Calidad y limpieza de datos

Definiciones y aspetos teóricos

Néstor Beltrán

Basada en:
Saraiva & Koch, 2016.
Koch, 2017.

- La presión ejercida sobre la biodiversidad del planeta continuará en incremento y el estado de la biodiversidad seguirá deteriorándose.

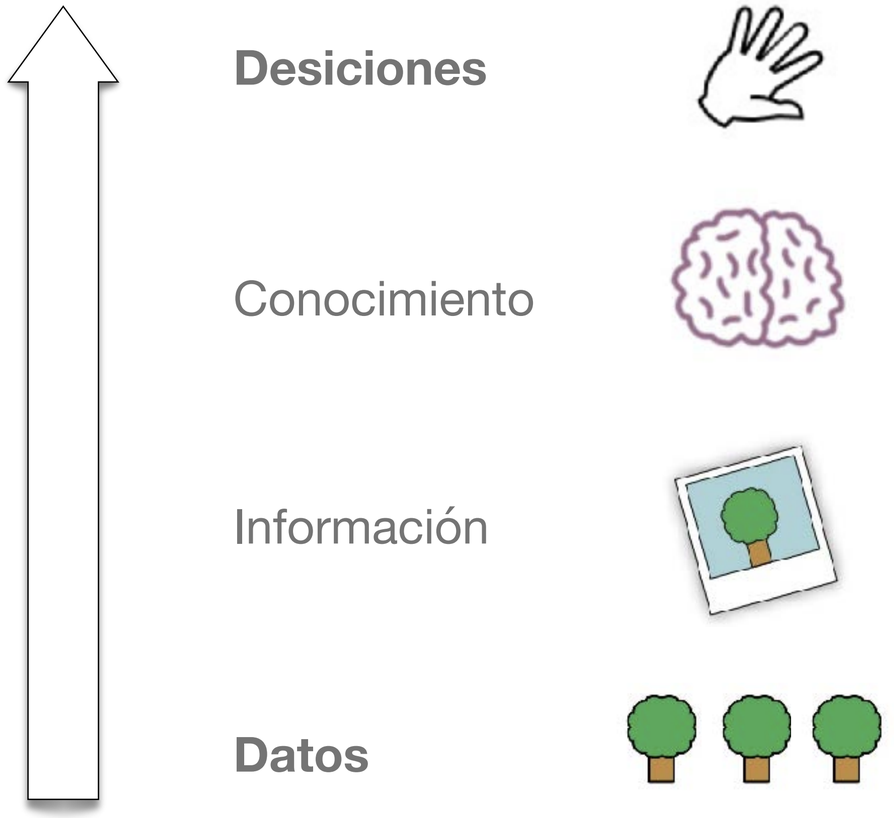
- Han ocurrido avances pero, en la mayoría de los casos, estos no han sido suficientes para alcanzar las metas del 2020.

Perdida global de la biodiversidad

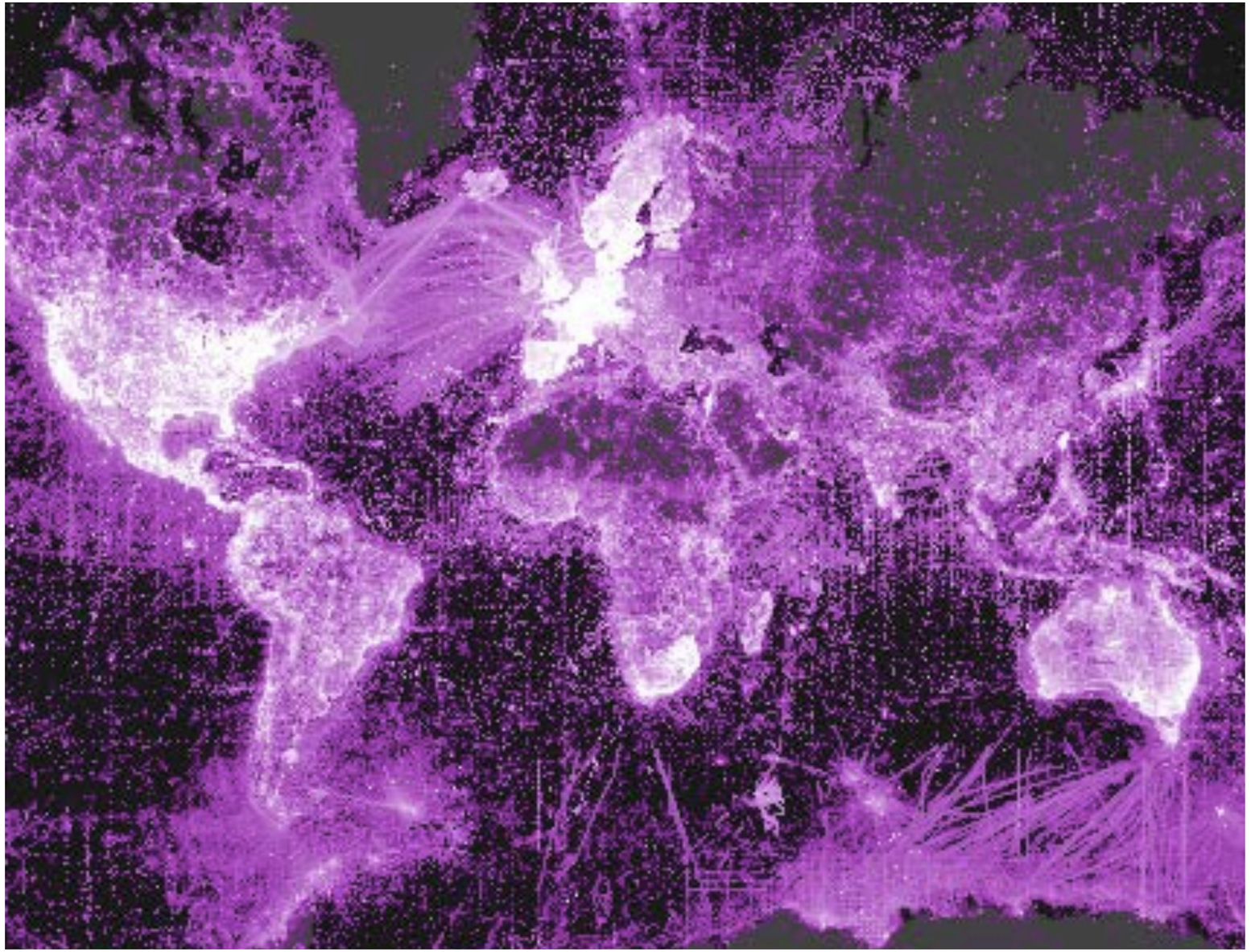
Plan Estratégico para la Diversidad Biológica 2011-2020 y las Metas Aichi

Compartir datos, desarrollar **indicadores** y **medidas,** fomentando la **generación** y **uso** de información científica. **Para sostener** la nueva plataforma intergubernamental científico-normativa sobre diversidad biológica y servicios de los ecosistemas (IPBES)

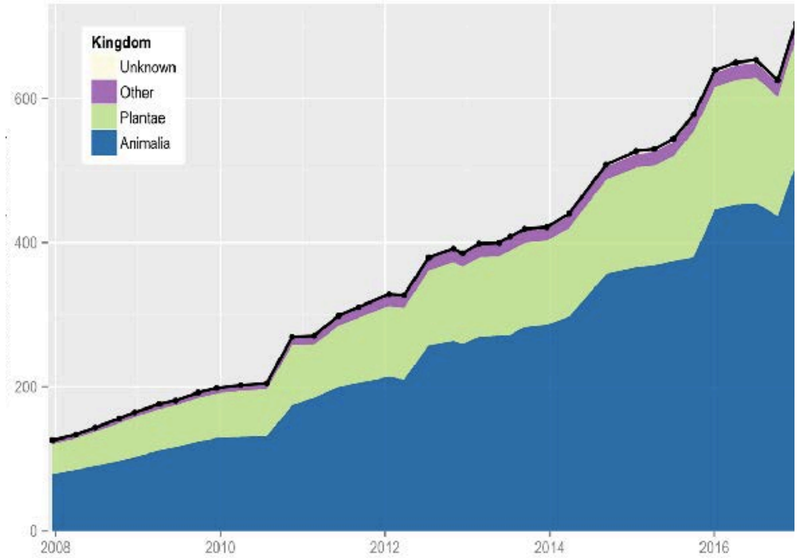
Debemos ir de los datos a la toma de decisiones



En este momento tenemos una gran cantidad de información



GBIF - Global Biodiversity Information facility



Red mundial de registros biológicos

727.113.682

Registros

35.164

Datasets

1203

Publicadores

Sin embargo, existen vacíos de información

- ¿Cuántas y cuáles especies tenemos?
- ¿Cuál es el tamaño poblacional y las dinámicas?
- ¿Cuál es su distribution espacial y temporal?
- ¿Cómo afectan las condiciones bióticas y abióticas?

¡Necesitamos más!

- Regiones poco estudiadas o representadas.
- Trabajo de campo y laboratorio.
- Apoyo y financiación a las colecciones biológicas.

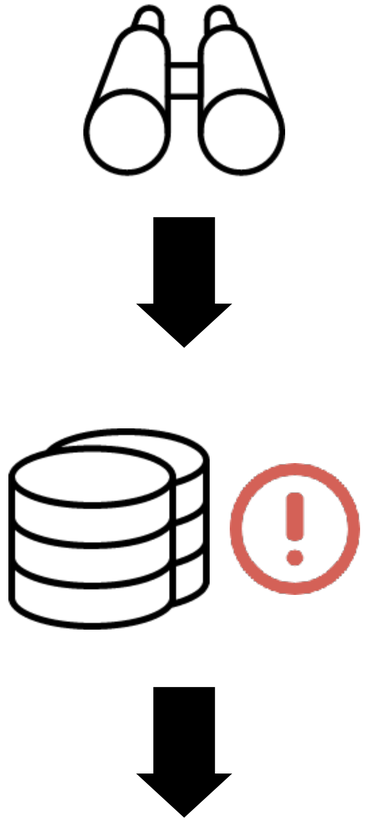
¡Necesitamos más!

Sacar provecho a los datos existentes

- Una gran cantidad no está disponible:
    - No digitalizado, no compartido
    - Difícil de acceder
    - **Problemas de calidad**

Basura entra - Basura sale

- **Problemas de calidad:** conllevan a resultados de mala calidad: análisis, decisiones, etc.
- **Los problemas surgen de:** toma de datos, digitalización, falta de metadatos, ausencia de estándares.
- **Hay mucho por hacer:** limpieza de datos (corrección), prevención y políticas de calidad de datos.

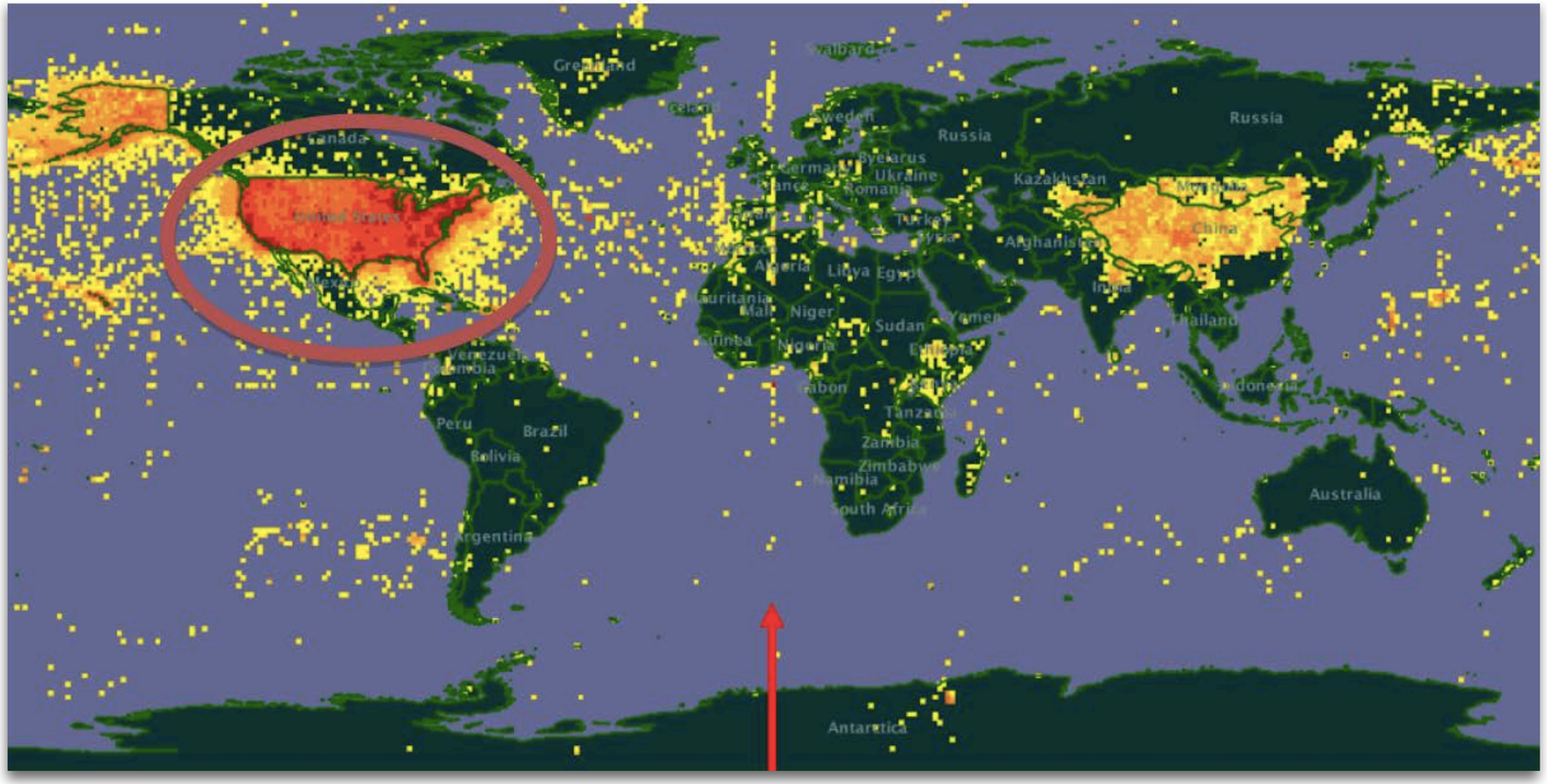


☑ Artículos científicos

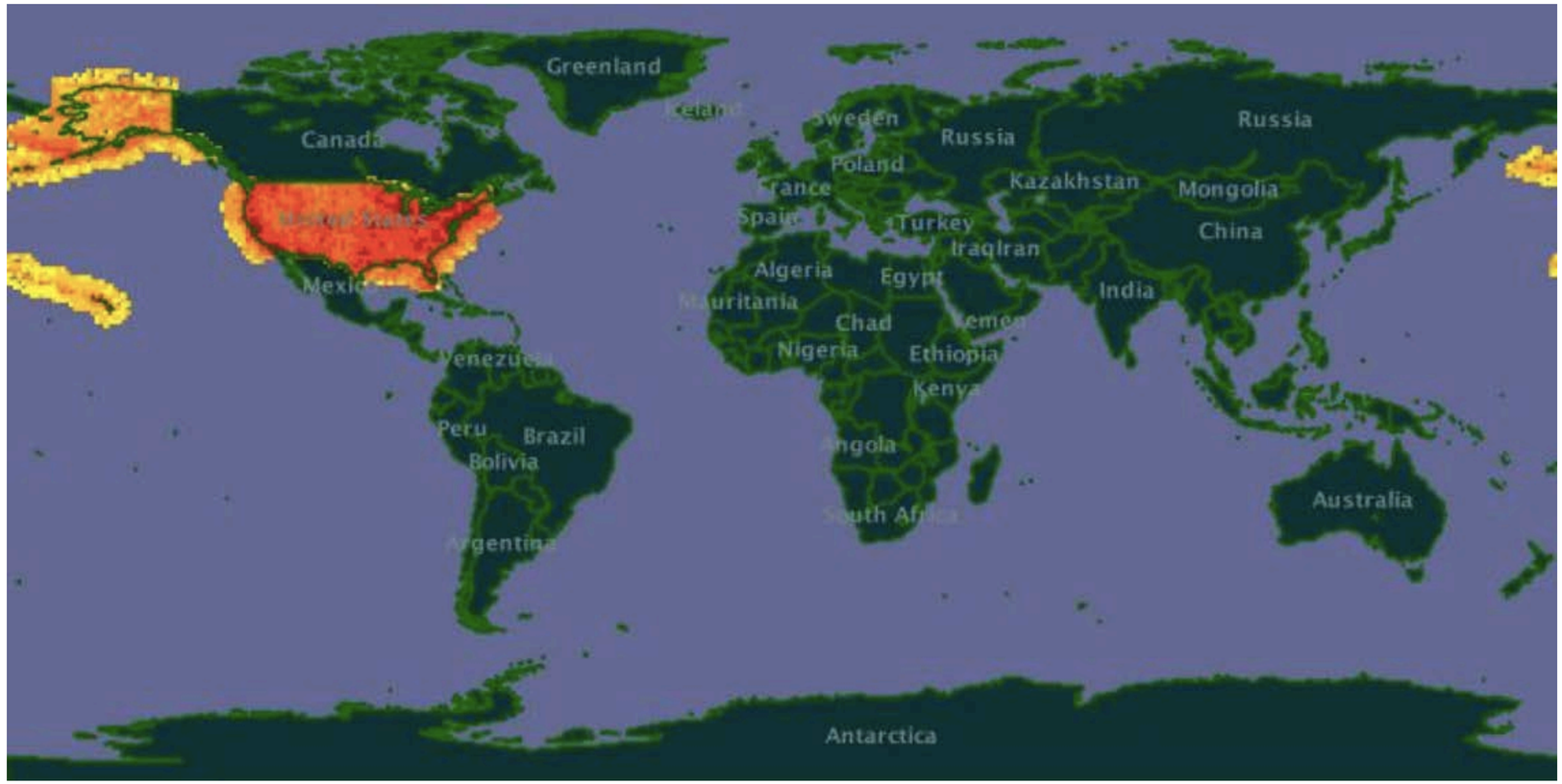
☑ Modelamiento y análisis

☑ Políticas de conservación

Ejemplo


Russia
Kazakhstan
China
Australia
Antarctica
Brazil
Peru
Bolivia
Sudan
Niger
Zambia
Zimbabwe
Ukraine
Belarus
Romania
Sweden
Thailand

Ejemplo


Greenland
Canada
Sweden
Russia
Poland
Kazakhstan
Mongolia
China
Turkey
Iraq
Iran
Algeria
Egypt
India
Chad
Nigeria
Ethiopia
Kenya
Peru
Brazil
Bolivia
Australia
Antarctica

La calidad de datos afecta indicadores, análisis y políticas

Red List Index - IUCN
Especies amenazadas -IUCN
Reporte de estado y tendencia - IAvH

Calidad de datos

Algunos conceptos

- **Información**: *morfè* (forma) / *éidos* (concepto)
  - Dar forma a la esencia de algo
  - Es la **representación** de la realidad
  - La realidad es diferente de la “representación de la realidad”

Algunos conceptos

- Existe una **brecha** entre la **representación de la realidad** y la **realidad misma**, la cual se puede medir en ciertas dimensiones:
  - Completitud
  - Precisión
  - Consistencia
  - Exactitud
  - Etc.

Calidad de datos

Definición #1

Los datos tienen calidad si la información derivada de estos representa correctamente el mundo real (hechos).

Información

Datos

Mundo real

Ejemplo

Dato1: *Saguinus*

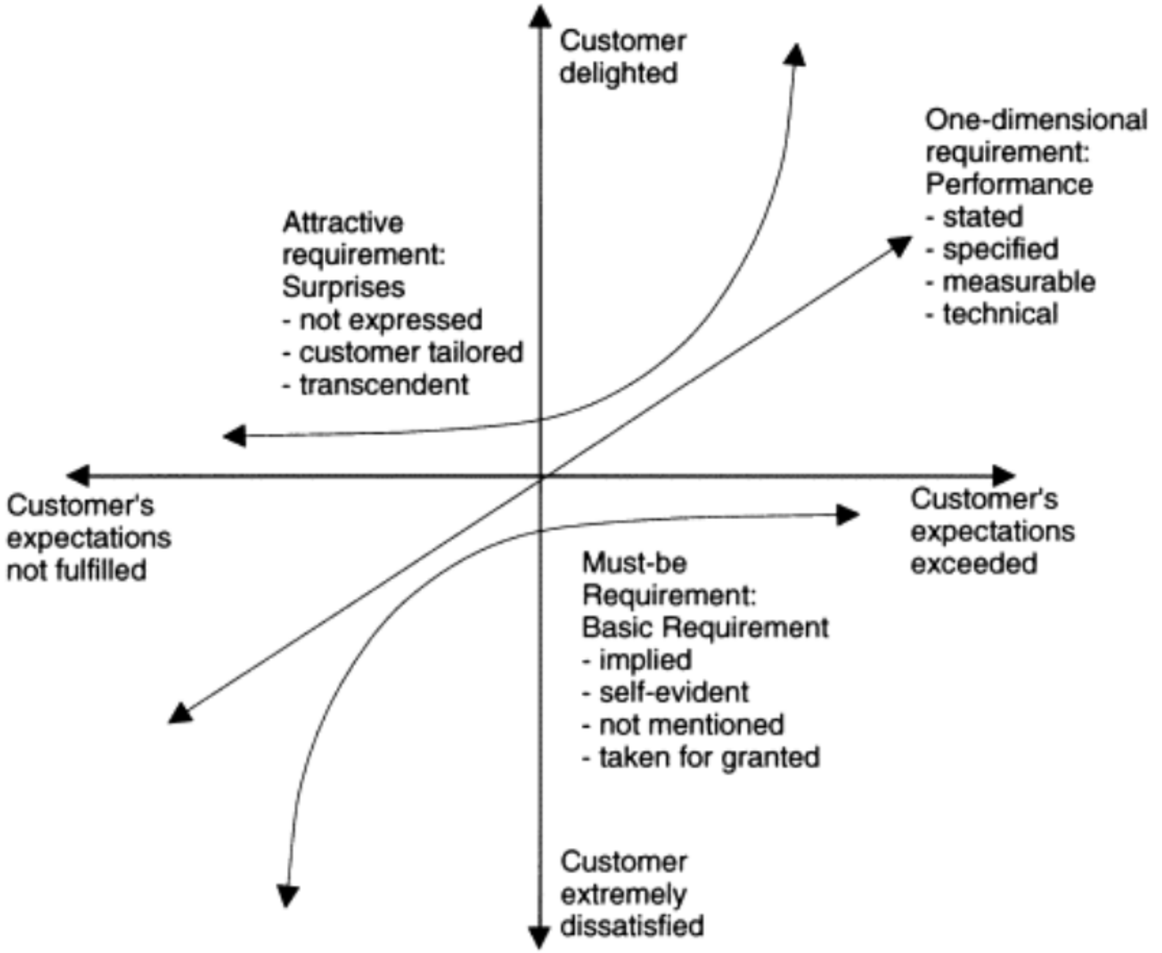
Dato 2: Mico tití

¿tienen la misma calidad?

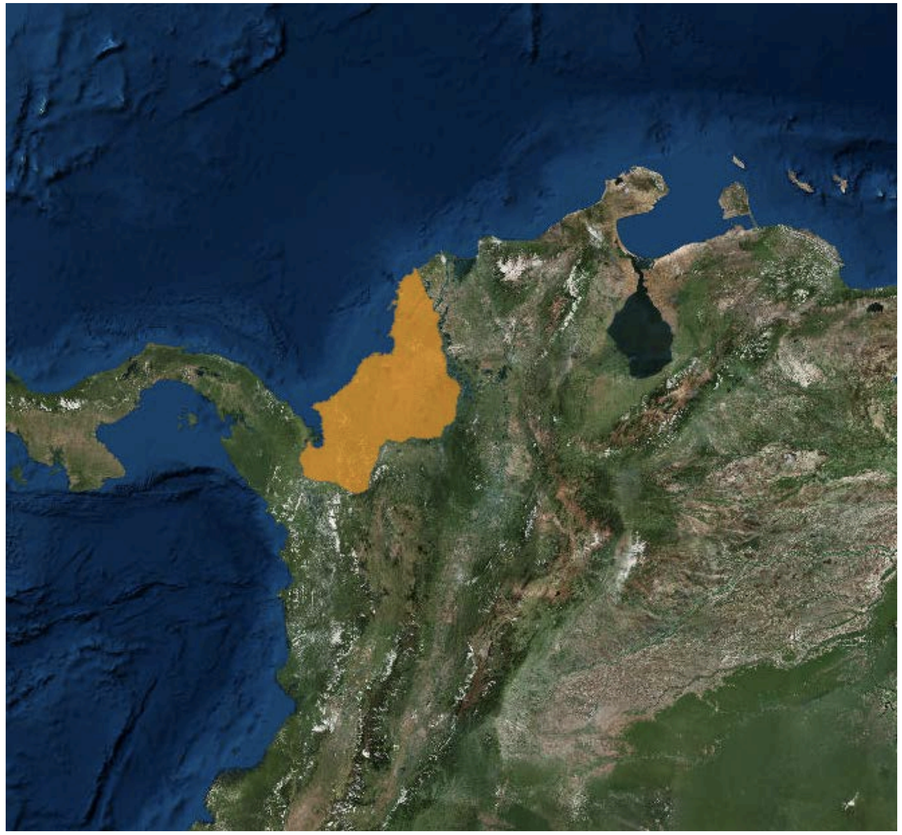
Calidad de datos

Definición #2

Satisfacción del consumidor. Si un consumidor está satisfecho con un servicio producto, este servicio o producto tiene calidad para este consumidor.



Ejemplo



Requerimiento: el dato debe tener nombre científico y debe ser suministrado a nivel de especie

Nombre: *Saguinus*
Categoria*: Genero*

¿este dato tiene calidad?

¿puede ser usado en el estudio de la distribución de primates en Suramérica?

Calidad de datos

Definición #3

Usabilidad. Un dato tiene calidad si es adecuado para ser usado. Si el dato no sirve para el propósito del que lo usa, **puede ser útil para otros**.

Algunos conceptos

La calidad de datos es un concepto **idiosincrásico**

”La idiosincracia es algo distintivo y propio de un individuo”

Definir **calidad de datos** es similar a definir qué es bonito, bueno divertido o valioso.

Algunos conceptos

La palabra clave y la definición mas aceptada:

Usabilidad de los datos

Calidad en relación a un propósito.

- *Modelos de distribución.*

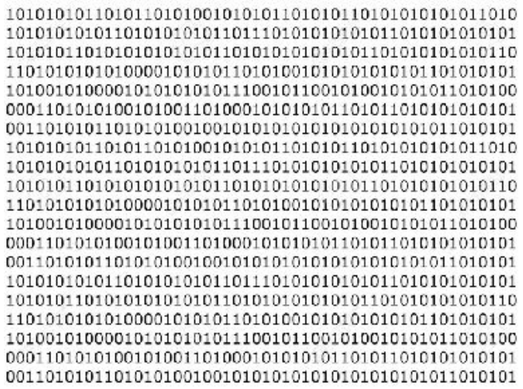
Para cada propósito existe un tipo de datos.

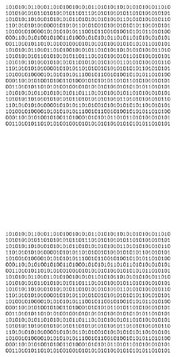
- *Con coordenadas y nombres de especies.*

Para cada tipo de dato existen atributos a cumplir.

- *Completitud, consistencia, precisión, exactitud, etc.*

Evaluación de la calidad y **manejo** de la calidad





Datos no aptos para el uso

Datos aptos para el uso

Evaluación de la calidad

El objetivo es identificar los **problemas** que degradan la calidad de una **dimension**

particular en un **dominio** especifico

Los datos tienen calidad cuando no hay problemas que **degradan** la misma

Evaluación de la calidad

El objetivo es identificar los **problemas** que degradan la calidad de una **dimension**

particular en un **dominio** especifico

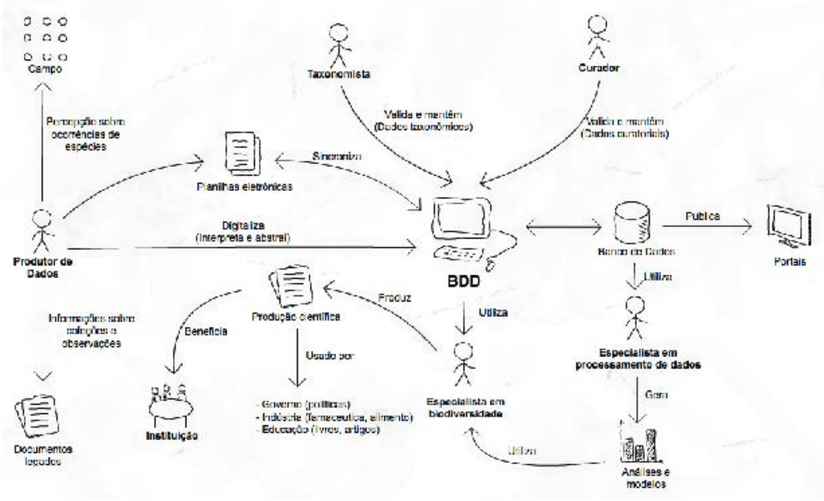
Los datos tienen calidad cuando no hay problemas que **degradan** la misma

Ejemplo: ¿Cuáles son los problemas que degradan la precisión de los datos geoespaciales?

Evaluación de la calidad

dominio

Son las **clases** de los datos que involucran la representación de un **aspecto** del mundo real.

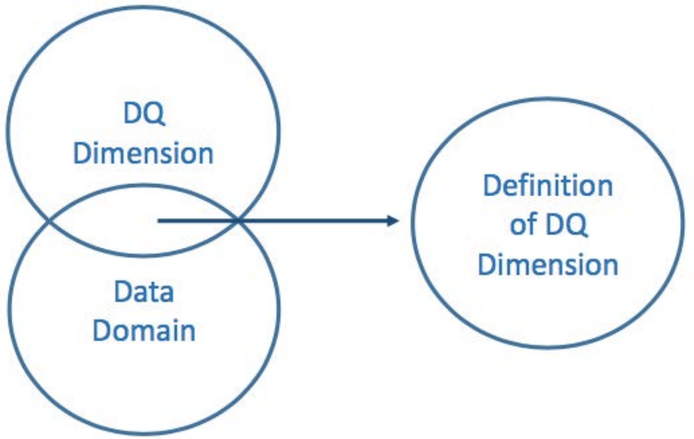


Ejemplo: ¿Cuáles son los problemas que degradan la precisión de los datos **geoespaciales**?

Evaluación de la calidad

dimension

Siendo la calidad de datos un concepto **multidimensional**, una dimensión es el **aspecto medible** de la calidad del dato.



Ejemplo: ¿Cuáles son los problemas que degradan la **precisión** de los datos geoespaciales?

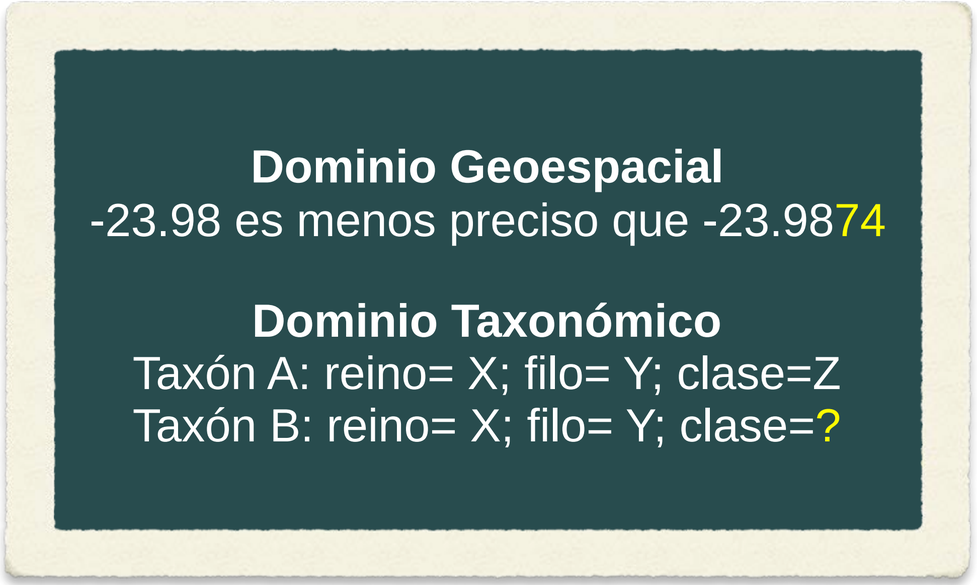
Evaluación de la calidad

dimension

Siendo la calidad de datos un concepto

multidimensional, una dimensión es el **aspecto**

medible de la calidad del dato.



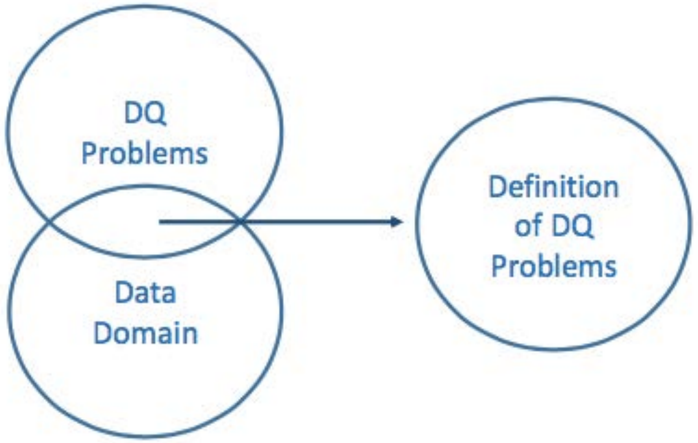
Ejemplo: ¿Cuáles son los problemas que degradan la **precisión** de los datos geoespaciales?

Evaluación de la calidad

problemas

Todo lo que pueda **degradar** la calidad para una o

mas **dimensiones**.



Ejemplo: ¿Cuáles son los **problemas** que degradan la precisión de los datos geoespaciales?

Evaluación de la calidad

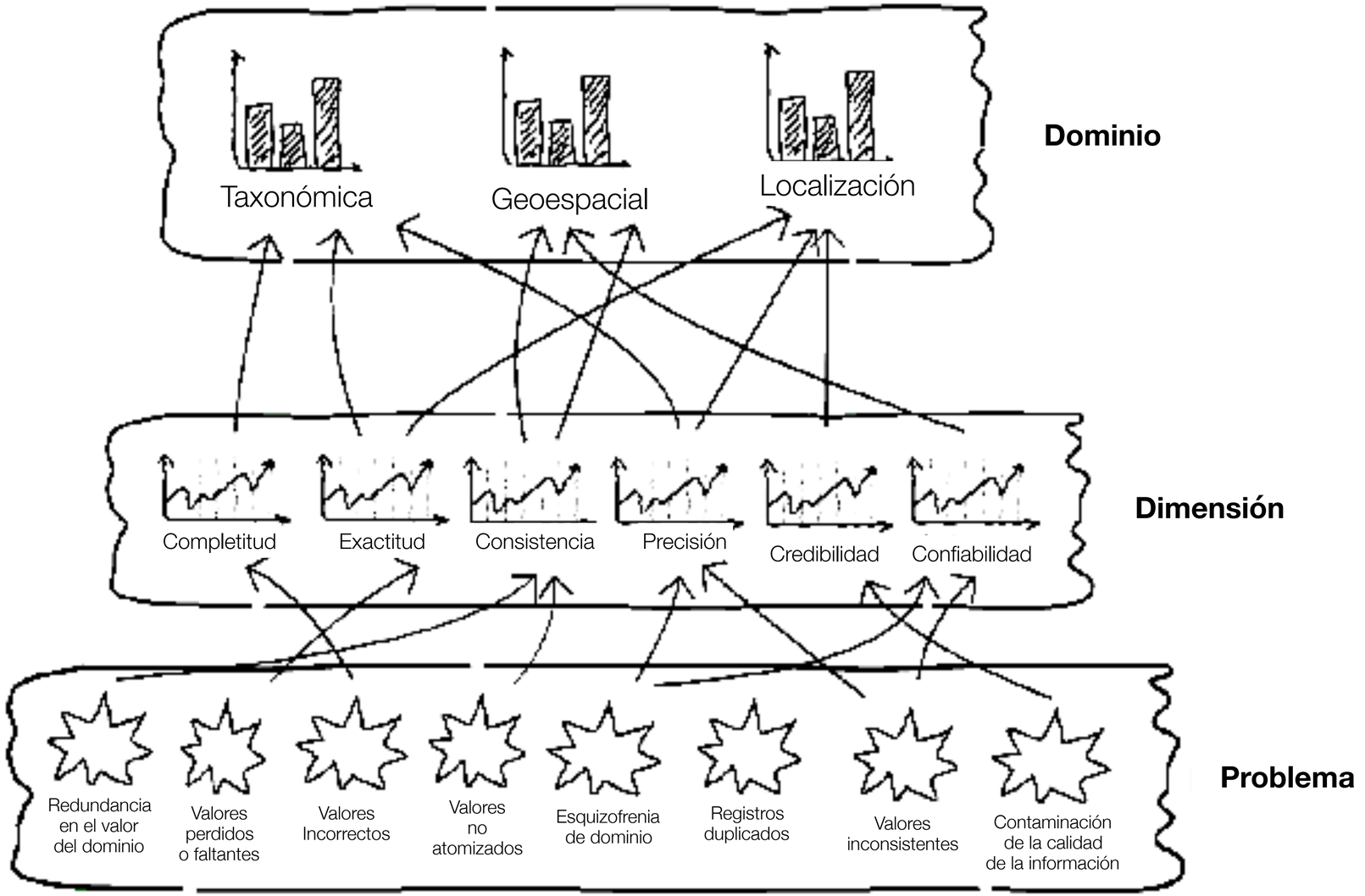
problemas

Todo lo que pueda **degradar** la calidad para una o mas **dimensiones**.

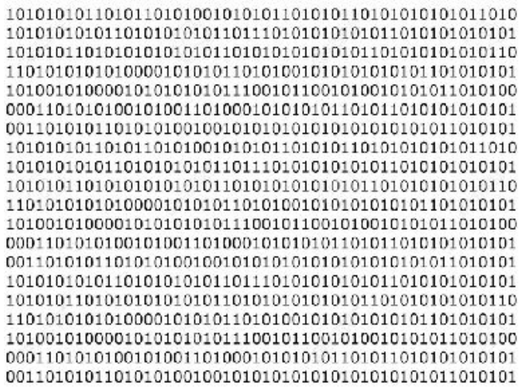
- **Domain value redundancy**: Ex.: Brazil x Brasil, São Paulo x Sao Paulo.
- **Missing data value**: absence of latitude or longitude.
- **Incorrect data values**: mispelling errors.
- **Nonatomic data values**: Ex.: São Paulo, SP.
- **Domain schizophrenia**: latitude and longitude equal "0" (zero) when they are not known.
- **Duplicate occurrences**: More than one record representing the same fact.
- **Inconsistent data values**: Ex.: lat=43; long=44; country=Brazil.
- **Information quality contamination**: create a new data based on an already existent wrong data.

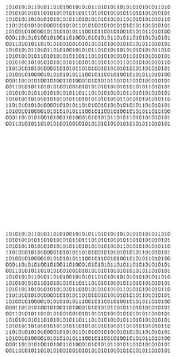
Ejemplo: ¿Cuáles son los **problemas** que degradan la precisión de los datos geoespaciales?

Evaluación de la calidad



Evaluación de la calidad y **manejo** de la calidad





Datos no aptos para el uso

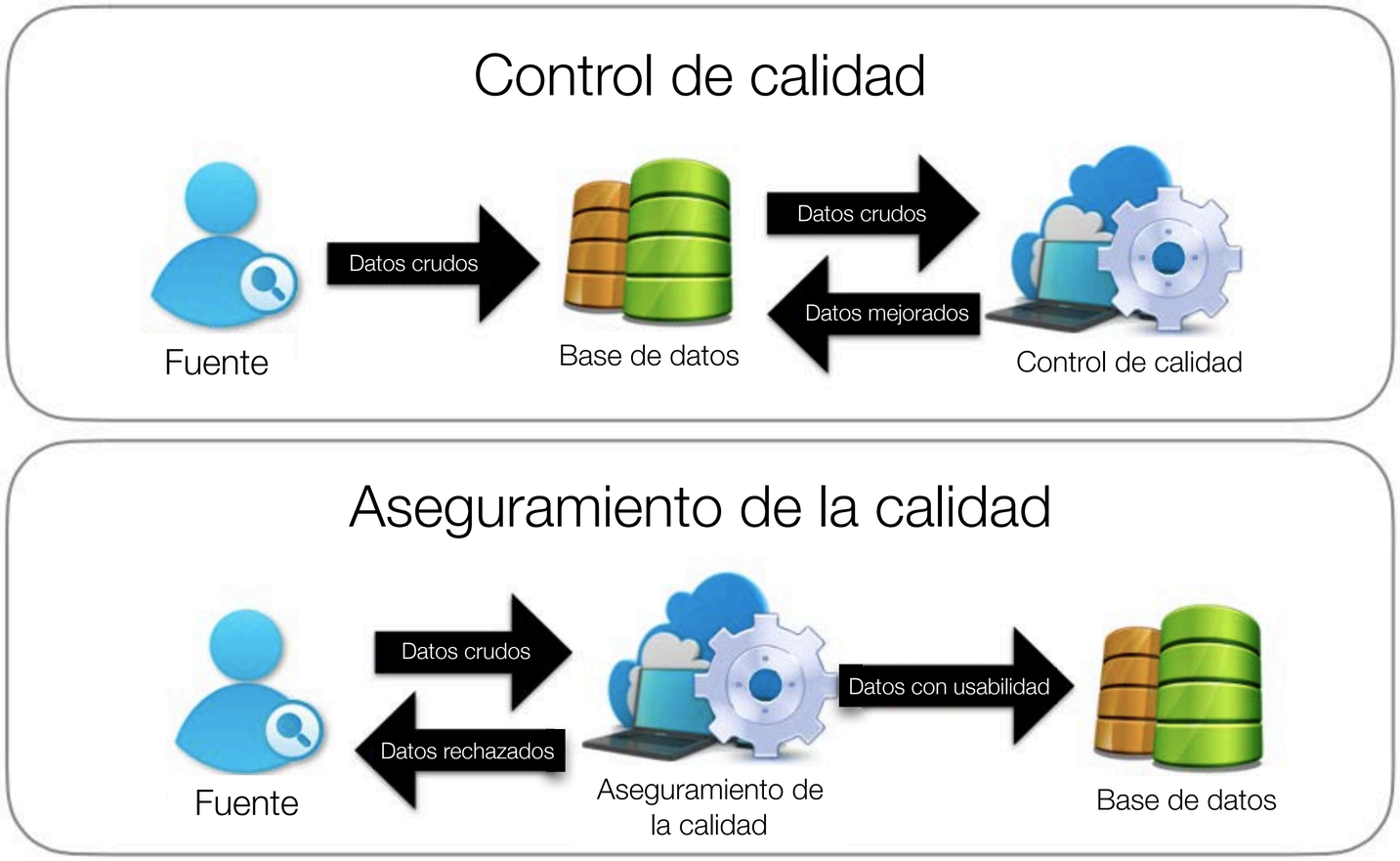
Datos aptos para el uso

Manejo de la calidad

El objetivo es **evitar** los **problemas** que degradan la calidad de una **dimension**

Los datos tienen calidad cuando están **libres** de defectos

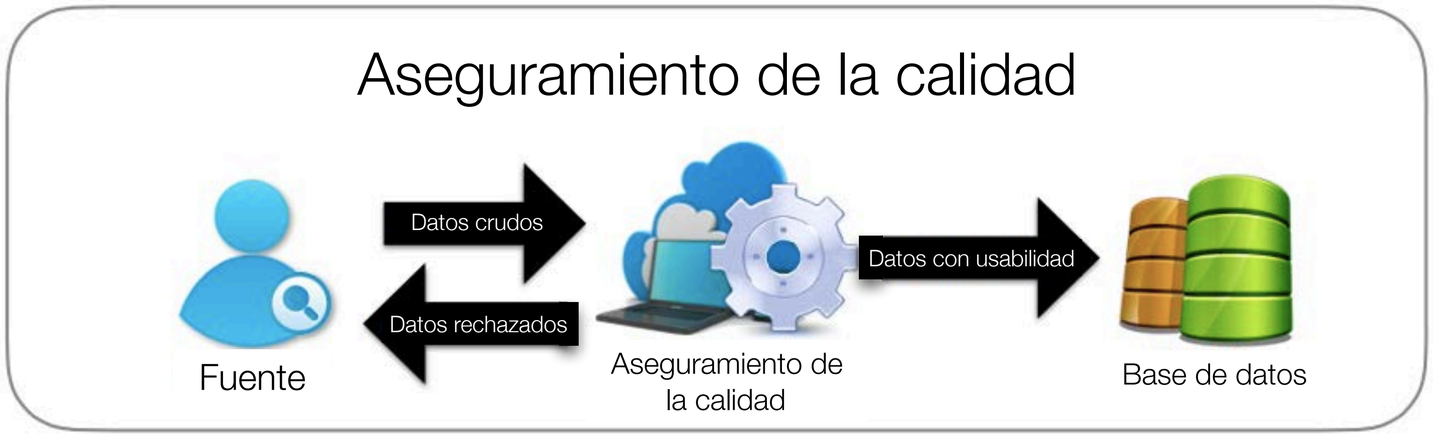
Manejo de la calidad



Koch, 2017

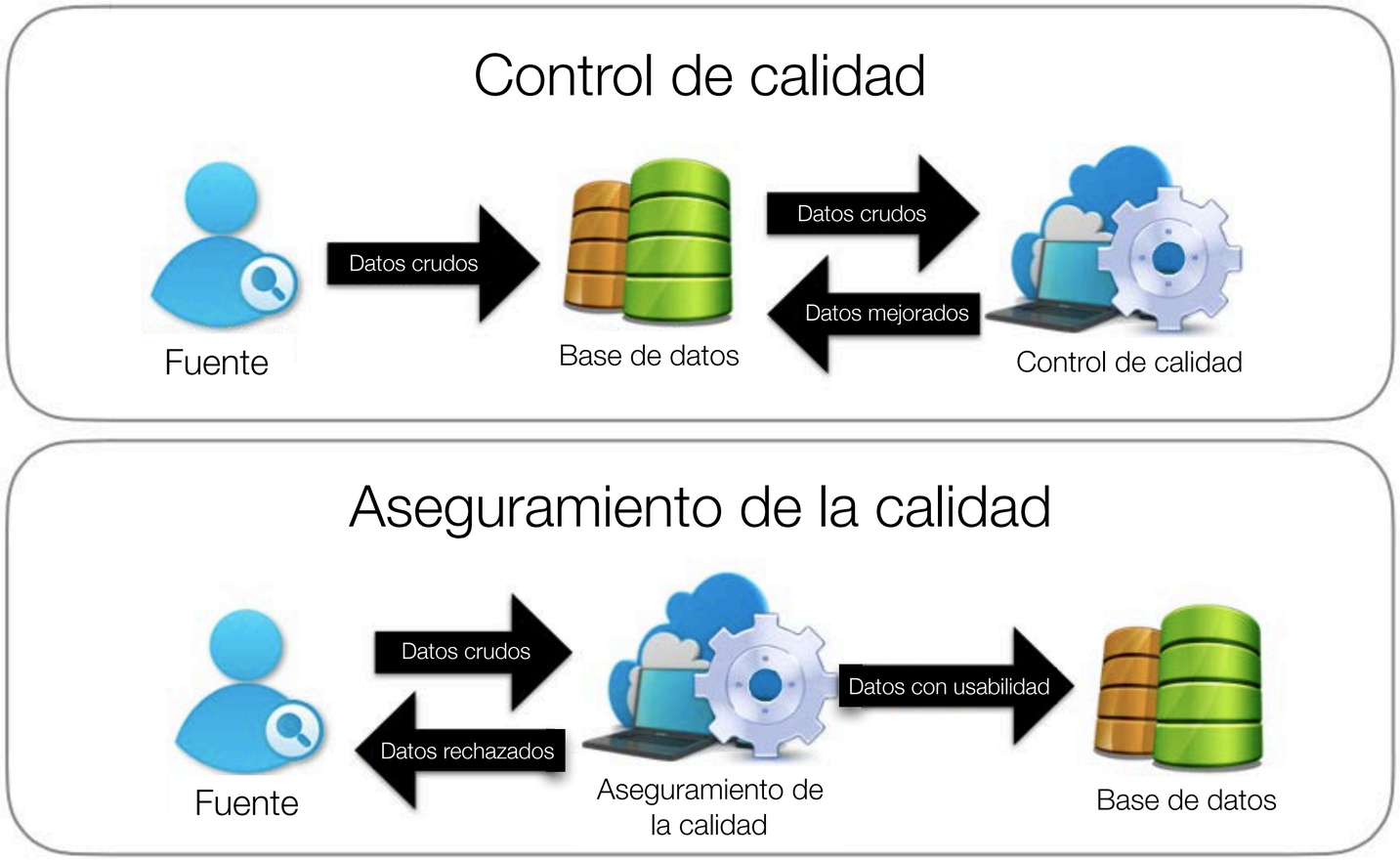
Manejo de la calidad

- Siempre es mejor **prevenir**: se debe evitar el error incluso antes de la misma construcción de los datos.
- Detección - corrección - documentación



Koch, 2017

Manejo de la calidad



Koch, 2017

Limpieza de datos

Es un proceso que mejora la calidad a través de

la corrección de los errores detectados

- Determinar el tipo
- Buscar e identificar
- Corregir
- Documentar **todo**
- Modificar la toma y manejo para prevenir futuros errores

Ejercicio practico

Limpieza de datos

Tipos de errores

Errores técnicos

- Completitud
  ¿Todos los elementos están presentes?
- Rangos y limites
  ¿La latitud está entre -90 y 90)
- Tipo de dato
  ¿El elemento Fecha tiene fechas o texto?
- Formato de los datos
  ¿Las medidas cumplen el formato?

Tipos de errores

Errores de consistencia

- Taxonómicos
  ¿Tiene una especie reportada el genero y el epíteto especifico?

- Continuidad
  ¿Existe una línea temporal clara de las fechas de recolección?

- Valores atípicos
  ¿hay alturas mayores a 6.962 m.s.n.m en Argentina?

- Geográficos
  ¿Están las coordenadas dentro de la localidad o región identificada?

Calidad y limpieza de datos

Definiciones y aspetos teóricos

Néstor Beltrán

Basada en:
Saraiva & Koch, 2016.
Koch, 2017.