Exploring the cancer data

Hadley Wickham & Heike Hofmann

Cancer data

- Cancer incidence and mortality (+ population), broken down by:
  - 52 states (!!)
  - 2 sexes (male, female)
  - 3 races (black, hispanic, white)
  - 26 sites
  - 6 years (1999–2004)

Goals

- Practice data handling skills
- Practice asking interesting questions
- Investigate spatial and temporal patterns:
  - Time series plots
  - Chloropleth (map) plots
- Learn how to aggregate data with reshape

Your turn

- Load cancer.csv into R
- Look at the data. Is there anything strange?
- Calculate incidence and mortality rates (per 100,000 people) and add them to the data.frame
- Create a subset for Iowa
- Start exploring the data with plots

Questions

- Why have I only included counts (and not rates) in this dataset? (Hint: what happens if we combine states or races or sexes?)
- What's the difference between incidence and mortality? Is there anything unusual about the values?

Time series

- Easy part: use `geom="line"`
- Hard part: what is a line?
  - Need someway to identify the combination of measurements that constitutes a line
  - `group=state:site:sex:race`

```
qplot(year, mrate, data=iowa)
qplot(year, mrate, data=iowa, geom="line")
```

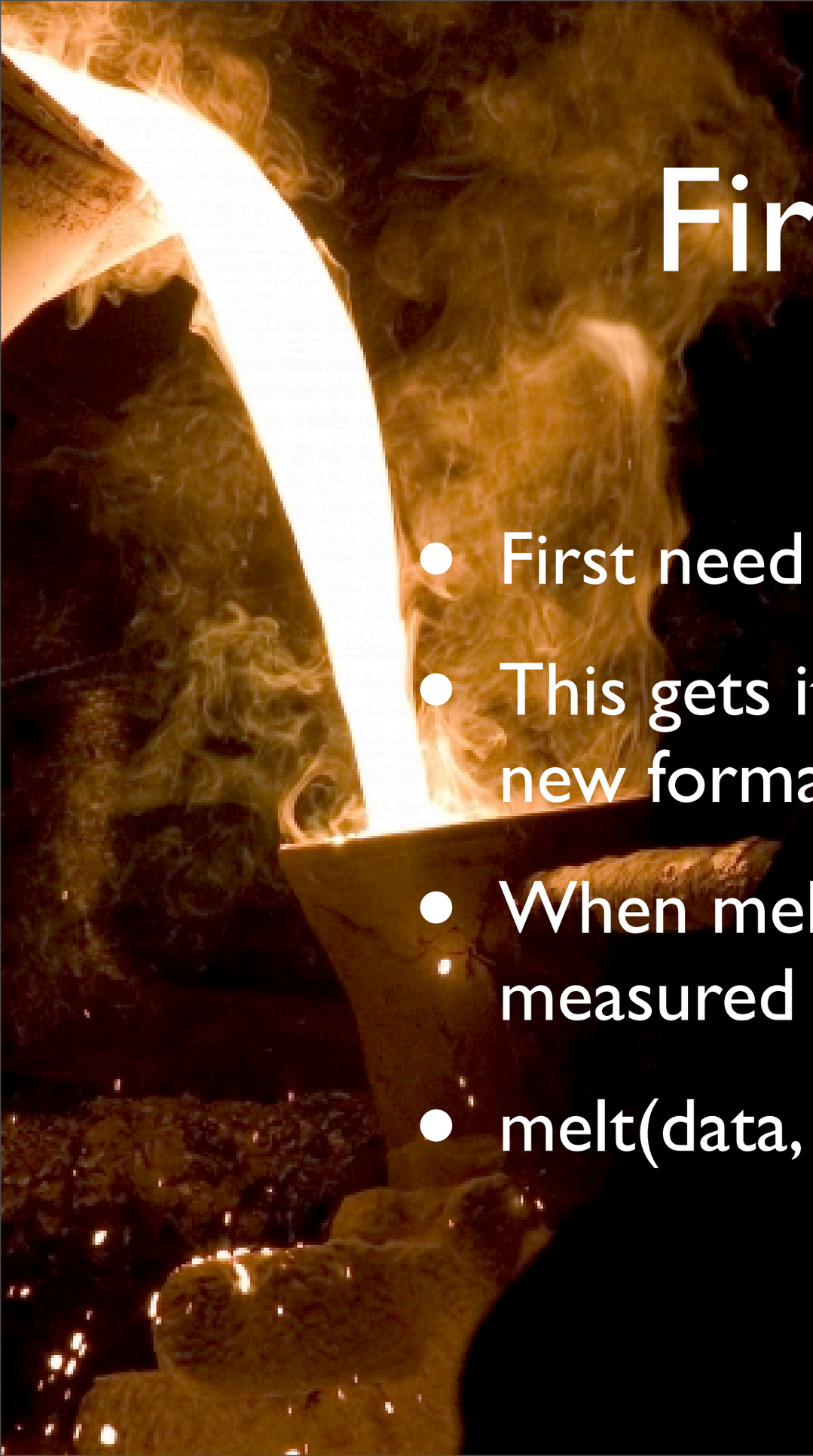
```
qplot(year, mrate, data=iowa, group =
site:race:sex, geom="line")
```

Your turn

- Experiment with drawing time series with both the Iowa subset and all the data
- Use facetting and colours to hunt for interesting features of the data

Aggregating data

- If we're not interested in a certain variable, we need to be able to aggregate of it - e.g. ignoring state and site to look at overall trends over time
- Many different ways to do this in R, but we're going to focus on one:
`library(reshape)`



First, `melt`

- First need to "melt" the data
- This gets it in a form useful for "casting" into new formats
- When melting, you need to specify the measured variables and the id variables
- melt(data, measure.var=c(1,2,3), id.var=5)

Then, cast

- Just like pivot tables and facetting plots
- Row variables, column variables, and a summary function (sum, mean, max, etc)
- `cast(molten, row ~ col, summary)`
- `cast(molten, row1 + row2 ~ col, summary)`
- `cast(molten, row ~ . , summary)`
- `cast(molten, . ~ col, summary)`

Example

```
library(reshape)
cancerm <- melt(cancer, id = 1:5)
cast(cancerm, race ~ variable, sum)
cast(cancerm, sex ~ variable, sum)
cast(cancerm, state ~ variable, sum)
```

Our first function

```
rates <- function(df) {
  transform(df,
    irate = incidence / population * 100000,
    mrate = mortality / population * 100000
  )
}

site_rates <- rates(cast(cancerm, site ~
variable, sum))
```

Cancer data

```
site_rates <- rates(
  cast(cancerm, site ~ variable, sum)
)

qplot(irate, site,
data=site_rates, xlim=c(0, NA))

qplot(irate, reorder(site, irate),
data=site_rates, xlim=c(0, NA))
```

Your turn

- Investigate the distribution of rates by state, race, and year
- Investigate the distribution of rates by site AND sex (hint: site + sex ~ variable)
- Investigate the change of rates over time, broken down by state or site. Visualise with a time series plot

Chloropleth maps

- How can we show the spatial distribution of cancer rates?
- What exactly is a map?

```
states <- read.csv("states.csv")

qplot(x, y, data=states, geom="path",
group=state)
qplot(x, y, data=states, geom="polygon",
group=state)

map_rates <- merge(states, state_rates,
by="state")

qplot(x, y, data=map_rates, group=state,
fill=irate, geom="polygon")
qplot(x, y, data=map_rates, group=state,
fill=mrate / irate, geom="polygon")
```

Your turn

- Load states.csv into R
- Summarise the cancer data at the state level. Combine with the states data and plot.
- Can you find a cancer with a clear geographic trend? (Hint: Use `cast` to produce a summary by state and site, and `subset` to pull out a single site)
- Extra: look at http://had.co.nz/ggplot2/scale_gradient.html and experiment with different colour schemes