What is tidy R?

Jonas Schöley

jschoeley@health.sdu.dk

Department of Public Health
University of Southern Denmark

Let’s write a program

Here’s 7,826 life-tables...

...fit a (Gompertz) curve to each life-table...

...and make a scatter plot of all estimated a and b parameters.

Base R

```r
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total counts, drop NAs
hmd_sub <-
  na.omit(subset(hmd_counts, age >= 30 & age < 80 & sex != 'Total'))
# split the data by sex, country and year
hmd_split <-
  split(hmd_sub, list(hmd_sub$sex, hmd_sub$country, hmd_sub$period),
        drop = TRUE)
# run a linear regression on each subset
hmd_regress <-
  lapply(hmd_split,
         function (lt) glm(round(nDx, 0) ~ I(age-30) + offset(log(nEx)),
                           family = 'poisson', data = lt))
# extract the coefficients from each regression model
hmd_coef <- t(sapply(hmd_regress, coef))
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coef[,2],
     main = 'Gompertz correlation', xlab = 'a', ylab = 'b')
```

Base R

```
# load data
load('data/hmd/hmd_counts.RData')
```

```
# A tibble: 1,304,694 x 7
   country sex    period   age    nx   nDx    nEx
   <chr>   <chr>   <int> <int> <int> <dbl>  <dbl>
 1 AUS     Female   1921     0     1 3842.  64052.
 2 AUS     Female   1921     1     1  719.  59619.
 3 AUS     Female   1921     2     1  330.  57126.
 4 AUS     Female   1921     3     1  166.  57484.
 5 AUS     Female   1921     4     1  190.  58407.
 6 AUS     Female   1921     5     1  149.  59220.
 7 AUS     Female   1921     6     1  150.  60386.
 8 AUS     Female   1921     7     1  109.  60179.
 9 AUS     Female   1921     8     1   81.0 58548.
10 AUS     Female   1921     9     1   78.0 56919.
# ... with 1,304,684 more rows
```

```
rop NAs
 < 80 & sex != 'Total'))
ountry, hmd_sub$period),
(age-30) + offset(log(nEx)),
', data = lt))
ion model
hmd_coef <- t(sapply(hmd_regress, coef))
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coef[,2],
     main = 'Gompertz correlation', xlab = 'a', ylab = 'b')
```

Base R

```
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total counts, drop NAs
hmd_sub <-
  na.omit(subset(hmd_counts, age >= 30 & age < 80 & sex != 'Total'))
```

```
# A tibble: 391,300 x 7
   country sex    period   age    nx   nDx    nEx
   <chr>   <chr>   <int> <int> <int> <dbl>  <dbl>
 1 AUS     Female   1921    30     1  183. 46315.
 2 AUS     Female   1921    31     1  148. 45239.
 3 AUS     Female   1921    32     1  197. 44581.
 4 AUS     Female   1921    33     1  213. 43609.
 5 AUS     Female   1921    34     1  201. 42276.
 6 AUS     Female   1921    35     1  180. 41148.
 7 AUS     Female   1921    36     1  199. 39935.
 8 AUS     Female   1921    37     1  212. 38196.
 9 AUS     Female   1921    38     1  238. 36662.
10 AUS     Female   1921    39     1  195. 35875.
# ... with 391,290 more rows
```

```
$country, hmd_sub$period),

I(age-30) + offset(log(nEx)),
on', data = lt))
ssion model

'a', ylab = 'b')
```

Base R

```r
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total count
hmd_sub <-
  na.omit(subset(hmd_counts, age >= 30 &
# split the data by sex, country and year
hmd_split <-
  split(hmd_sub, list(hmd_sub$sex, hmd_sub$country, hmd_sub$period),
        drop = TRUE)
# run a linear regression on each subset
hmd_regress <-
  lapply(hmd_split,
         function (lt) glm(round(nDx, 0)
                           family = 'poi
# extract the coefficients from each reg
hmd_coef <- t(sapply(hmd_regress, coef))
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coef[,2],
     main = 'Gompertz correlation', xlab
```

```
10 CHE     Male    1883    39     1   211 17442.
# ... with 40 more rows

$Female.DNK.1883
# A tibble: 50 x 7
   country sex     period   age    nx   nDx    nEx
   <chr>   <chr>    <int> <int> <int> <dbl>  <dbl>
 1 DNK     Female    1883    30     1  117. 14746.
 2 DNK     Female    1883    31     1  117. 14145.
 3 DNK     Female    1883    32     1  116. 13936.
 4 DNK     Female    1883    33     1  115. 14086.
 5 DNK     Female    1883    34     1  113. 13174.
 6 DNK     Female    1883    35     1  110. 12447.
 7 DNK     Female    1883    36     1  107. 12466.
 8 DNK     Female    1883    37     1  105. 12122.
 9 DNK     Female    1883    38     1  104. 12121.
10 DNK     Female    1883    39     1  103. 12048.
# ... with 40 more rows

$Male.DNK.1883
# A tibble: 50 x 7
   country sex    period   age    nx   nDx    nEx
   <chr>   <chr>   <int> <int> <int> <dbl>  <dbl>
 1 DNK     Male     1883    30     1  87.9 13708.
```

Base R

```r
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total
hmd_sub <-
  na.omit(subset(hmd_counts, age >=
# split the data by sex, country ar
hmd_split <-
  split(hmd_sub, list(hmd_sub$sex,
        drop = TRUE)
# run a linear regression on each subset
hmd_regress <-
  lapply(hmd_split,
         function (lt) glm(round(nDx, 0) ~ I(age-30) + offset(log(nEx)),
                           family = 'poisson', data = lt))
# extract the coefficients from eac
hmd_coef <- t(sapply(hmd_regress, c
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coe
     main = 'Gompertz correlation'
```

```
$Female.DNK.1883

Call:  glm(formula = round(nDx, 0) ~ I(age - 30) + offset(log(nEx)),
    family = "poisson", data = lt)

Coefficients:
(Intercept)  I(age - 30)
   -5.43680      0.06109

Degrees of Freedom: 49 Total (i.e. Null);  48 Residual
Null Deviance:      5742
Residual Deviance: 432.6        AIC: 779.5

.1883

Call:  glm(formula = round(nDx, 0) ~ I(age - 30) + offset(log(nEx)),
    family = "poisson", data = lt)

   -5.37060      0.06467

Degrees of Freedom: 49 Total (i.e. Null);  48 Residual
Null Deviance:      6136
Residual Deviance: 95.36        AIC: 444.8

 [ reached getOption("max.print") -- omitted 6826 entries ]
```

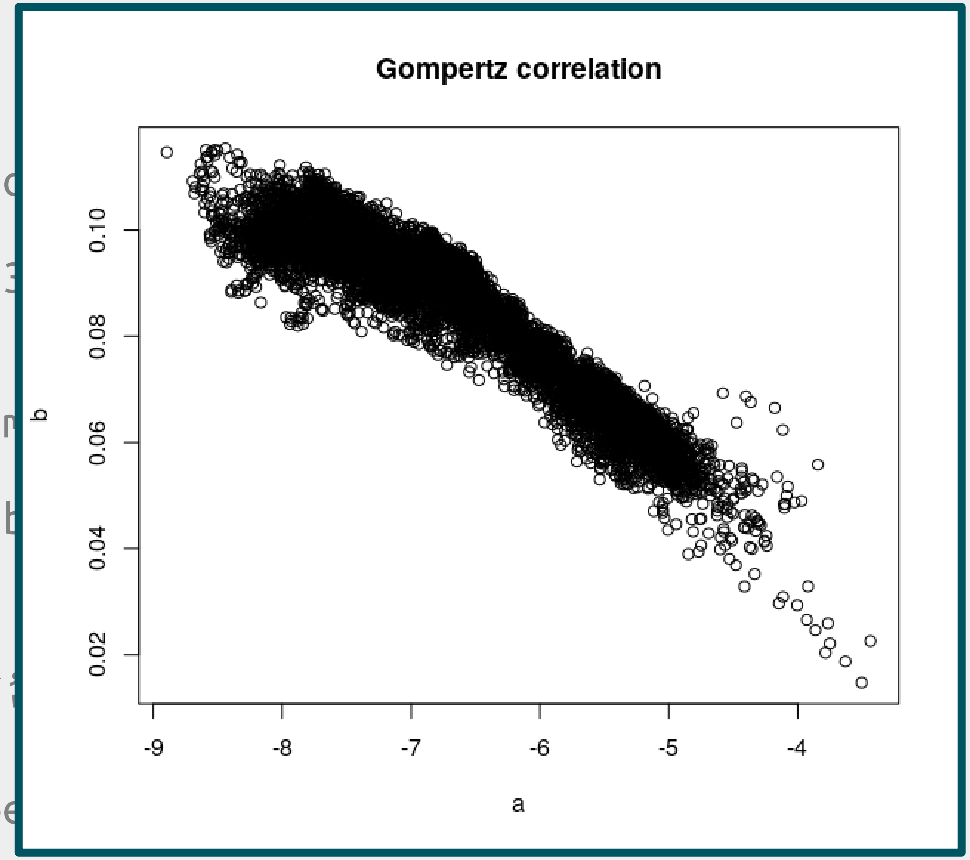
Base R

```r
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total counts
hmd_sub <-
  na.omit(subset(hmd_counts, age >= 30 & a
# split the data by sex, country and year
hmd_split <-
  split(hmd_sub, list(hmd_sub$sex, hmd_sub
        drop = TRUE)
# run a linear regression on each subset
hmd_regress <-
  lapply(hmd_split,
         function (lt) glm(round(nDx, 0)
                           family = 'pois
# extract the coefficients from each regression model
hmd_coef <- t(sapply(hmd_regress, coef))
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coef[,2],
     main = 'Gompertz correlation', xlab = 'a', ylab = 'b')
```

```
Female.BEL.1857        -5.117261  0.05780343
Male.BEL.1857          -5.233241  0.06181491
Female.DNK.1857        -5.214574  0.06009715
Male.DNK.1857          -5.130916  0.06186015
Female.FRATNP.1857     -5.154605  0.06011134
Male.FRATNP.1857       -5.239940  0.06353478
Female.GBR_SCO.1857    -5.146907  0.05534004
Male.GBR_SCO.1857      -5.003335  0.05541752
Female.GBRTENW.1857    -5.069002  0.05545477
Male.GBRTENW.1857      -5.035500  0.05710080
Female.ISL.1857        -5.584053  0.05890721
Male.ISL.1857          -4.889947  0.04974989
Female.NLD.1857        -4.872800  0.05569208
Male.NLD.1857          -4.917034  0.05910703
Female.NOR.1857        -5.343450  0.05897105
Male.NOR.1857          -5.186390  0.05689471
Female.SWE.1857        -5.013773  0.05955975
Male.SWE.1857          -4.784690  0.05668819
Female.BEL.1858        -5.144409  0.06077933
Male.BEL.1858          -5.225792  0.06393606
 [ reached getOption("max.print") -- omitted 7326 rows ]
```

Base R

```
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total co
hmd_sub <-
  na.omit(subset(hmd_counts, age >= 3
# split the data by sex, country and
hmd_split <-
  split(hmd_sub, list(hmd_sub$sex, hm
        drop = TRUE)
# run a linear regression on each sub
hmd_regress <-
  lapply(hmd_split,
         function (lt) glm(round(nDx,
                           family = '
# extract the coefficients from each
hmd_coef <- t(sapply(hmd_regress, coe
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coef[,2],
     main = 'Gompertz correlation', xlab = 'a', ylab = 'b')
```



Base R

```
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total counts, drop NAs
hmd_sub <-
  na.omit(subset(hmd_counts, age >= 30 & age < 80 & sex != 'Total'))
# split the data by sex, country and year
hmd_split <-
  split(hmd_sub, list(hmd_sub$sex, hmd_sub$country, hmd_sub$period),
        drop = TRUE)
# run a linear regression on each subset
hmd_regress <-
  lapply(hmd_split,
         function (lt) glm(round(nDx, 0) ~ I(age-30) + offset(log(nEx)),
                           family = 'poisson', data = lt))
# extract the coefficients from each regression model
hmd_coef <- t(sapply(hmd_regress, coef))
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coef[,2],
     main = 'Gompertz correlation', xlab = 'a', ylab = 'b')
```

Tidy R

```r
library(tidyverse)
# load data
load('data/hmd/hmd_counts.RData')

hmd_counts %>%
  # select ages 30 to 80, drop total counts
  filter(age >= 30, age < 80, sex != 'Total') %>%
  # drop NAs
  drop_na() %>%
  # for each period...
  group_by(period, country, sex) %>%
  # ...run a Poisson regression of deaths versus age
  do(lm = glm(round(nDx, 0) ~ I(age-30) + offset(log(nEx)),
              family = 'poisson', data = .)) %>%
  # extract the regression coefficients
  mutate(a = coef(lm)[1], b = coef(lm)[2]) %>%
  # plot a versus b coefficients and label with year
  ggplot() +
  geom_point(aes(x = a, y = b), shape = 1, size = 3) +
  labs(title = 'Gompertz correlation')
```

Assignment

```r
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total counts, drop NAs
hmd_sub <-
  na.omit(subset(hmd_counts, age >= 30 & age < 80 & sex != 'Total'))
# split the data by sex, country and year
hmd_split <-
  split(hmd_sub, list(hmd_sub$sex, hmd_sub$country, hmd_sub$period),
        drop = TRUE)
# run a linear regression on each subset
hmd_regress <-
  lapply(hmd_split,
         function (lt) glm(round(nDx, 0) ~ I(age-30) + offset(log(nEx)),
                           family = 'poisson', data = lt))
# extract the coefficients from each regression model
hmd_coef <- t(sapply(hmd_regress, coef))
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coef[,2],
     main = 'Gompertz correlation', xlab = 'a', ylab = 'b')
```

Pipes

```r
library(tidyverse)
# load data
load('data/hmd/hmd_counts.RData')

hmd_counts %>%
  # select ages 30 to 80, drop total counts
  filter(age >= 30, age < 80, sex != 'Total') %>%
  # drop NAs
  drop_na() %>%
  # for each period...
  group_by(period, country, sex) %>%
  # ...run a Poisson regression of deaths versus age
  do(lm = glm(round(nDx, 0) ~ I(age-30) + offset(log(nEx)),
              family = 'poisson', data = .)) %>%
  # extract the regression coefficients
  mutate(a = coef(lm)[1], b = coef(lm)[2]) %>%
  # plot a versus b coefficients and label with year
  ggplot() +
  geom_point(aes(x = a, y = b), shape = 1, size = 3) +
  labs(title = 'Gompertz correlation')
```

Various data structures

Dataframe

List

Matrix

```
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total counts, drop NAs
hmd_sub <-
  na.omit(subset(hmd_counts, age >= 30 & age < 80 & sex != 'Total'))
# split the data by sex, country and year
hmd_split <-
  split(hmd_sub, list(hmd_sub$sex, hmd_sub$country, hmd_sub$period),
        drop = TRUE)
# run a linear regression on each subset
hmd_regress <-
  lapply(hmd_split,
         function (lt) glm(round(nDx, 0) ~ I(age-30) + offset(log(nEx)),
                           family = 'poisson', data = lt))
# extract the coefficients from each regression model
hmd_coef <- t(sapply(hmd_regress, coef))
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coef[,2],
     main = 'Gompertz correlation', xlab = 'a', ylab = 'b')
```

A single data structure: The Dataframe

Dataframe

```
library(tidyverse)
# load data
load('data/hmd/hmd_counts.RData')

hmd_counts %>%
  # select ages 30 to 80, drop total counts
  filter(age >= 30, age < 80, sex != 'Total') %>%
  # drop NAs
  drop_na() %>%
  # for each period...
  group_by(period, country, sex) %>%
  # ...run a Poisson regression of deaths versus age
  do(lm = glm(round(nDx, 0) ~ I(age-30) + offset(log(nEx)),
              family = 'poisson', data = .)) %>%
  # extract the regression coefficients
  mutate(a = coef(lm)[1], b = coef(lm)[2]) %>%
  # plot a versus b coefficients and label with year
  ggplot() +
  geom_point(aes(x = a, y = b), shape = 1, size = 3) +
  labs(title = 'Gompertz correlation')
```

Various indexing styles

NSE

List index

Matrix index

```
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total counts, drop NAs
hmd_sub <-
  na.omit(subset(hmd_counts, age >= 30 & age < 80 & sex != 'Total'))
# split the data by sex, country and year
hmd_split <-
  split(hmd_sub, list(hmd_sub$sex, hmd_sub$country, hmd_sub$period),
        drop = TRUE)
# run a linear regression on each subset
hmd_regress <-
  lapply(hmd_split,
         function (lt) glm(round(nDx, 0) ~ I(age-30) + offset(log(nEx)),
                           family = 'poisson', data = lt))
# extract the coefficients from each regression model
hmd_coef <- t(sapply(hmd_regress, coef))
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coef[,2],
     main = 'Gompertz correlation', xlab = 'a', ylab = 'b')
```

A single indexing style: Non-standard evaluation

NSE

```
library(tidyverse)
# load data
load('data/hmd/hmd_counts.RData')

hmd_counts %>%
  # select ages 30 to 80, drop total counts
  filter(age >= 30, age < 80, sex != 'Total') %>%
  # drop NAs
  drop_na() %>%
  # for each period...
  group_by(period, country, sex) %>%
  # ...run a Poisson regression of deaths versus age
  do(lm = glm(round(nDx, 0) ~ I(age-30) + offset(log(nEx)),
              family = 'poisson', data = .)) %>%
  # extract the regression coefficients
  mutate(a = coef(lm)[1], b = coef(lm)[2]) %>%
  # plot a versus b coefficients and label with year
  ggplot() +
  geom_point(aes(x = a, y = b), shape = 1, size = 3) +
  labs(title = 'Gompertz correlation')
```

Important information stored in row names

```
# load data
load('data/hmd/hmd_counts.RData')

# select ages 30 to 80, drop total counts, drop NAs
hmd_sub <-
  na.omit(subset(hmd_counts, age >= 30 & age < 80 & sex != 'Total'))
# split the data by sex, country and year
hmd_split <-
  split(hmd_sub, list(hmd_sub$sex, hmd_sub$country, hmd_sub$period),
        drop = TRUE)
# run a l
hmd_regre
  lapply(
                                    ge-30) + offset(log(nEx)),
                                     data = lt))
# extract                           n model
hmd_coef
# plot a versus b coefficients
plot(x = hmd_coef[,1], y = hmd_coef[,2],
     main = 'Gompertz correlation', xlab = 'a', ylab = 'b')
```

	(Intercept)	I(age - 30)
Female.SWE.1751	-4.954454	0.05379309
Male.SWE.1751	-4.834564	0.05463634
Female.SWE.1752	-5.109355	0.05332976
Male.SWE.1752	-4.973304	0.05313042
Female.SWE.1753	-5.176551	0.05545279
Male.SWE.1753	-4.946853	0.05258433

Every variable in its own column

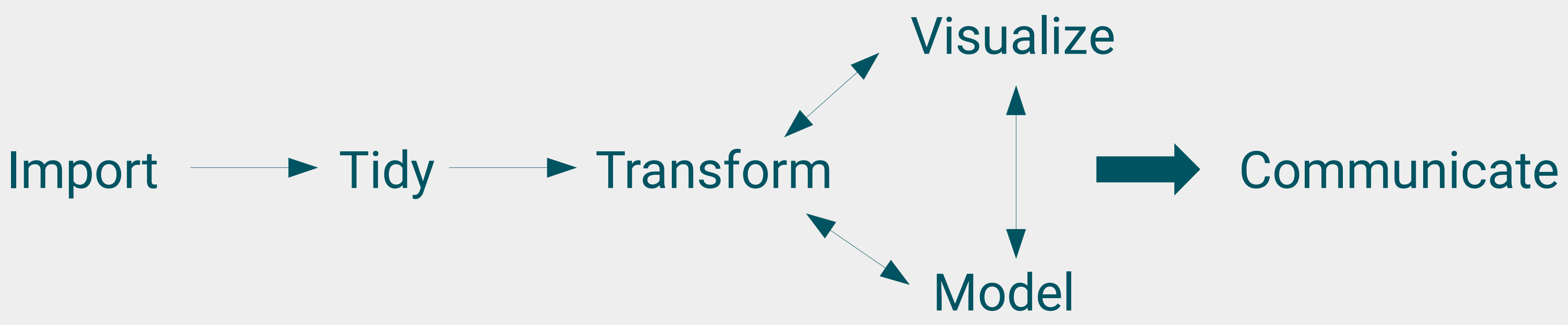
```
library(tidyverse)
# load data
load('data/hmd/hmd_counts.RData')

hmd_counts %>%
  # select ages 30 to 80, drop total
  filter(age >= 30, age < 80, sex !=
  # drop NAs
  drop_na() %>%
  # for each period...
  group_by(period, country, sex) %>%
  # ...run a Poisson regression of d
  do(lm = glm(round(nDx, 0) ~ I(age-
              family = 'poisson', da
  # extract the regression coefficie
  mutate(a = coef(lm)[1], b = coef(lm)[2]) %>%
  # plot a versus b coefficients and label with year
  ggplot() +
  geom_point(aes(x = a, y = b), shape = 1, size = 3) +
  labs(title = 'Gompertz correlation')
```

```
# A tibble: 7,826 x 6
   period country sex    lm             a      b
    <int> <chr>   <chr>  <list>     <dbl>  <dbl>
 1   1751 SWE     Female <S3: glm> -4.95 0.0538
 2   1751 SWE     Male   <S3: glm> -4.83 0.0546
 3   1752 SWE     Female <S3: glm> -5.11 0.0533
 4   1752 SWE     Male   <S3: glm> -4.97 0.0531
 5   1753 SWE     Female <S3: glm> -5.18 0.0555
 6   1753 SWE     Male   <S3: glm> -4.95 0.0526
 7   1754 SWE     Female <S3: glm> -5.11 0.0561
 8   1754 SWE     Male   <S3: glm> -4.82 0.0528
 9   1755 SWE     Female <S3: glm> -5.03 0.0551
10   1755 SWE     Male   <S3: glm> -4.82 0.0523
# ... with 7,816 more rows
```

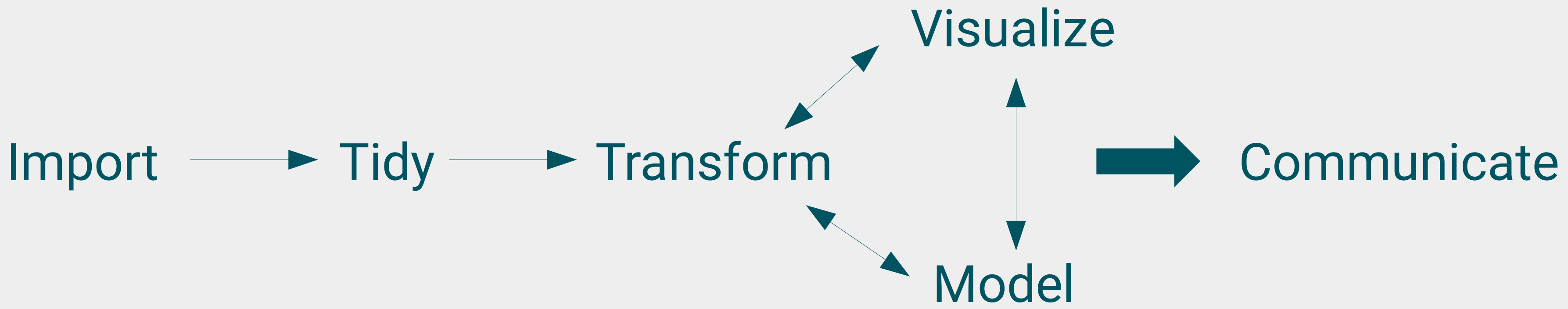
Tidy principles

Readability **Modularity** **Consistency**

A typical data analysis workflow



The tidyverse



github.com/jschoeley/ced18-tidyr

Jonas Schöley

jschoeley@health.sdu.dk

Twitter: @jschoeley