Czyszczenie Danych

Piotr Guzik, Lucjan Janowski, Krzysztof Rusek

October 17, 2017

Outline

Wartości odstająca

Kod w R.

Dziwne wartości

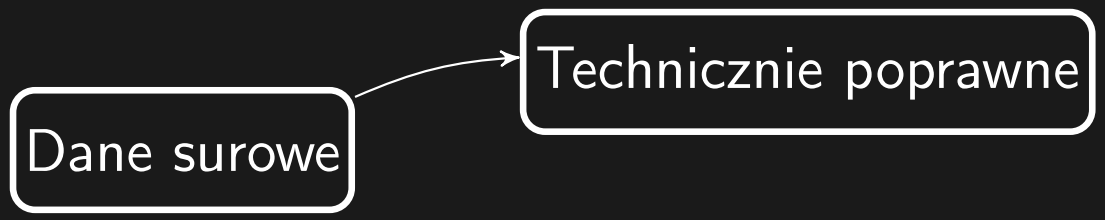
- Analiza, czy wartość jest możliwa?
- Usunięcie
- Usunięcie, jeżeli nie chcemy analizować takich wartości
- Zwielokrotnienie, jeżeli jesteśmy zainteresowani takimi przypadkami
- Korekta analizy

Outline

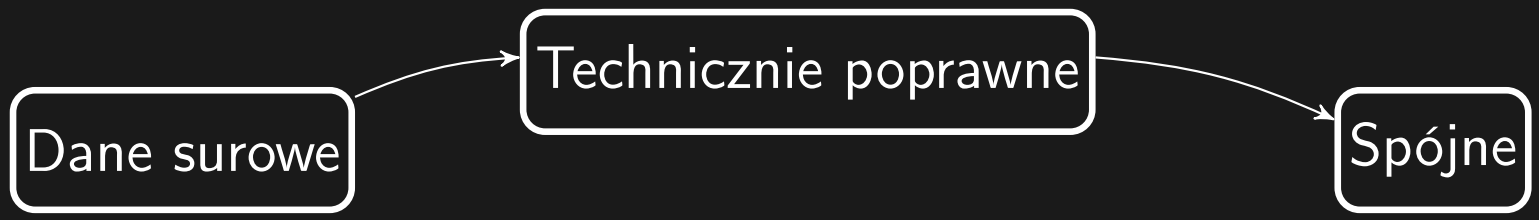
Przepływ danych

Dane surowe

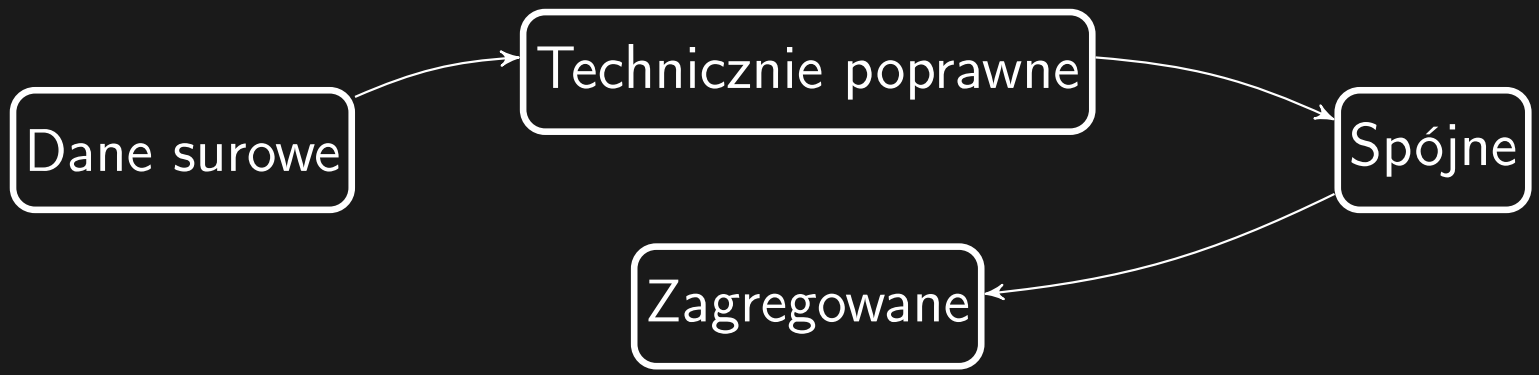
Przepływ danych



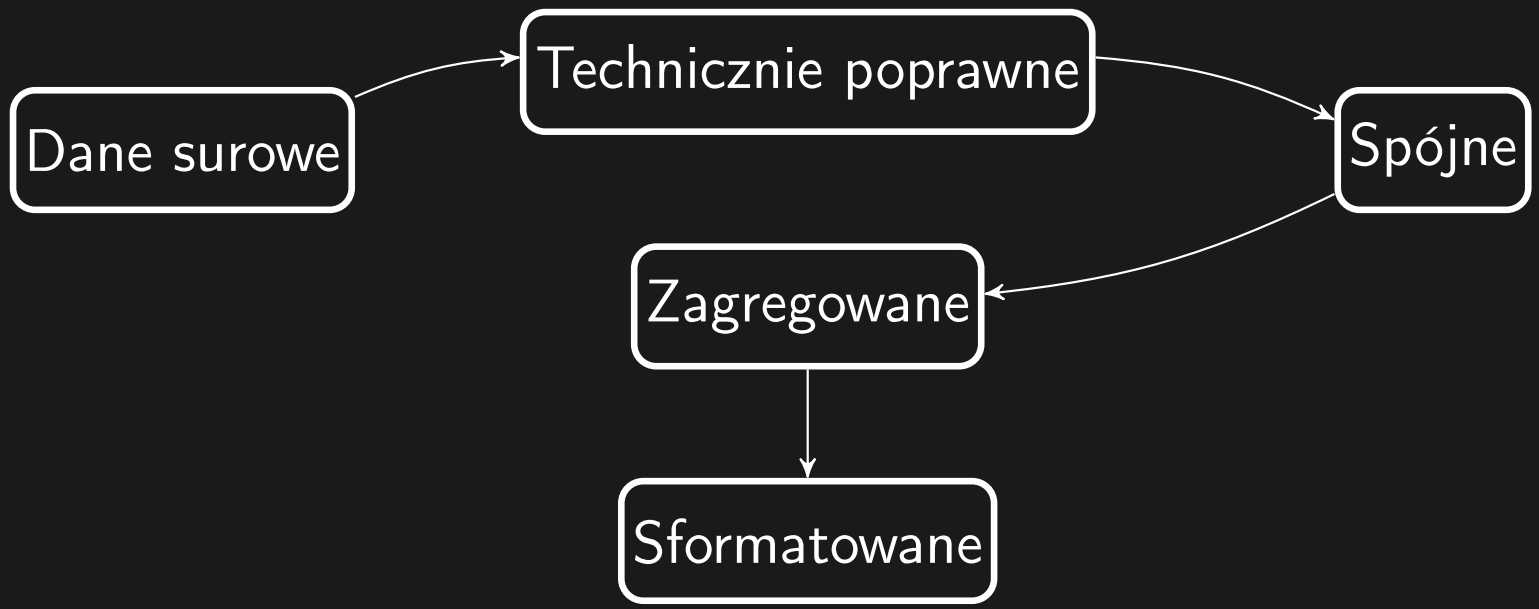
Przepływ danych



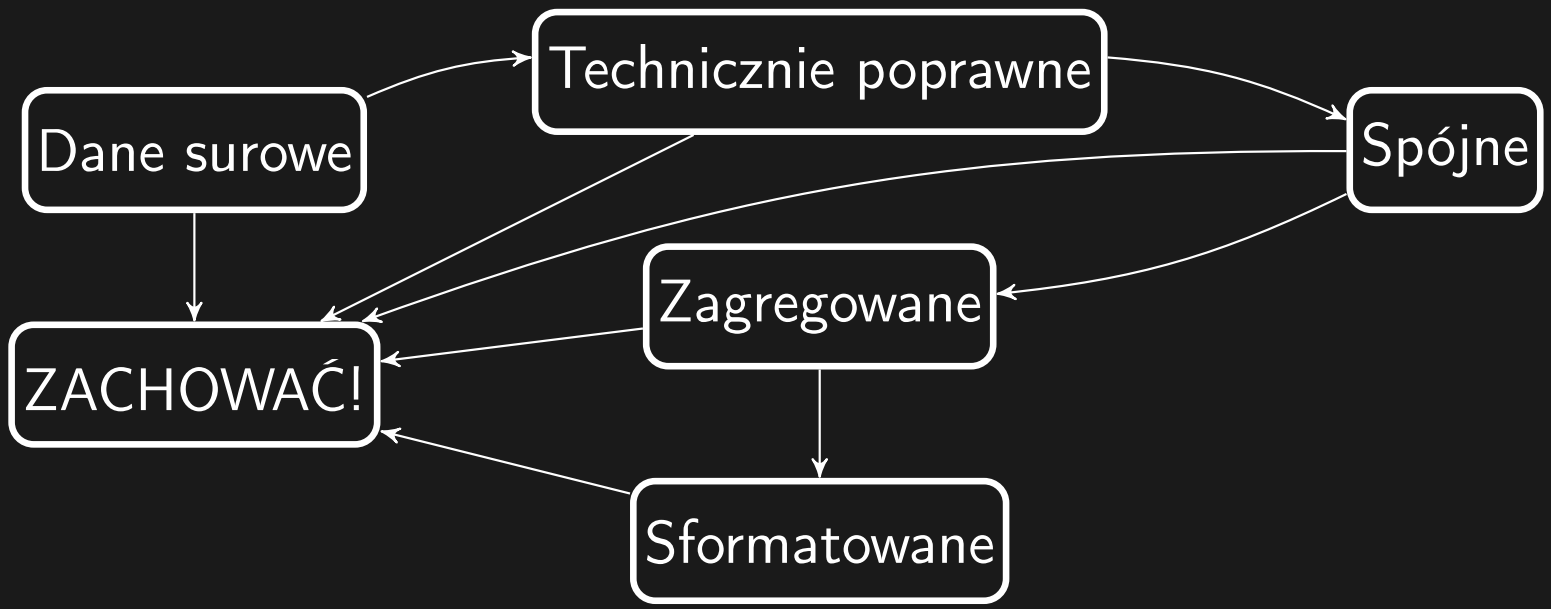
Przepływ danych



Przepływ danych



Przepływ danych



Użyteczność i wiarygodność

Trzeba rozróżnić:

- Trafność: siła predykcji danej zmiennej
- Wiarygodność: jaka jest szansa, że nasza zmienna posiada właściwą wartość, taką jak w rzeczywistości

Wynikowa użyteczność zmiennej to połączenie trafności i wiarygodności.

Outline

Cel

Cel: osiągnąć technicznie poprawne dane. Co to znaczy:

- Każda wartość jaką mamy w systemie może zostać przypisana do konkretnej zmiennej

Cel

Cel: osiągnąć technicznie poprawne dane. Co to znaczy:

- Każda wartość jaką mamy w systemie może zostać przypisana do konkretnej zmiennej
- Każda zmienna posiada odniesienie do rzeczywistości

Cel

Cel: osiągnąć technicznie poprawne dane. Co to znaczy:

- Każda wartość jaką mamy w systemie może zostać przypisana do konkretnej zmiennej
- Każda zmienna posiada odniesienie do rzeczywistości
- Typ zmiennej jest dobrany tak, aby reprezentować rzeczywistość

Cel

Cel: osiągnąć technicznie poprawne dane. Co to znaczy:

- Każda wartość jaką mamy w systemie może zostać przypisana do konkretnej zmiennej
- Każda zmienna posiada odniesienie do rzeczywistości
- Typ zmiennej jest dobrany tak, aby reprezentować rzeczywistość
- Każda wartość dla danej zmiennej ma odpowiedni typ

Cel

Cel: osiągnąć technicznie poprawne dane. Co to znaczy:

- Każda wartość jaką mamy w systemie może zostać przypisana do konkretnej zmiennej
- Każda zmienna posiada odniesienie do rzeczywistości
- Typ zmiennej jest dobrany tak, aby reprezentować rzeczywistość
- Każda wartość dla danej zmiennej ma odpowiedni typ

Dla R oznacza to, że dane są przechowywane w ramce w której każda kolumna jest poprawnie nazwana i ma odpowiedni typ.

Specjalne wartości

Not available

```
NA
is.na
```

NULL

```
NULL
is.null
```

Infinity

```
Inf
```

Not a number

```
NaN
is.nan
```

Czytanie z pliku

Jeżeli mamy szczęście wystarczy:

```
read.cvs
```

wraz z konwersją typów, o której później.
Jak plik nie jest tak regularny to musimy użyć:

```
readLines()
grepl()
strsplit()
```

Do tego może się przydać równoległość obliczeń z pakietu:

```
parallel
```

Typy zmiennych

Jeżeli dane są już przygotowane i wydają się poprawne zawsze należy je przetestować z wykorzystaniem funkcji:

```
head
str
summary
```

Typy danych

Gotowa ramka danych powinna mieć przypisane typy, można to zrobić z wykorzystaniem funkcji:

```
as.numeric    as.logical
as.integer    as.factor
as.character  as.ordered
```

Osobnym typem jest data, tu trzeba indywidualnie analizować dane, pmocna może być biblioteka:

```
lubridate
dmy   myd   ydm   mdy   dym   ymd
```

Przetestowanie posiadanych typów można zrobić z wykorzystaniem funkcji:

```
sapply(ramkaDanych, class)
```

Outline

Cel

Cel: otrzymać dane nadające się do analizy. Co to znaczy, że wszystkie:

- wartości specjalne (NA, NULL, Inf, NaN)
- "oczywiste" błędy
- wartości odstające

są usunięte, wyjaśnione lub zamienione (kolejność alfabetyczna, zawsze trzeba wyjaśnić!).

Cel

Cel: otrzymać dane nadające się do analizy. Co to znaczy, że wszystkie:

- wartości specjalne (NA, NULL, Inf, NaN)
- "oczywiste" błędy
- wartości odstające

są usunięte, wyjaśnione lub zamienione (kolejność alfabetyczna, zawsze trzeba wyjaśnić!).
To jest najbardziej ludzka część analizy ...

Cel

Cel: otrzymać dane nadające się do analizy. Co to znaczy, że wszystkie:

- wartości specjalne (NA, NULL, Inf, NaN)
- "oczywiste" błędy
- wartości odstające

są usunięte, wyjaśnione lub zamienione (kolejność alfabetyczna, zawsze trzeba wyjaśnić!).
To jest najbardziej ludzka część analizy ...
Spójność może być testowana wewnątrz zmiennej (10^9km), pomiędzy zmiennymi (wiek 3 lata i status małżeński: tak), pomiędzy bazami danych (liczba użytkowników korzystających z onetu jest znacząco różna dla pomiaru z Gdańska i Przemyśla).

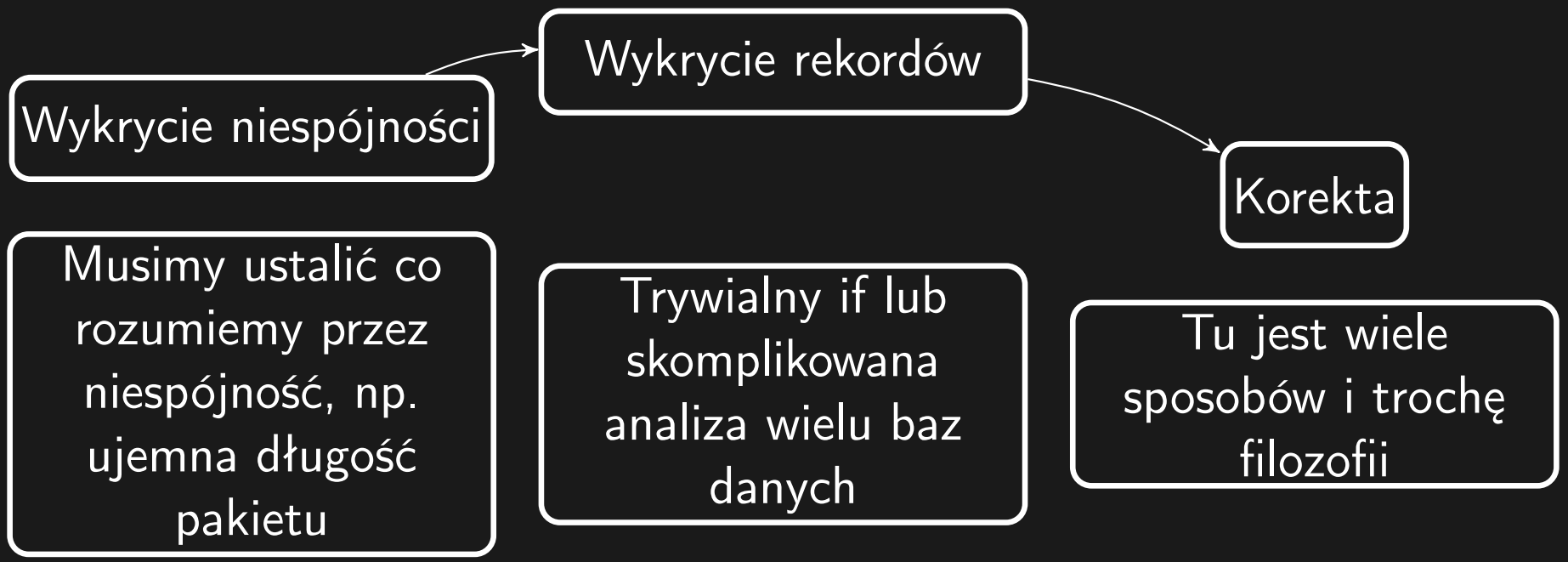
Przepływ danych

Wykrycie niespójności

Musimy ustalić co rozumiemy przez niespójność, np. ujemna długość pakietu

Przepływ danych

Przepływ danych



Wartości specjalne

Wykrywanie w R jest proste. Famy funkcje:

```
is.finite
is.na
```

Wiele funkcji statystycznych posiada opcję usunięcia braku danych:

```
sum(length, na.rm=TRUE)
```

Pamiętajmy, że wartości tych nie należy kodować. Brak typu aplikacji czy NA dla typu aplikacji to dwie różne wartości! Pamiętajmy, że NA może być zakodowane w danych. Dla przykładu, jeżeli aplikacja jest nie rozpoznana wpisujemy 0. W takiej sytuacji powinniśmy zamienić 0 na NA.

Wartości odstające

Temat rzeka. Trzeba być bardzo uważnym żeby nie usunąć wartości istotnych.

```
boxplot.stats(x, coef = 2)
```

Wartości odstające

Temat rzeka. Trzeba być bardzo uważnym żeby nie usunąć wartości istotnych.
Podstawowy sposób to wykres pudełkowy.

$$x_{0.25} - r(x_{0.75} - x_{0.25}) \leftrightarrow x_{0.75} + r(x_{0.75} - x_{0.25}) \qquad (1)$$

gdzie $r = 1.5$, choć dokładna implementacja w R jest nieco inna.

```
boxplot.stats(x, coef = 2)
```

Wartości odstające

Temat rzeka. Trzeba być bardzo uważnym żeby nie usunąć wartości istotnych.
Podstawowy sposób to wykres pudełkowy.

$$x_{0.25} - r(x_{0.75} - x_{0.25}) \leftrightarrow x_{0.75} + r(x_{0.75} - x_{0.25}) \quad (1)$$

gdzie $r = 1.5$, choć dokładna implementacja w R jest nieco inna.

```
boxplot.stats(x, coef = 2)
```

Zauważmy, że powyższy sposób nie zadziała dla rozkładów typu Gamma i wartości bliskich 0. Jeżeli x ma tylko wartości dodatnie i spodziewamy się wartości odstających bliskich zeru to może wykorzystać metodę Hiridoglou i Berthelot daną równaniem:

$$h(x) = \max(\frac{x}{x_m}, \frac{x_m}{x}) \geq r \quad (2)$$

gdzie x_m to mediana. Implementacja w pliku R.

Błędy oczywiste

Błędy oczywiste to takie które przeczą logice.

- Ujemna długość pakietu
- Pakiet IP o długości miliona bajtów
- Pakiet UDP posiadający w nagłówku wartość ACK
- Pakiet z polem ACK, który został rozpoznany jako pakiet UDP

Takie błędy powinny być sprawdzone. Dobrym sposobem jest pakiet

```
editrules
```

Korekcja błędów

Ponownie można znaleźć wiele sposobów na korektę błędów. Pomocny może być pakiet:

`deducorrect`

- Najlepiej znaleźć poprawną wartość w oparciu o inne wartości
- Wstawianie średniej. Dla telekomunikacji bardzo niebezpieczne
- Pewien model typu $y = ax$ gdzie x jest wartością znaną
- Wartość losowa wylosowana spośród innych wartości x
- Wykorzystanie sąsiadów KNN
- Co wymyślicie

Na koniec trzeba wrócić do testowania oczywistych błędów bo dodane zmienne mogą przeczyć pewnym regułą. Możemy np. wylosować rozmiar pakietu inny niż spotykany w tej sieci.

Czytanie danych z WWW

R