Engineering data analysis

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair

Department of Statistics / Rice University

June 2011

1. What is data analysis?
2. Why use a programming language?
3. Why use R?
4. Why use DSLs within R?
5. Case study: Mexico mortality

Data analysis is the process by which data becomes understanding, knowledge and insight

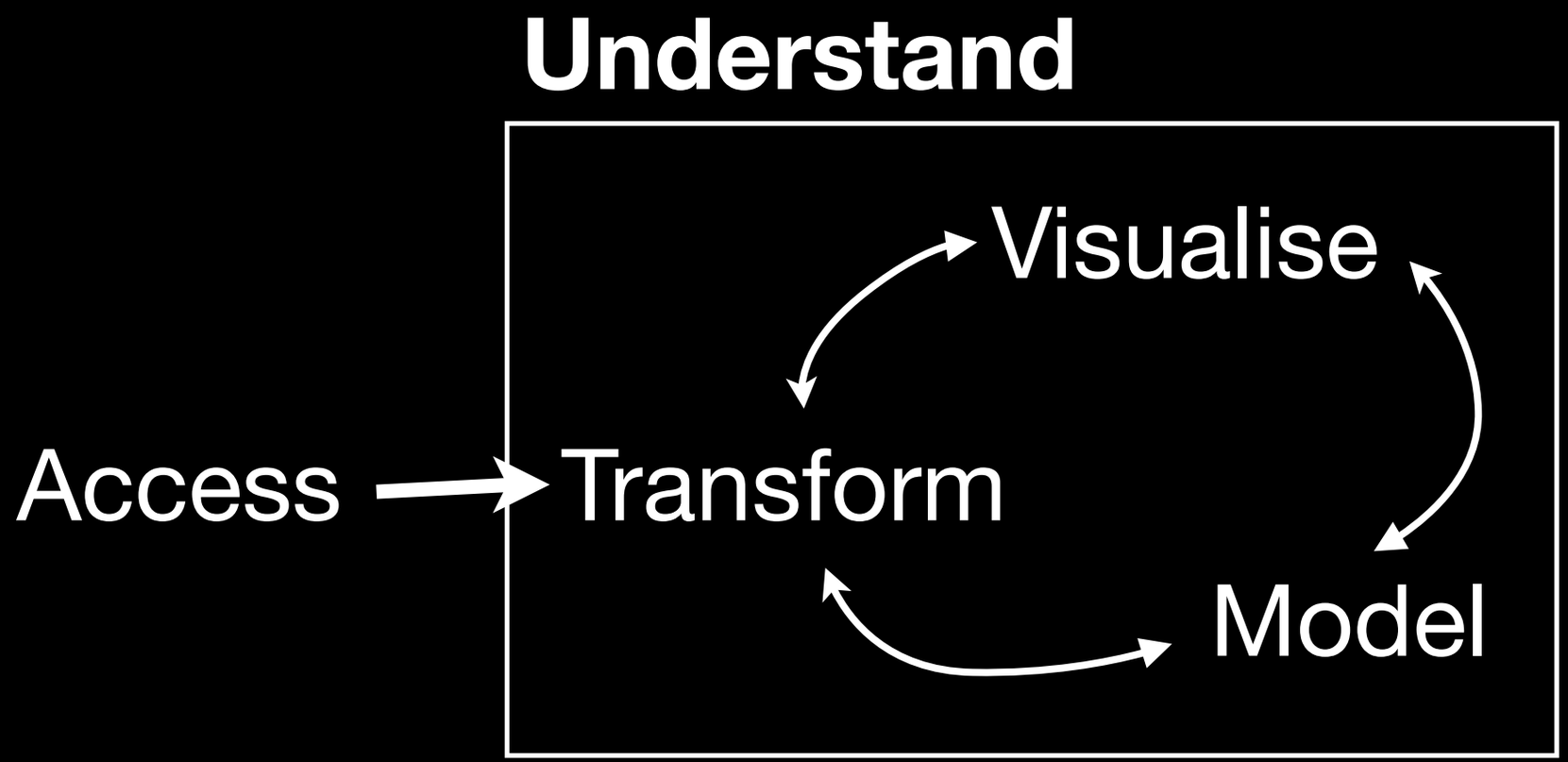
Data analysis is the process by which data becomes understanding, knowledge and insight

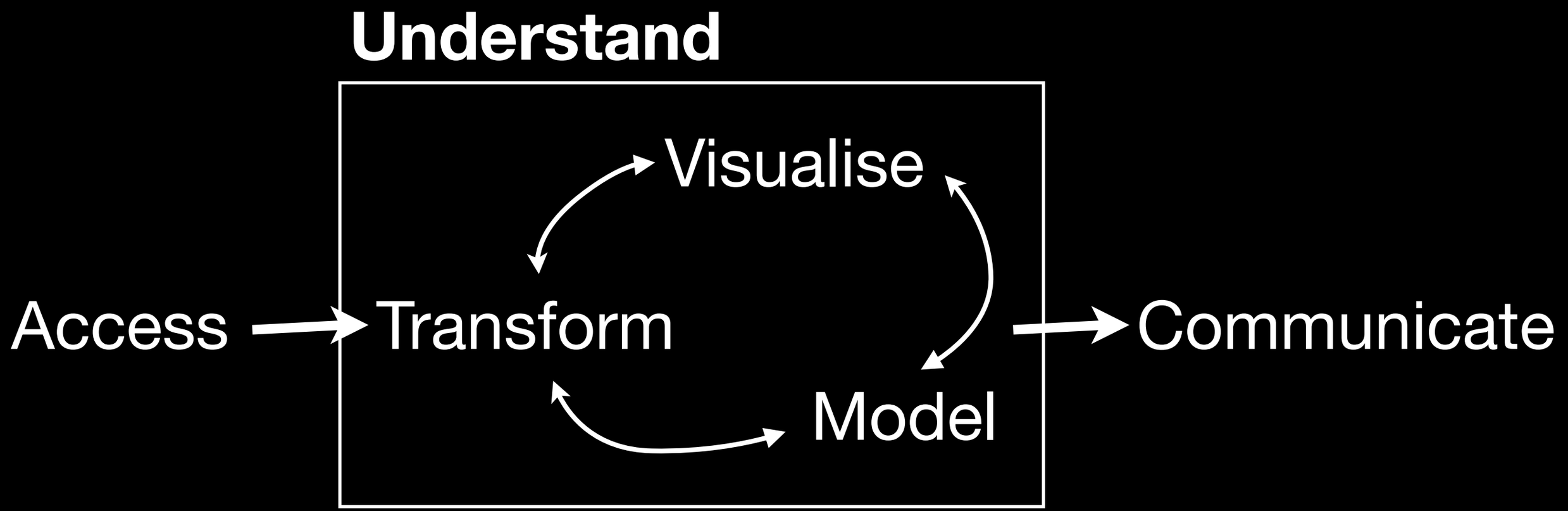
Data analysis is the process by which data becomes understanding, knowledge and insight

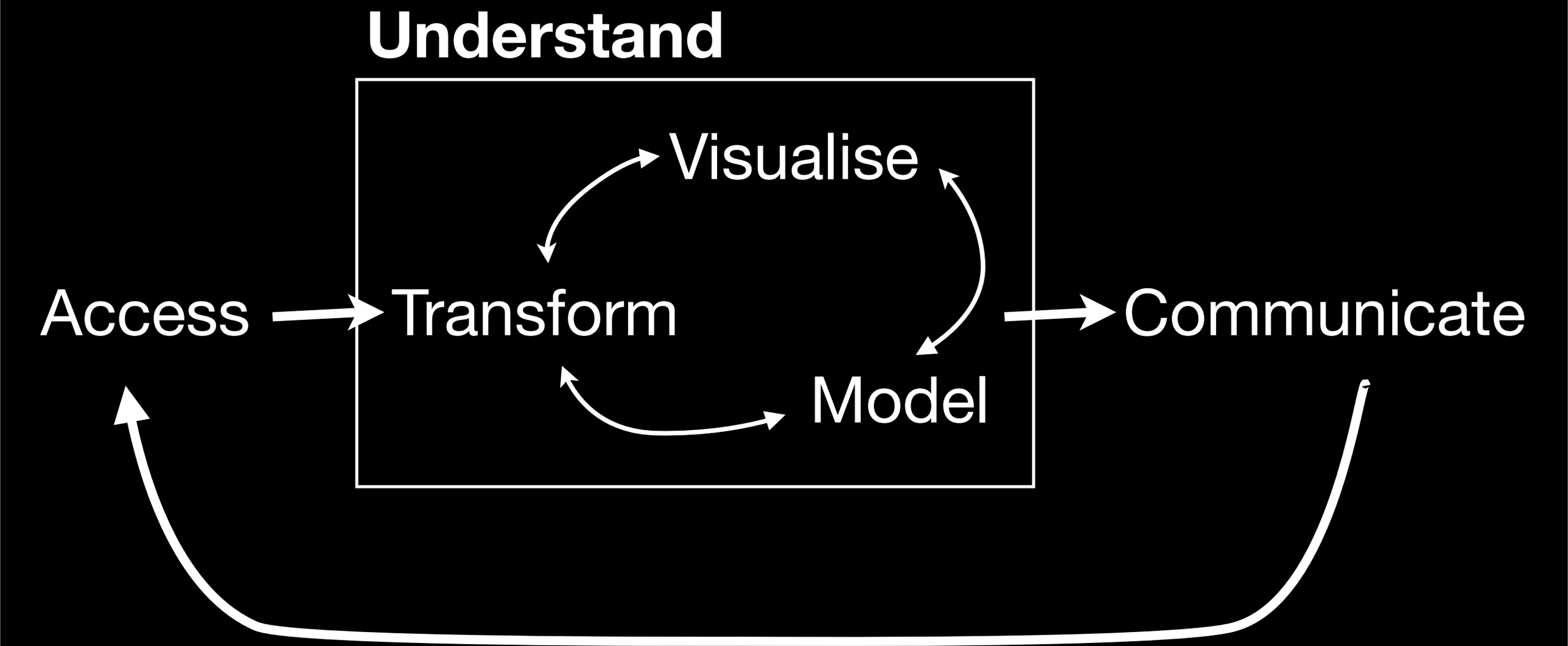
Access

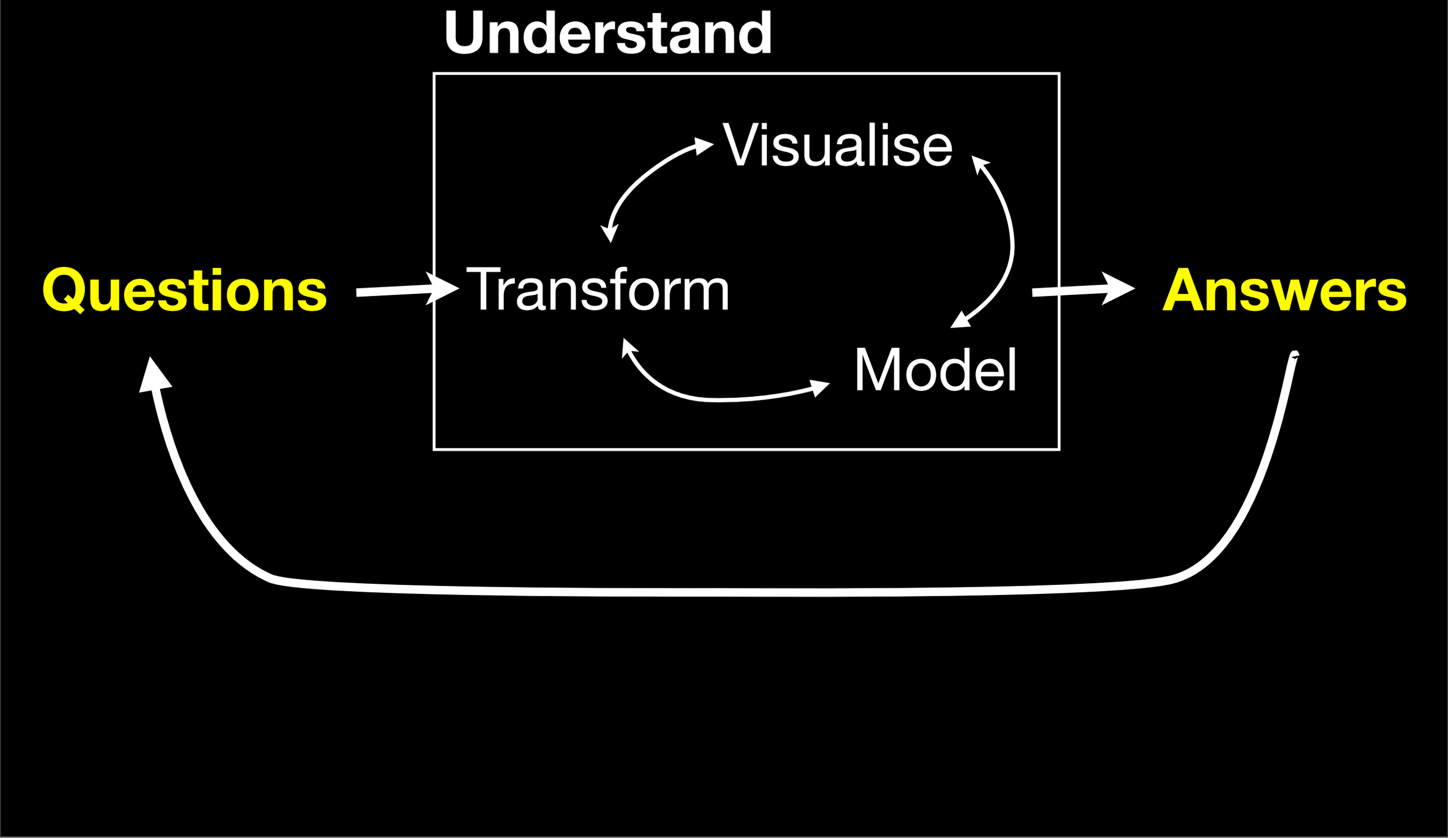
Understand

Access →


Understand
Visualise
Access
Transform
Model

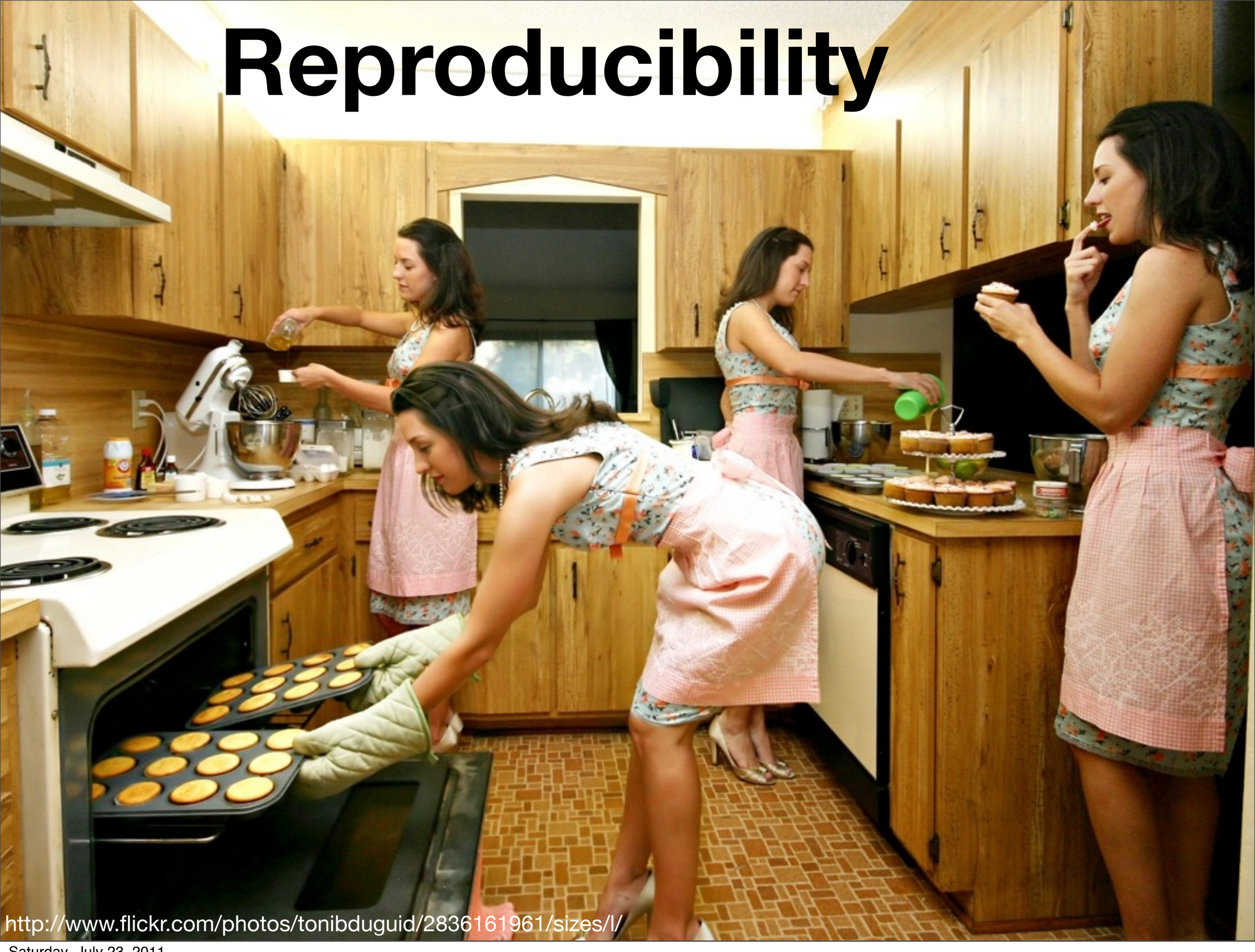

Understand
Visualise
Access
Transform
Communicate
Model


Understand
Visualise
Access
Transform
Communicate
Model


Understand
Visualise
Questions
Transform
Answers
Model

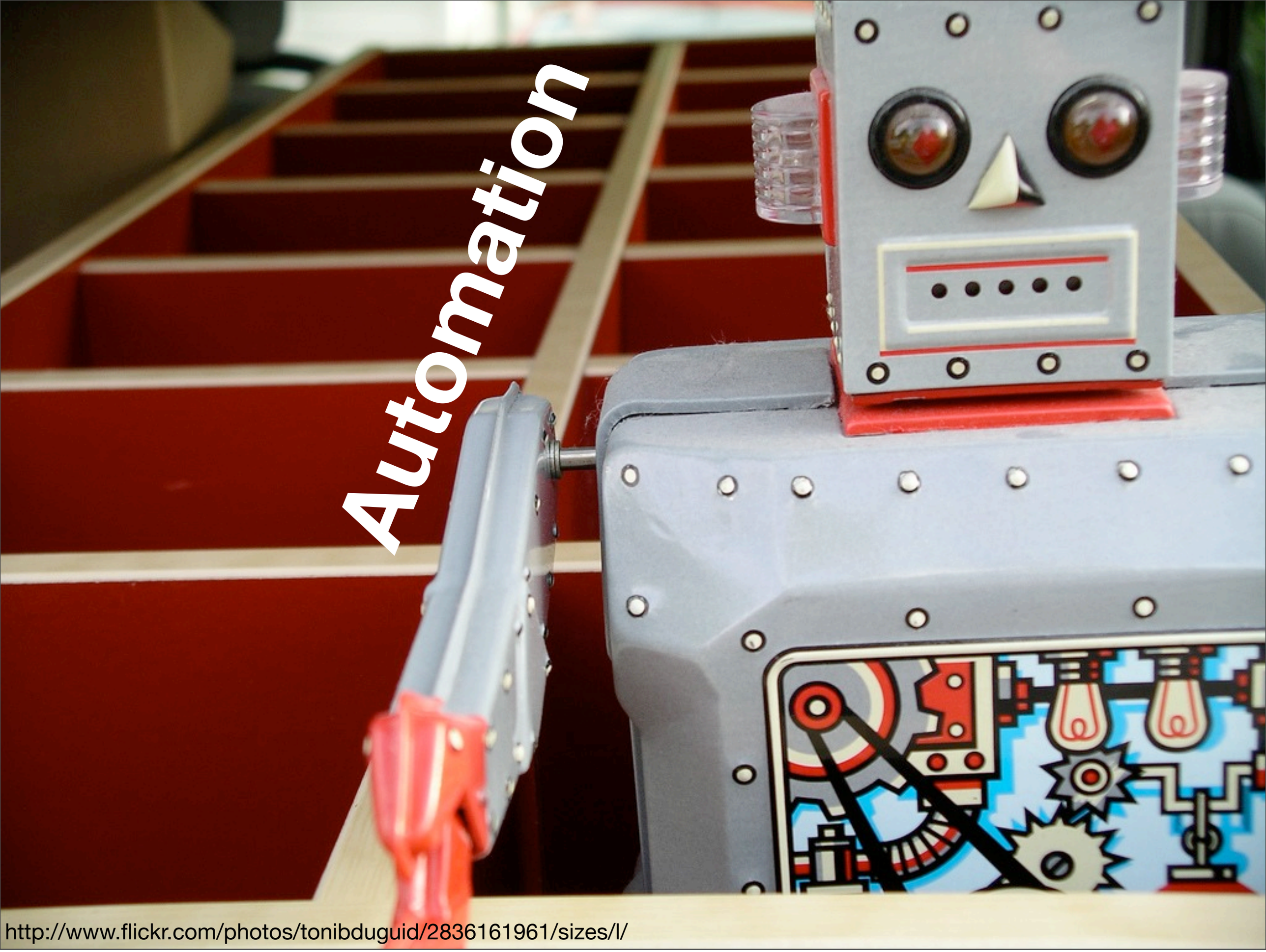
Why program?

Reproducibility



http://www.flickr.com/photos/tonibduguid/2836161961/sizes/l/

Automation



http://www.flickr.com/photos/tonibduguid/2836161961/sizes/l/

Just text

```
# Load data and create smaller subsets
tb <- read.csv("tb.csv")
tb2008 <- subset(tb, year == 2008)

# Choropleth map ------------------------------------------------------------
borders <- read.csv("world-borders.csv")
choro <- merge(tb2008, borders, by = "iso2")
choro <- choro[order(choro$order), ]

qplot(long, lat, data = choro, fill = cut_number(rate, 5), geom = "polygon", group =
group) + scale_fill_brewer("Rate", pal = "Blues")

# Bubble maps ---------------------------------------------------------------
centres <- read.csv("world-centres.csv")
bubble <- merge(centres, tb2008, by = "iso2")

world_coord <- coord_map(xlim = c(-180, 180), ylim = c(-50, 70))

# This is basically what a choropleth is showing us
qplot(long, lat, data = bubble, size = area, colour = rate) +
  scale_area(to = c(2, 25), legend = FALSE) +
  world_coord

# More traditional options
qplot(long, lat, data = bubble, size = rate) + world_coord
qplot(long, lat, data = bubble, size = log10(pop), colour = rate) +
  world_coord

# Even better if we add world boundaries
ggplot(bubble, aes(long, lat)) +
  geom_polygon(data = borders, aes(group = group)) +
  geom_point(aes(colour = rate)) +
  coord_map()
ggsave("world-4.png", width = 8, height = 6, dpi = 128)

# Works better if we tweak aesthetics
ggplot(bubble, aes(long, lat)) +
  geom_polygon(data = borders, aes(group = group), colour = "grey70",
```

Communication

http://www.flickr.com/photos/altemark/337248947/sizes/l/


Learning curve

Why R?

```
SEXP applyClosure(SEXP call, SEXP op, SEXP arglist, SEXP rho, SEXP suppliedenv)
{
    SEXP body, formals, actuals, savedrho;
    volatile  SEXP newrho;
    SEXP f, a, tmp;
    RCNTXT cntxt;

    /* formals = list of formal parameters */
    /* actuals = values to be bound to formals */
    /* arglist = the tagged list of arguments */

    formals = FORMALS(op);
    body = BODY(op);
    savedrho = CLOENV(op);

    /*  Set up a context with the call in it so error has access to it */

    begincontext(&cntxt, CTXT_RETURN, call, savedrho, rho, arglist, op);

    /*  Build a list which matches the actual (unevaluated) arguments
        to the formal paramters.  Build a new environment which
        contains the matched pairs.  Ideally this environment sould be
        hashed.  */

    PROTECT(actuals = matchArgs(formals, arglist, call));
    PROTECT(newrho = NewEnvironment(formals, actuals, savedrho));

    /*  Use the default code for unbound formals.  FIXME: It looks like
        this code should preceed the building of the environment so that
        this will also go into the hash table.  */

    /* This piece of code is destructively modifying the actuals list,
       which is now also the list of bindings in the frame of newrho.
       This is one place where internal structure of environment
       bindings leaks out of envir.c.  It should be rewritten
       eventually so as not to break encapsulation of the internal
       environment layout.  We can live with it for now since it only
       happens immediately after the environment creation.  LT */
```

Open source

Community

http://www.flickr.com/photos/ianlayzellphotographs/3977042044


Prickly
http://www.flickr.com/photos/meantux/367751359

Runs anywhere

http://www.flickr.com/photos/jonlucas/204213732

Build it yourself

http://www.flickr.com/photos/wwworks/2473052504

http://www.flickr.com/photos/54945394@N00/2987214939

Connectivity

http://www.flickr.com/photos/billy64/2226377312

Programming infrastructure

http://www.flickr.com/photos/rbrwr/121511103/

Domain specific languages

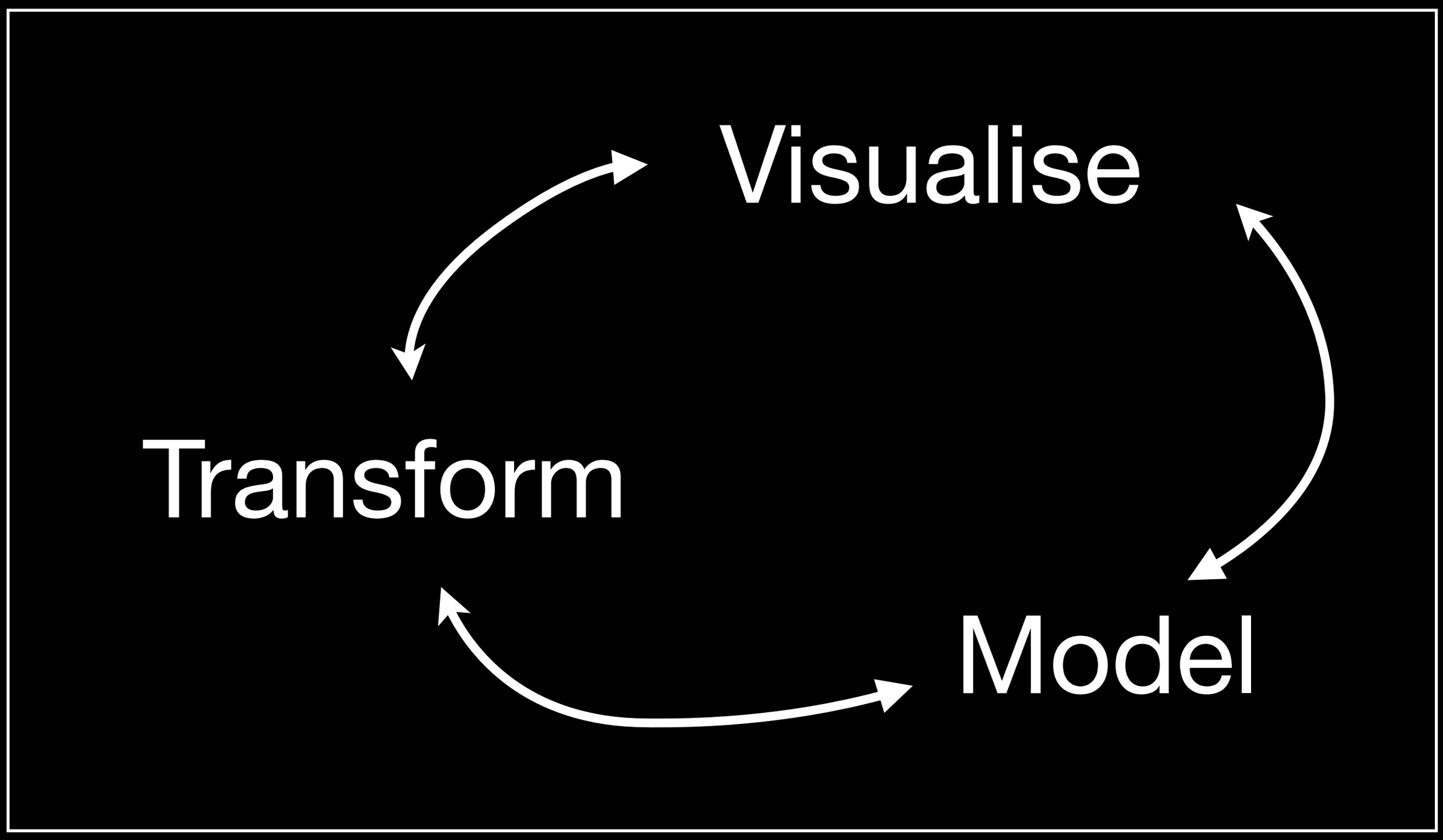
"If any number of magnitudes are each the same multiple of the same number of other magnitudes, then the sum is that multiple of the sum."

Euclid, ~300 BC

“If any number of magnitudes are each the same multiple of the same number of other magnitudes, then the sum is that multiple of the sum.”

Euclid, ~300 BC

$$ab + ac = a(b + c)$$


Visualise
Transform
Model

Model

```
y ~ x

y ~ x1 + x2

y ~ x1 * x2
y ~ x1 + x2 + x1:x2

y ~ s(x)

cbind(y1, y2) ~ x1 * x2
...
```

Visualise

```
ggplot(data, aes(x = var1, y = var2, colour = var3) +
  geom_point() +
  geom_smooth()
```

Transform

subset
mutate
arrange
summarise

*

by operator (ddply)

+

join
match_df

Case study

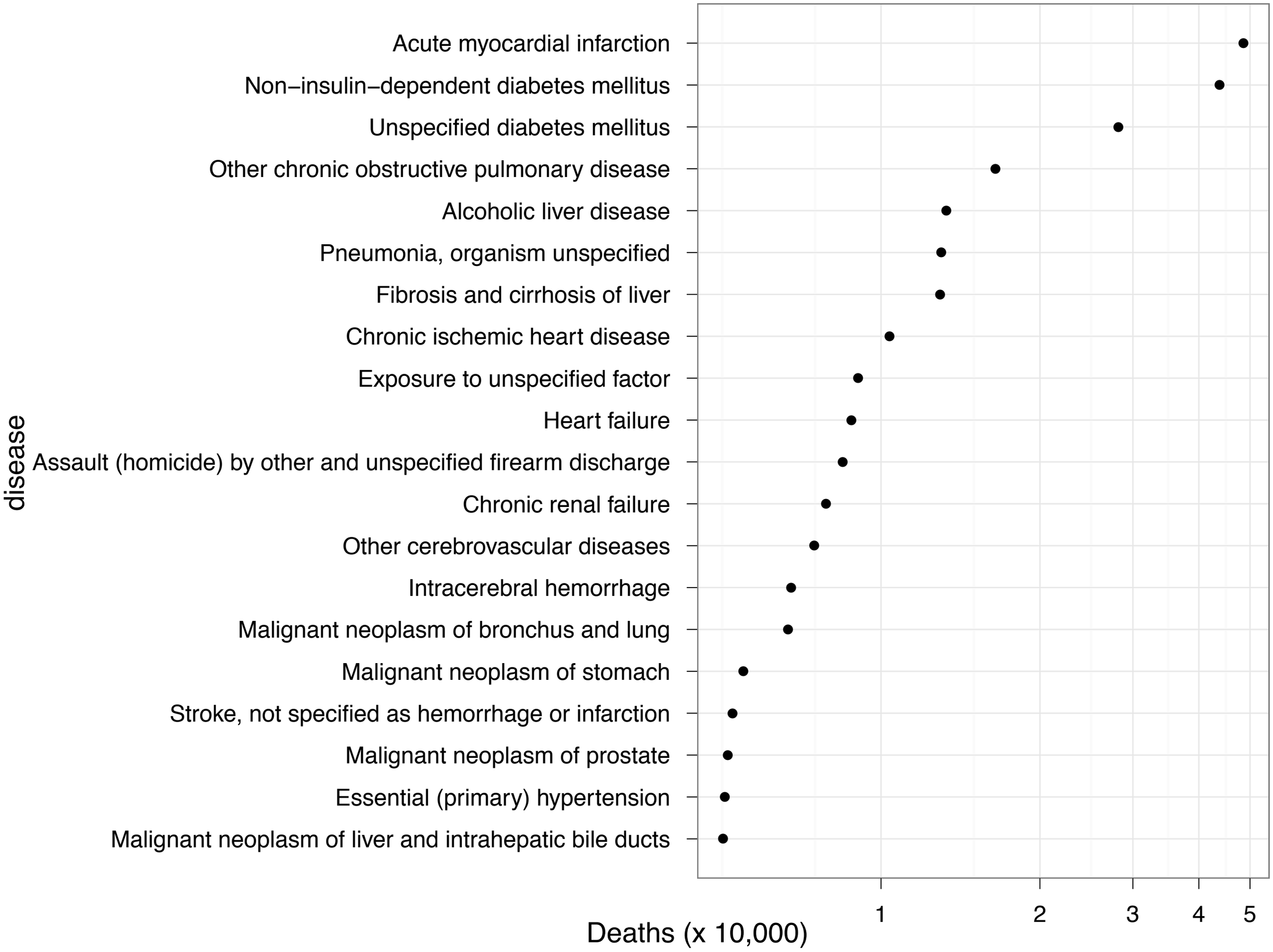
Motivation

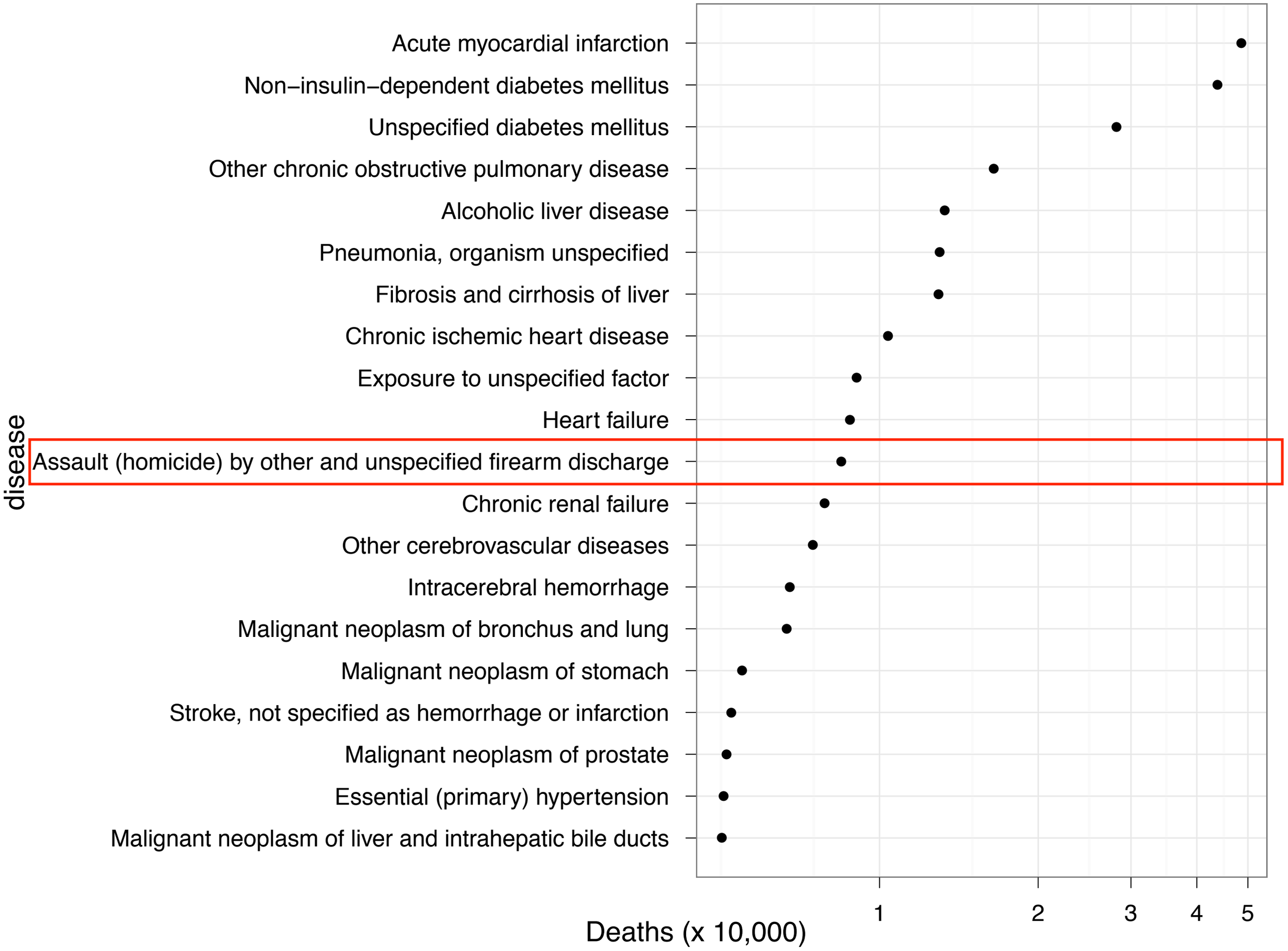
Data: Individual data on all 532,355 deaths in Mexico in 2008.

Variables: cod, hod, dod, location, dob, marital status, job, ...

Question: How do DSLs help us understand this data?

Cause of death

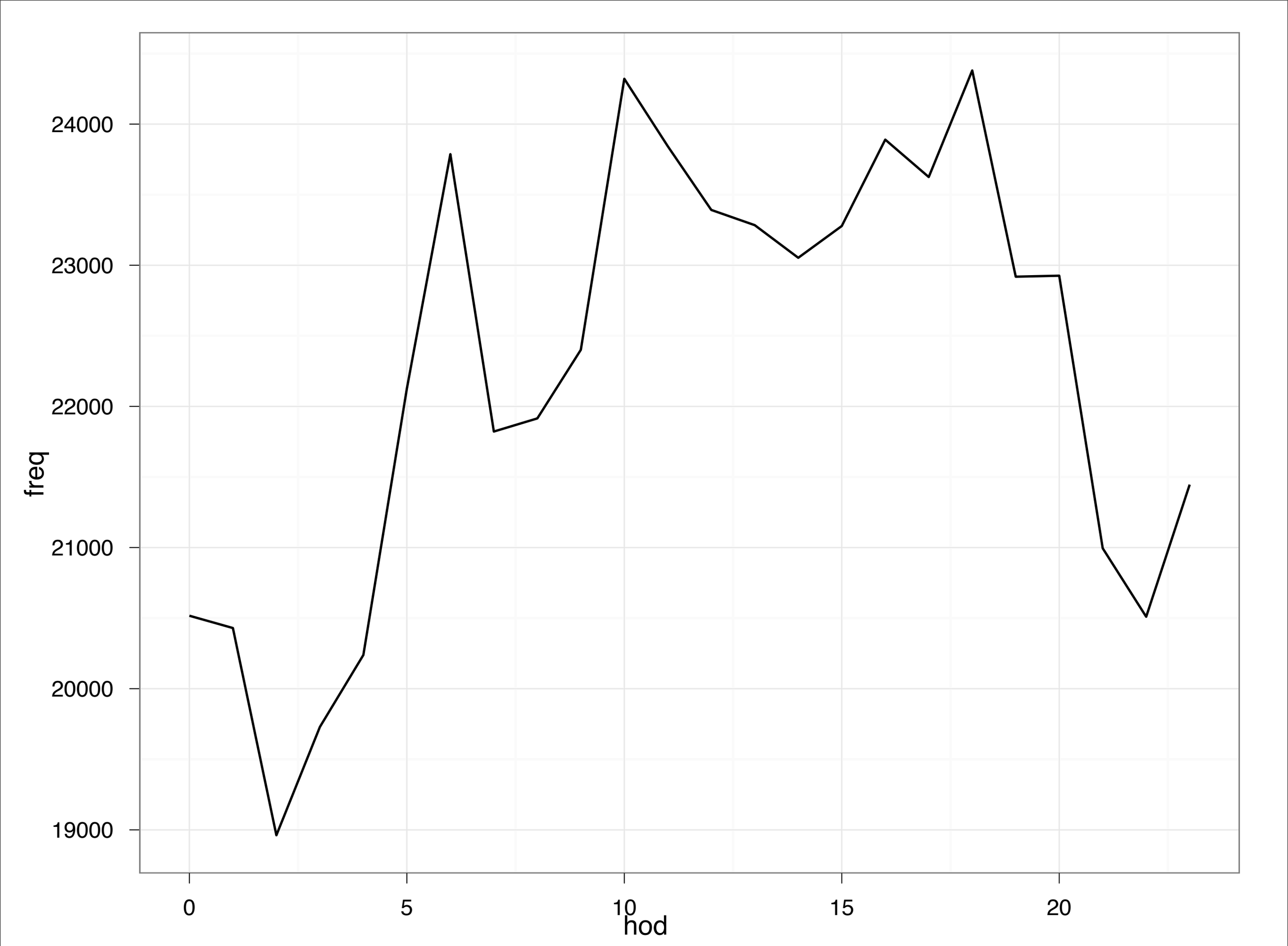

Acute myocardial infarction
Non–insulin–dependent diabetes mellitus
Unspecified diabetes mellitus
Other chronic obstructive pulmonary disease
Alcoholic liver disease
Pneumonia, organism unspecified
Fibrosis and cirrhosis of liver
Chronic ischemic heart disease
Exposure to unspecified factor
Heart failure
Assault (homicide) by other and unspecified firearm discharge
Chronic renal failure
Other cerebrovascular diseases
Intracerebral hemorrhage
Malignant neoplasm of bronchus and lung
Malignant neoplasm of stomach
Stroke, not specified as hemorrhage or infarction
Malignant neoplasm of prostate
Essential (primary) hypertension
Malignant neoplasm of liver and intrahepatic bile ducts
disease
1
2
3
4
5
Deaths (x 10,000)


Acute myocardial infarction
Non–insulin–dependent diabetes mellitus
Unspecified diabetes mellitus
Other chronic obstructive pulmonary disease
Alcoholic liver disease
Pneumonia, organism unspecified
Fibrosis and cirrhosis of liver
Chronic ischemic heart disease
Exposure to unspecified factor
Heart failure
Assault (homicide) by other and unspecified firearm discharge
Chronic renal failure
Other cerebrovascular diseases
Intracerebral hemorrhage
Malignant neoplasm of bronchus and lung
Malignant neoplasm of stomach
Stroke, not specified as hemorrhage or infarction
Malignant neoplasm of prostate
Essential (primary) hypertension
Malignant neoplasm of liver and intrahepatic bile ducts
disease
1
2
3
4
5
Deaths (x 10,000)

```r
library(ggplot2)
library(plyr)
load("deaths.rdata")

cause <- arrange(count(deaths, "cod"), desc(freq))
cause <- join(cause, codes)

ggplot(cause, aes(x = freq, y = disease) +
  geom_point() +
  scale_x_log10()
```

```r
top20 <- head(cause, 20)

ggplot(top20, aes(freq / 1e5,
                  reorder(disease, freq)) +
  geom_point() +
  scale_y_reverse(drop = T) +
  scale_x_log10("deaths (x 10,000)", breaks = 1:5)
```

Time of death


freq
24000
23000
22000
21000
20000
19000
0
5
10
15
20
hod

```
deaths$hod[deaths$hod == 99] <- NA
deaths$hod[deaths$hod == 24] <- NA

hod <- subset(count(deaths, "hod"), !is.na(hod))
ggplot(hod, aes(x = hod, y = freq)) +
  geom_line()
```

Assault (homicide) by other and unspecified firearm discharge

Assault (homicide) by sharp object

Drowning and submersion while in natural water

Exposure to unspecified electric current

Motor– or nonmotor–vehicle accident, type of vehicle unspecified

Pedestrian injured in other and unspecified transport accidents

Traffic accident of specified type but victim's mode of transport unknown

Unspecified drowning and submersion

prop

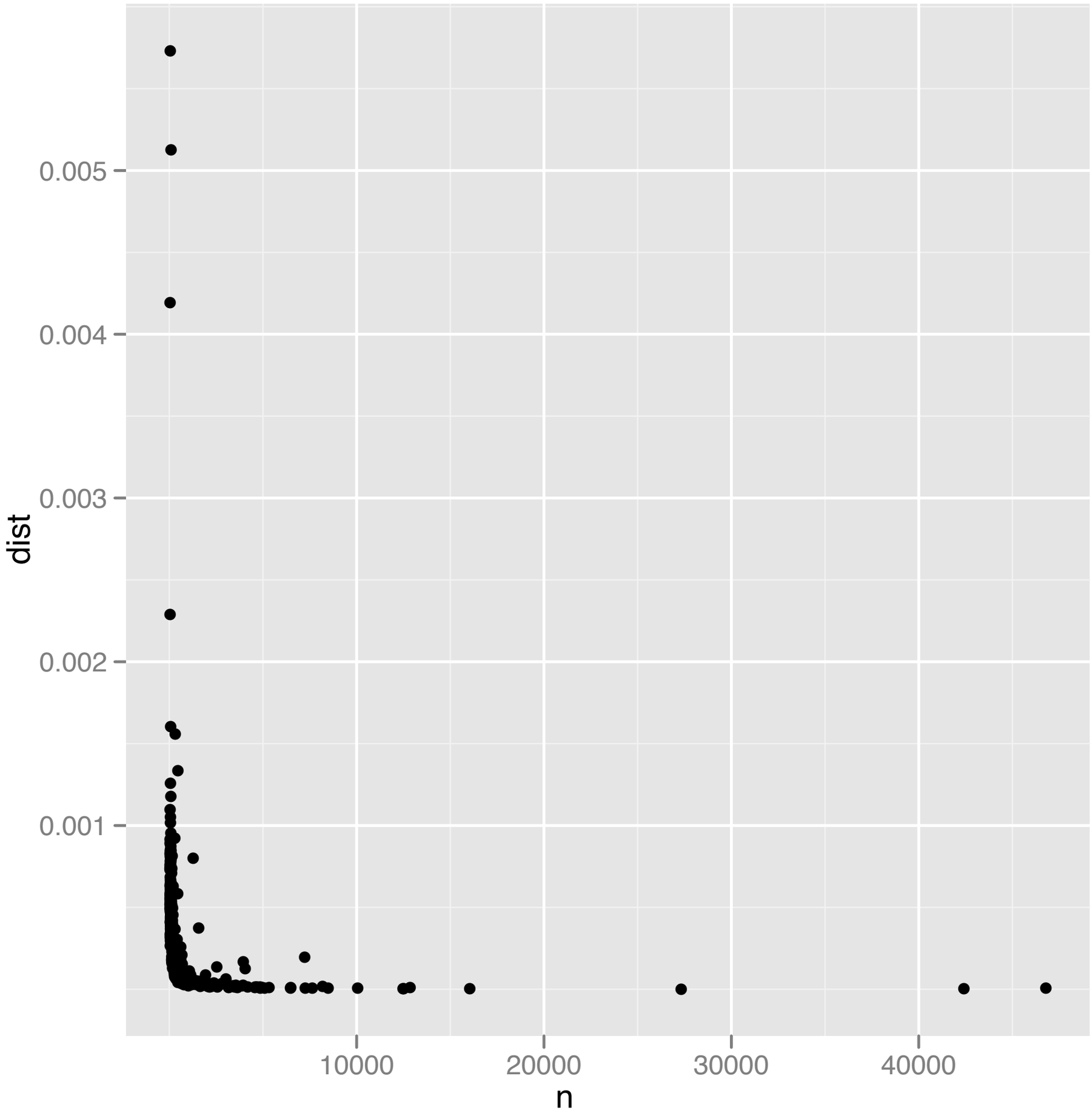
0.10
0.08
0.06
0.04
0.02

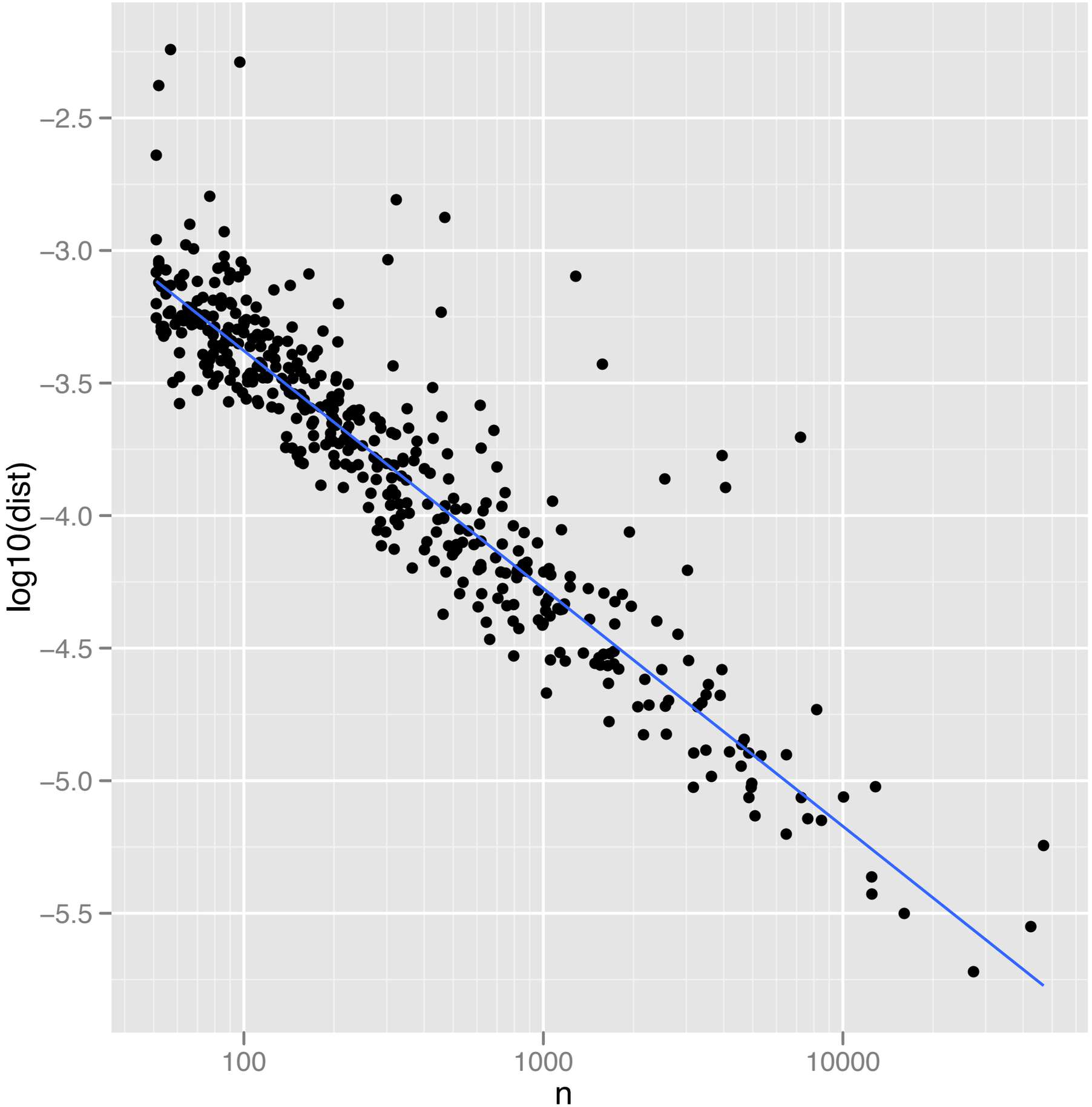
5
10
15
20

hod

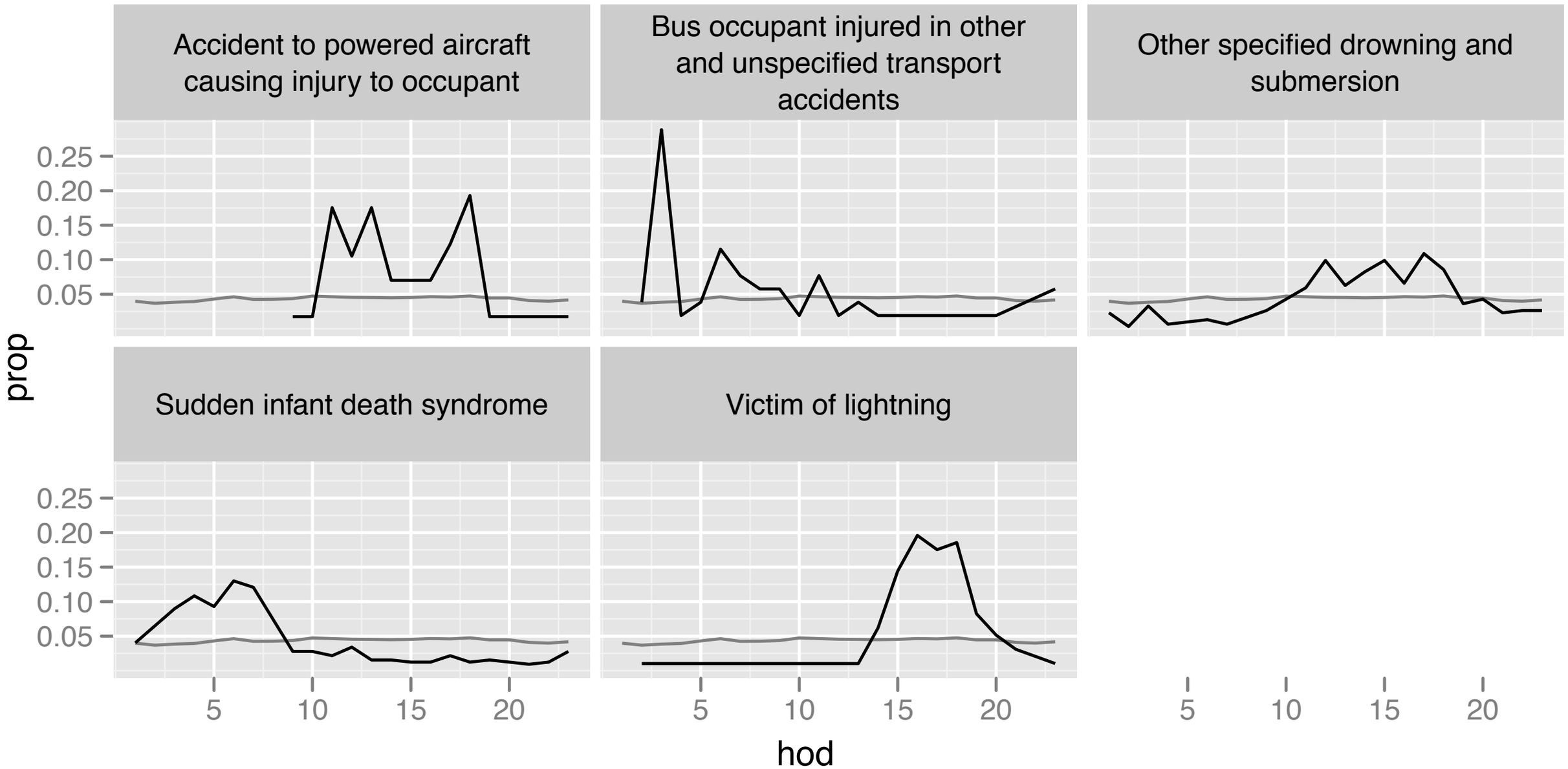
```r
# Compute deaths by hour by cause, and the
# proportion dying at each hour
hod2 <- count(deaths, c("cod", "hod"))
hod2 <- ddply(hod2, "cod", mutate,
  prop = freq / sum(freq))

# Compute for overall abundance
overall <- ddply(hod2, "hod", summarise,
  freq_all = sum(freq))
overall <- mutate(overall,
  prop_all = freq_all / sum(freq_all))

# Combine the time
hod2 <- join(overall, hod2, by = "hod")
```

```r
# Find outliers
devi <- ddply(hod2, "cod", summarise,
   n = sum(freq),
   dist = mean((prop - prop_all)^2))
devi <- subset(devi, n > 50)
```


dist
0.005
0.004
0.003
0.002
0.001
10000
20000
30000
40000
n

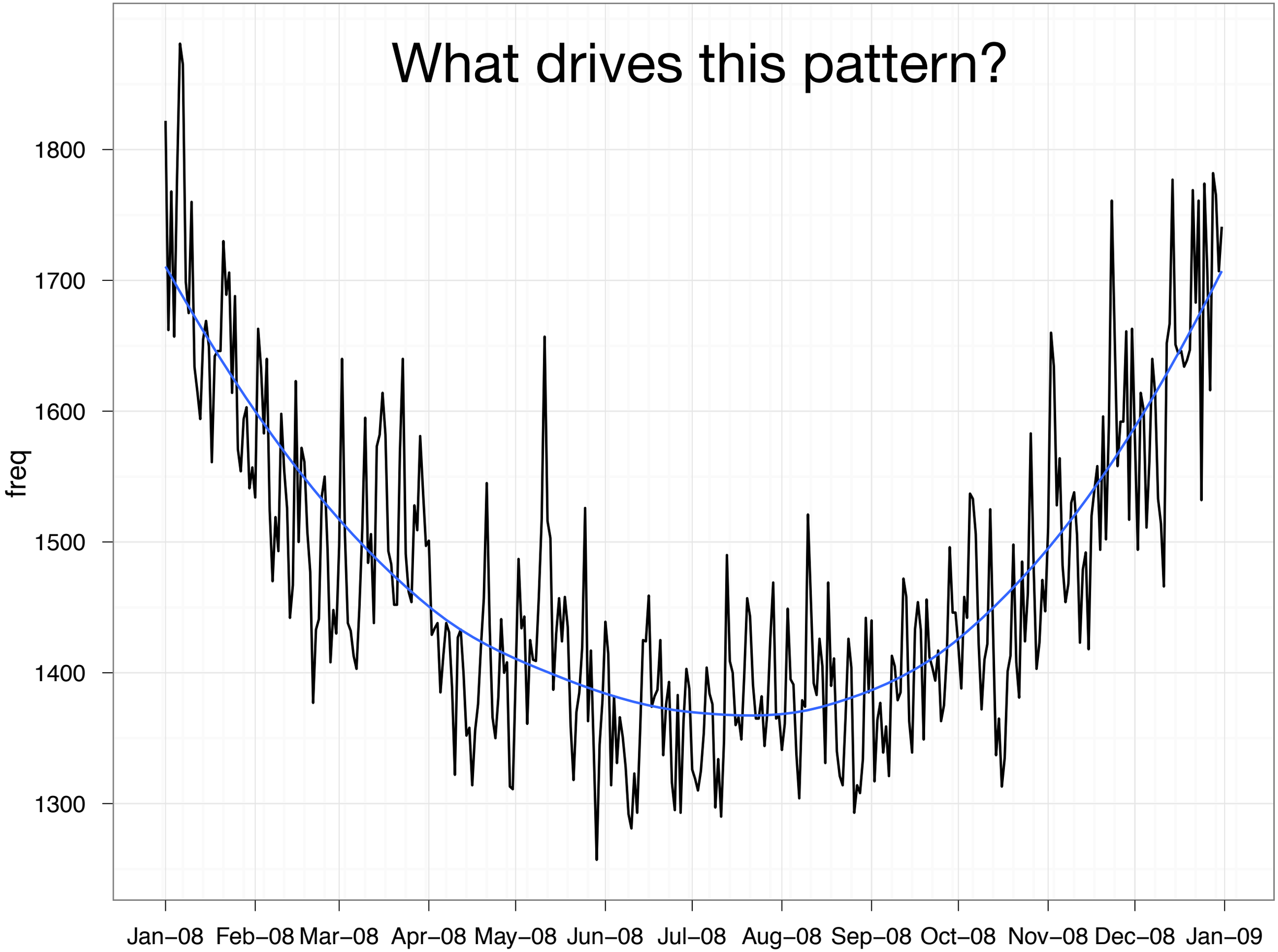

log10(dist)
−2.5
−3.0
−3.5
−4.0
−4.5
−5.0
−5.5
100
1000
10000
n

```
devi$resid <- resid(rlm(log(dist) ~ log(n),
  data = devi))
# ...

unusual <- subset(devi, resid > 1.5)
hod_unusual_big <- match_df(hod2,
  subset(unusual, n > 350))
hod_unusual_sml <- match_df(hod2,
  subset(unusual, n <= 350))
```

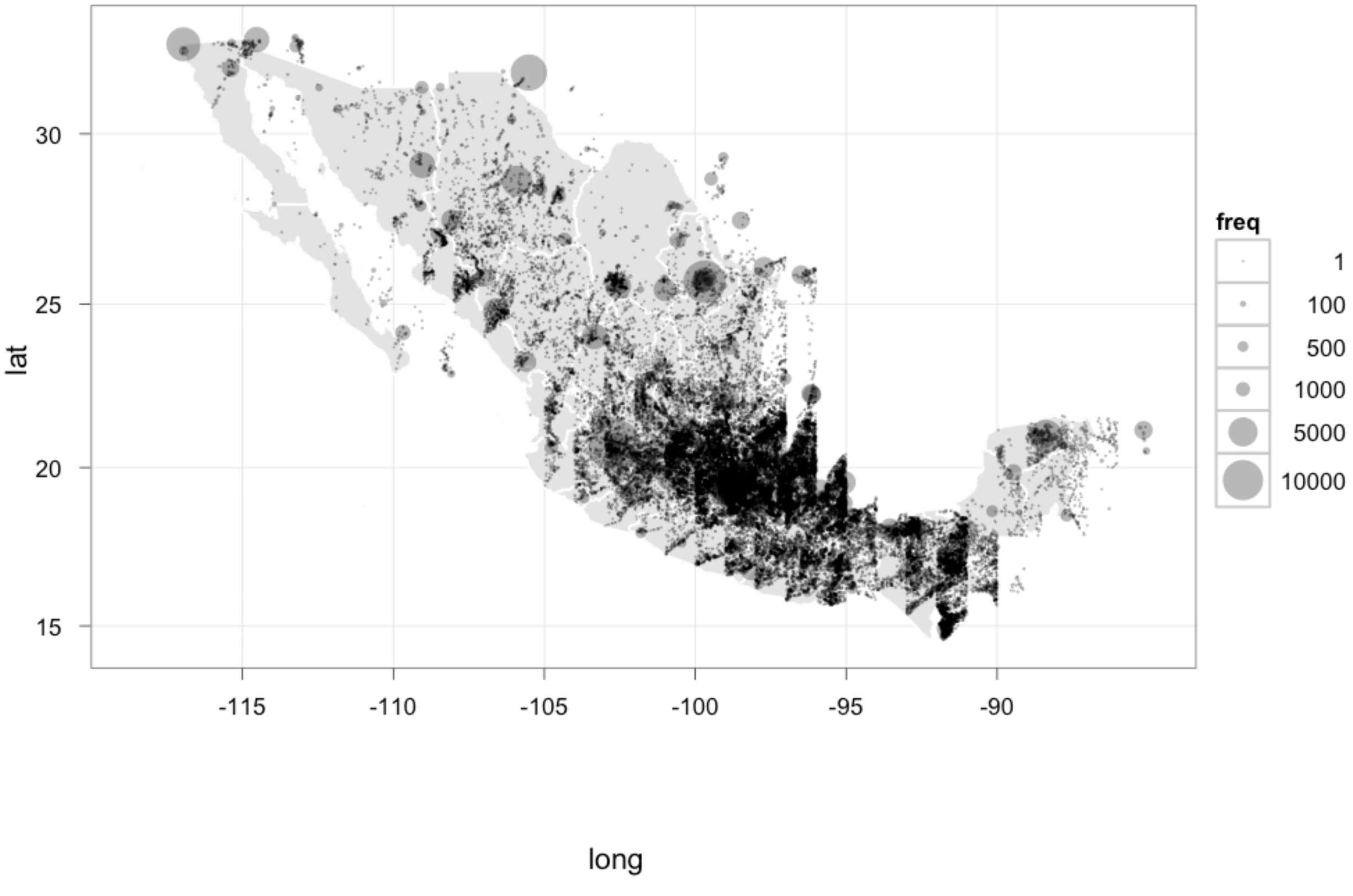

Accident to powered aircraft causing injury to occupant
Bus occupant injured in other and unspecified transport accidents
Other specified drowning and submersion
Sudden infant death syndrome
Victim of lightning
prop
0.25
0.20
0.15
0.10
0.05
5
10
15
20
hod

Challenge

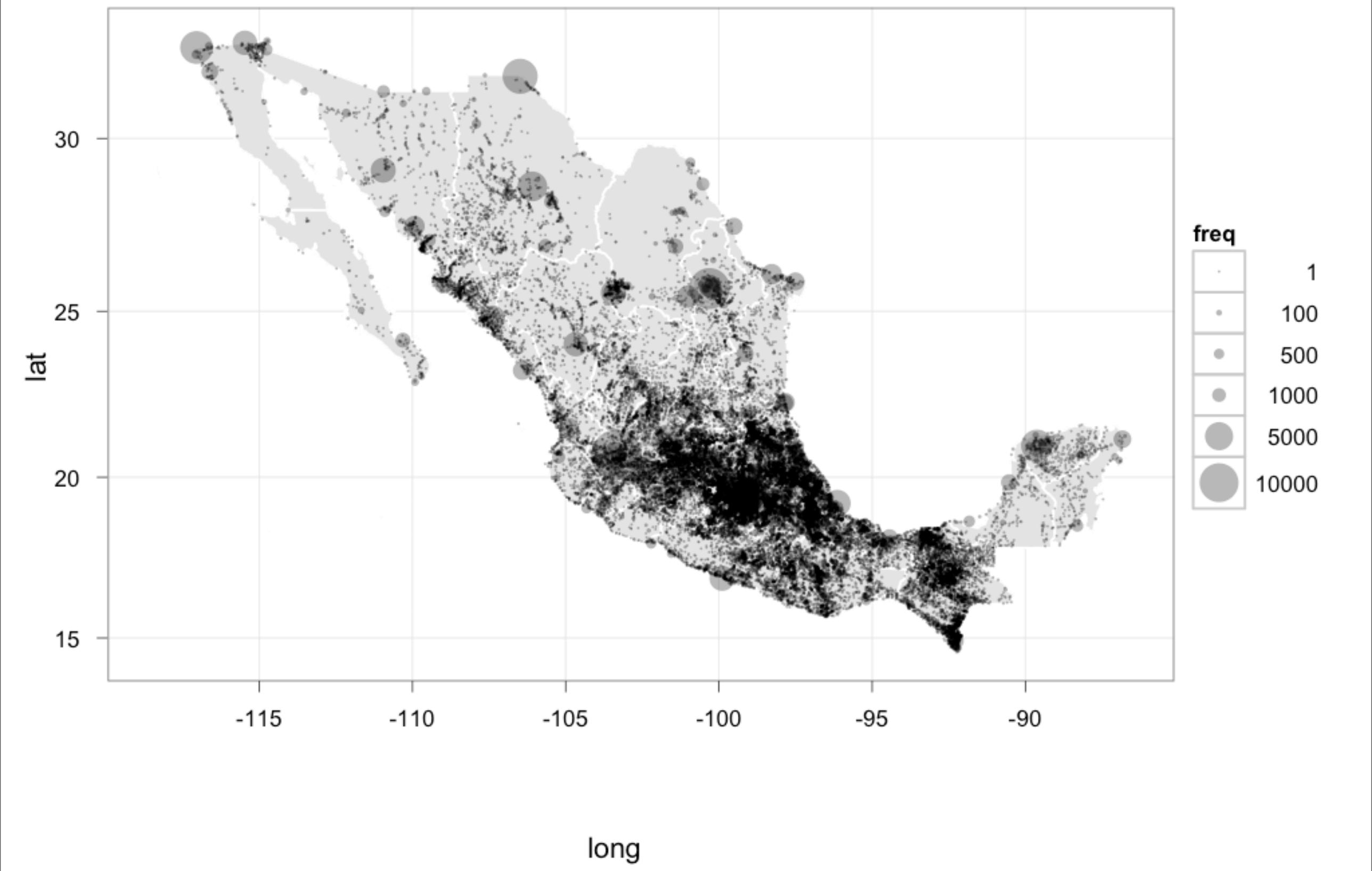
What drives this pattern?



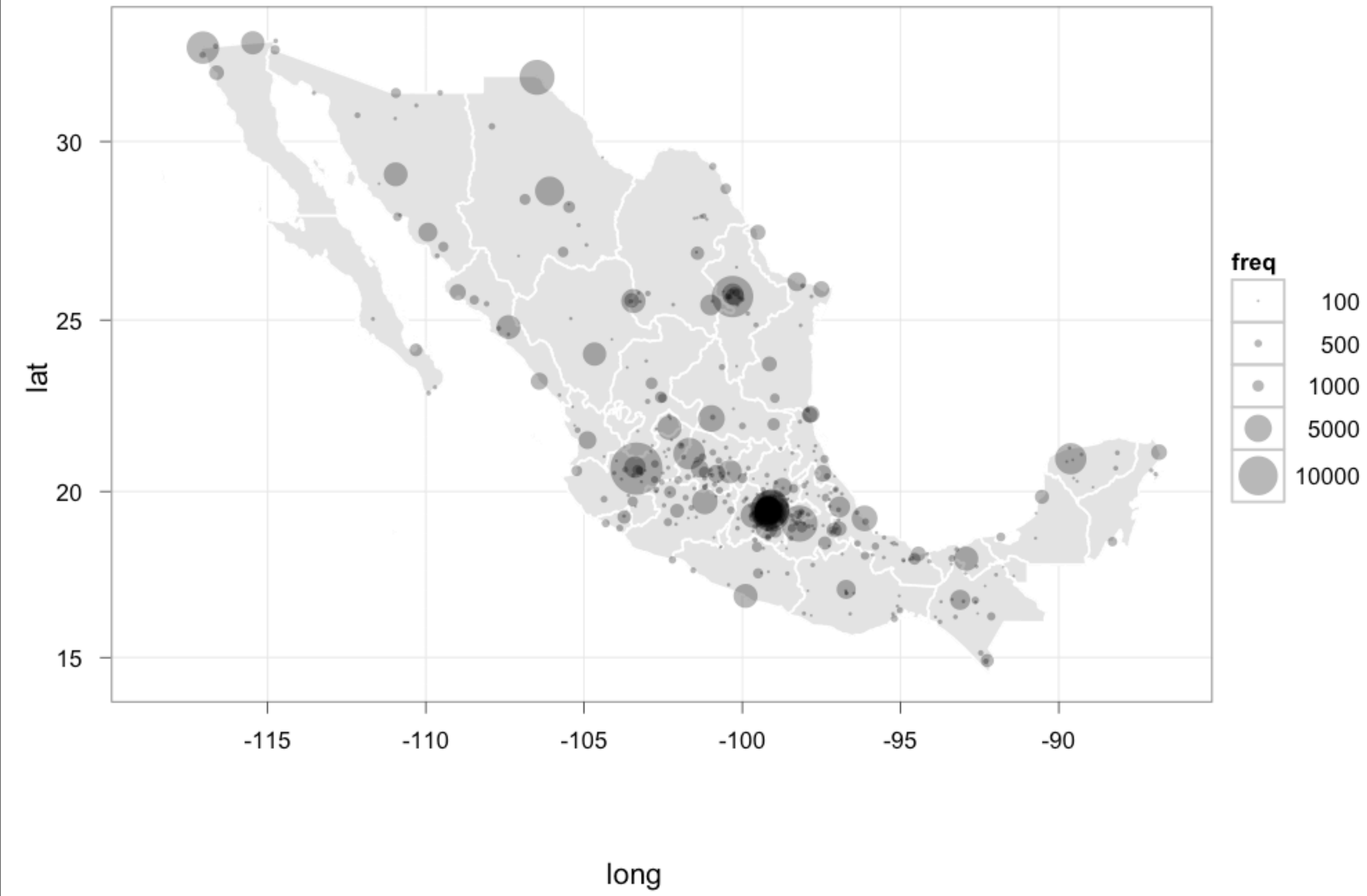
First need location:



New data source



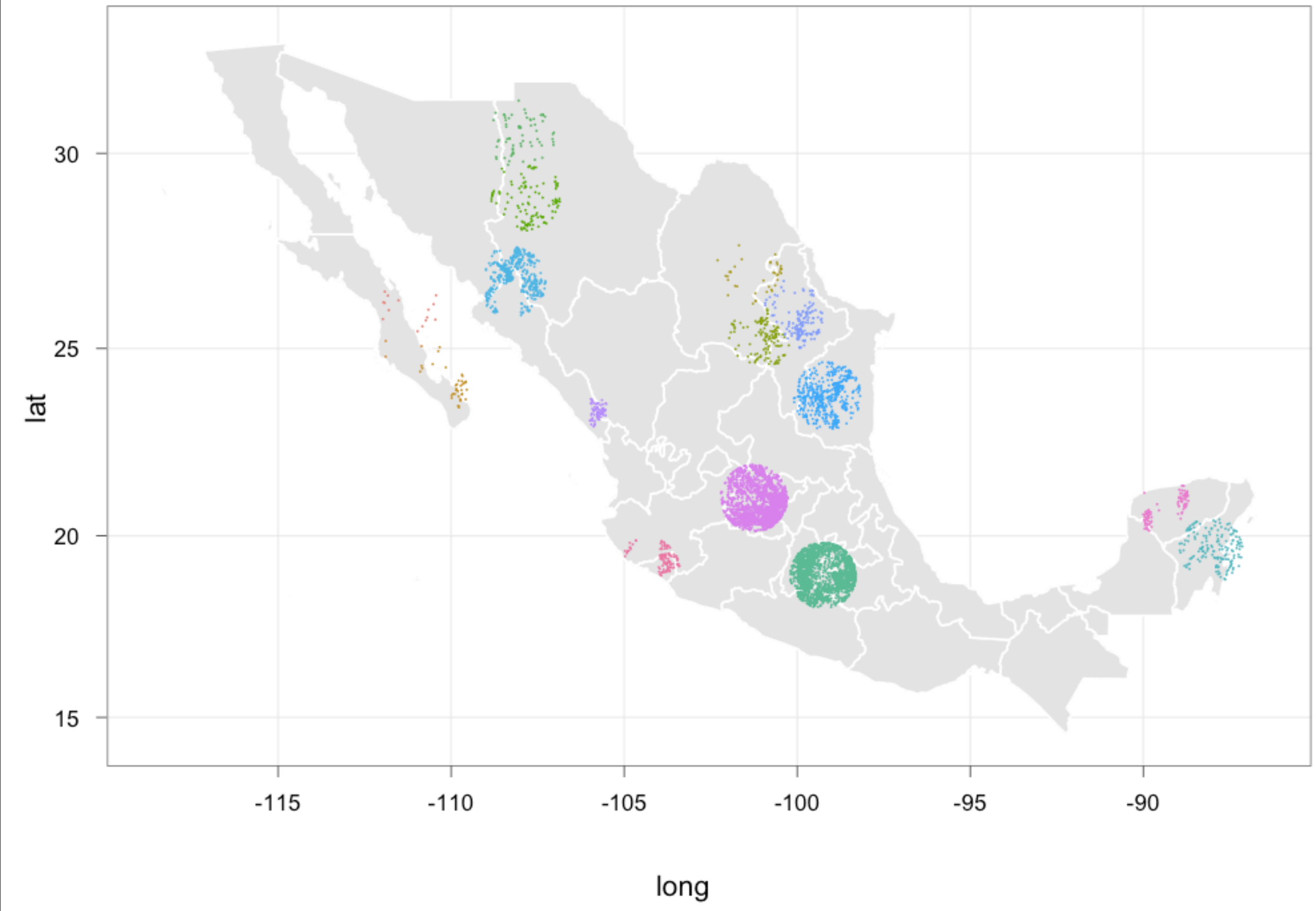
Only locations with >100 deaths



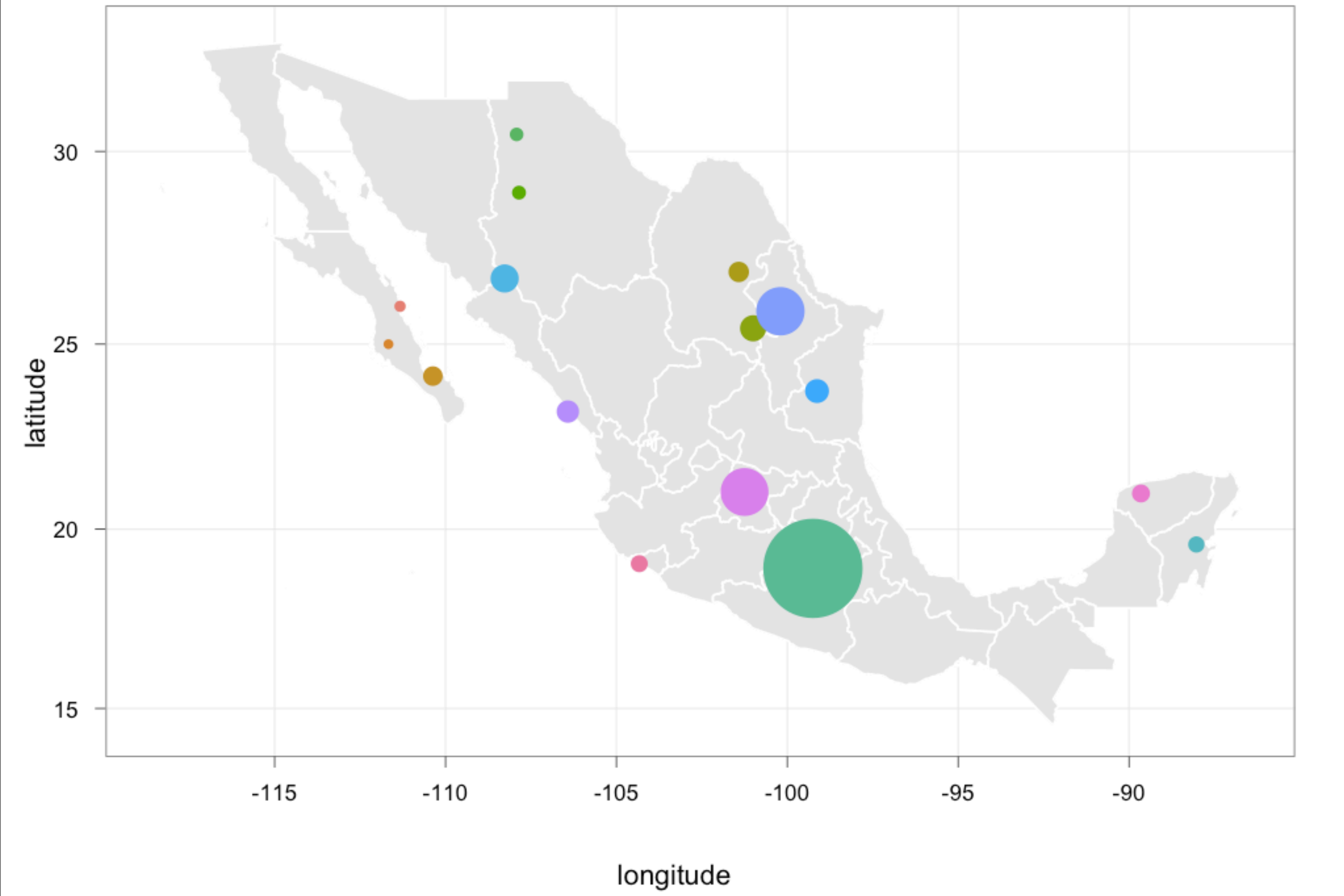
```
locs <- count(deaths, c("long", "lat"))
ggplot(locs, aes(long, lat)) +
  geom_polygon(aes(group = group), data = states,
    colour = "white", fill = "grey90") +
  geom_point(aes(size = freq, order = freq),
    alpha = 1/3, to = c(0.1, 6)) +
scale_area(breaks = c(1, 100, 500, 1000, 5000, 10000),
  to = c(0.5, 10)) +
coord_map()
```

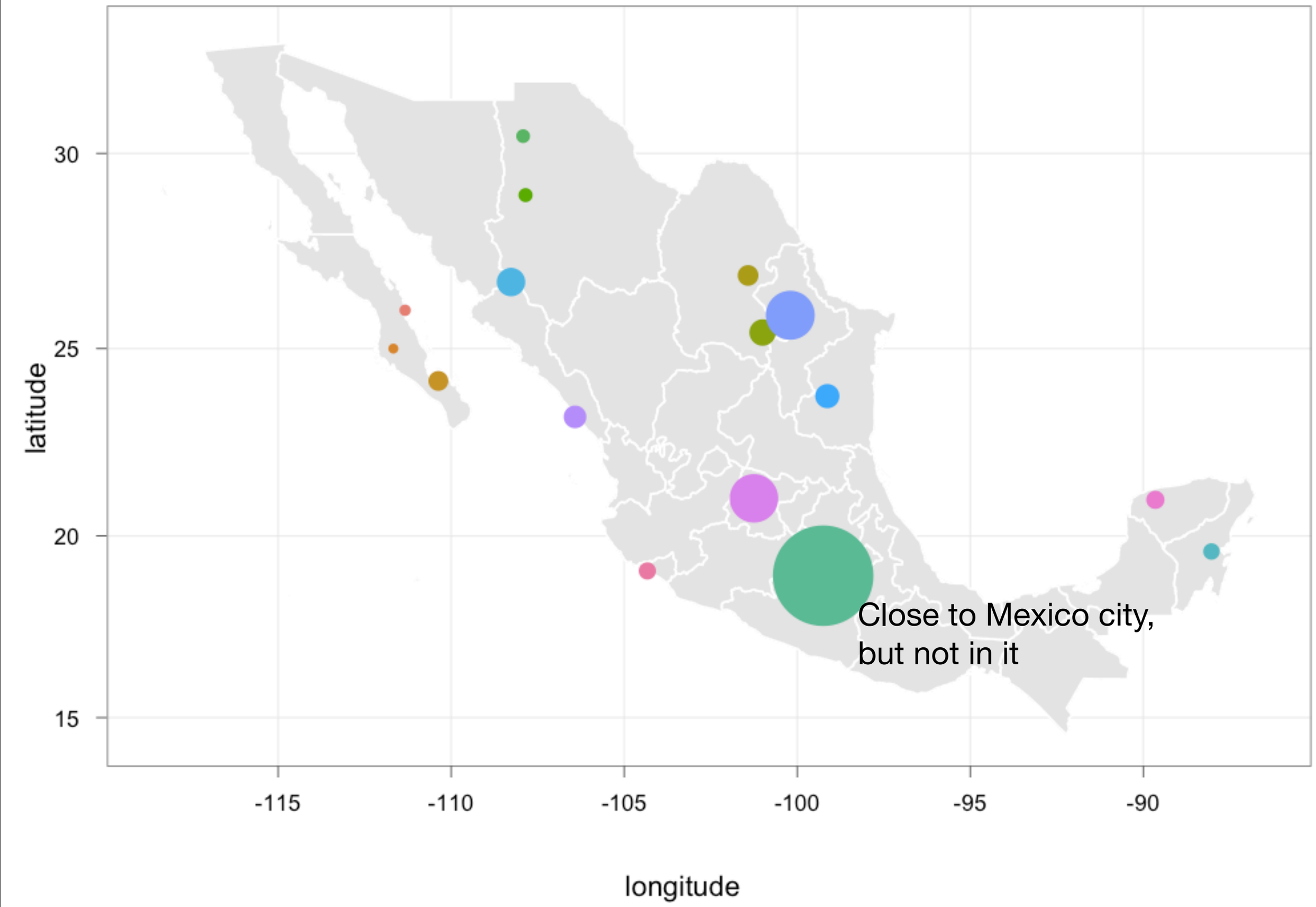
Hours of pain and suffering ...

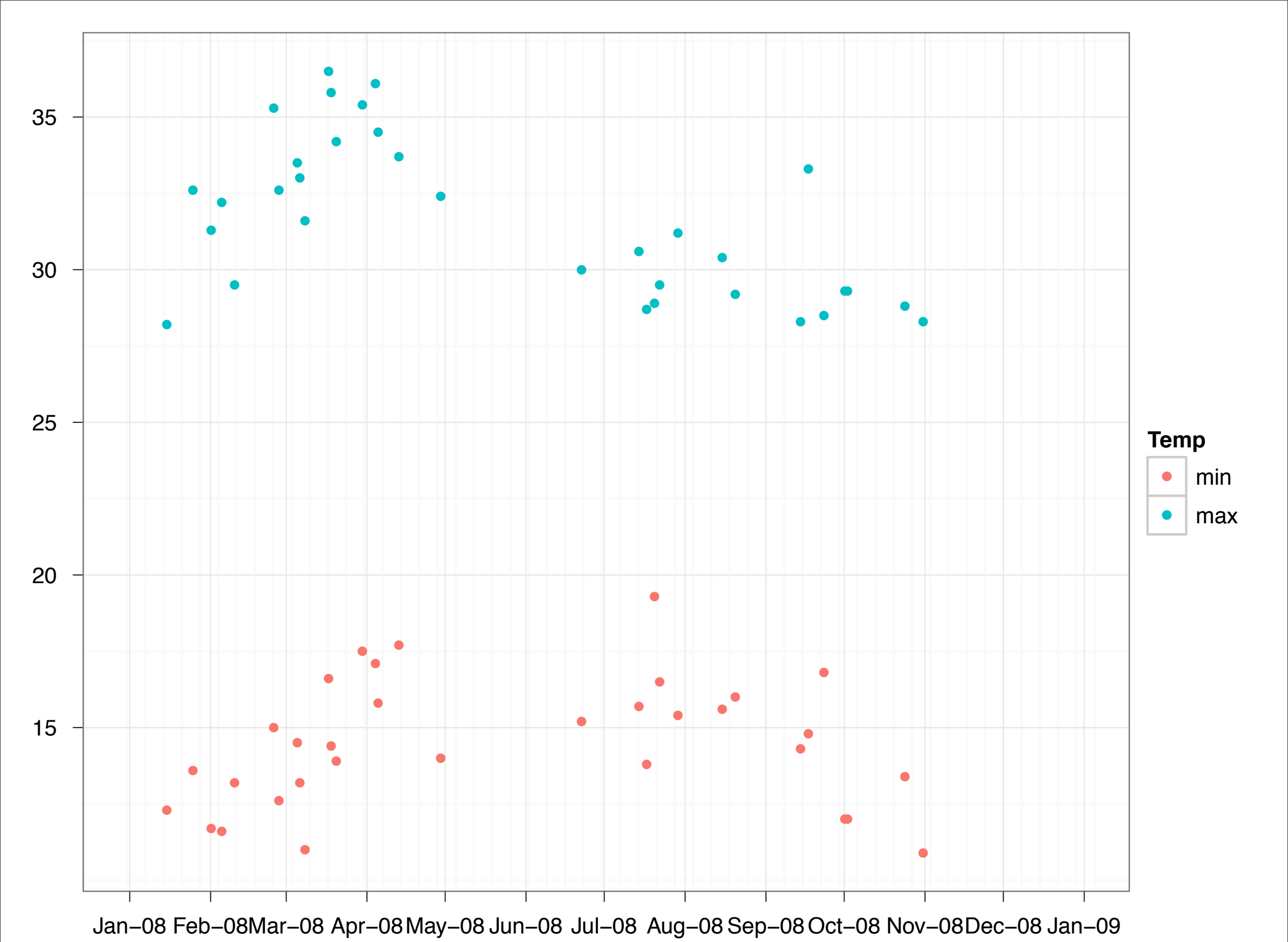
Locations within 50km of a weather station

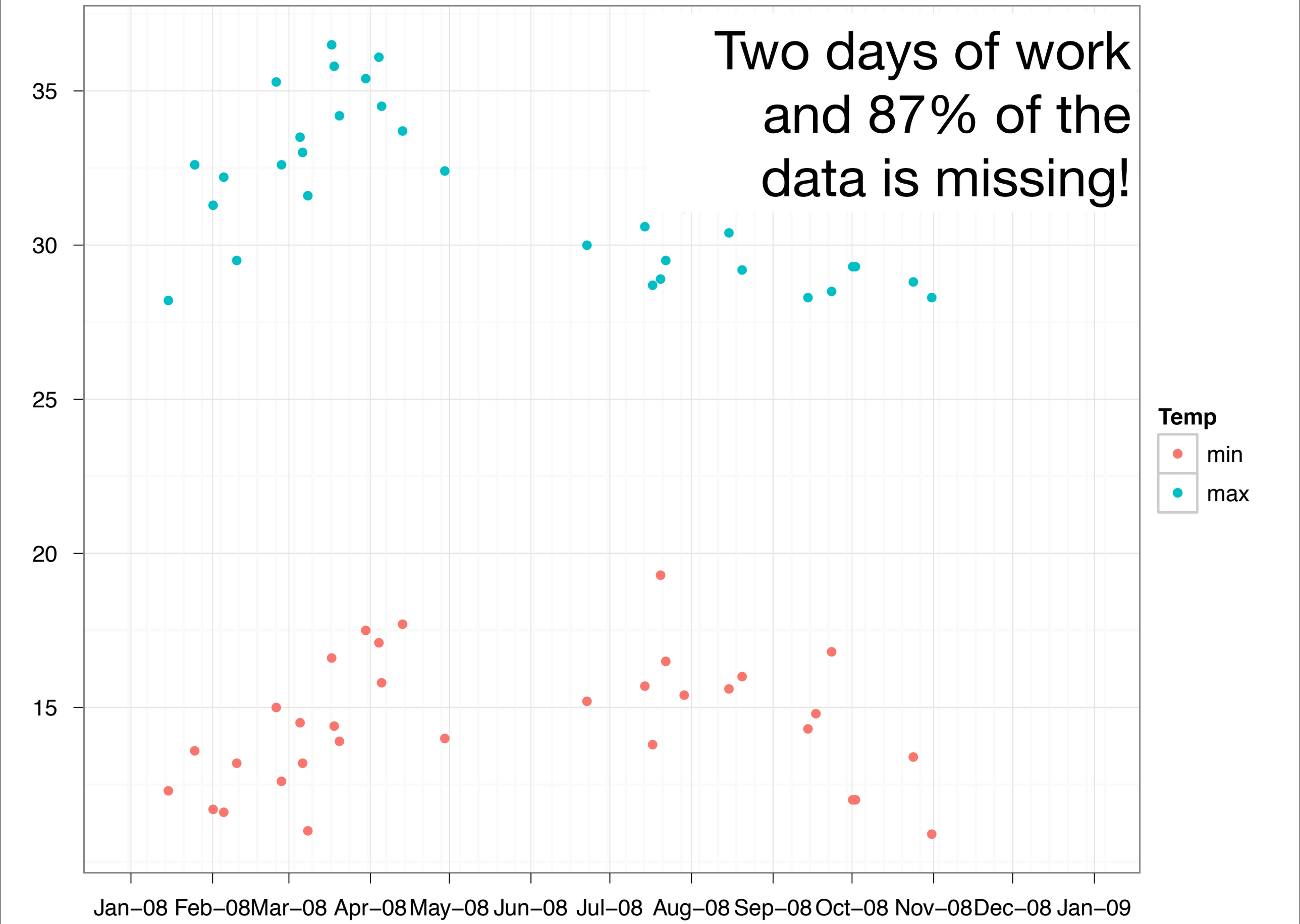


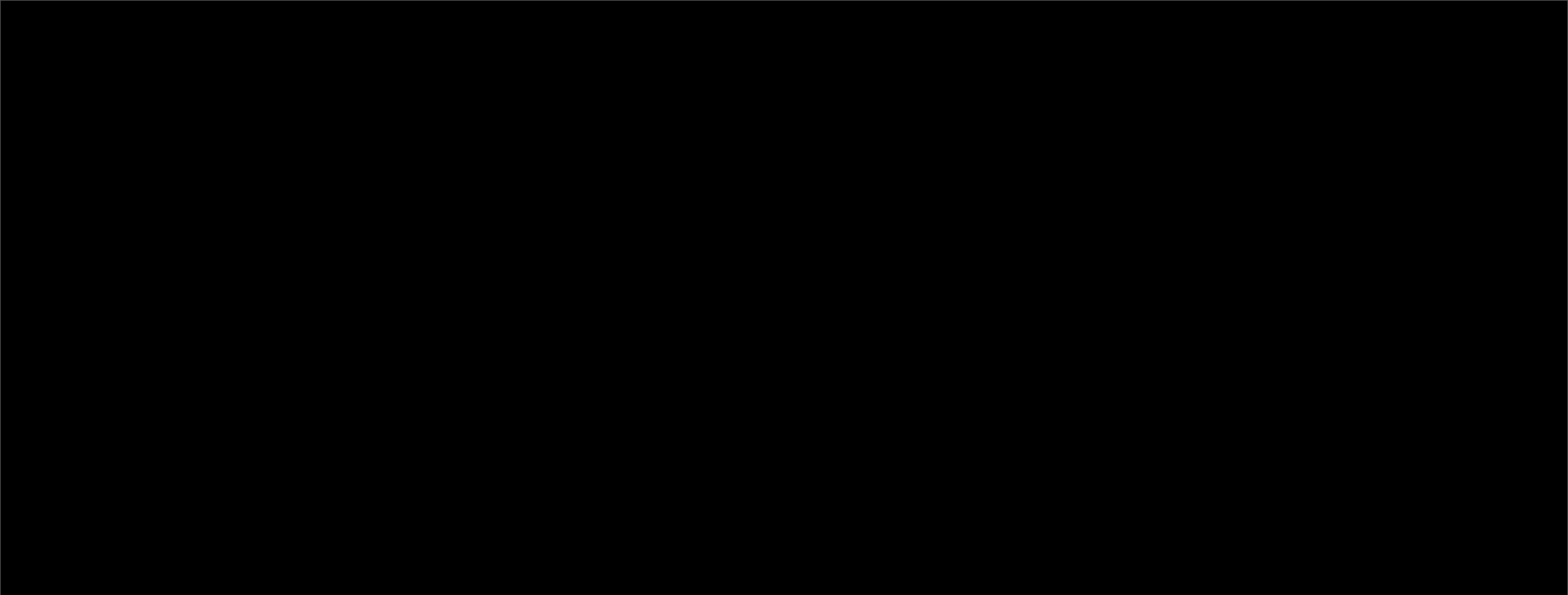
Saturday, July 23, 2011

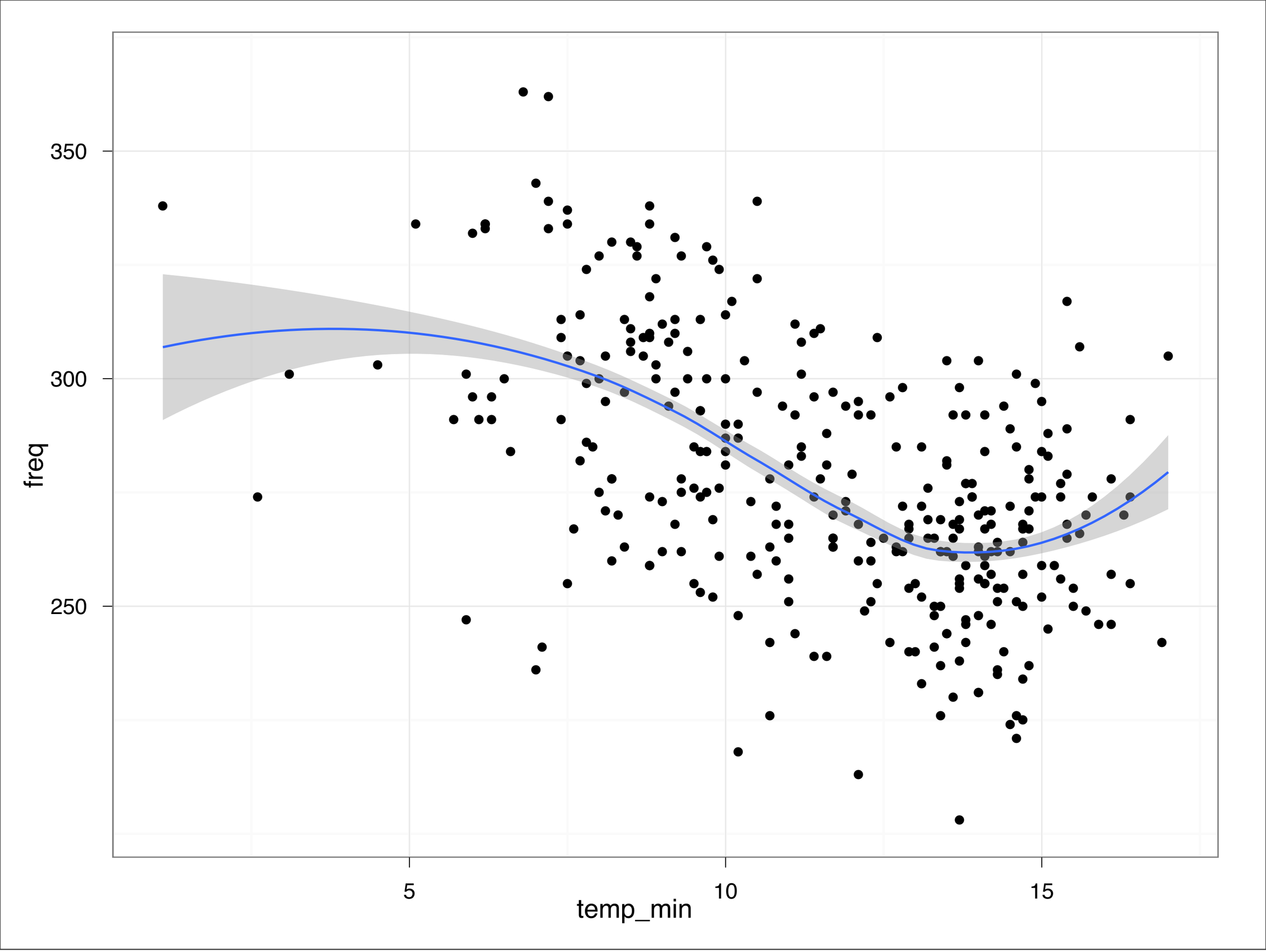

30
25
20
15
-115
-110
-105
-100
-95
-90
latitude
longitude

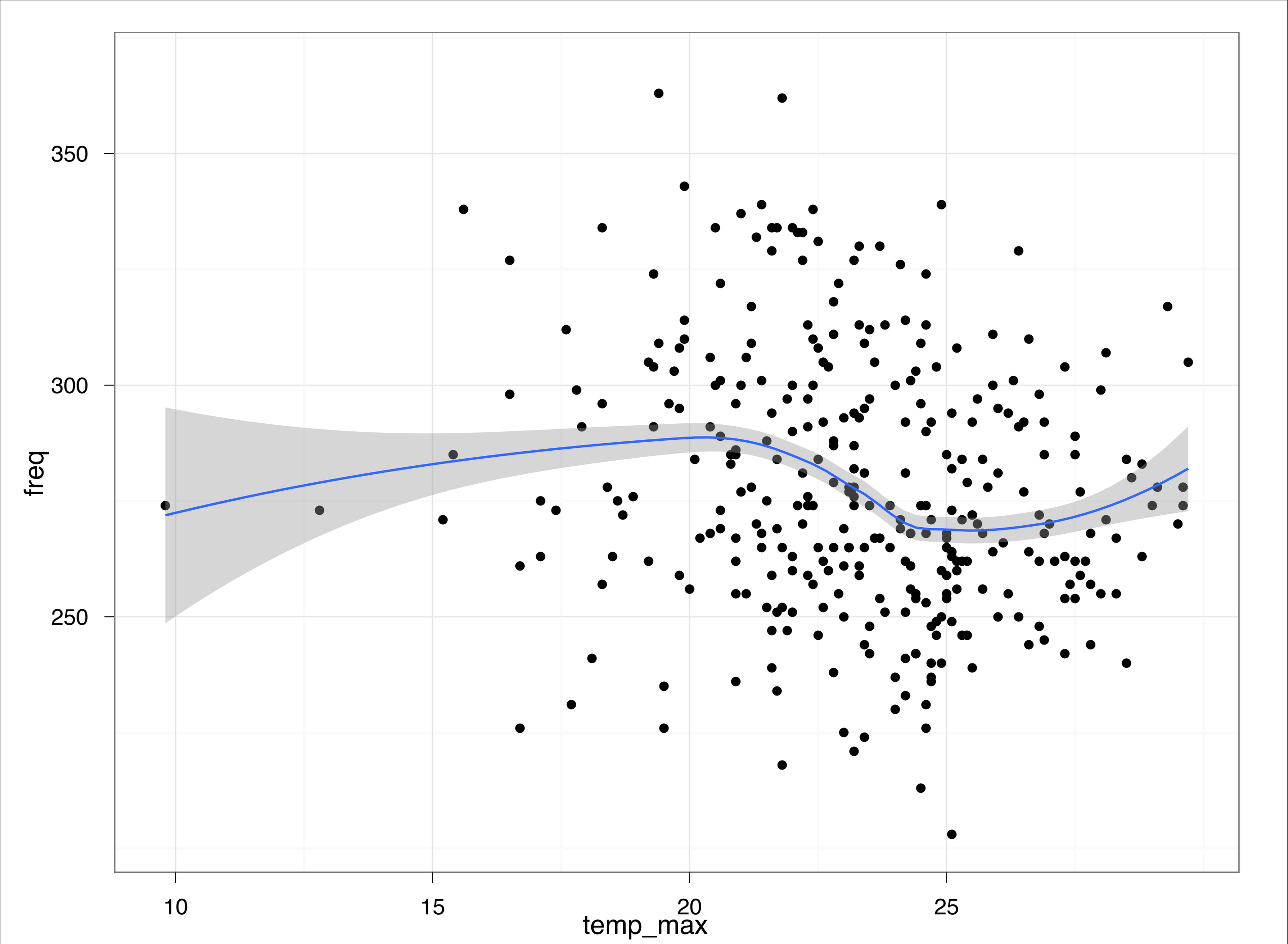

30
25
20
15
latitude
-115
-110
-105
-100
-95
-90
longitude
Close to Mexico city,
but not in it

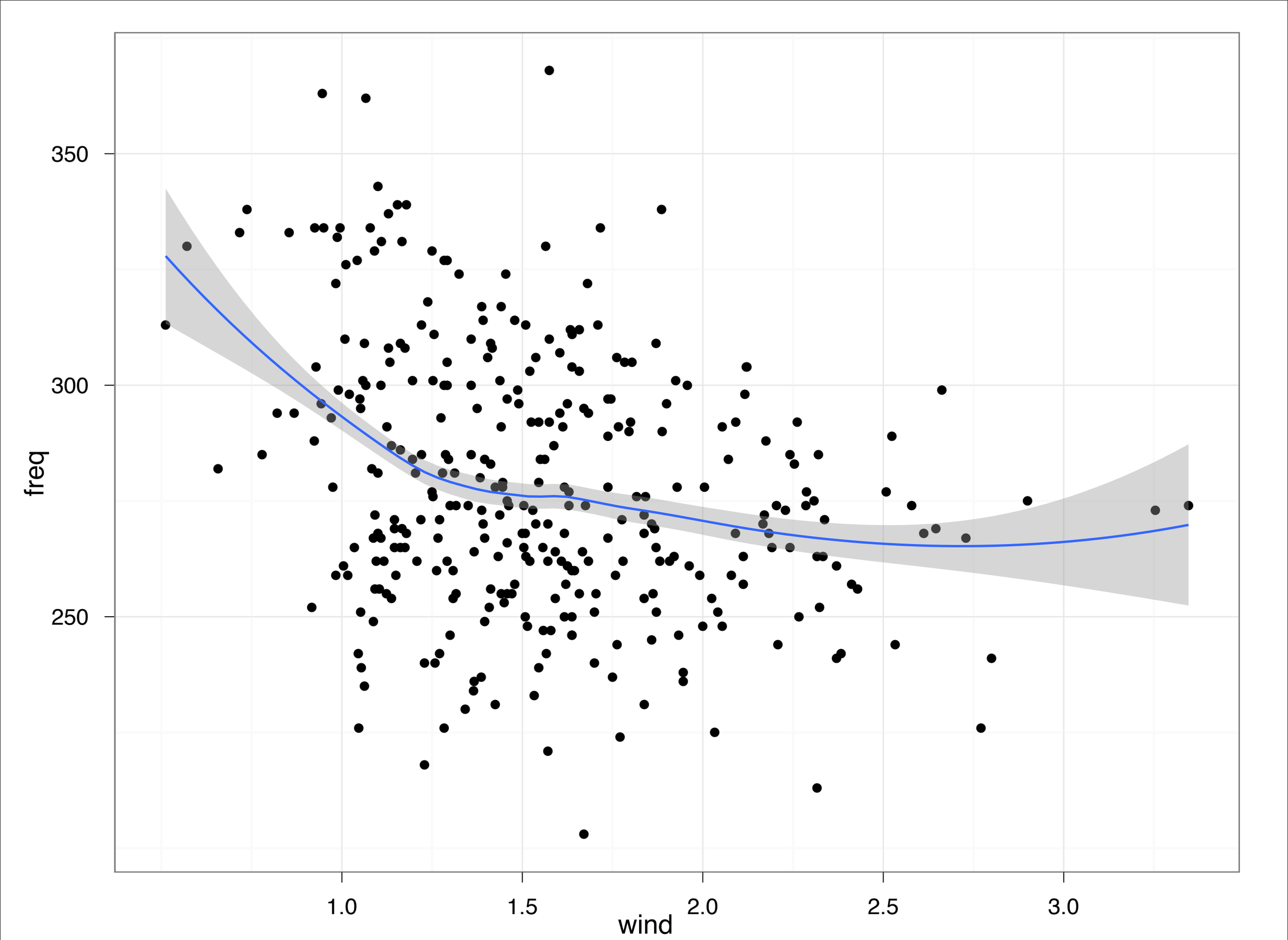

35
30
25
20
15
Jan–08
Feb–08
Mar–08
Apr–08
May–08
Jun–08
Jul–08
Aug–08
Sep–08
Oct–08
Nov–08
Dec–08
Jan–09
Temp
min
max


Two days of work
and 87% of the
data is missing!
35
30
25
20
15
Jan–08
Feb–08
Mar–08
Apr–08
May–08
Jun–08
Jul–08
Aug–08
Sep–08
Oct–08
Nov–08
Dec–08
Jan–09
Temp
min
max




350
300
250
freq
5
10
15
temp_min


350
300
250
freq
10
15
20
25
temp_max


350
300
250
freq
1.0
1.5
2.0
2.5
3.0
wind

Acute myocardial infarction

Alcoholic liver disease

Chronic ischemic heart disease

Fibrosis and cirrhosis of liver

Non−insulin−dependent diabetes mellitus

Other cerebrovascular diseases

Other chronic obstructive pulmonary disease

Pneumonia, organism unspecified

Unspecified diabetes mellitus

prop

0.008

0.006

0.004

0.002

5

10

15

temp_min

```
ggplot(daily, aes(temp_min, prop)) +
  geom_point(alpha = 1/3) +
  geom_smooth(se = F, size = 1) +
  facet_wrap(~ disease2)
```

Conclusions

A programming language gives you: reproducibility, automation, communication, but has a learning curve.

R gives you: freedom, a community, connectivity, building blocks, but the community can be prickly and it is slow (relative to other languages).

Thoughtful DSLs should make it easier to solve common data analysis problems.

Office hours

MTV-1098-1-Gwydir

3-4pm

hadley@rice.edu