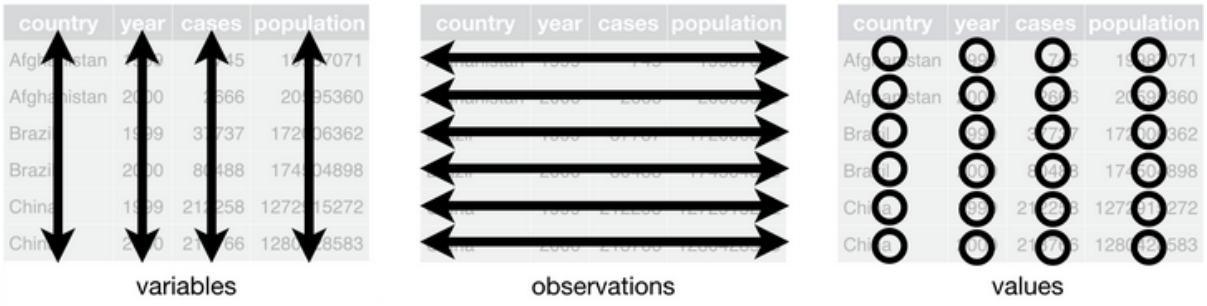
BÀI THỰC HÀNH 2: TÌM HIỂU VỀ TIDY DATA VÀ XÂY DỰNG TIDY DATA

1. Tidy data là gì

Theo Hadley Wickham, tidy data có các đặc tính sau:

- Mỗi biến (variable) được đặt trong một cột (column).
- Mỗi dữ liệu (observation) được đặt trong một dòng (row).
- Mỗi giá trị (value) của một biến được đặt trong một ô (cell).
- Dòng đầu của file mô tả tên các biến (hay thuộc tính). Tên thuộc tính phải có ý nghĩa.



Hình 1: Minh hoạ về Tidy data

Các thành phần của Tidy data:

Bảng 1: Mô tả các thành phần của Tidy data

Thành phần	**Chức năng**
Code book	Mô tả thông tin tổng quát về bộ dữ liệu, gồm các thông tin sau: 1. Tên bộ dữ liệu, chức năng bộ dữ liệu. 2. Nguồn thu thập, cách thức thu thập. 3. Thông tin về các thuộc tính: số lượng, tên từng thuộc tính. 4. Thông tin cho từng thuộc tính: tên thuộc tính - tên biến dữ liệu, ... 5. Thông tin tác giả: tên tác giả/tổ chức, email, ...
Instruction list	Các để tạo ra một Tidy data hoàn chỉnh từ dữ liệu thô. Có 2 cách: - Thủ công: Dùng word mô tả các bước làm chi tiết.

	- Tự động: Viết script thực thi.
Raw data	Dữ liệu thô
Tidy data	Dữ liệu đã xử lý.

2. Tạo tidy data cho bộ dữ liệu.

Bộ dữ liệu: **Diabets Dataset** Nguồn: https://archive.ics.uci.edu/ml/datasets/diabetes

Code book:

Code book được mô tả trong file **Data-Code**.

Bảng 2: CODE Book mô tả bộ dữ liệu.

Thông tin	Nội dung
Tên bộ dữ liệu	Diabetes patient records
Nguồn thu thập và cách thức thu thập	Từ 2 nguồn: 1. **Máy đo tự động**: có đồng hồ bấm giờ tự động, sẽ ghi lại chính xác thời gian tại lúc đo số liệu. 2. **Thu thập bằng tay**: Sử dụng bản ghi giấy (paper record), giờ được định sẵn vào các khung giờ: sáng (8:00), trưa (12:00), chiều (18:00) và tối (22:00).
Số thuộc tính	4
Thông tin tên các thuộc tính	**Date**: Ngày thu thập, định dạng: MM-DD-YYYY **Time**: Giờ thu thập, định dạng: XX:YY (24 giờ). **Code**: Mã code theo danh sách sau: 33 = Regular insulin dose 34 = NPH insulin dose 35 = UltraLente insulin dose 48 = Unspecified blood glucose measurement 57 = Unspecified blood glucose measurement 58 = Pre-breakfast blood glucose measurement 59 = Post-breakfast blood glucose measurement

	60 = Pre-lunch blood glucose measurement 61 = Post-lunch blood glucose measurement 62 = Pre-supper blood glucose measurement 63 = Post-supper blood glucose measurement 64 = Pre-snack blood glucose measurement 65 = Hypoglycemic symptoms 66 = Typical meal ingestion 67 = More-than-usual meal ingestion 68 = Less-than-usual meal ingestion 69 = Typical exercise activity 70 = More-than-usual exercise activity 71 = Less-than-usual exercise activity 72 = Unspecified special event **Value**: Gía trị thu thập được.
Thông tin tác giả	kahn@informatics.WUSTL.EDU (Internet) or 70333,34 (CompuServe)

Raw data:

Raw data gồm tập hợp các file: **data-01, data-02, ... data-70**.

Tidy data:

Tidy data sẽ được lưu lại thành file: **diabets.csv**.

Instruction list:

Code R

```
rm(list=ls())

# Lay danh sach cac fil ra
myFiles <- list.files(path="data/", pattern="data-")

k = TRUE
# Tien hanh doc tung file
for (f in myFiles) {
  if (k==TRUE) {
    file <- read.csv(paste("data/", f, sep=""), sep="\t",
header = FALSE)
```

```
    k = FALSE
  }
  else {
    file <- rbind(file, read.csv(paste("data/", f,
sep=""), sep="\t", header = FALSE))
  }
}

dataset <- file
variables <- c("Date", "Time", "Code", "Value")

# Dat ten cho cot trong bo du lieu
colnames(dataset) <- variables

write.csv(dataset, file = "diabet.csv")
```

3. Bài tập.

Bài 1: Hiện thực lại các thành phần của tidy data ở phần 2.

Bài 2: Mô tả các thành phần của tidy data cho các bộ dataset sau:

a) Iris: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/

b) Bank Marketing: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

c) Car Evaluation: https://archive.ics.uci.edu/ml/machine-learning-databases/car/

d) Wines Data Set: https://archive.ics.uci.edu/ml/datasets/Wine

Nộp bài: Nộp Bài 2.

Các nội dung cần nộp: Ứng với mỗi bộ dataset thì tạo một folder khác nhau, trong mỗi folder chứa các thành phần sau:

- File PDF mô tả chi tiết các thành phần: code book, raw data, tidy data, instruction list như mục 2
- File instruction list (file code R hoặc python).
- Raw data.
- File tidy data, lưu ở định dạng csv.

Nén lại và đặt tên theo cú pháp **<MSSV>_<Họ tên>_BT2.rar**. Nộp qua course (Giảng viên sẽ tạo submission sau).

Chúc tất cả các bạn học tốt