CALIDAD DE DATOS

Diego Rosselli

Objetivos

- Introducción Calidad Datos 1

- Estrategia aseguramiento calidad 2

- Lecciones aprendidas 3

Objetivos

- Introducción Calidad Datos 1

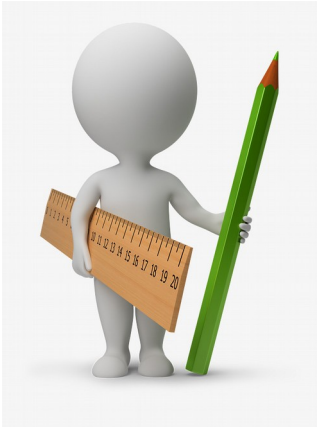
- Estrategia aseguramiento calidad 2

- Lecciones aprendidas 3

Introducción

Calidad:

- La percibimos
- La definimos
- La medimos



Calidad datos: ¿algo nuevo?

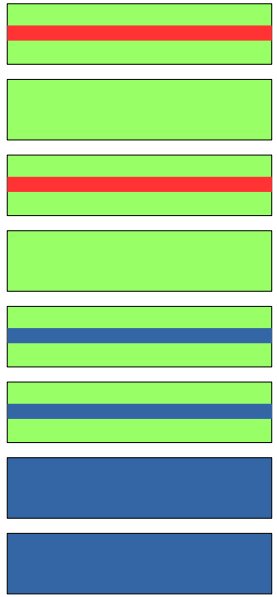
Reporte

A	B	C
1	1	2
2	1	1

F	G	H
a	c	d

D	E
1	1
4	2

Datos



Gestión de calidad datos

Adecuación al uso

- Calle, número puerta
- Calle, esquina, destinatario
- Plano

Calidad - Valor - Expectativas

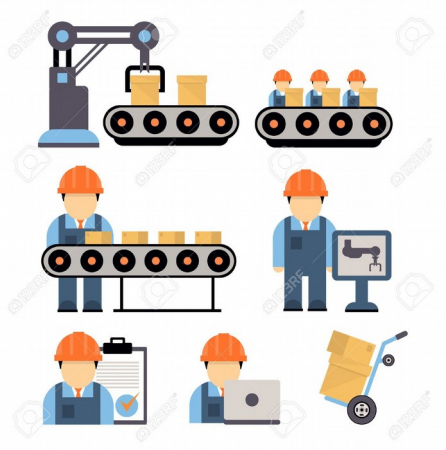
Puede ser correcto, preciso, actualizado y **no colaborar** con las **necesidades** del usuario.

Calidad <> Precisión

Puede no alcanzar o por el contrario superar las **expectativas** del usuario.

Causas de problemas de calidad

- Producción: ingreso manual, sensores, unificar diferentes fuentes



Causas de problemas de calidad

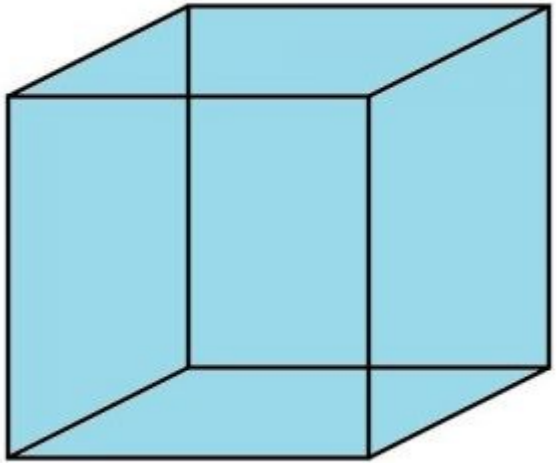
- Almacenamiento: ausencia de formato común, diseño inadecuado de base de datos

Causas de problemas de calidad

- Utilización: actualización, cambios de requerimientos, errores de interpretación

Multi-dimensión de la calidad de datos

- Completitud
- Unicidad
- Exactitud
- Consistencia
- Actualidad
- Otras: ISO/IEC 25012, ISO 8000



Dimensión Unicidad

CI	Nombre	Calle	Nro.	F.Nto.	Edad
1	José	Calle A	1	A	30
11	José	Calle A	1	A	30
1	José	Calle A	1	A	30
3	María	Calle B	2	B	28
4	Juana	Calle C	3	C	29

Unicidad: Medición y Registro

CI	Nombre	Calle	Calidad de cada línea			
1	José	Calle A	1	A	30	0
11	José	Calle A	1	A	30	0
1	José	Calle A	1	A	30	0
3	María	Calle B	2	B	28	1
4	Juana	Calle C	3	C	29	1

Unicidad: Medición y Registro

1

CI	Nombre	Calle				Calidad de cada línea
1	José	Calle A	1	A	30	3
11	José	Calle A	1	A	30	3
1	José	Calle A	1	A	30	3
3	María	Calle B	2	B	28	1
4	Juana	Calle C	3	C	29	1

Dimensión Completitud

CI	Nombre	Calle	Nro.	F.Nto.	Edad
1	José	Calle A	1	A	30
2	Pepe			C	
4	Juana	Calle B	2		20
3	María	Calle C		D	40

Falta

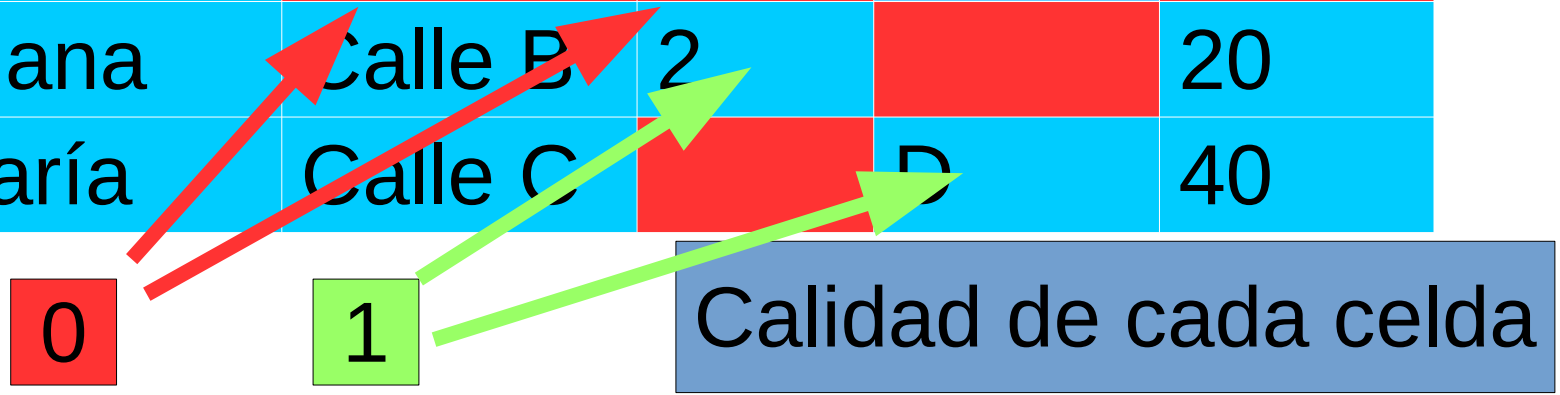
Inferir

NC

Estimar

Completitud: Medición y Registro

CI	Nombre	Calle	Nro.	F.Nto.	Edad
1	José	Calle A	1	A	30
2	Pepe			C	
4	Juana	Calle B	2		20
3	María	Calle C		D	40



Dimensión Completitud

CI	Nombre	Calle	Nro.	F.Nto.	Edad
1	José	Calle A	1	A	30
2	Pepe			C	
4	Juana	Calle B	2		20
3	María	Calle C		D	40

Falta

Inferir

NC

Estimar

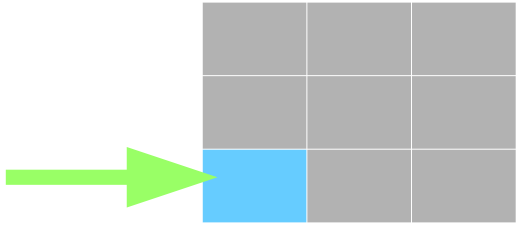
Dimensión Correctitud sintáctica

CI	Nombre	Calle	Nro.	F.Nto.	Edad
1	José	Calle A	1	A	30
2	Pepe	Calle D	A#	C	30
4	Juana	Calle B	B&	B	40
3	María	Calle C	3	D	40

Entero > 0

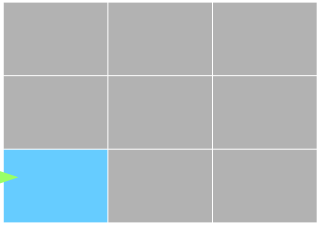
Especificación de una métrica

Nombre	M1C.SintácticaNumPuerta
Valores correctos	Entero > 0
Tipo resultado	{0,1}
Granularidad	Celda
Método medición	VC → 1 VNC → 0

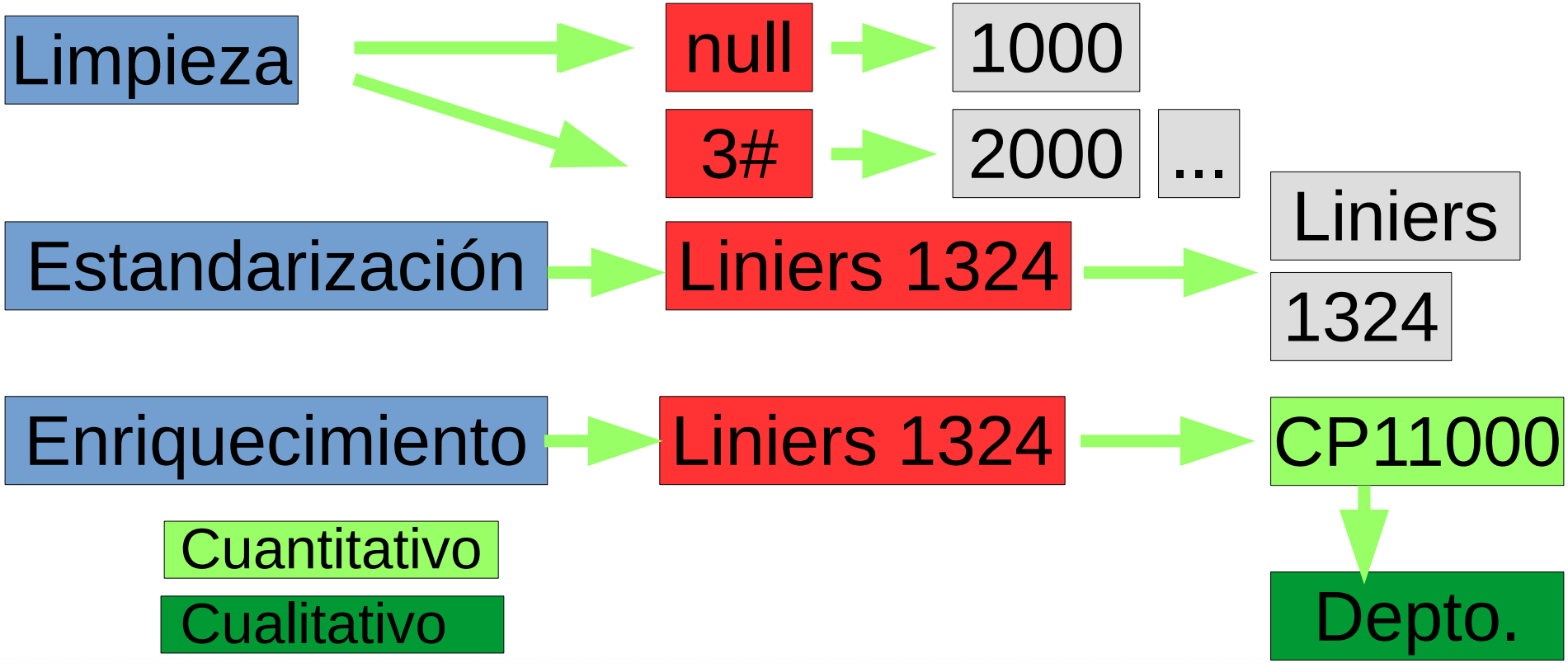


Especificación de una métrica

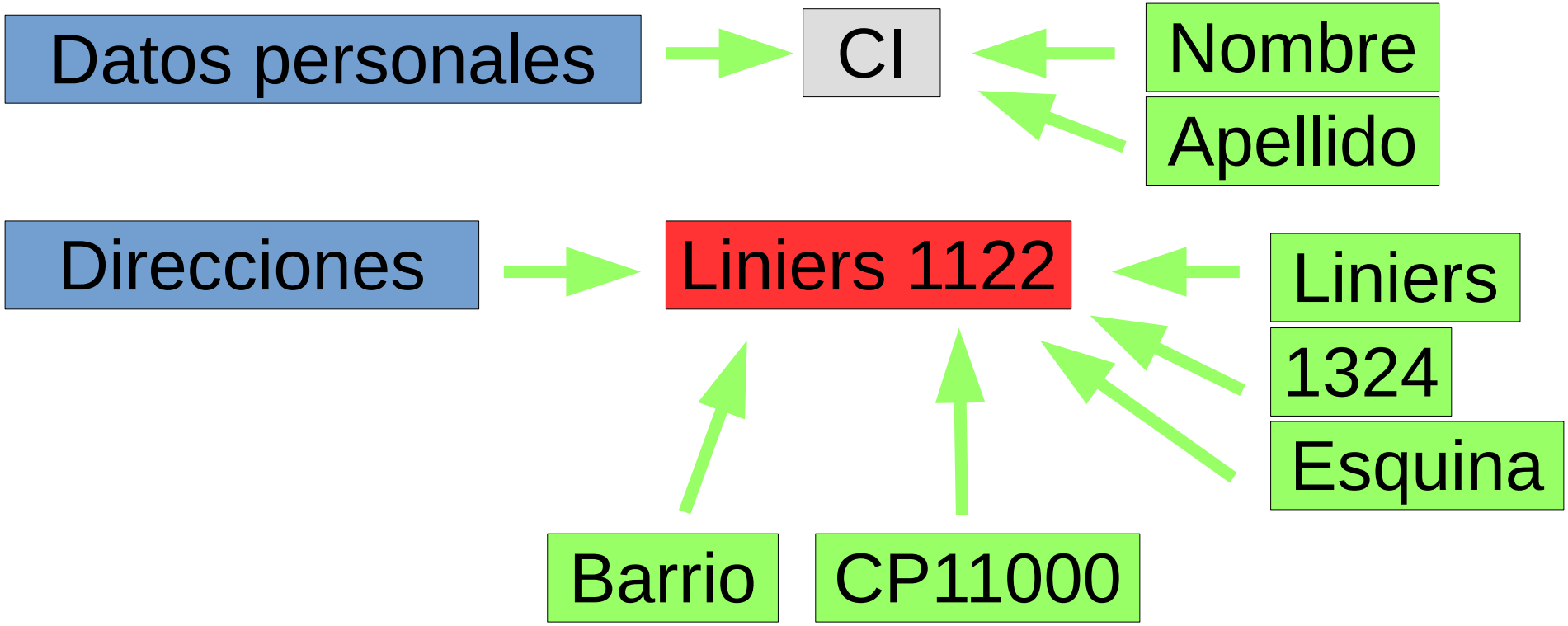
Nombre	M1C.SintácticaNomEmp
Valores correctos	(A-Z)(a-z)+
Tipo resultado	{0,1}
Granularidad	Celda
Método medición	VC → 1 VNC → 0



Corrección de errores y mejoras



Prevención de errores: Catálogos



Prevención de errores: Estándares

País	Departamento	Municipio	Localidad

Liniers 1324

Barrio | Bloque | Nombre vialidad

... ... | Sección | Torre | Número puerta

Componentes de una Dirección uy

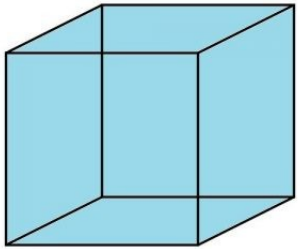
Tipo vialidad

Letra puerta

Nombre inmueble

Resumen

- Principales conceptos, y dimensiones



- Especificación métrica

- Corrección de errores

- Prevención de errores

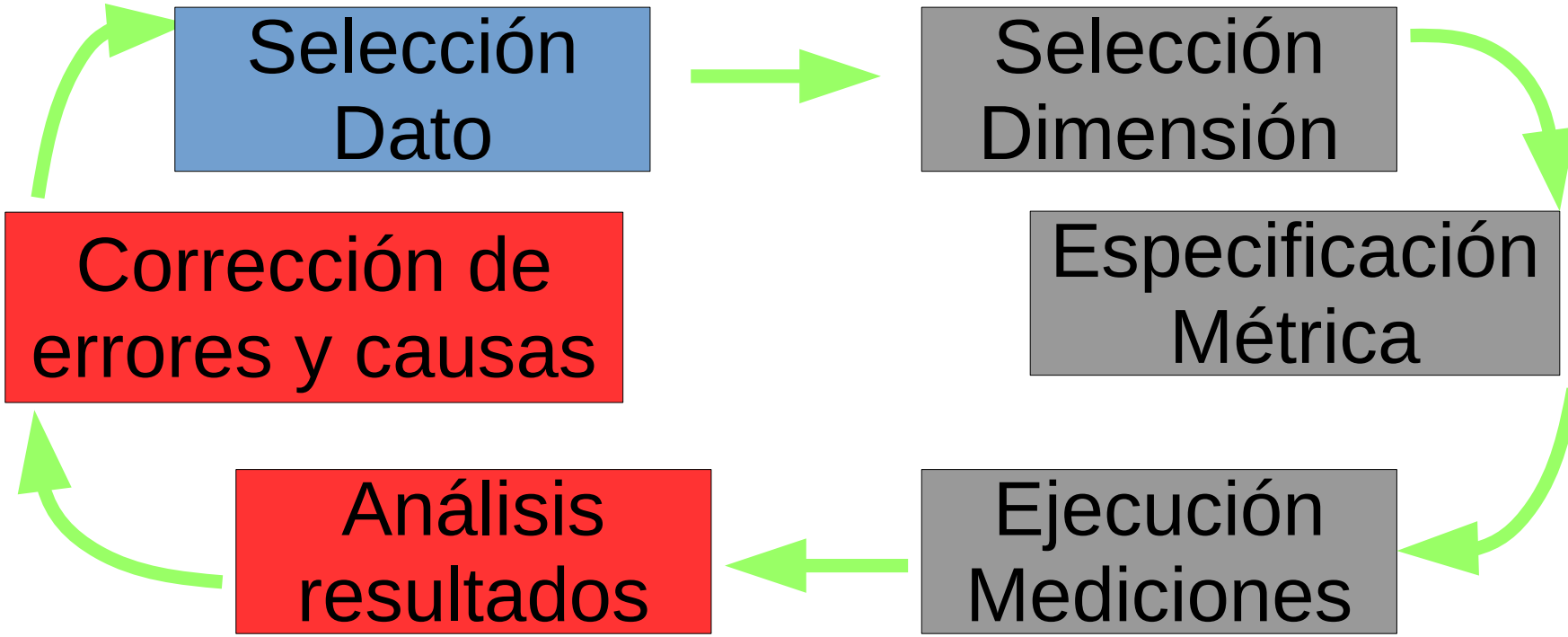
Objetivos

- Introducción Calidad Datos 1

- **Estrategia aseguramiento calidad** 2

- Lecciones aprendidas 3

Mediciones de calidad

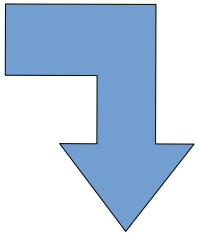


Aseguramiento de calidad

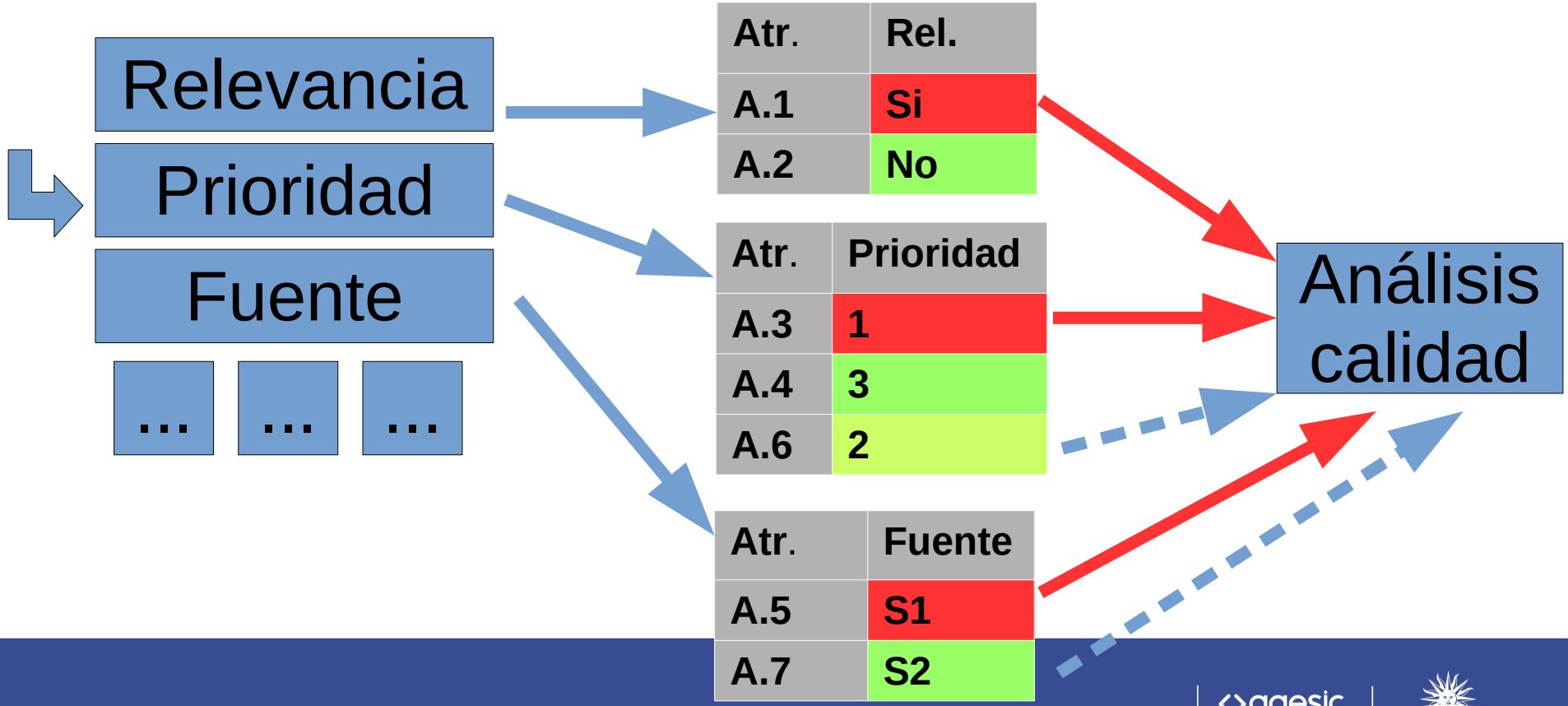
Dimensiones

Métricas

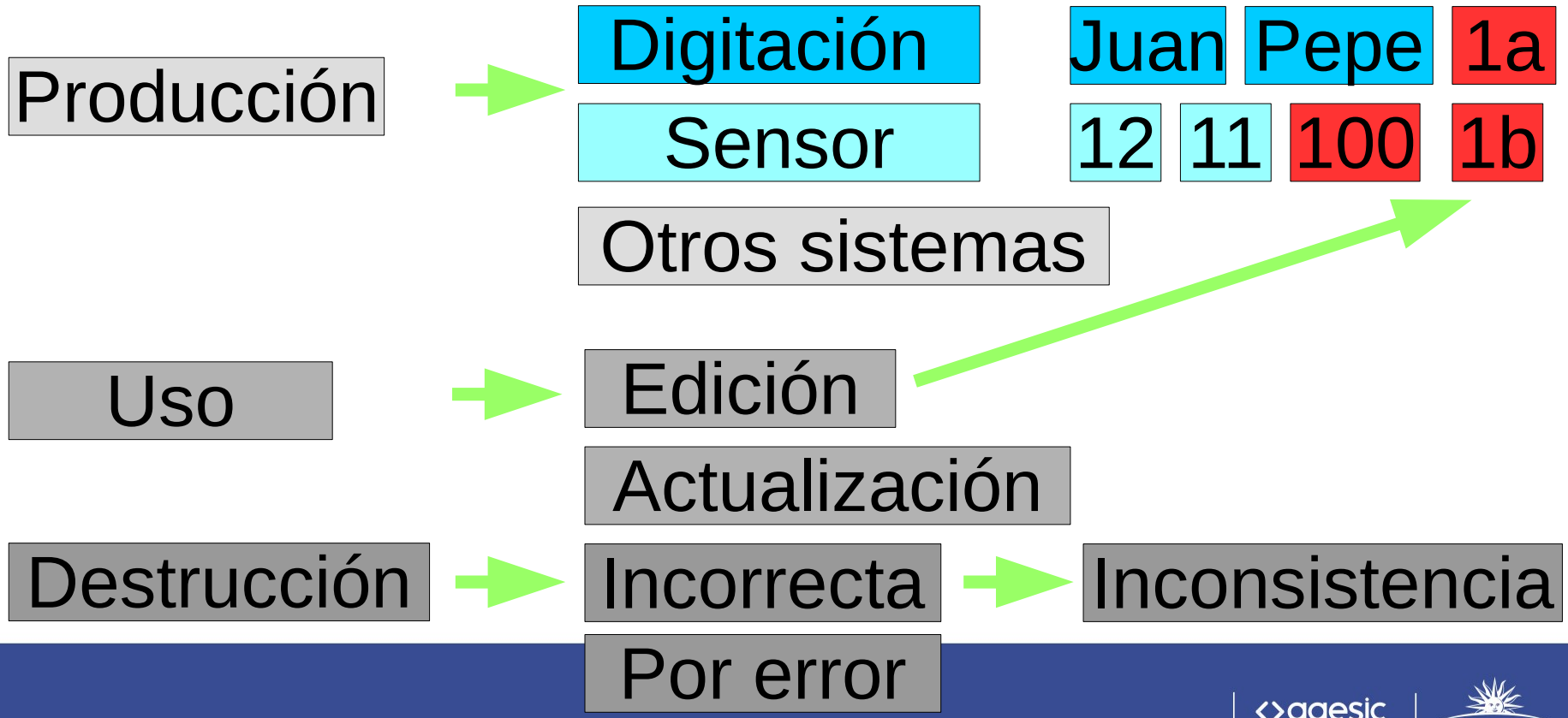
Mediciones



Clasificación datos



Ciclo de vida del dato



Procesos

Propios de los datos

- Negocio
- Origen
- Actualización
- Integración

Analizar

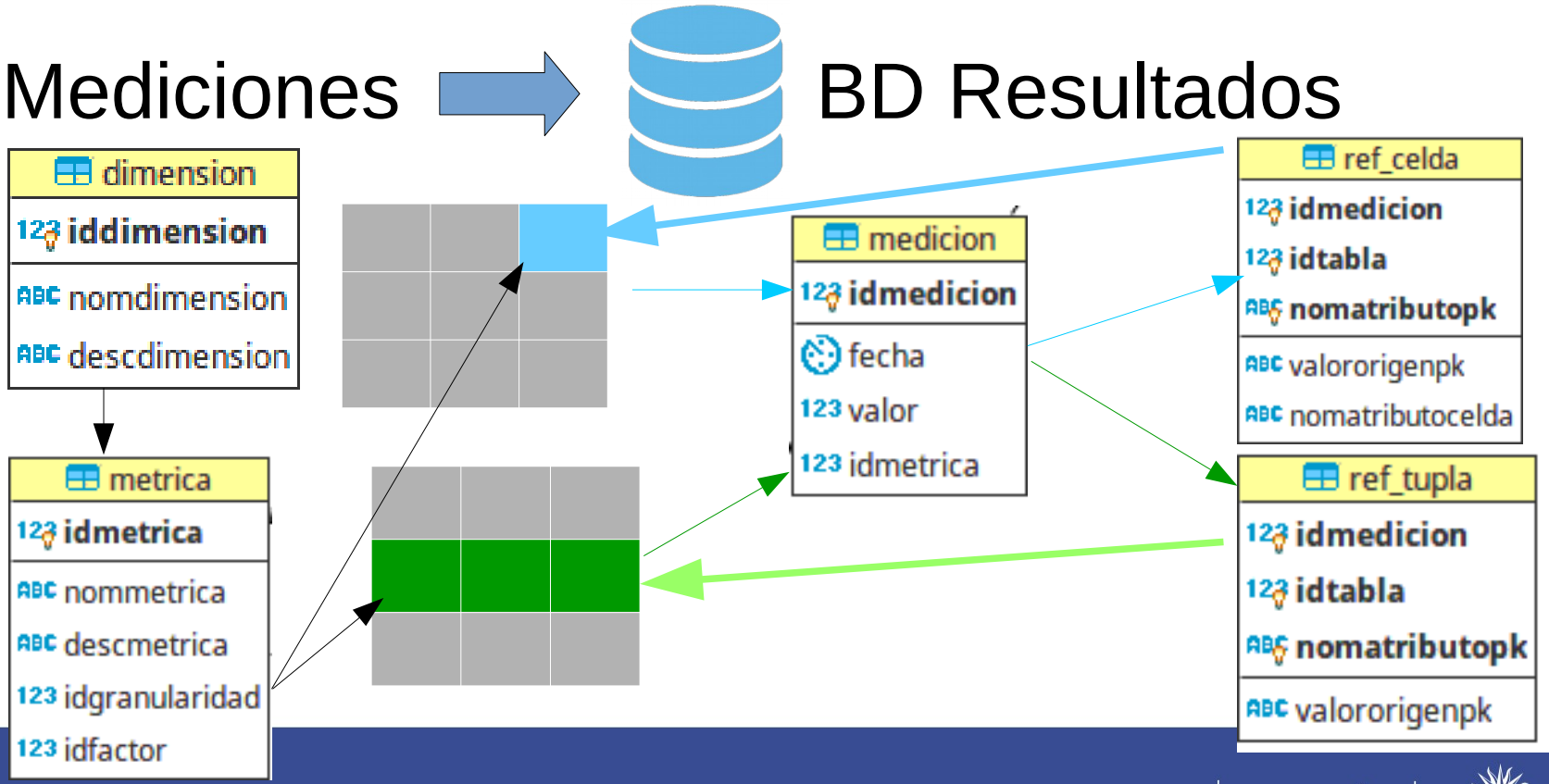
Análisis de calidad

- Medición
- Corrección

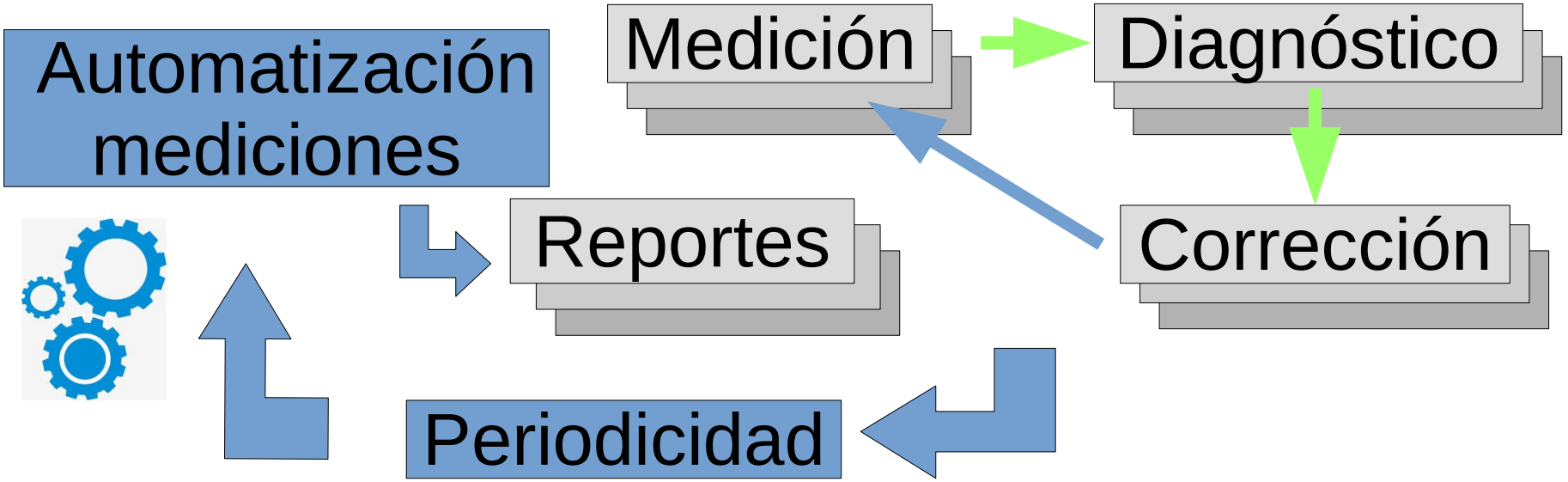
Construir

Almacenamiento resultados

- Mediciones → BD Resultados



Mediciones periódicas



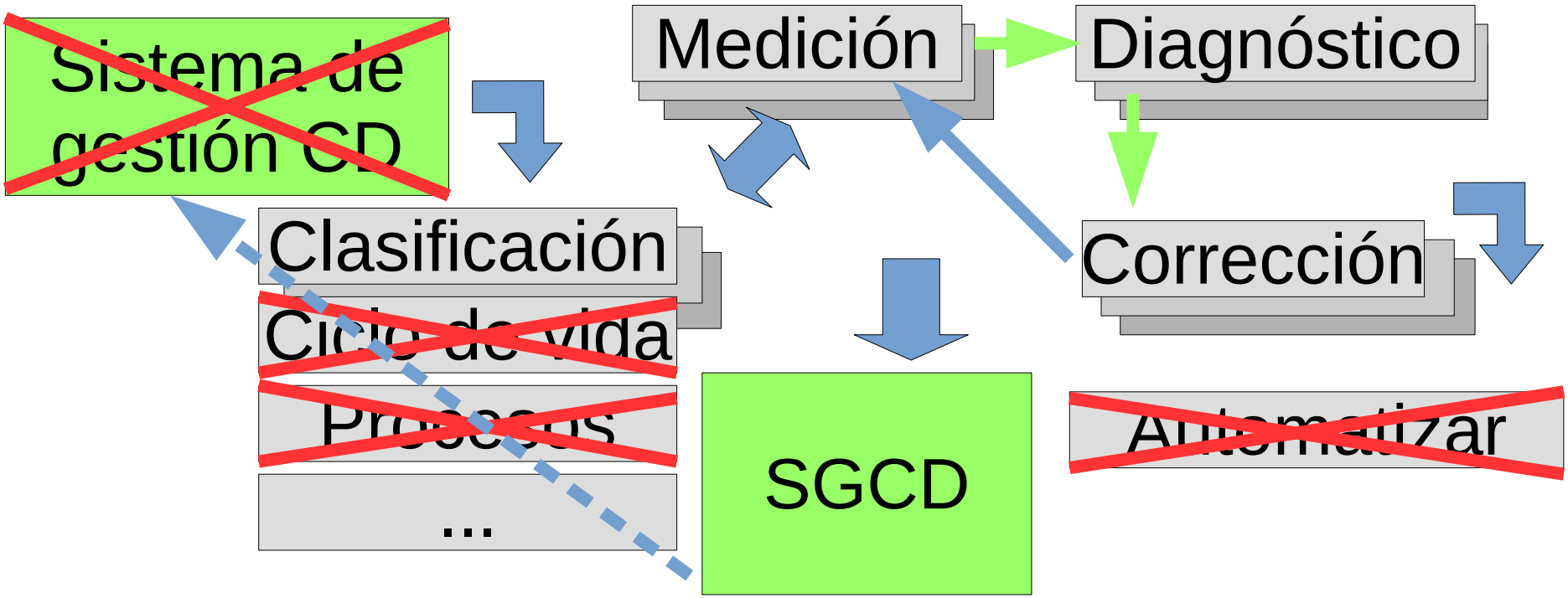
Aseguramiento calidad

Clasificación de datos	Mediciones de calidad
Ciclo vida del dato	Automatización mediciones
Procesos	Personas-Roles
Prevención	Almacenamiento resultados

Sistema de gestión CD

.. .. ..

Estrategia de comienzo



Objetivos

- Introducción Calidad Datos 1

- Estrategia aseguramiento calidad 2

- **Lecciones aprendidas** 3

Herramientas

- Mediciones con Pentaho PDI
  - Simplifica implementación
  - Permite automatizar ejecución
  - Libre uso

Correcciones

- Esfuerzo importante de análisis y corrección de datos y errores

alterar sistemas y procesos

alterar BD

Datos generados en mediciones

CI	Att1	Att2
1	A	10
2	B	20
1	C	30
3	D	40

celda → 12 mediciones

línea → 4 mediciones

Datos generados en mediciones

CI	Att1	Att2
1	A	10
2	B	20
1	C	30
3	D	40

58 → 80.000.000

2.800.000

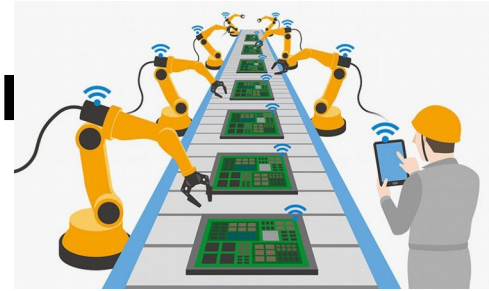
3 → 7.000.000

Clasificación

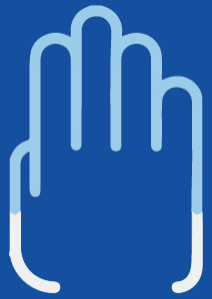
Granularidad

Automatización y Paneles

- Automatización de medicion para sucesivas ejecuciones



- Diseñar panel de resultados para mostrar resultados a usuarios



Espacio de intercambio

jornadas tecnológicas

MUCHAS GRACIAS

POR DUDAS, CONSULTAS O SUGERENCIAS

JORNADASTECNOLOGICAS@AGESIC.GUB.UY