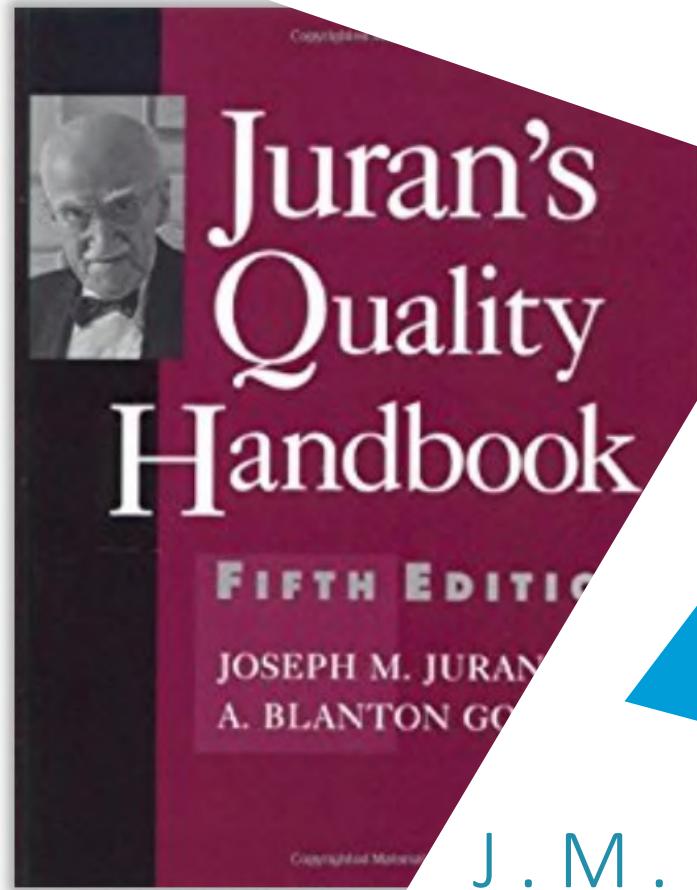


2 – Calidad de los datos

ALWAYS overrated



J. M. JURÁN
A. B. GODFREY

“

Un conjunto de datos es de calidad si da cobertura suficiente a todos sus usos previstos en las operaciones, la toma de decisiones y la planificación

”

Calidad de datos

El concepto hace referencia al **estado** de piezas de información, ya sea **cualitativa** o **cuantitativa**.

Normalmente se asume alta calidad si **representa correctamente** las construcciones del mundo real a las que representa.

A medida que aumentan las fuentes de datos, el concepto de **consistencia** se hace más importante, más allá de su adecuación a cualquier tarea específica.

Calidad de datos

La ISO 9000:2015 define **calidad** como:

Grado en el que las características inherentes de un objeto satisfacen los requisitos.

De aquí se puede derivar la definición de **calidad de datos**:

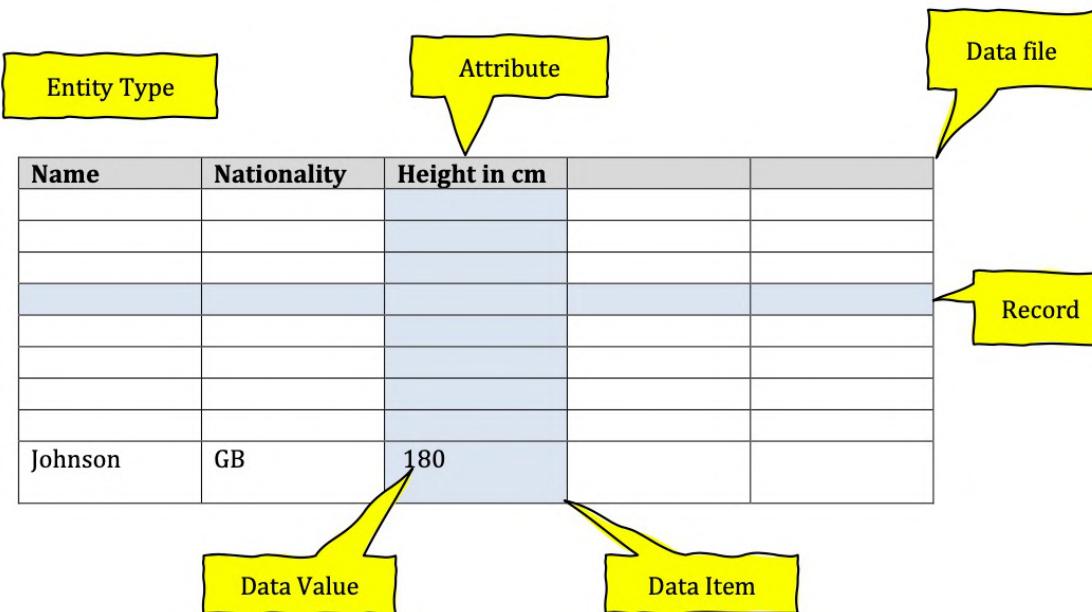
*Grado en el que las **dimensiones** de los datos satisfacen los requisitos.*

Calidad de datos: definiciones relacionadas

Concepto de datos:

Forma en que los datos son estructurados y organizados en un sistema de información.

Ejemplos: conjunto de datos, fichero, atributo, valor



Calidad de datos: definiciones relacionadas

Dimensión de datos:

*Una característica medible de un **concepto de datos**.*

Ejemplos: precisión, completitud

Combinación

Una dimensión y un concepto de datos deben estar **lógicamente relacionados**.

No tiene sentido hablar de la **precisión** de un **fichero**.

Dimension	Data concept
Accuracy	Data values
Completeness	Records
Completeness	Data values
Referential integrity	Data files (tables)

Calidad de datos

Las opiniones sobre la calidad de los datos pueden ser **subjetivas** y **diferir**.
Incluso hablando del mismo conjunto de datos para las mismas tareas.

Cuando este es el caso, la **gobernanza** de datos se usa para **conseguir acuerdos** sobre las definiciones y estándares de calidad a aplicar.

La **limpieza** de datos (*data cleansing*), incluyendo estandarización, puede ser **necesaria** para lograr la **calidad deseada**.

Problema con la calidad de datos

Datos de pobre calidad ponen en riesgo a la organización.

Pueden llevar a:

- malas decisiones
- clientes insatisfechos
- consumidores de datos insatisfechos
- multas por incumplimientos
- costos ocultos (repetir trabajos)
- mala reputación
- empleados insatisfechos
- falta de operatividad

Calidad de los datos

DAMA International

Un posible modelo



Compleitud
Completeness. La proporción de datos almacenados, respecto del potencial “100% completo”

Exactitud
Accuracy. En nivel de correctitud con que los datos describen el objeto o evento de “mundo real”

Consistencia
Consistency. No debe haber diferencia si comparamos 2 o más representaciones de un objeto con su definición

Unicidad
Uniqueness. No debe haber datos repetidos (de acuerdo a cómo se identifiquen las cosas o personas)

Puntualidad
Timeliness. El nivel en que los datos representan la realidad, desde el correspondiente punto de vista temporal

Validez
Validity. Respetan la sintaxis (formato, tipo, rango) de su definición

Calidad de datos: Completitud

Definición	Medida en que los datos requeridos están en la base de datos
Referencia	Reglas de negocio que definen lo que representa un “100% completo”
Medida	Una medida de la ausencia de blancos (nulos o strings vacíos), o la presencia de strings no vacíos

Calidad de datos: Completitud (II)

La falta de completitud puede afectar:

- filas (muestra sesgada o casos no representados)
- columnas (valores perdidos)

Preguntas:

- ¿Los datos están incompletos? ¿Algunos valores vienen vacíos?
- ¿Hay datos ausentes? ¿Están todos los que son?
- ¿Los que faltan pueden provocar un sesgo?



Calidad de datos: Validez

Definición	Si los datos están bien formados respecto de la definición de los mismos
Referencia	Base de datos, metadatos o reglas de documentación de los tipos permitidos (texto, entero, real, etc.) y su rango (mínimo, máximo) o valores permitidos (nominales o categóricos)
Medida	Comparación entre los datos y los metadatos o documentación relativa a cada ítem

Calidad de datos: Validez (II)

Por ejemplo formatos de emails, teléfonos, fechas, etc.

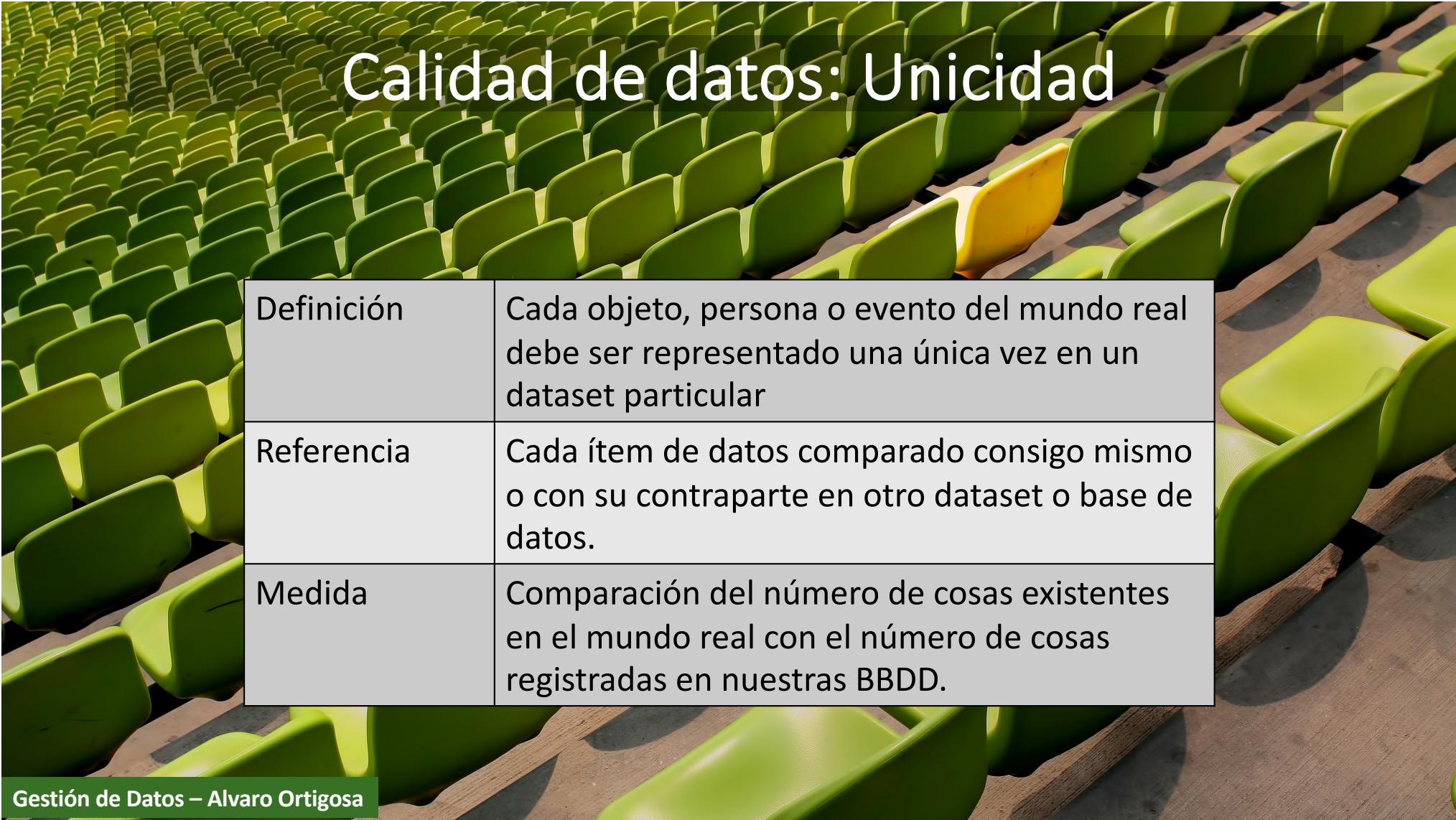
Número de casos limitados en los categóricos.

Muchas veces comprobable con expresiones regulares o estadísticas de frecuencia.

Una vez identificados los datos inválidos, caemos en un problema de completitud.

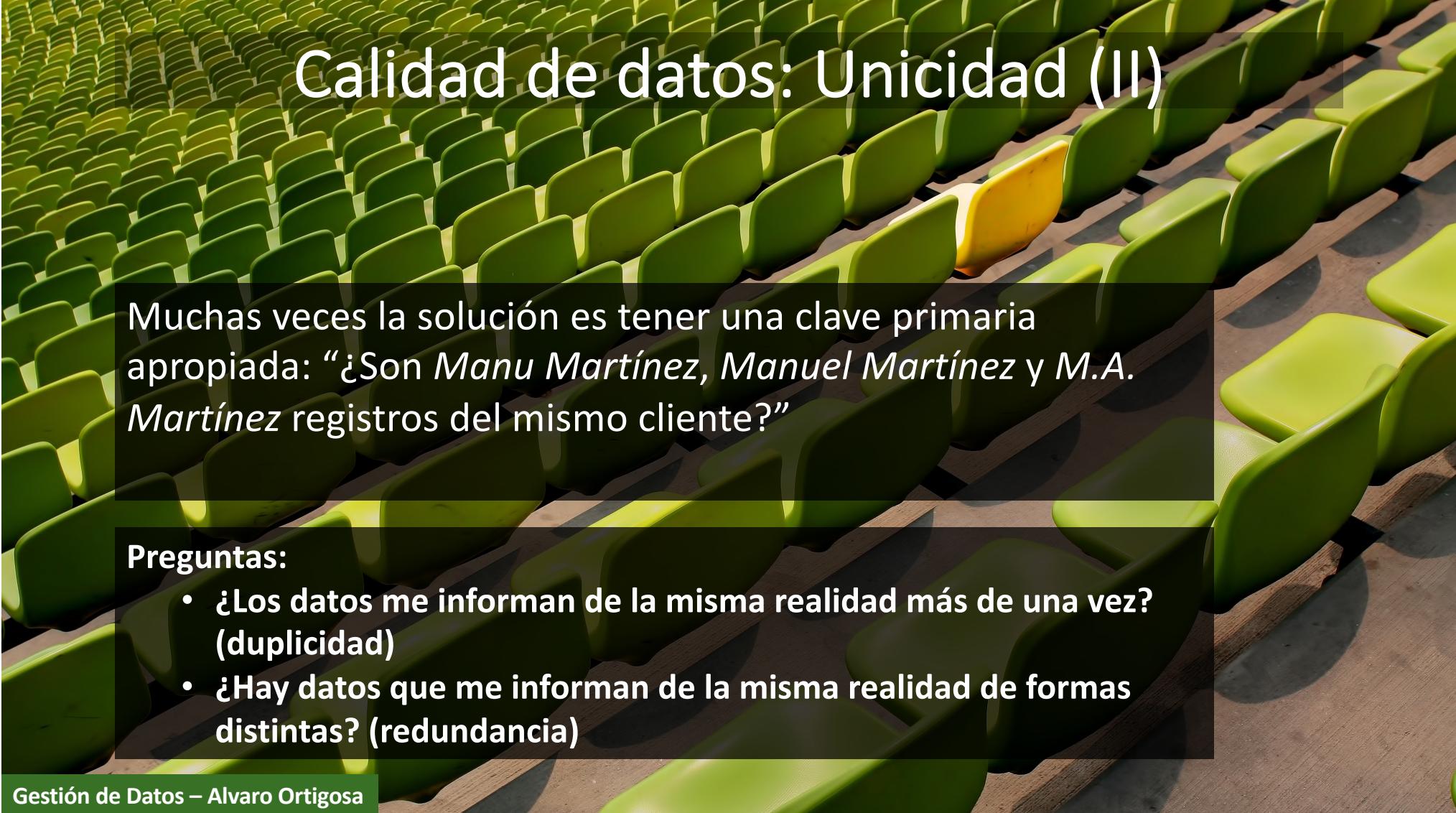
Preguntas:

- **¿Existe más de una forma de representar estos datos? ¿En cual de ellas vienen representados?**
- **¿Existen rangos de valores a los que deben adecuarse? ¿Se cumple?**
- **¿Existen valores atípicos (*outliers*)?**



Calidad de datos: Unicidad

Definición	Cada objeto, persona o evento del mundo real debe ser representado una única vez en un dataset particular
Referencia	Cada ítem de datos comparado consigo mismo o con su contraparte en otro dataset o base de datos.
Medida	Comparación del número de cosas existentes en el mundo real con el número de cosas registradas en nuestras BBDD.



Calidad de datos: Unicidad (II)

Muchas veces la solución es tener una clave primaria apropiada: “*¿Son Manu Martínez, Manuel Martínez y M.A. Martínez registros del mismo cliente?*”

Preguntas:

- **¿Los datos me informan de la misma realidad más de una vez? (duplicidad)**
- **¿Hay datos que me informan de la misma realidad de formas distintas? (redundancia)**

Calidad de datos: Consistencia

Definición	Grado de similitud al comparar 2 o más representaciones del mismo mismo objeto
Referencia	El ítem de dato comparado consigo mismo en otro data set o base de datos
Medida	Análisis de patrones o frecuencia

Calidad de datos: Consistencia

Efecto secundario de la redundancia, y por lo tanto relacionado con unicidad

Preguntas:

- ¿Los datos me informan de realidades incompatibles entre sí?
- ¿Los datos me informan de realidades que son incompatibles con la fuente autorizada?

Calidad de datos: Puntualidad

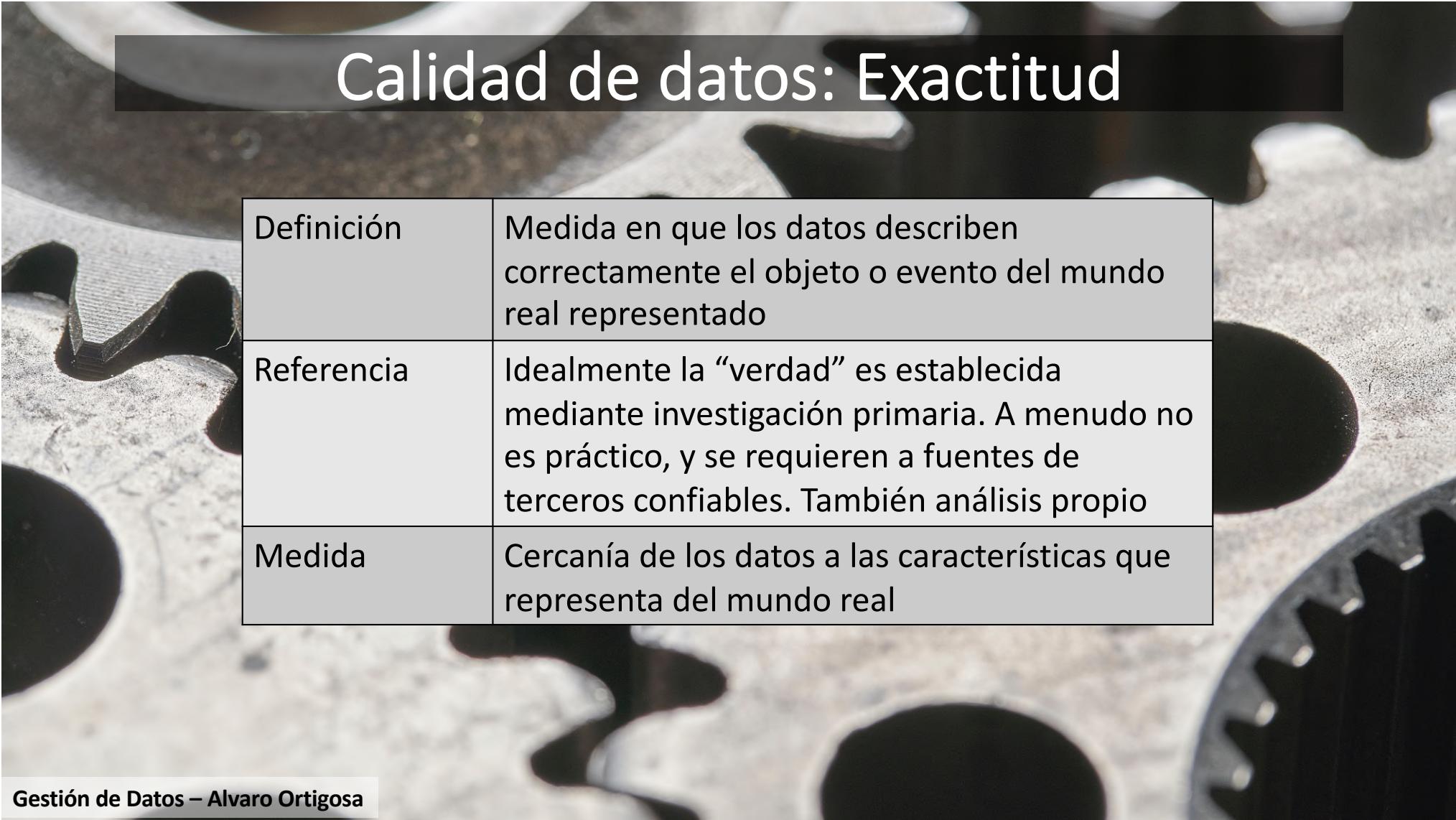
Definición	Nivel de actualidad de los datos
Referencia	Momento en que ocurre el evento registrado en el mundo real
Medida	Diferencia de tiempo

Calidad de datos: Puntualidad

Muchas veces la decisión es hacer las actualizaciones en (casi) tiempo real o en procesamiento por lotes
→ eficiencia Vs. inmediatez

Preguntas:

- ¿Los datos son vigentes?
- ¿A qué momento temporal se refieren?
- ¿En qué momento temporal se calcularon?
- ¿Cómo afecta la variable tiempo a la realidad representada?



Calidad de datos: Exactitud

Definición	Medida en que los datos describen correctamente el objeto o evento del mundo real representado
Referencia	Idealmente la “verdad” es establecida mediante investigación primaria. A menudo no es práctico, y se requieren a fuentes de terceros confiables. También análisis propio
Medida	Cercanía de los datos a las características que representa del mundo real

Calidad de datos: Exactitud (II)

Cada paso desde el mundo real al dataset debe preservar la esencia del original.

Especial atención a la toma de datos

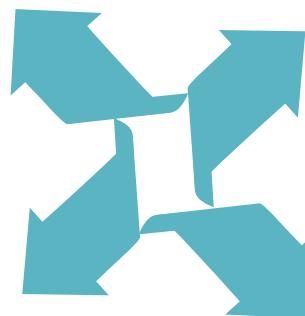
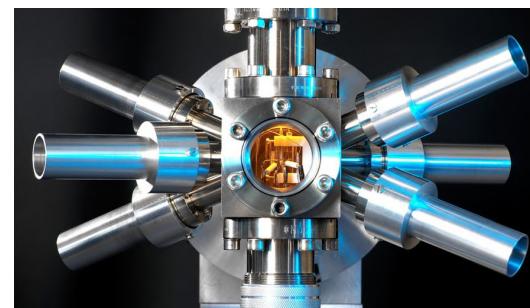
Demuestra la importancia de análisis básico y perfilado para entender los datos

Preguntas:

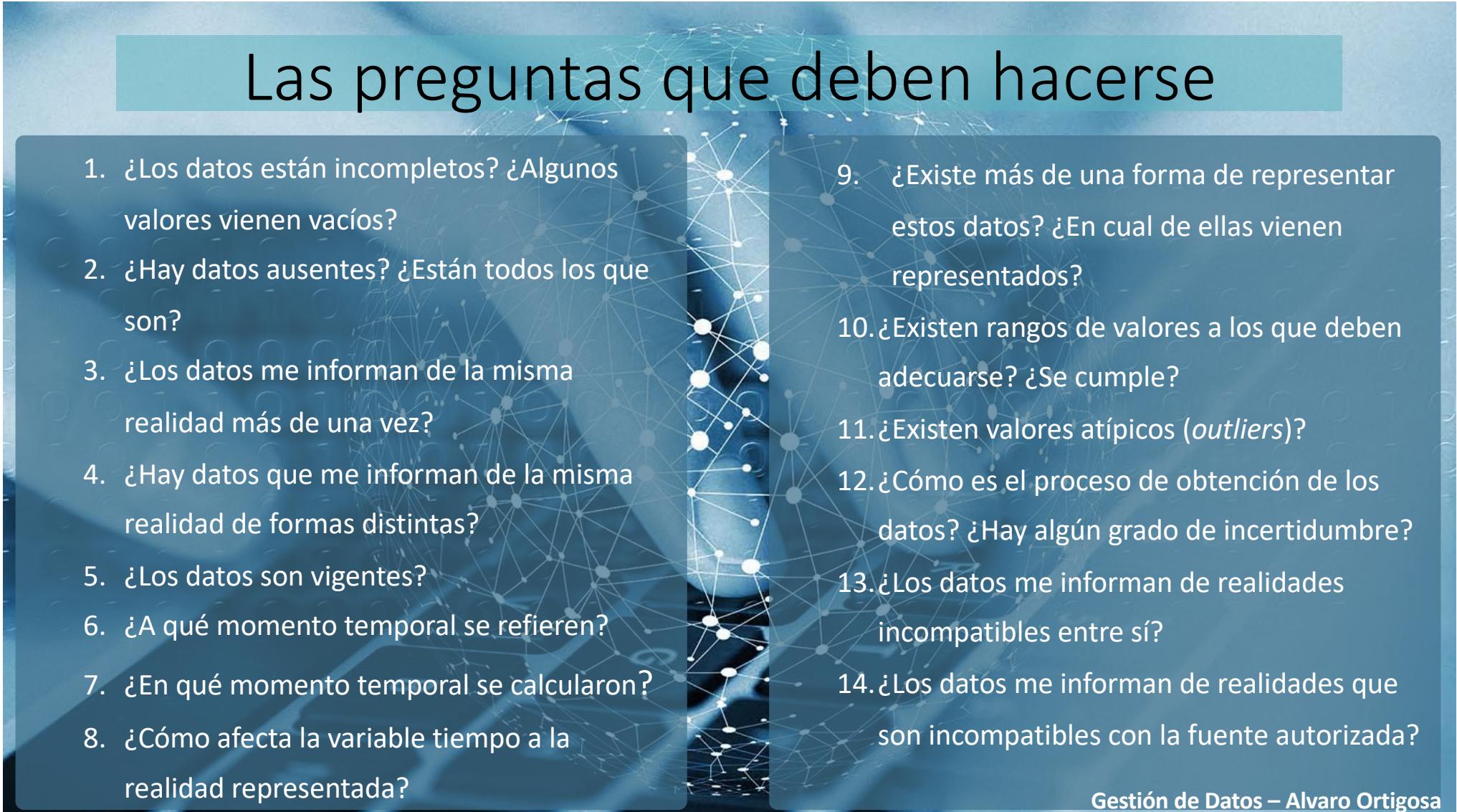
- ¿Cómo es el proceso de obtención de los datos? ¿Hay algún grado de incertidumbre?
- Ojo, también (¡y especialmente!) en datos cualitativos.



Relacionado: Precisión



Las preguntas que deben hacerse

- 
1. ¿Los datos están incompletos? ¿Algunos valores vienen vacíos?
 2. ¿Hay datos ausentes? ¿Están todos los que son?
 3. ¿Los datos me informan de la misma realidad más de una vez?
 4. ¿Hay datos que me informan de la misma realidad de formas distintas?
 5. ¿Los datos son vigentes?
 6. ¿A qué momento temporal se refieren?
 7. ¿En qué momento temporal se calcularon?
 8. ¿Cómo afecta la variable tiempo a la realidad representada?
 9. ¿Existe más de una forma de representar estos datos? ¿En cual de ellas vienen representados?
 10. ¿Existen rangos de valores a los que deben adecuarse? ¿Se cumple?
 11. ¿Existen valores atípicos (*outliers*)?
 12. ¿Cómo es el proceso de obtención de los datos? ¿Hay algún grado de incertidumbre?
 13. ¿Los datos me informan de realidades incompatibles entre sí?
 14. ¿Los datos me informan de realidades que son incompatibles con la fuente autorizada?

Las preguntas que suelen hacerse



- ¿Me fío del proveedor de los datos?
 - Sí, claro, faltaría más → **No miro nada.**
- Si el análisis sale mal por problemas de calidad de datos ¿Puedo echarle la culpa a otro?
 - Sí, *no es mi culpa* si la “materia prima” es mala → **No miro nada.**
- Preferimos “confiar” porque...
 - Analizar la calidad cuesta tiempo y esfuerzo; y ...
 - ... podemos descubrir cosas que no queremos descubrir

Las preguntas que suelen hacerse



- ¿Me fío del proveedor de los datos?
 - Sí, claro, faltaría más → **No miro nada.**
- Si el análisis sale mal por problemas de calidad de datos ¿Puedo echarle la culpa a ...?
 - Sí, no es mi culpa si la "inerzia" es mala → **No miro nada.**
- Preferimos confiar porque...
 - Analizar la calidad cuesta tiempo y esfuerzo; y ...
 - ... podemos descubrir cosas que no queremos descubrir

Posibles dimensiones a considerar

- Informe reciente del capítulo danés de *Dana international* describe **60 dimensiones** de calidad de los datos.
- Obviamente **imposible** considerar todas en un proyecto.
- Debe **seleccionar las relevantes** (para nuestro proyecto).
 - Determinar si una dimensión **contribuye** lo suficiente para los **objetivos de negocio**

Fuentes de materiales

- DAMA International (<https://www.dama.org/>)
- How to Select the Right Dimensions of Data Quality. Dana Netherlands.
- J.M. Juran, A.B. Godfrey. Juran's Quality Handbook
- Towards Data Science
 - The Six Dimensions of Data Quality — and how to deal with them (<https://towardsdatascience.com/the-six-dimensions-of-data-quality-and-how-to-deal-with-them-bdcf9a3dba71>).
- David Taieb. Data Analysis with Python.
- Fotos: la mayoría de Unsplash (<https://unsplash.com/>)