Tidy Data

Hadley Wickham

#dataCleaning
#dataTidying
#relationalDatabases
#Rstats

Journal of Statistical Software
(2014)
VOL: 59 NO: 1
PAGES: 1 - 23

INTRODUCTION.

TIDYING
80%.
PREPARING
~~Data Analysis~~
is CLEANING

analysis!

relational algebra

Relational Database

DOMAIN PROBLEM

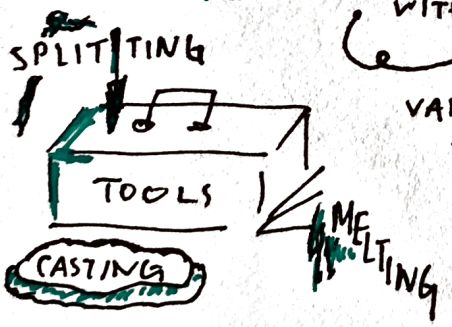
TIDY DATA PRINCIPLES + TOOLS

NOT Logics of Data.

- facilitate exploration & initial analysis
- simplify development of data analysis tools that work well together.

Translated for Statisticians
"SQL"
"COMPSCI"
"RELATIONAL DBs"

TIDYING messy DATA SETS



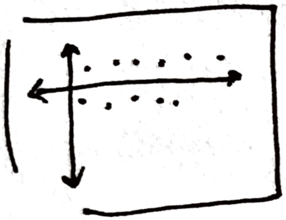
5 COMMON PROBLEMS with messy data sets

- COLUMNS = VALUES, NOT VAR.
- MULTIPLE VALUES IN 1 COL.
- MULTIPLE OBSERVATIONAL UNITS IN 1 TABLE
- VARIABLES BOTH ⇒ ROWS & COLUMNS
- SINGLE OBSERVATION UNIT STORED IN MULTIPLE TABLES

DEFINING TIDY DATA: A standard way of mapping meaning to structure. requires satisfying 3 conditions:

1. VARIABLE ⇒ COLUMN
2. OBSERVATION ⇒ ROW
3. UNIT TYPE ⇒ TABLE

"... every messy dataset is messy in its own way."



Structure : physical layout

Tidy Data

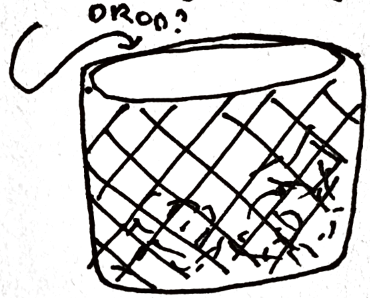
Links

semantics: ~ meaning.

var 1: ~ person with name ~ first & ~ last ... ~

missing values

DROP?



IS IT AN OBS. THAT SHOULD HAVE BEEN MADE?

- YES! ↓ DON'T DROP
- NO IT'S IMPOSSIBLE TO MEASURE ↓ BIN IT!

DATA SET: collection of VALUES (numbers, strings)

VARIABLE: all values that measure the same UNDERLYING ATTRIBUTE (age, year, date, height, age)

OBSERVATION: all values measure on the SAME UNIT < person, date, location

ORDER

→ columns → Fixed variables → MEASURED variables

↓ sort rows by fixed (1, 2, 3 ...)