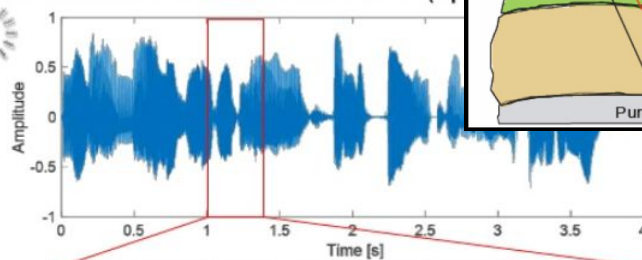
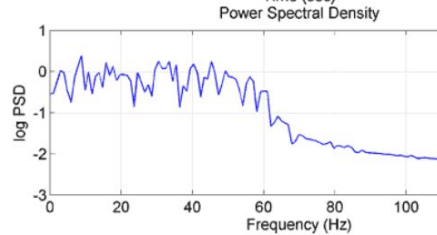
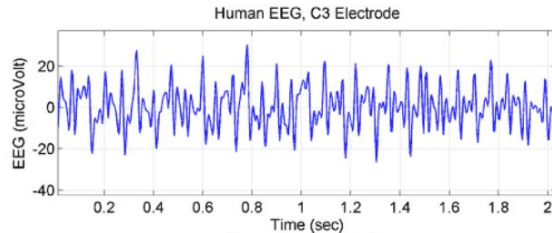
Datos meteorológicos



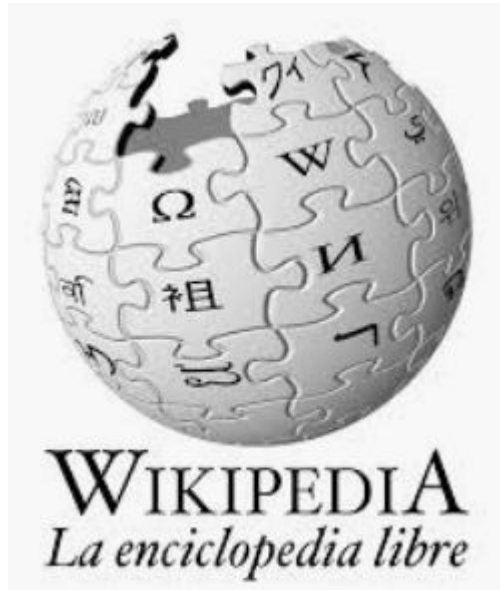
Disponibilidad de datos - Imágenes



Disponibilidad de datos - Señales



Disponibilidad de datos - Texto



Admission (2004500000)

Immunization

Substance: 2004500000
Type: Sub
Priority review: No
Review status: Pending
Date of birth: 08/07/1984
Sex: Male
Blood group: AB

Date: 08/07/2004
Type: Teleregion
Medicine: Public health immunization
Doseage: 0.5 mg, 0.5
Way: IM
Refuse date: 08/07/2004
Application type: Sub-consultation
Application by: Sub-consultation

Search: Immunization (immunization)
Please enter search keyword:
[Search]

Top 10 Outlets
Teleregion [1] [2] [3] [4] [5] [6] [7] [8] [9] [10]



Disponibilidad de datos - Estructurados



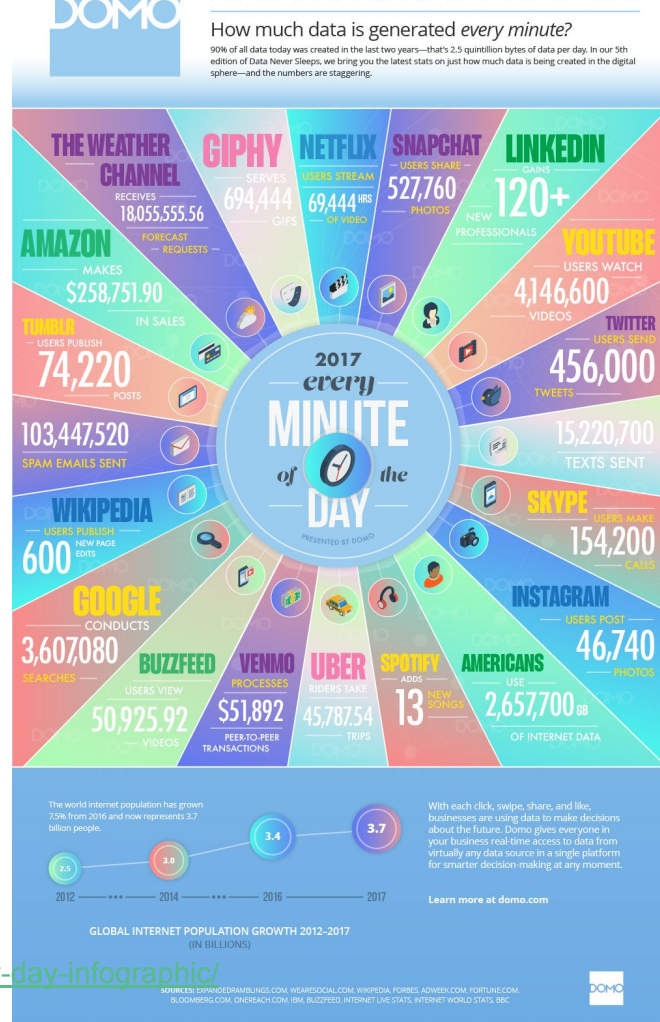
Internet de las cosas (IoT-Internet of things)



Kuva 1. Internet of Things. Lähde: Huffington Post

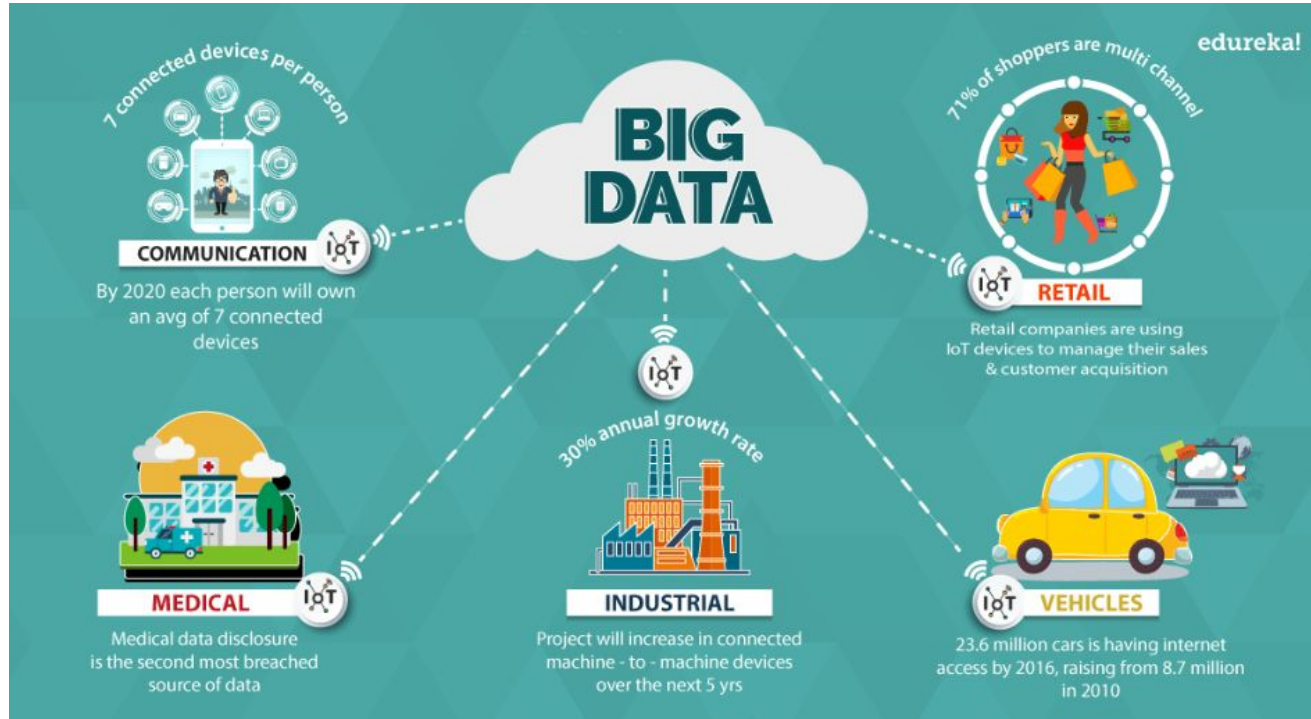
Disponibilidad de Datos

-Datos generados por minuto

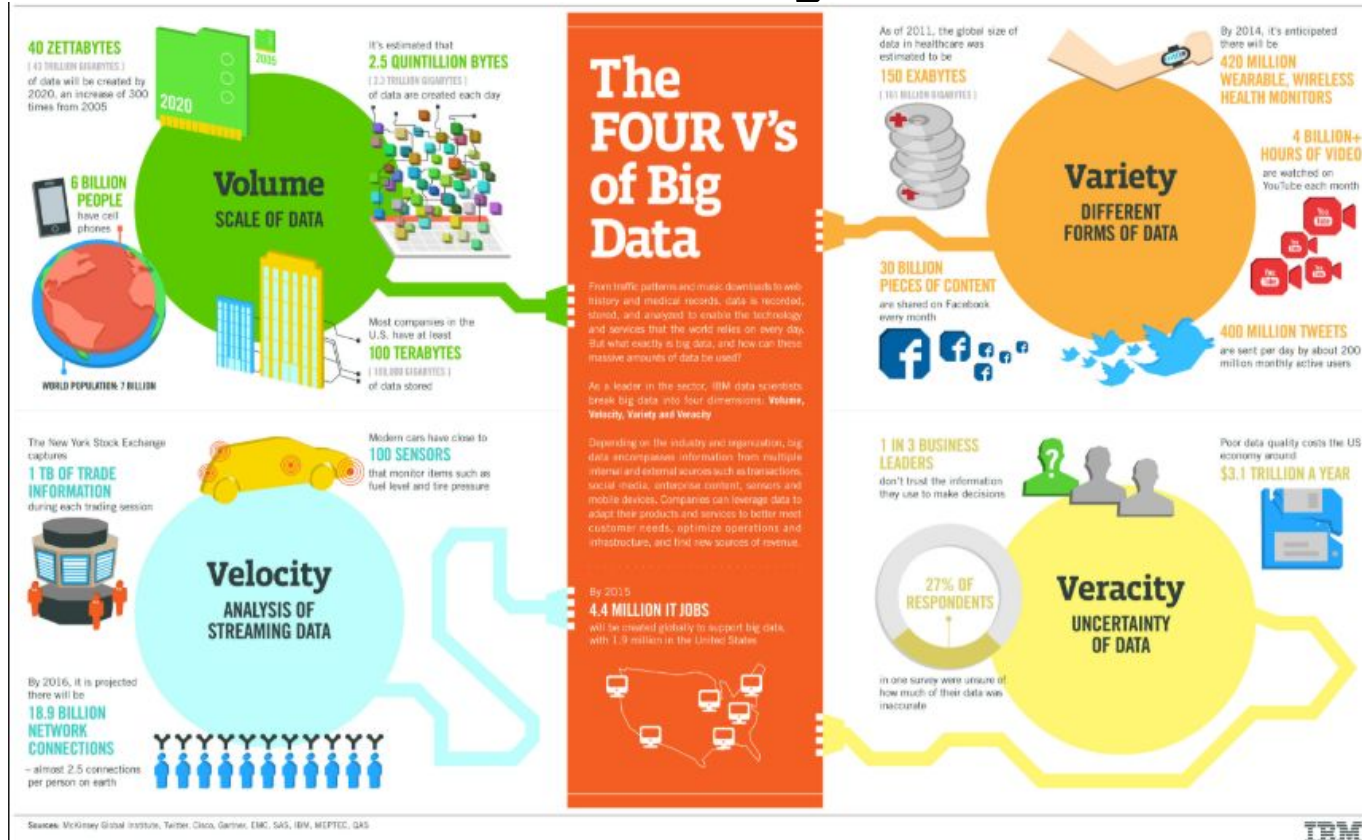


<https://techstartups.com/2018/05/21/how-much-data-do-we-create-every-day-infographic/>

Disponibilidad de datos - Big Data



Disponibilidad de datos - Big Data



Digital Footprint



Digital footprint activa

datos dejados online intencionalmente. Ej:

- fotos
- posteos

Digital footprint pasiva

datos que se dejan online de forma no intencional. Ej.

- hábitos de web browsing
- historia de búsquedas

Digital Footprint - Ejemplos

Google Timeline



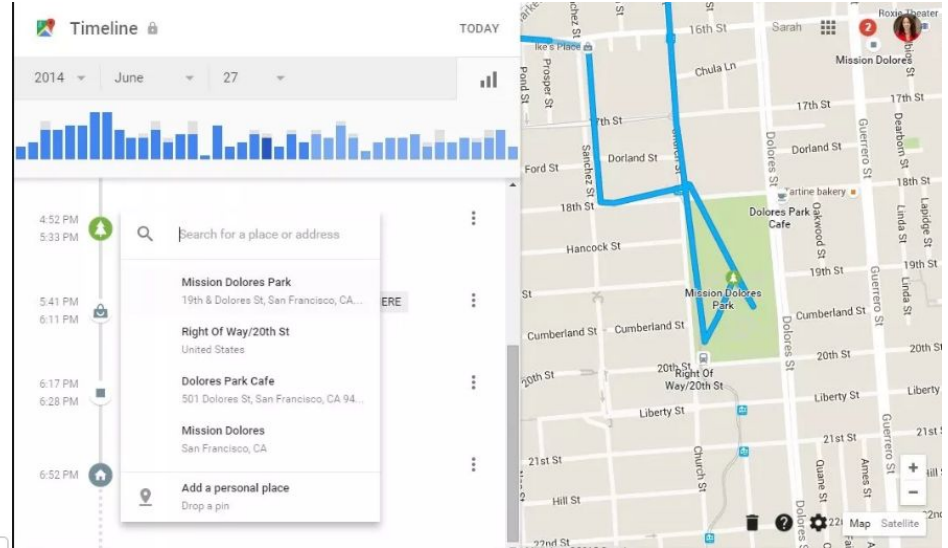
Your activity in timeline



51 km walked this month



47 hours spent in a vehicle
this month



Digital footprint - Ejemplos



Experimento de control de emociones

- **Alteración de news feed** para estudiar efectos de posts positivos y negativos.
- **Remoción de posteos positivos y negativos** a distintos grupos.
- Posteos de usuarios acordes a posts que se les mostraban.

Índice

- Qué cambió
- **Disponibilidad de datos**
- Calidad de datos
- Algunos atributos de calidad. Sesgos y datos abiertos
- Aspectos éticos
- Algunas definiciones

Datos

- estructurados
- no estructurados
- semiestructurados

Datos estructurados:

- numéricos
- ordinales (tienen un orden). Ej.:
 - status socioeconómico,
 - severidad de una patología
 - insuficiente, aprobado, sobresaliente
- categóricos (no tienen orden). Ej.:
 - ubicación: urbano, suburbano, rural
 - educación: pública, privada

Calidad de datos

Ediciones anteriores

Clarín.com » Edición Viernes 14.02.2003 » Economía » **Un banco debe pagar \$ 120.000 por**

FIGURAR EN UN LISTADO DE INCUMPLIDORES ARRUIÑO UN NEGOCIO
Un banco debe pagar \$ 120.000 por incluir mal a un cliente en Veraz

Daniel Gutman

El Banco Río lo incluyó en las listas negras de deudores de la Organización Veraz y solicitó al Banco Central que lo inhabilitara. Pero todo era un error, porque no había existido ningún incumplimiento. El cliente hizo juicio y obtuvo una sentencia de Cámara a su favor. Hasta ahí, un caso igual a muchos otros que ha habido en los últimos años. Lo novedoso es que la Cámara en lo Comercial acaba de establecer la que seguramente sea la indemnización más alta en este tipo de casos: el Banco Río deberá pagarle 120.000 pesos a su ex cliente.

A esa cifra deberán sumársele **los intereses** a la tasa activa del Banco Nación desde la fecha de inhabilitación en los registros del Central, que es mayo de 1996, lo que **llevaría la indemnización a más de medio millón de pesos**, según los abogados del demandante.

La importancia de la indemnización —según se explicó en el fallo— tiene que ver con que el damnificado **es un empresario que estaba en pleno proceso de ampliación de sus negocios**.

El hombre, dueño de una confitería, estaba construyendo un edificio en la avenida Cruz en el cual **pensaba instalar una concesionaria de autos**, además de una confitería y salón de fiestas en la planta alta. Sin embargo, en mayo de 1996 quedó sin posibilidad de obtener crédito y operar con cheques, por lo que **la obra y sus proyectos quedaron inconclusos**.

Así, la Sala B de la Cámara —en un voto de la jueza María de Díaz Cordero, al que adhirió Enrique Butty— aplicó el concepto de "pérdida de chance". Es decir, el Río deberá indemnizar al empresario porque **lo privó de una oportunidad de ganar dinero**.

Las pruebas presentadas y la trayectoria de Eloy Domínguez Álvarez convencieron a la jueza de que él "tenía intención de culminar con la construcción del edificio y ampliar sus negocios" y de que lo hubiera hecho "de no haber existido la arbitraria y errónea decisión adoptada por la entidad bancaria demandada".

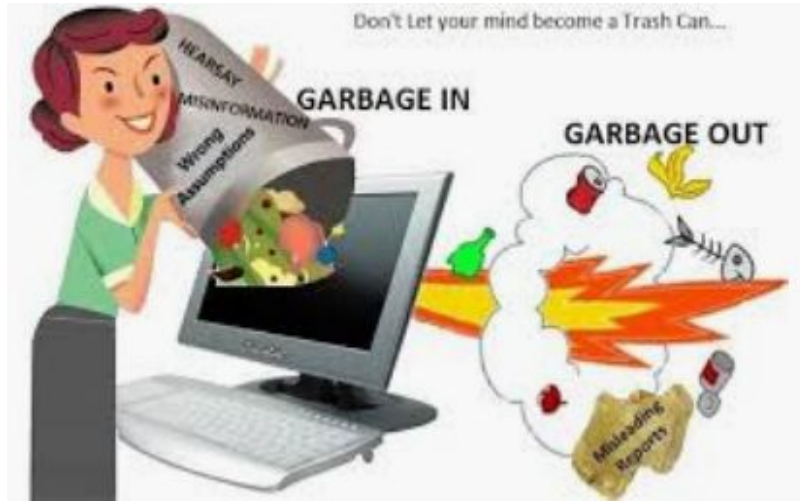
Calidad de datos

Definición:

- “Un dato o conjunto de datos X es de mayor (mejor) calidad que un dato o conjunto de datos Y si X satisface las necesidades del usuario mejor que Y” [Redman, 1996]
- “Satisfacer de manera consistente las expectativas de los usuarios” [English, 1999]

Definición subjetiva

Calidad de datos



Consecuencias:

- descreimiento
- insatisfacción de clientes
- costos innecesarios
- impacto en toma de decisiones
- ...

Calidad de datos

Ejemplos:

- procesos masivos que reparan un dato, pero no reconstruyen información relacionada
- misma información cargada en distintos sistemas
- valores predeterminados



Depende de:

- calidad de software
 - usabilidad
 - interfaz (obligatoriedad de carga)
- definición de procesos asociados a los datos
- diseño de base de datos
- capacitación
-

Calidad de Datos

Qué se requiere?

- datos completos
- datos oportunos (timeliness) y vigentes
- datos consistentes y correctos**
- datos en cantidad adecuada
- datos disponibles/accesibles (ej. medicina), open data.
- datos seguros y privados (protección de datos personales)

Atributos de calidad

Podemos definir la calidad a partir de atributos, por ej:

- completitud
- relevancia
- vigencia
- disponibilidad
- confiabilidad
- consistencia
- corrección
- seguridad/privacidad

Atributos de calidad

- **completitud**
 - están presentes todos los valores para representar la realidad
 - están presentes todas las instancias existentes en el mundo real
- **relevancia**
 - los datos son relevantes para representar la realidad
- **vigencia**
 - los datos se mantienen actualizados con la frecuencia adecuada
- **disponibilidad**
 - los datos están accesibles
- **confiabilidad**
 - se puede considerar que los datos representan información verídica

Atributos de calidad

- consistencia
 - no hay contradicciones entre distintos datos almacenados
- corrección
 - los datos representan la situación real
- seguridad/privacidad
 - los datos cumplen con los requerimientos de privacidad adecuados de acuerdo a la reglamentación nacional / criterios éticos
 - los datos son sólo accesibles por los usuarios autorizados

Calidad de datos

- No existen datos perfectos
- Es necesario priorizar las calidades deseadas

Índice

- Qué cambió
- Disponibilidad de datos
- Calidad de datos
- **Algunos atributos de calidad. Sesgos y datos abiertos**
- Algunas definiciones
- Conocimientos necesarios

Calidad de datos- Correctitud / Bias

Machine learning muchas veces entrenado con **grandes conjuntos de datos anotados** (estudiantes, crowdsourcing).

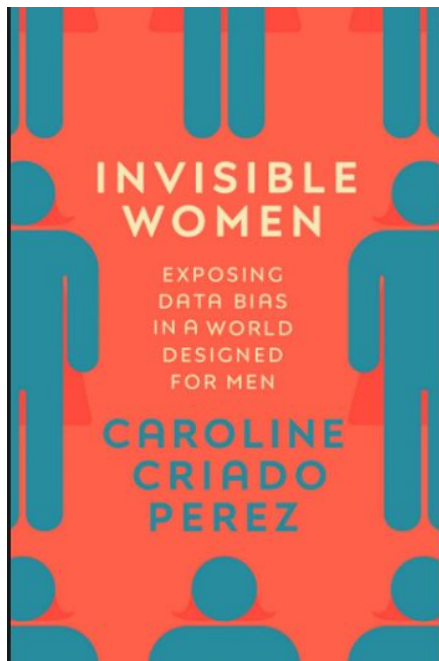
Datos a veces tomados a partir de scrapping:

- origen de los datos
- calidad de queries / consultas

Posibilidad de **bias étnicos, de género**

Una de las causas: suele ser **costoso obtener datos y poder etiquetarlos**

Calidad de datos- Correctitud / Bias



MUNDO

De los estantes muy altos a la mayor probabilidad de morir en un choque: el mundo medido para los hombres pone en peligro a las mujeres

La brecha de datos entre géneros, a la que aludieron Bill y Melinda Gates en su reciente carta anual, se explica con una enorme cantidad de información en "Mujeres invisibles", el nuevo libro de la británica Caroline Criado Perez

Infobae 2/03/2019

<https://www.infobae.com/america/mundo/2019/03/02/de-los-estantes-muy-altos-a-la-mayor-probabilidad-de-morir-en-un-choque-el-mundo-medido-para-los-hombres-pone-en-peligro-a-las-mujeres/>

Calidad de datos- Correctitud / Bias

- **Hombre de referencia** para estandarizar investigaciones científicas, tecnológicas y comerciales
 - género: masculino
 - raza: blanca
 - peso: 70 kg.
 - altura: 1,70 m.
 - edad: 25 - 30.
- Fórmula para **determinar la temperatura estándar en la oficina** desarrollada **a partir de la tasa metabólica de descanso del hombre promedio**. Sobrestima la tasa metabólica femenina en un 35%
- Síntomas de un ataque al corazón:
 - en hombres: dolor en el pecho
 - en mujeres: dolor de estómago, náuseas y disnea
- Sistemas de reconocimiento de voz de automóviles, sólo escucha a hombres. Ford Focus 2012

Calidad de datos- Correctitud / Bias



New York Times 17/03/2019

<https://www.nytimes.com/paidpost/loreal-fondation/gender-in-the-world-of-science.html>

Calidad de datos- Correctitud / Bias



COMMENT • 18 JULY 2018

AI can be sexist and racist — it's time to make it fair

Computer scientists must identify sources of bias, de-bias training data and develop artificial-intelligence algorithms that are robust to skews in the data, argue James Zou and Londa Schiebinger.

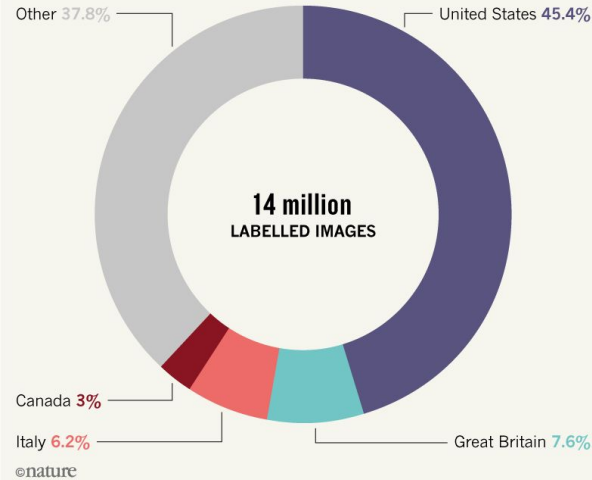
<https://www.nature.com/articles/d41586-018-05707-8>

Calidad de datos- Correctitud / Bias



IMAGE POWER

Deep neural networks for image classification are often trained on ImageNet. The data set comprises more than 14 million labelled images, but most come from just a few nations.





Computer vision: ImageNet:

- 45 % de EEUU (4% de población mundial),
- 3% de China e India: (36% de población mundial)


Calidad de datos- Correctitud / Bias

Spanish ▾



dijo que compraría
harina

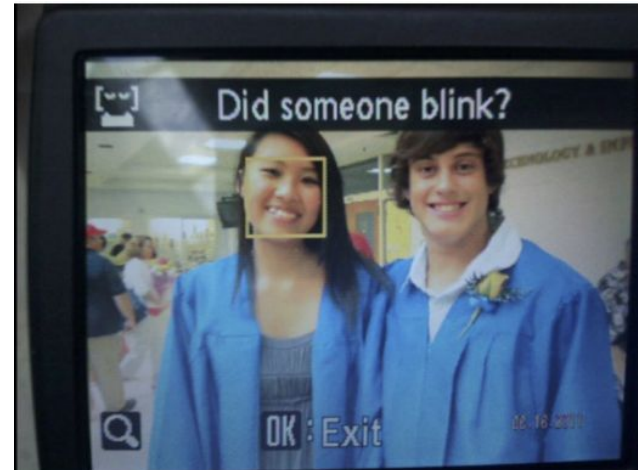
English ▾



He said he would
buy flour

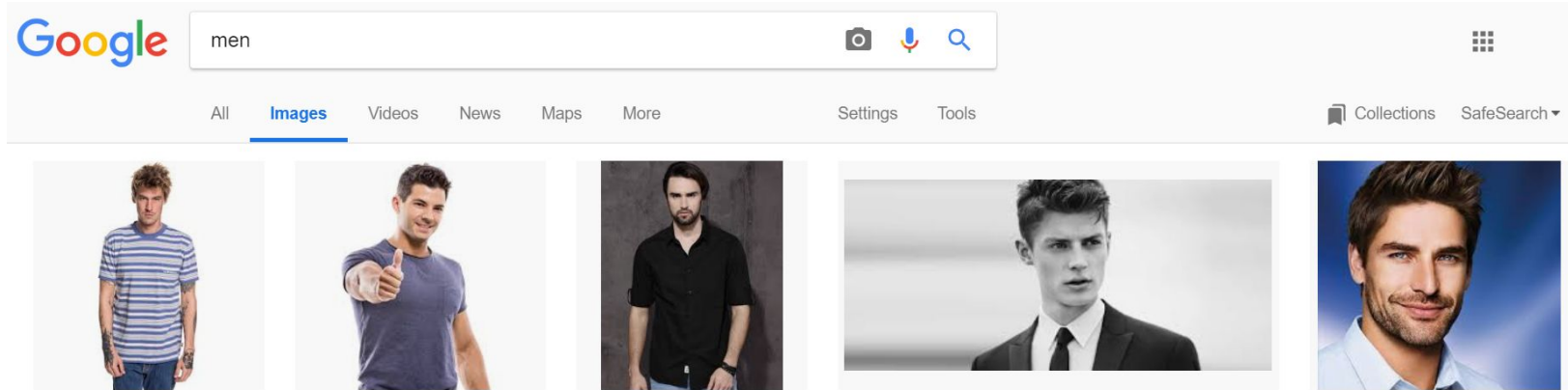
[Open in Google Translate](#)

[Feedback](#)

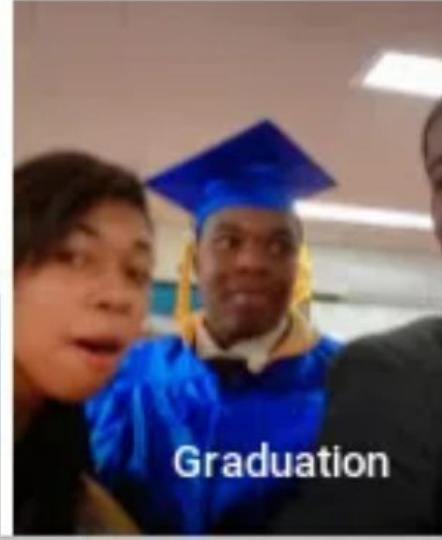


CámarasNikon

Calidad de datos- Correctitud / Bias



Calidad de datos- Correctitud / Bias



Google photos

Calidad de datos- Correctitud / Bias

Análisis de género

Medicina:

- **modelo para desarrollo de drogas es el hombre.** drogas metabolizan distinto en hombres y en mujeres. Costo: vidas y dinero. En los 90s 10 drogas fueron retiradas del mercado norteamericano por tener riesgos de muerte. 8 de ellos mayor riesgo para mujeres.
- **enfermedad cardiovascular.** históricamente de hombres, pruebas clínicas en hombres. mujeres mal o no diagnosticadas, menos cirugías. En las últimas dos décadas se vio que afecta distinto a las mujeres que a los hombres. Principal causa de muerte de mujeres.

“We need to teach scientists the true effects of gender analysis so that they can create research that works for everyone.”

Dr. Londa Schiebinger, professor of history of science and director of Gendered Innovations, Stanford

Calidad de datos- Correctitud / Bias

Más ejemplos. Debido a datos y técnicas

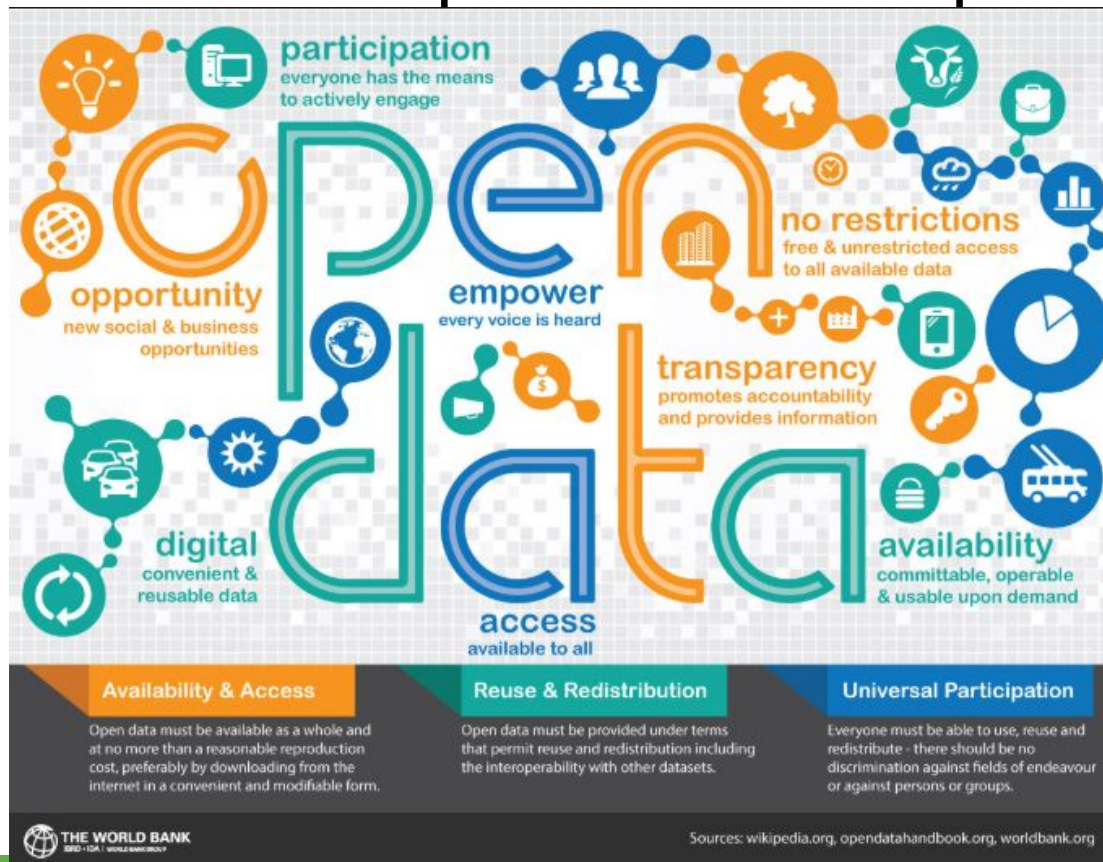
- identificación de cancer de piel a partir de fotografías.
 - Entrenamiento: 129,450 imágenes, 60% tomadas de Google Images
 - sólo 5% corresponden a individuos de piel oscura
- En general, si en el **conjunto de entrenamiento es desbalanceado** (ej. un grupo de individuos aparece más frecuentemente que otro), se optimizarán los resultados para ese grupo de individuos, ya que incrementa la exactitud del sistema

Calidad de datos- Correctitud / Bias

“Every training data set should be accompanied by information on how the data were collected and annotated. If data contain information about people, then summary statistics on the geography, gender, ethnicity and other demographic information should be provided (see ‘Image power’). If the data labelling is done through crowdsourcing, then basic information about the crowd participants should be included, alongside the exact request or instruction that they were given.”

AI can be sexist and racist — it’s time to make it fair. James Zou, Londa Schiebinger

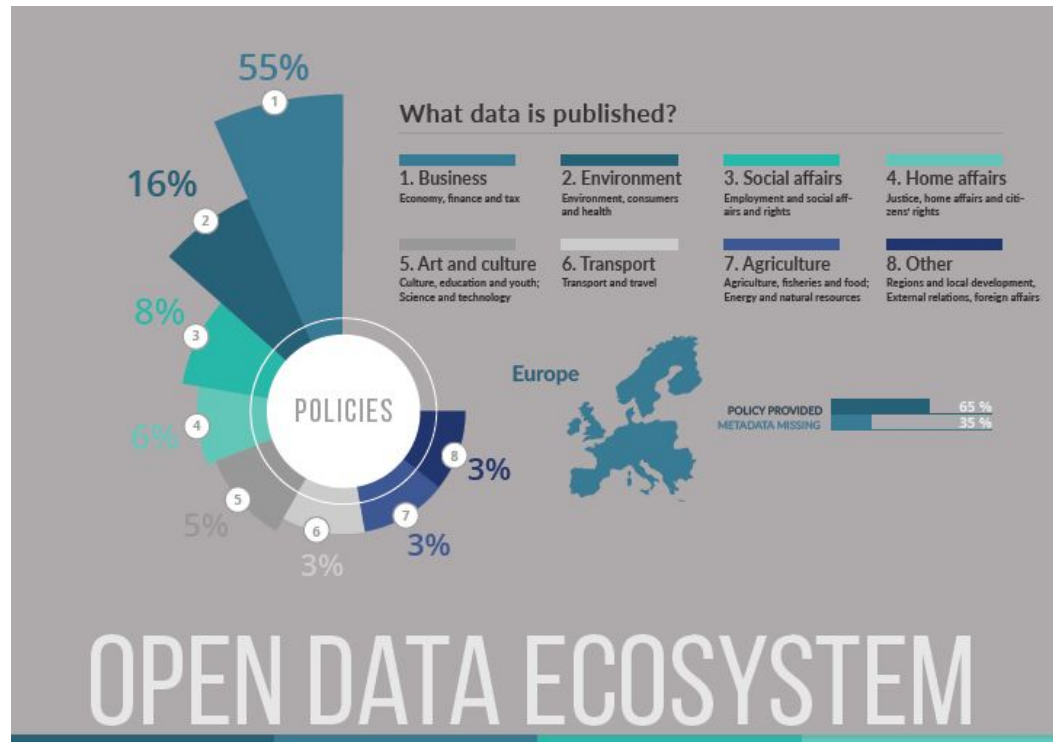
Calidad de datos - Disponibilidad - Open data



Calidad de datos - Disponibilidad - Open data

A summary visualisation of the Open Data Ecosystem

(Source: OpenDataMonitor, 2016).



<https://opendataincubator.eu/odine-stars-on-the-future-of-open-data/>

Índice

- Qué cambió
- Disponibilidad de datos
- Calidad de datos
- Algunos atributos de calidad. Sesgos y datos abiertos
- **Ética**
- Data science

Problemas éticos

Big data ethics: sistematizar, defender, recomendar actitudes éticas en relación a los datos, en particular a los datos personales.

Mucha más información disponible



Problemas éticos

Privacidad:

Quién debería controlar acceso a los datos?

Ownership:

Quién es el dueño de los datos, cuáles son las obligaciones de quienes generan y usan datos?

Reputación

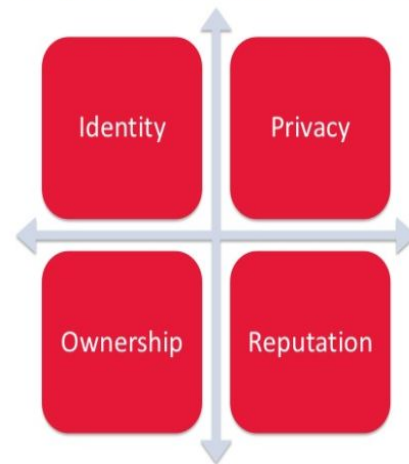
Cómo podemos determinar que los datos son confiables?

Identidad

La identidad de a quién se refieren los datos debería permanecer privada



Four Aspects of Big Data Ethics



Problemas éticos

The ethics of Big Data:

Balancing economic benefits and ethical questions of Big Data in the EU policy context

STUDY



European Economic and Social Committee

<https://www.eesc.europa.eu/resources/docs/qe-02-17-159-en-n.pdf>

Problemas éticos

<https://www.nytimes.com/2019/03/01/business/ethics-artificial-intelligence.html>

The New York Times

Is Ethical A.I. Even Possible?

Problemas éticos - Algunas instituciones

The Institute for Ethical AI & Machine Learning

The Institute for Ethics and Emerging Technologies

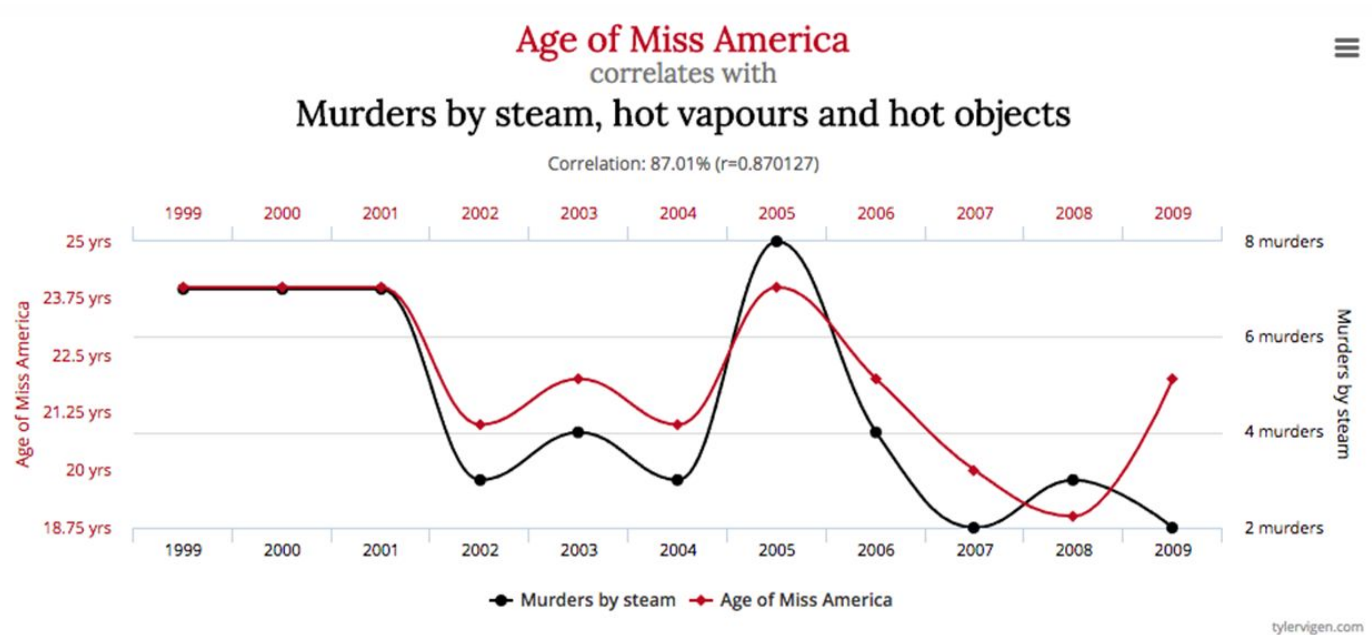
Problemas éticos

Los sistemas son creados por humanos, que pueden tener sesgos.

Posibles problemas en aplicación de técnicas de ML

- Datos:** mala calidad (no disponibilidad, cantidad inadecuada, mal balanceo, no actuales, formato, sesgos, no correctitud, no consistencia), errores en su estructuración, mala interpretación (ej, correlación no implica causalidad)
- Algoritmos:** parametrización, mala elección
- Capacidad de procesamiento:** inadecuada
- Etica**

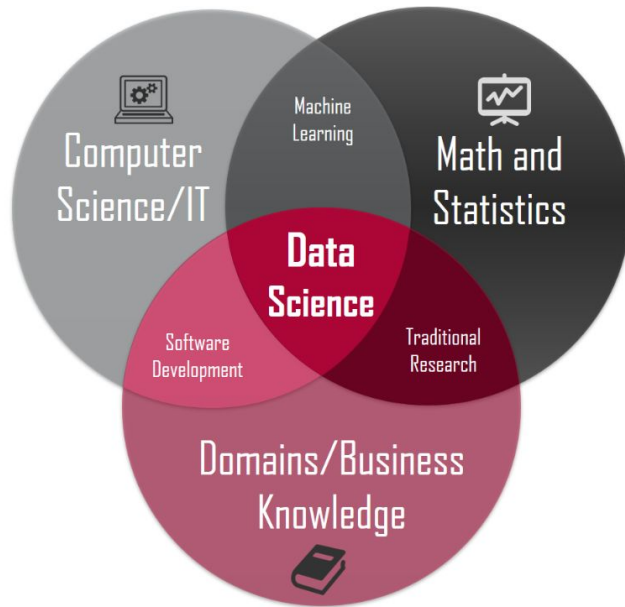
Correlación no implica causalidad



Índice

- Qué cambió
- Disponibilidad de datos
- Calidad de datos
- Algunos atributos de calidad. Sesgos y datos abiertos
- **Aspectos éticos**
- Algunas definiciones

Algunas definiciones - Datascience



Para aplicar técnicas de machine learning correctamente, debe haber interacción con conocedores del dominio:

- social scientists y expertos en humanidades
- medicina
- género
- leyes
- ...

¿Qué habilidades requiere un data scientist?

- recolectar datos,
- saber interpretar los datos,
- organizar, resumir y analizar datos,
- sacar conclusiones válidas
- ...

Bibliografía

Artículos:

. Nature. AI can be sexist and racist — it's time to make it fair.

<https://www.nature.com/articles/d41586-018-05707-8>

. Datos generados por minuto:

<https://techstartups.com/2018/05/21/how-much-data-do-we-create-every-day-infographic/>

. De los estantes muy altos a la mayor probabilidad de morir en un choque (...).

<https://www.infobae.com/america/mundo/2019/03/02/de-los-estantes-muy-altos-a-la-mayor-probabilidad-de-morir-en-un-choque-el-mundo-medido-para-los-hombres-pone-en-peligro-a-las-mujeres/>ht

Gender in the world of science

<https://www.nytimes.com/paidpost/loreal-fondation/gender-in-the-world-of-science.html>