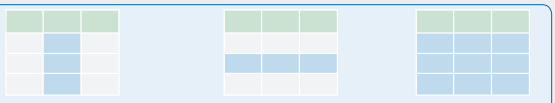
Common Threats to Tidy Data

There are lots of ways to make data untidy.
These examples are illustrative, not exhaustive.

Inspired by the Tidyverse:
vita.had.co.nz/papers/tidy-data.pdf



TIDY

MESSY

Merged Cells

region	cholera cases	
	2012	2014
region1	2,500	9,000
region2	7,000	100

combined cells →

region	year	cholera_cases
region1	2012	2,500
region2	2012	7,000
region1	2014	9,000
region2	2014	100

project	region
project1	Afar, Oromia
project2	Oromia, Somali, SNNP
project3	SNNP, Afar, Amhara, Gambela

multiple values per cell →

project	region
project1	Afar
project1	Oromia
project2	Oromia
...	...

No Unique ID

project	region
project1	Afar
project1	Oromia
project2	Oromia
...	...

→

activity_id	project	region
001	project1	Afar
002	project1	Oromia
003	project2	Oromia
...	...	...

Variable Names Not Meaningful

region	variable1	variable2
region1	2,500	4,000
region2	7,000	4,100

→

region	cholera	TB
region1	2,500	4,000
region2	7,000	4,100

Variable Names Contain Measurements

region	2012cholera	2014cholera
region1	2,500	9,000
region2	7,000	100

→

region	year	cholera_cases
region1	2012	2,500
region2	2012	7,000
region1	2014	9,000
region2	2014	100

Inconsistent Data

different formats
names misspelled

project	start_date	region	funding
project1	14-04-13	Afar	25,000,000
project2	2016-12-25	Affar	30000000
project3	01/01	Benishangul-Gumuz	$75M
project3	01012014	Benishangul Gumuz	€68M

duplicate values (?)
special characters
values in different units

→

project	start_date	region	funding_USD
project1	2014-04-13	Afar	25,000,000
project2	2016-12-25	Afar	30,000,000
project3	2014-01-01	Benishangul-Gumuz	75,000,000

Missing Values Not Explicit

region	TB_cases	cholera_cases
region1	4,200	2,500
region2	1,250	7,000
region3		9,000
region4		

is this missing or zero?

→

region	TB_cases	cholera_cases
region1	4,200	2,500
region2	1,250	7,000
region3	NA	9,000
region4	0	NA

Data Not Computer Readable

	project status
project 1	
project 2	
project 3	
project 4	

LEGEND
- good
- kinda okay
- really not okay

→

project	status
project1	kinda okay
project2	good
project3	kinda okay
project4	really not okay

Undocumented, Vague Data

project	region	funding
project1	5	25,000,000
project2	1	10,000,000
project3	7	15,000,000

code not defined
units not specified

→

project	region	funding_USD
project1	Afar	25,000,000
project2	Oromia	10,000,000
project3	Somali	15,000,000

Tim Essam (tessam@usaid.gov @StataRGIS) & Laura Hughes (lhughes@usaid.gov @flaneuseks)

July 2016
GEOCENTER