



Ordenamiento y limpieza de datos

Segunda Parte

Mg. Yanina Bellini Saibene - INTA Anguil

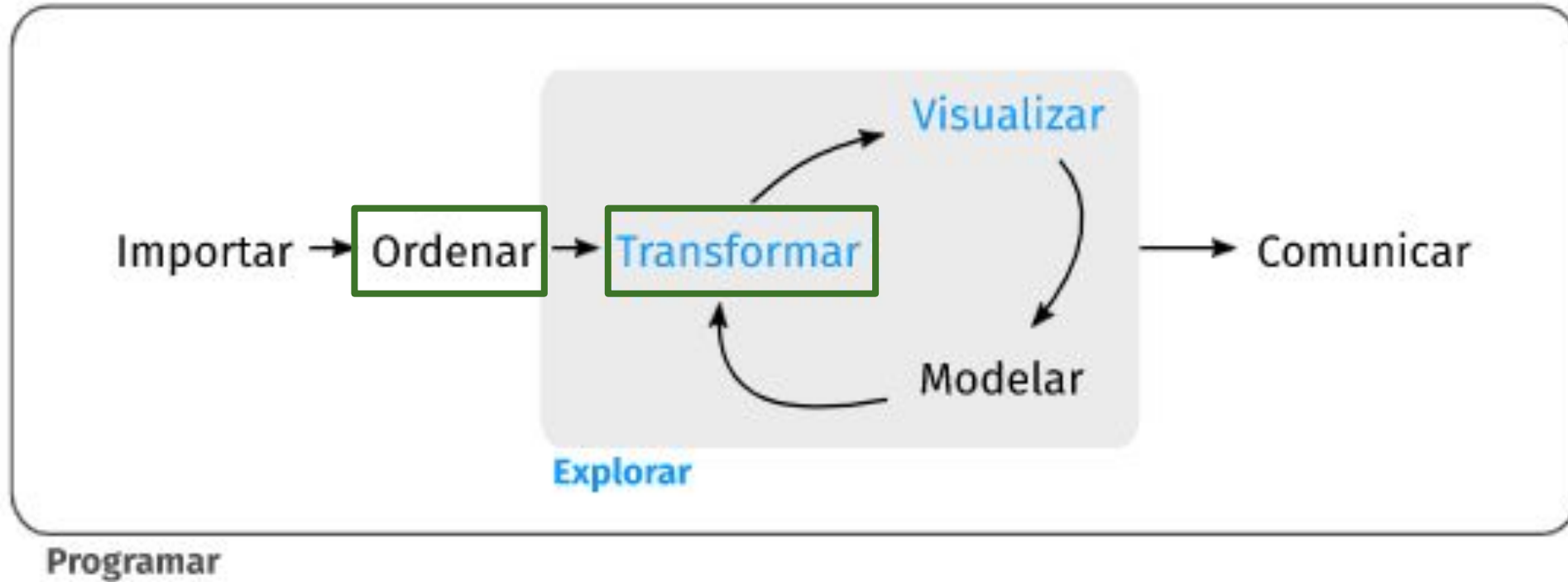
Dra. María Florencia D'Andrea - IRB - CNIA



Instituto Nacional
de Tecnología Agropecuaria

Un lenguaje para ciencia de datos

Ordenar datos



Un lenguaje para ciencia de datos
Ordenar datos

Datos usables

Datos abiertos

Datos útiles

Datos limpios

Datos masivos

Datos ordenados

Datos útiles



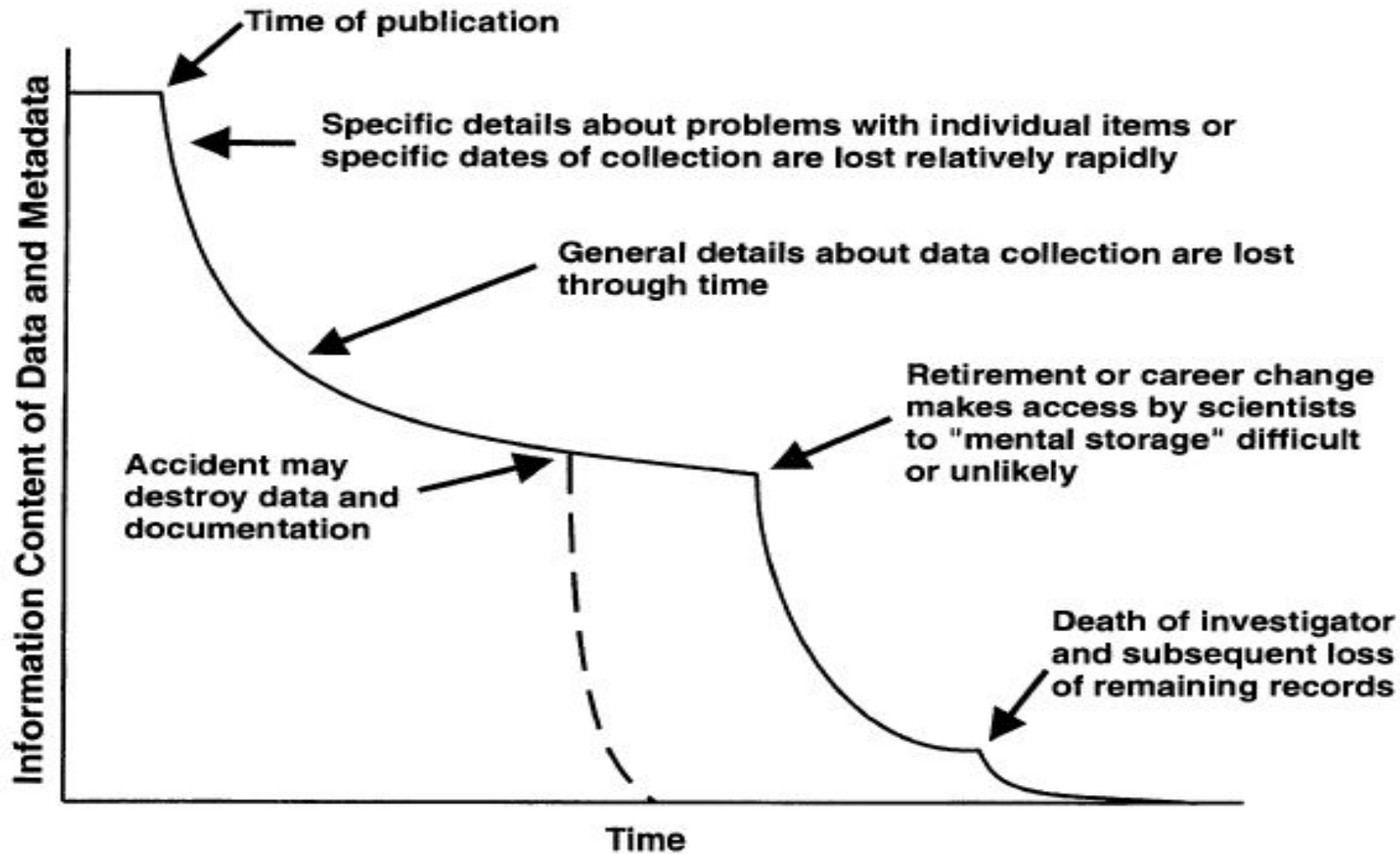
The value of storing volumes of data depends on our ability to extract useful reports, spot interesting events and trends, support decisions and policy based on statistical analysis and inference, and exploit the data to achieve business, operational, or scientific goals.

*Usama Fayyad ,Gregory Piatetsky -Shapiro ,and
Padhraic Smyth*

Datos usables: Metadatos y Licencias

- Los metadatos representan un **conjunto de instrucciones** o documentación que describe el *contenido, contexto, calidad, estructura y accesibilidad* de un recurso.
- Las **licencias de uso** permiten determinar **como se pueden usar estos datos y como reconocer a los autores**.

Datos...metadatos...para qué tanto dato?



Ejemplo de la degradación normal de la información asociada a los datos y metadatos a través del tiempo.

La actualización de la tecnología y los accidentes pueden eliminar el acceso a los datos y sus metadatos en cualquier momento.

Datos ordenados y datos limpios

El 80% del tiempo del análisis de datos se utiliza en el proceso de **limpieza y preparación** de los datos.

Esta tarea se realiza **varias veces** durante el análisis de los datos

**Datos ordenados (Tidy Data):
estructuración de conjuntos de datos
para facilitar el análisis.**

Principios de Tidy Data

CULTIVAR	Días a floración	Altura (cm)	Vuelco (%)	Densidad (pl/ha)	Humedad de grano	Rendimiento de granos (kg/ha)	Aceite (%)
ACA 203 CL	85	181	0	48554	6.1	2719	43.6
ACA 861	85	166	0	47521	6.1	2319	51.8
ACA 869	87	189	3	45455	6.0	2300	54.0

Observación

Variable ó Atributo

1. Cada **variable** es una **columna**.
2. Cada **observación** es una **fila**.
3. Cada **tipo de unidad de observación** forma una **tabla**.

Síntomas comunes de datos desordenados

- ▶ Los encabezados de columna son valores, no nombres de variables.
- ▶ Múltiples variables se almacenan en una columna.
- ▶ Las variables se almacenan tanto en filas como en columnas.
- ▶ Múltiples tipos de unidades de observación se almacenan en la misma tabla.
- ▶ Una sola unidad de observación se almacena en varias tablas.

Síntomas comunes de datos desordenados

Los encabezados de columna son valores, no nombres de variables


DEPART	CABECERA	SUP_JURI	AVENA	CEBADA	CENTENO	TRIGO	GIRASOL	MAIZ
CHICALCO	LA PASTORIL	9117	0	0	0	0	0	0
LIMAY MAHUIDA	LIMAY MAHUIDA	9985	0	0	0	0	0	0
CHALILEO	SANTA ISABEL	8917	0	0	0	0	0	0
HUCAL	BERNASCONI	6047	3363	182	219	12606	1289	226
LOVENTUE	VICTORICA	9235	208	39	77	1256	603	337
RANCUL	RANCUL	4933	2679	413	1614	12135	33910	14696

DEPART	CABECERA	SUP_JURI	CULTIVO	SUPERFICIE
CHICALCO	LA PASTORIL	9117	AVENA	0
CHICALCO	LA PASTORIL	9117	CEBADA	0
CHICALCO	LA PASTORIL	9117	CENTENO	0
CHICALCO	LA PASTORIL	9117	TRIGO	0
LOVENTUE	VICTORICA	9235	GIRASOL	0
CHICALCO	LA PASTORIL	9117	MAIZ	0

Síntomas comunes de datos desordenados

Múltiples variables se almacenan en una columna

CULTIVAR	Días a floración	Días a madurez	Altura (cm)	Vuelco (%)
ACA 203 CL (ACA)	85	122	181	0
ACA 861 (ACA)	85	122	166	0
Aguara 6 (ADVANTA)	85	126	174	0
CACIQUE 312 CL (EL CENCERRO)	90	127	161	1
KWS 480 CL (KWS)	90	127	164	1
LG 56.78 CLP (LIMAGRAIN)	88	127	188	4



CULTIVAR	EMPRESA	Días a floración	Días a madurez	Altura (cm)	Vuelco (%)
ACA 203 CL	ACA	85	122	181	0
ACA 861	ACA	85	122	166	0
Aguara 6	ADVANTA	85	126	174	0
CACIQUE 312 CL	CENCERRO	90	127	161	1
KWS 480 CL	KWS	90	127	164	1
LG 56.78 CLP	LIMAGRAIN	88	127	188	4

Síntomas comunes de datos desordenados

Las variables se almacenan tanto en filas como en columnas.

id	año	mes	elemento	1	2	3	4	5	6	7	8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—

id	fecha	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

Síntomas comunes de datos desordenados

Múltiples tipos de unidades de observación se almacenan en la misma tabla.

ID	PROVIN	CAP_PROV	DEPART	CABECERA	SUP_JURI	AVENA	CEBADA	CENTENO
42063	LA PAMPA	SANTA ROSA	CHICALCO	LA PASTORIL	9117	0	0	0
42091	LA PAMPA	SANTA ROSA	LIMAY MAHUIDA	LIMAY MAHUIDA	9985	0	0	0
42049	LA PAMPA	SANTA ROSA	CHALILEO	SANTA ISABEL	8917	0	0	0
42077	LA PAMPA	SANTA ROSA	HUCAL	BERNASCONI	6047	3363	182	219



Los datos de provincia y capital de la provincia se repite por cada departamento

- Tabla Provincias
- Tabla Departamentos
- Tabla CultivosXDeptos

Síntomas comunes de datos desordenados

Múltiples tipos de unidades de observación se almacenan en la misma tabla.

ID	PROVIN	DEPART	CABECERA	SUP_JURI	CULTIVO	SUPERFICIE		CEBADA	CENTENO
42063	LA PAMPA	SA	CHICALCO	LA PASTORIL	9117	AVENA	0	0	0
42091	LA PAMPA	SA	CHICALCO	LA PASTORIL	9117	CEBADA	0	0	0
42049	LA PAMPA	SA	CHICALCO	LA PASTORIL	9117	CENTENO	0	0	0
42077	LA PAMPA	SA	CHICALCO	LA PASTORIL	9117	TRIGO	0	0	0
		SA	LOVENTUE	VICTORICA	9235	GIRASOL	0	363	182
		SA	CHICALCO	LA PASTORIL	9117	MAIZ	0		219

Los datos de provincia y capital de la provincia se repite por cada departamento

Tabla Provincias

Tabla Departamentos

Tabla CultivosXDeptos

Síntomas comunes de datos desordenados

Múltiples tipos de unidades de observación se almacenan en la misma tabla.

Tabla Provincias

PROVIN	CAP_PROV
LA PAMPA	SANTA ROSA

Tabla Departamentos

ID	PROVIN	DEPART	CABECERA	SUP_JURI
42063	LA PAMPA	CHICALCO	LA PASTORIL	9117
42091	LA PAMPA	LIMAY MAHUIDA	LIMAY MAHUIDA	9985
42049	LA PAMPA	CHALILEO	SANTA ISABEL	8917
42077	LA PAMPA	HUCAL	BERNASCONI	6047

Tabla CultivosXDeptos

DEPART	CULTIVO	SUPERFICIE
42077	AVENA	3363
42077	CEBADA	182
42077	CENTENO	219

Tidy Data

- ▶ Cuando se recolectan datos por primera vez, siempre es mejor pensar una estructura ordenada desde el inicio
- ▶ Cuando nos envían datos ya registrados, debemos analizar su estructura y generar una que sea ordenada
- ▶ La estructura ordenada hará la tarea de manejo de datos mucho más sencilla.

Ordenemos datos juntos

- ¿Esta tabla está ordenada (Tidy)?

ID del envío	3651	3655	3662	3663
Título	Telemetría L	Evaluación d	Desarrollo d	Caminos Rur
Resumen	Organizaci	La evapotran	Existe una br	En un país tar
Primer nombre (Autor 1)	Pablo	Mónica	Santiago	diego
Segundo Nombre (Autor 1)	Guillermo			gabriel
Apellidos (Autor 1)	Di Nanno	Bocco	Lombardo	giordano
País (Autor 1)	AR	AR	UY	AR
Filiación (Autor 1)	INTA - Instit	Facultad de C	Instituto Pla	26884654
Correo electrónico (Autor 1)	pablo.dinani	mbocco@gm	slombardo@	dgiordano@1
URL (Autor 1)			http://www.planagropecu	
Resumen biográfico (Autor 1)	Investigador en tecnologí	Agrónomo	Director de C	
Primer nombre (Autor 2)		Miguel	Federico	Maria
Segundo Nombre (Autor 2)				Beatriz
Apellidos (Autor 2)		Nolasco	Arias	Rodulfo
País (Autor 2)		AR	UY	AR
Filiación (Autor 2)		Facultad de C	Instituto Plan Agropecuari	
Correo electrónico (Autor 2)		mnolasconq	farias@plan	miriambrodu
URL (Autor 2)				
Resumen biográfico (Autor 2)			Desarrollado	Directora de
Primer nombre (Autor 3)		Silvina		Griselda
Segundo Nombre (Autor 3)				
Apellidos (Autor 3)		Sayago		Galeano

Todo bien...pero y R??

Data Wrangling with dplyr and tidyr

Cheat Sheet



Syntax - Helpful conventions for wrangling

dplyr::tbl_df(iris)

Converts data to tbl class. tbl's are easier to examine than data frames. R displays only the data that fits onscreen:

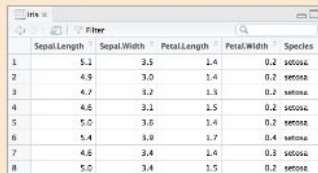
```
Source: local data frame [150 x 5]
  Sepal.Length Sepal.Width Petal.Length
1             5.1         3.5         1.4
2             4.9         3.0         1.4
3             4.7         3.2         1.3
4             4.6         3.1         1.5
5             5.0         3.6         1.4
...
Variables not shown: Petal.Width (dbl),
Species (fctr)
```

dplyr::glimpse(iris)

Information dense summary of tbl data.

utils::View(iris)

View data set in spreadsheet-like display (note capital V).



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

dplyr::%>%

Passes object on left hand side as first argument (or argument) of function on righthand side.

x %>% f(y) is the same as f(x, y)
y %>% f(x, , z) is the same as f(x, y, z)

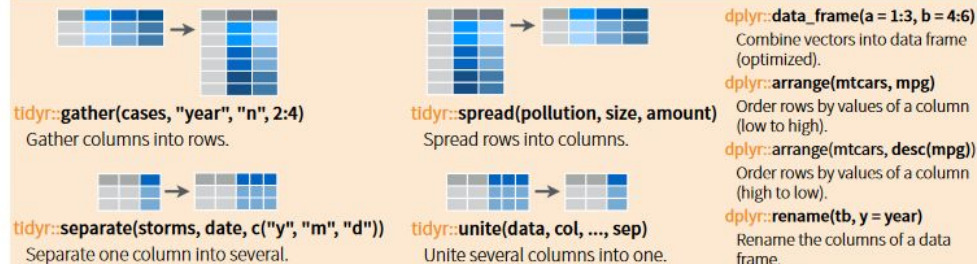
"Piping" with %>% makes code more readable, e.g.

```
iris %>%
  group_by(Species) %>%
  summarise(avg = mean(Sepal.Width)) %>%
  arrange(avg)
```

Tidy Data - A foundation for wrangling in R



Reshaping Data - Change the layout of a data set



Subset Observations (Rows)



dplyr::filter(iris, Sepal.Length > 7)

Extract rows that meet logical criteria.

dplyr::distinct(iris)

Remove duplicate rows.

dplyr::sample_frac(iris, 0.5, replace = TRUE)

Randomly select fraction of rows.

dplyr::sample_n(iris, 10, replace = TRUE)

Randomly select n rows.

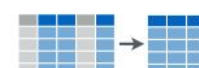
dplyr::slice(iris, 10:15)

Select rows by position.

dplyr::top_n(storms, 2, date)

Select and order top n entries (by group if grouped data).

Subset Variables (Columns)



dplyr::select(iris, Sepal.Width, Petal.Length, Species)

Select columns by name or helper function.

Helper functions for select - ?select

select(iris, contains(" "))
Select columns whose name contains a character string.

select(iris, ends_with("Length"))
Select columns whose name ends with a character string.

select(iris, everything())
Select every column.

select(iris, matches("t. "))
Select columns whose name matches a regular expression.

select(iris, num_range("x", 1:5))
Select columns named x1, x2, x3, x4, x5.

select(iris, one_of(c("Species", "Genus")))
Select columns whose names are in a group of names.

select(iris, starts_with("Sepal"))
Select columns whose name starts with a character string.

select(iris, Sepal.Length:Petal.Width)
Select all columns between Sepal.Length and Petal.Width (inclusive).

select(iris, -Species)
Select all columns except Species.

Fechas: **Lubridate**

<https://cran.r-project.org/web/packages/lubridate/vignettes/lubridate.html>

¡Manos a la obra!

R_inta_LC2_2019.R

- ▶ Cargar Tidyr
- ▶ Separar y unir columnas y filas
- ▶ Manejar valores vacíos
- ▶ Gather y Spread (pivot_wider y pivot_longer)



Almuerzo

Foto: gentiliza Mauro Lepore