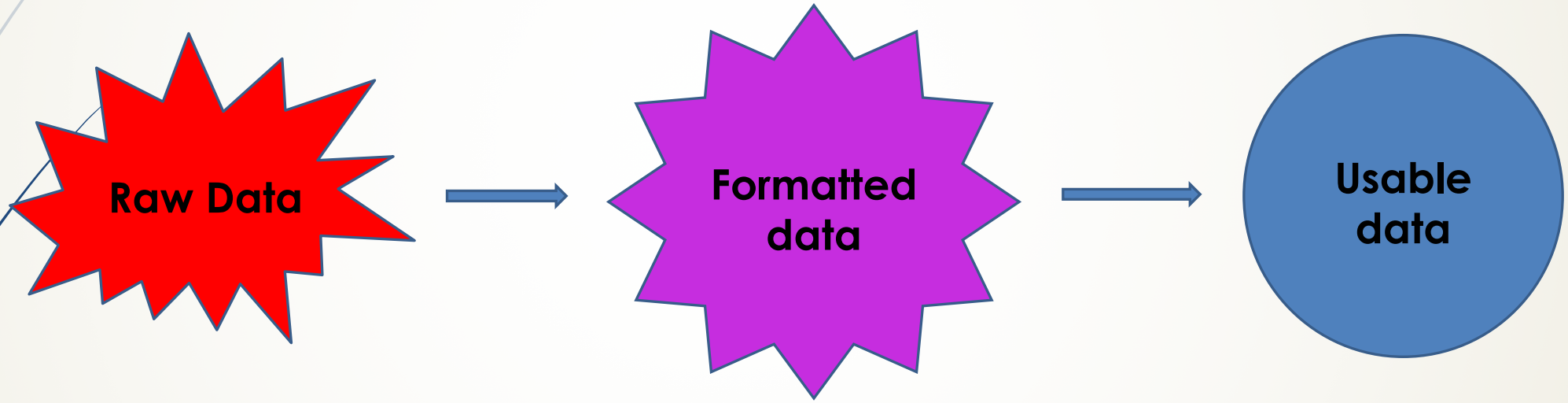
Data cleaning with R

Ecological and Epidemiological Modeling Madagascar (E2M2)
ValBio, Ranomafana
Fianarantsoa, Madagascar
03-14 January 2020

*Hafaliana **Christian** Ranaivoson*
Virology Unit, Intitut Pasteur de Madagascar
Mention of Zoology and Animal Biodiversity
Faculty of Sciences, Univeristy of Antananarivo

Data cleaning steps



Initial data

Technically correct

Consistent

Data cleaning with R

Formatted data,

Variable types and errors

```
> str(e2m2_FB)

 $ Id     : chr  "fb_1" "fb_2" "fb_3" "fb_4"…
 $ Sex    : chr  "f" "f" "f" "F" "f " ...
 $ Weight : num  34.8 36.1 36.5 36.6 38.9 ...
 $ Age    : chr "one" "6.65" "6.77" "seven" ...
 $ Date   : chr  "1/11/2015" "1/12/2015 "…
```

```
> as.factor(e2m2_FB$Sex)
Levels: f F f  m
> as.numeric(e2m2_FB$Age)
Warning message:
NAs introduced by coercion
> as.Date(e2m2_FB$Date,"%m/%d/%Y")
```

Categorical: Factor (n levels)
Continuous: Numeric (Range)
Time: Date (Range)
Binary: logic (T,F)
Missing Value: NA

as.factor(…)
as.Date(…) **"%Y-%m-%d"**
as.numeric(…)

Needed format -> Re-format
Value error -> Correct value
Missing error -> Handle NA values

Consistent data

Give a sense to your data

- Extreme values will affect the outcome
- Make a consistent link between each variable
- No cheating
- Cleaning is not inventing

Consistent data

What is wrong with this data frame?

Id	Site	Date	Sex	Weight	Rainfall	gmam	stes
bat_12	Site_1	41586	F	650	140mm	*NA*	3.90
bat_13	Site_1	41586	f	70	140mm	*NA*	3.89
bat_14	Site_2	41593	M	710	136mm	*NA*	10.05
bat_15	Site_2	41593	M	690	136mm	*NA*	10.02
bat_16	Site_2	41593	F	590	136mm	3.88	*NA*
bat_17	Site_2	41593	M	125	136mm	*NA*	9.95
bat_18	Site_2	41593	M	150	136mm	*NA*	9.64
bat_19	Site_2	41593	F	530	136mm	4.03	*NA*
bat_20	SITE_2	41593	M	130	136mm	*NA*	10.61
bat_21	Site_2	41594	F	640	136mm	4.26	*NA*
bat_22	Site_2	41594	F	590	136mm	4.18	*NA*
bat_23	Site_2	41594	M	145	136mm	*NA*	10.02
bat_23	Site_2	41594	F	520	136mm	3.97	*NA*
bat_25	Site_2	41594	M	150	136mm	*NA*	10.00
bat_26	Site_2	41596	f	650	136mm	4.10	*NA*
bat_27	Site_2	41596	F	165	136mm	4.14	*NA*
bat_28	Site_2	41596	M	130	136mm	*NA*	10.34

Conclusion

- Important step for good data interpretation
- Need a good comprehension of each variables
- Should be reproductible

Thank you all!