Datasets

1. Data structure
2. Data exploration
3. Practice
4. Other sources

Dataset structure

Cross-sectional data capture the characteristics of a sample of comparable units at a single point in time

- **Units** can be individual respondents, countries, firms...
- **Observations** vary by their characteristics, *not* by unit type

Time series capture repeated observations over time of either sampled or nonsampled units

- **Cross-sectional time series** (CSTS) capture fixed, nonsampled units at different time intervals
- **Longitudinal data** capture sampled units, called a cohort or panel, at different time intervals

Sample characteristics

Concepts

- Target and survey population
- Sampling frame and design
- Randomization component

We will only cover sampling weights.

Issues

- Undercoverage
- Sampling bias
- Unit nonresponse

Each issue affects representativeness.

Example: Industry Canada File Sharing Survey, 2006

Individual-level 'micro' data on illegal downloading practices among a random sample of the Canadian population aged 15+:

	id	prov	qregn	date	age	sex	download	q1	q2
1	1065	ON	Ontario	20060502	Less than 25 years old	Male	NON-DOWNLOADER	Yes	20
2	1129	AB	Alberta	20060423	Less than 25 years old	Female	NON-DOWNLOADER	No	.
3	1152	QC	Quebec	20060519	Less than 25 years old	Female	DOWNLOADER	No	.
4	1166	ON	Ontario	20060429	Less than 25 years old	Male	NON-DOWNLOADER	Yes	20
5	1191	ON	Ontario	20060423	25 years old or more	Female	NON-DOWNLOADER	Yes	20
6	1214	ON	Ontario	20060423	25 years old or more	Female	NON-DOWNLOADER	Don't Know/Refused	.
7	1215	QC	Quebec	20060422	Less than 25 years old	Female	NON-DOWNLOADER	Yes	10
8	1245	ON	Ontario	20060423	25 years old or more	Female	NON-DOWNLOADER	No	.
9	1266	BC	British Columbia	20060419	25 years old or more	Female	NON-DOWNLOADER	No	.
10	1315	QC	Quebec	20060430	25 years old or more	Male	NON-DOWNLOADER	No	.
11	1317	ON	Ontario	20060423	25 years old or more	Female	NON-DOWNLOADER	Don't Know/Refused	25
12	1643	ON	Ontario	20060423	25 years old or more	Female	DOWNLOADER	Don't Know/Refused	20

- **Layout:** one observation per row, one variable per column
- **Formats:** numeric, string, encoded (values/labels)
- **Missing data:** encoded as . (dot), counted as $+\infty$.

Exploring the documentation

Knowing the data is not an option. You naturally do not have to 'read' through the data itself, but you need to read everything else.

The **codebook** is essential to measurement:

- Data collection and measurement are publicly documented to allow for sceptical scrutiny of sources and method.
- The unit of continuous data, scale of ordinal data or categories of nominal data are given with their construction notes.

Requirements

Format — Stata Guide, Sections 5–8

- The dataset format is **.dta** — otherwise convert
- The data are **cross-sectional** — otherwise subset
- The columns hold **only variables** — otherwise reshape

Referencing

- Recommended **bibliographic citation** for the data source.
- Precise **unit of analysis** and number of observations *N*
- Summary of **sampling strategy**, citing documentation

Codebook example: measurement details

Quality of Government, 2011, p. 229:

wdi_fr **Fertility Rate (Births per Woman)**

(Time-series: 1960-2008, n: 8560, N: 189, $\overline{N}$: 175, $\overline{T}$: 45)
(Cross-section: 2000-2005 (varies by country), N: 189)

Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates. Sources: The United Nations Population Division's World Population Prospects, national statistical offices, Eurostat, Secretariat of the Pacific Community, US Census Survey, and household surveys conducted by national agencies, Macro International and the US Centers for Disease Control and Prevention.

In the cross-sectional version of the dataset, the variable `wdi_fr` was measured from 2000 to 2005 in $N = 189$ countries.

Codebook example: coding details

European Social Survey Round 4, 2008, p. 191:

# **rlgblge: Ever belonging to particular religion or denomination**	
Question	**All rounds:** Have you ever considered yourself as belonging to any particular religion or denomination?
Question number	**ESS1, ESS2:** C 11 **ESS3, ESS4:** C 19
Routing	**ESS1, ESS2:** If code 2, (7) or 8 at C9 **ESS3, ESS4:** If codes 2, 7 or 8 at C17
Comments	**ESS2:** Austria: Distributions differ from ESS round 1 due to change of wording. Finland: Data from Finland have been omitted from the international file. For further details please see item 46 in the Documentation Report.

Value	Label
1	Yes
2	No
6	Not applicable
7	Refusal
8	Don't know
9	No answer

The variable named `rlgblge`, coding for religious membership, takes six unique values, four of which code for missing values.

Data exploration

```
* Load a dataset; -clear- wipes previous data in memory.
use data/ess2008, clear
* Look for keywords in variable names and labels.
lookfor army homo
* The -lookfor_all- package searches across datasets.
lookfor_all health, dir(data)
* Describe all variables (shorthand: -d-).
describe
* Describe one variable (name and variable label).
d wrkstat
* Describe a variable in more detail.
codebook wrkstat
```

Data preparation

```
* Set up survey weights.
svyset psu [pweight=perweight], strata(strata)
* Study the missing values for a set of variables.
misstable pat age sex race religion income
* Keep only some observations.
keep if cntry == "FR"
* Drop observations for which religion is missing.
drop if mi(religion)
* Count the number of observations in the data.
count
```

Reminder

Use `help` when you need more details on any command.

Selections

Range: `in`

Observations have a row number `_n`, from `1` to sample size `_N`.

```
* List first ten observations.
list in 1/10
* List last ten observations.
list in -10/l
```

Condition: `if`

Many commands can be applied `if` a given condition is true.

```
* Delete observations if age is below 18.
drop if age < 18
* Keep only observations for survey year 2010.
keep if year == 2010
```

Conditions

Translate logical statements into full phrases.

‘and’, ‘or’

```
* Clone variable, except for values 0 and above 8.
clonevar party = partyid if partyid != 0 & partyid > 8
* Keep only a selection of countries.
keep if inlist(country, "FR", "DE", "IT", "SP")
```

‘missing’, ‘nonmissing’

```
* Delete if missing age, sex or marital status.
drop if mi(age, sex, married)
* Create a dummy for sex (female = 1, else = 0).
* Applied only to nonmissing observations.
gen female = (sex == 2) if !mi(sex)
```

Practice: NHIS dataset

$$\text{Body Mass Index} = \frac{\text{mass (kg)}}{(\text{height(m)})^2} = \frac{\text{mass (lb)} \times 703}{(\text{height(in)})^2}$$

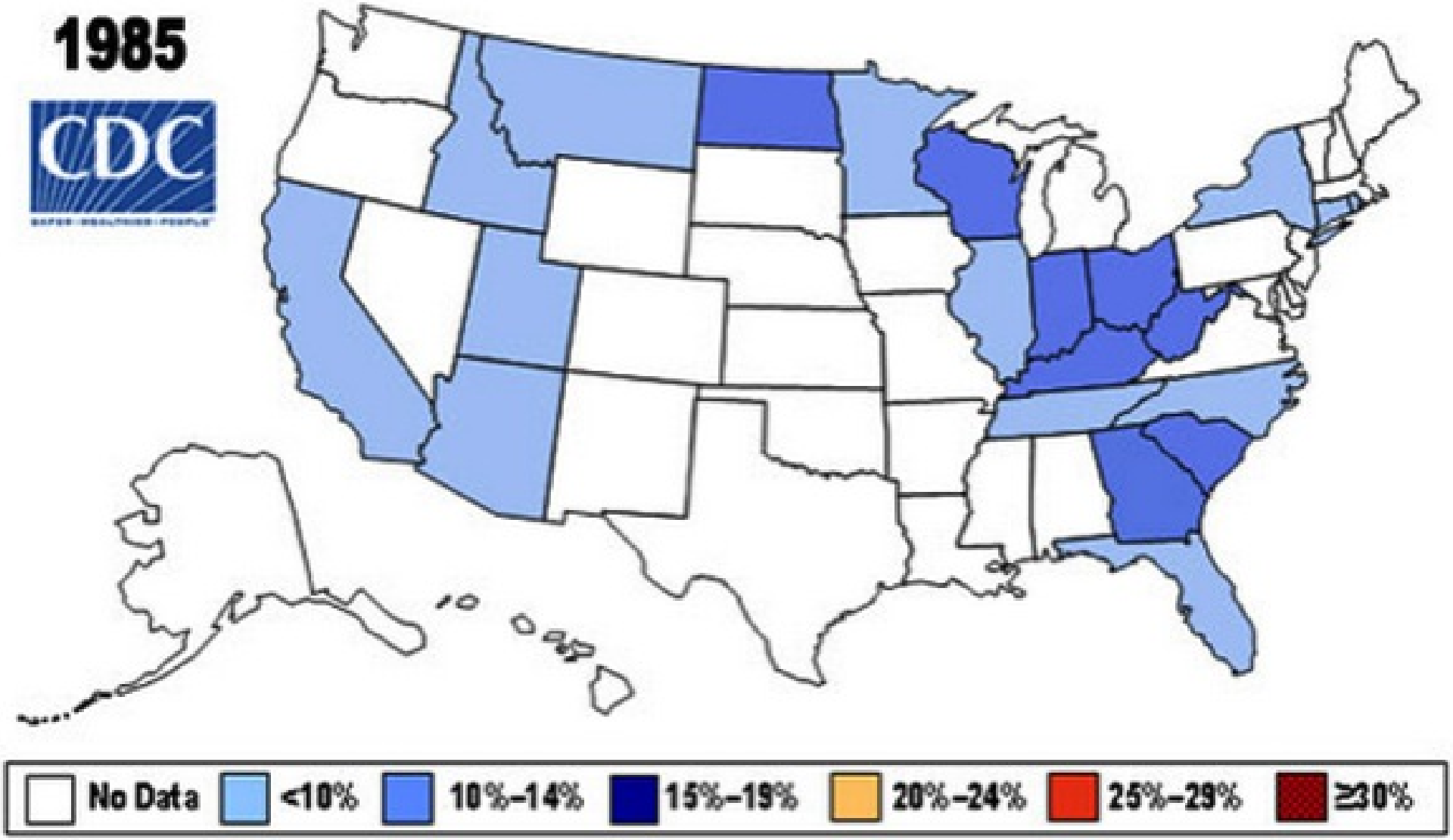
- For **normal weight** adults, $18.5 < \text{BMI} < 25$.
- For **overweight** adults, $25 \leq \text{BMI} < 30$.
- For **obese** adults, $\text{BMI} \geq 30$.

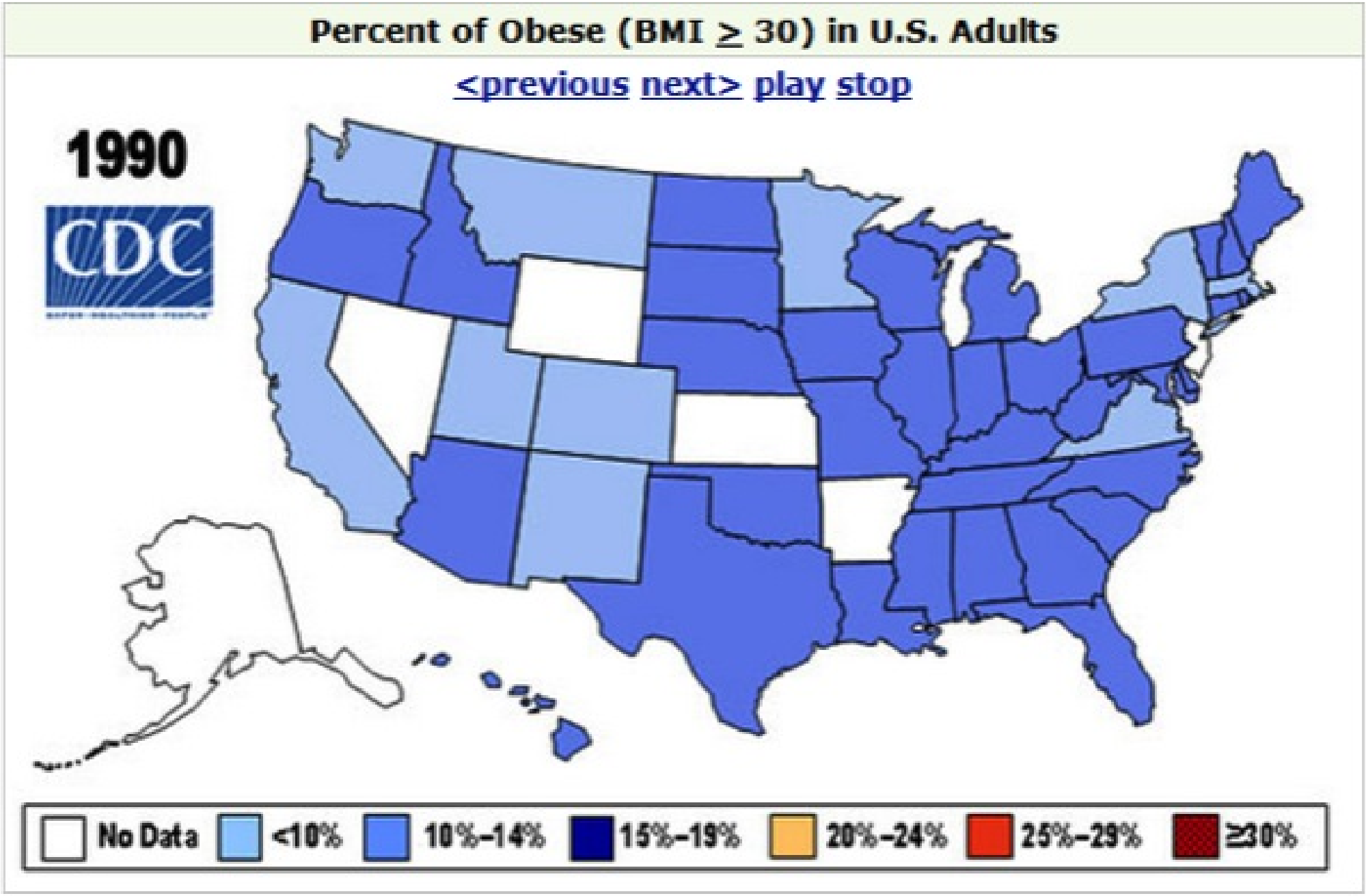
Data:

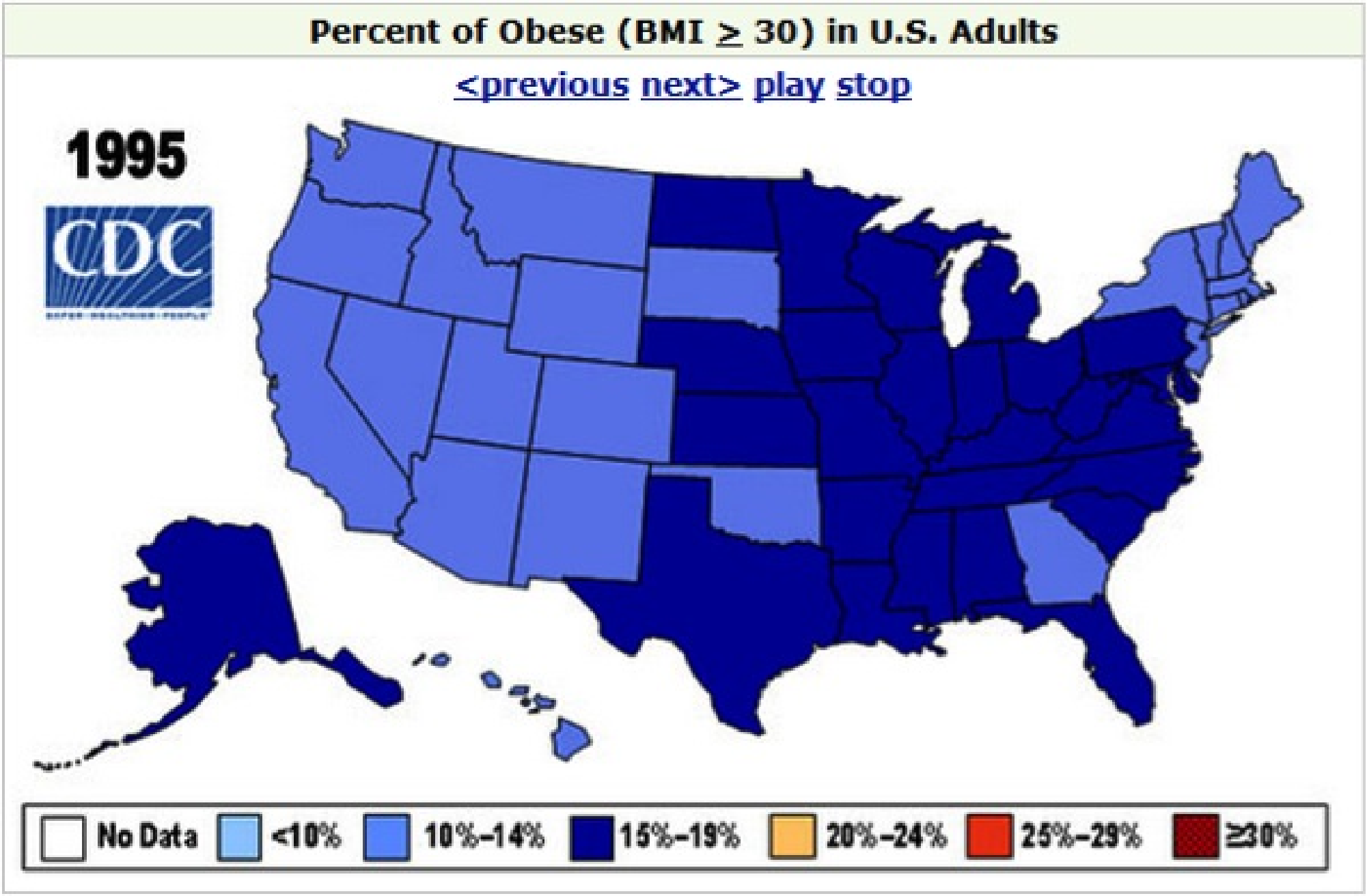
- National Health Interview Survey (NHIS)
- Sample: U.S. adult population, 2009

Percent of Obese (BMI ≥ 30) in U.S. Adults

<previous next> play stop

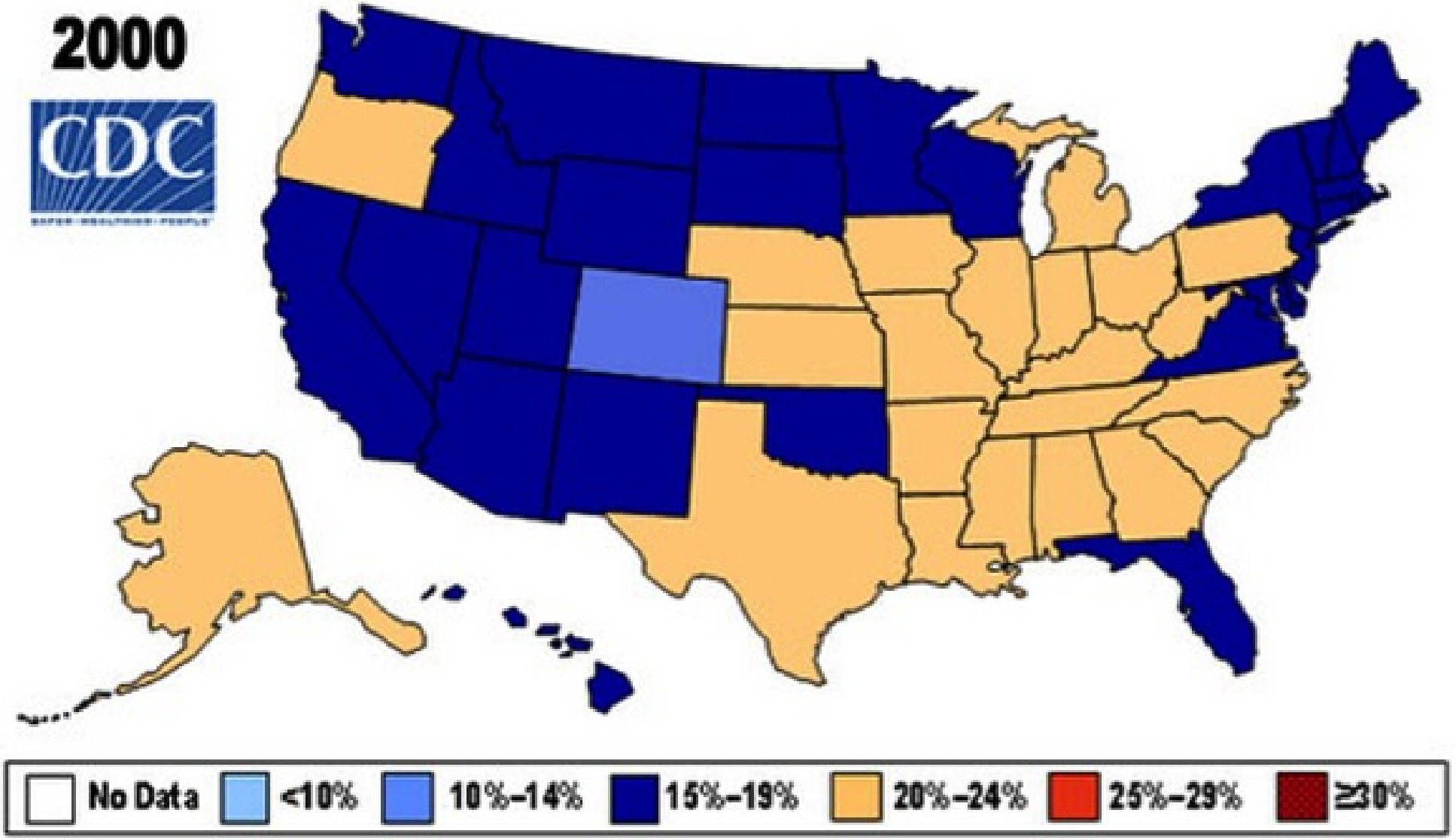



Percent of Obese (BMI ≥ 30) in U.S. Adults
<previous next> play stop
1990
CDC
No Data
<10%
10%–14%
15%–19%
20%–24%
25%–29%
≥30%


Percent of Obese (BMI ≥ 30) in U.S. Adults
<previous next> play stop
1995
CDC
No Data
<10%
10%–14%
15%–19%
20%–24%
25%–29%
≥30%

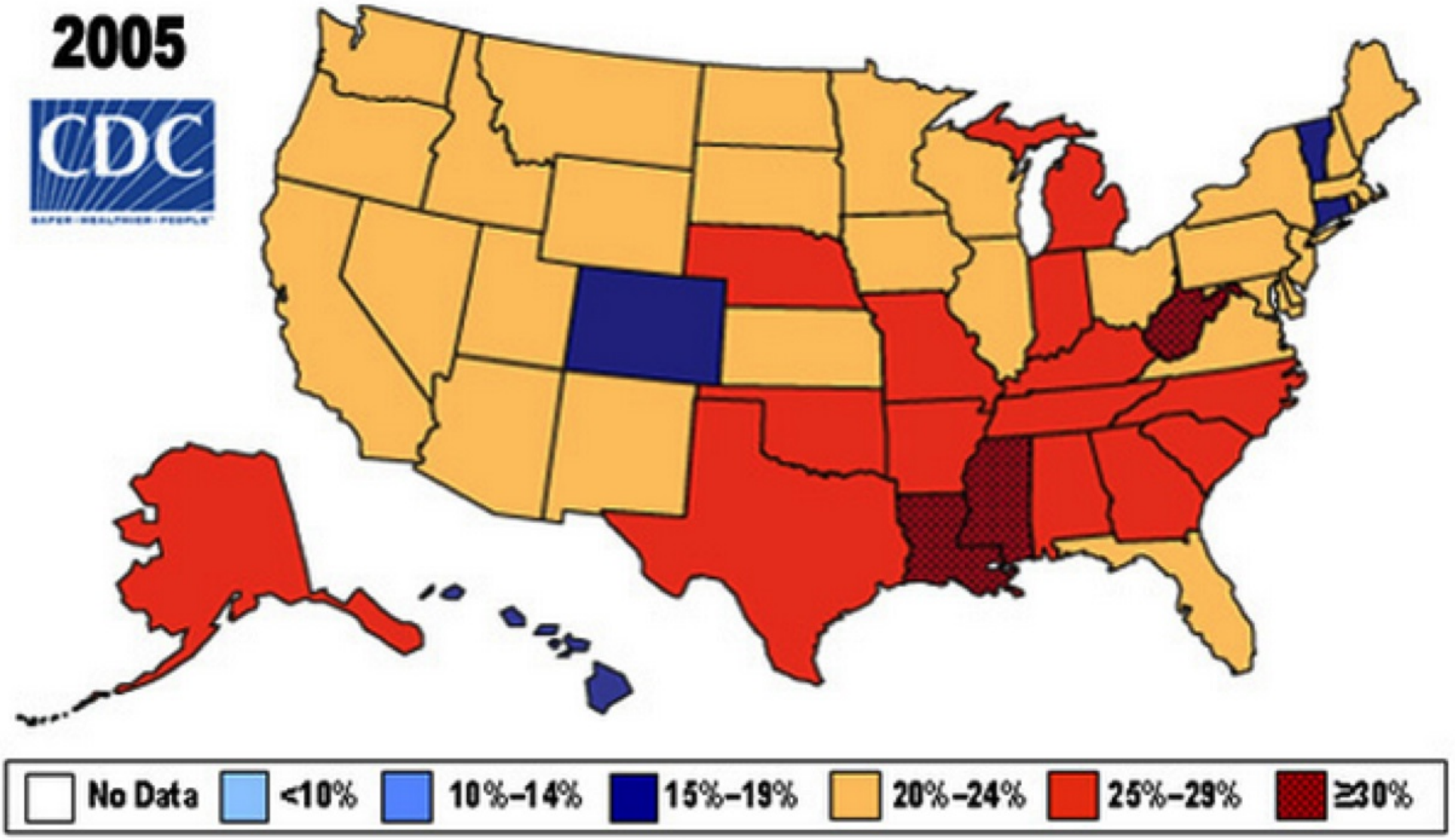
Percent of Obese (BMI ≥ 30) in U.S. Adults

<previous next> play stop



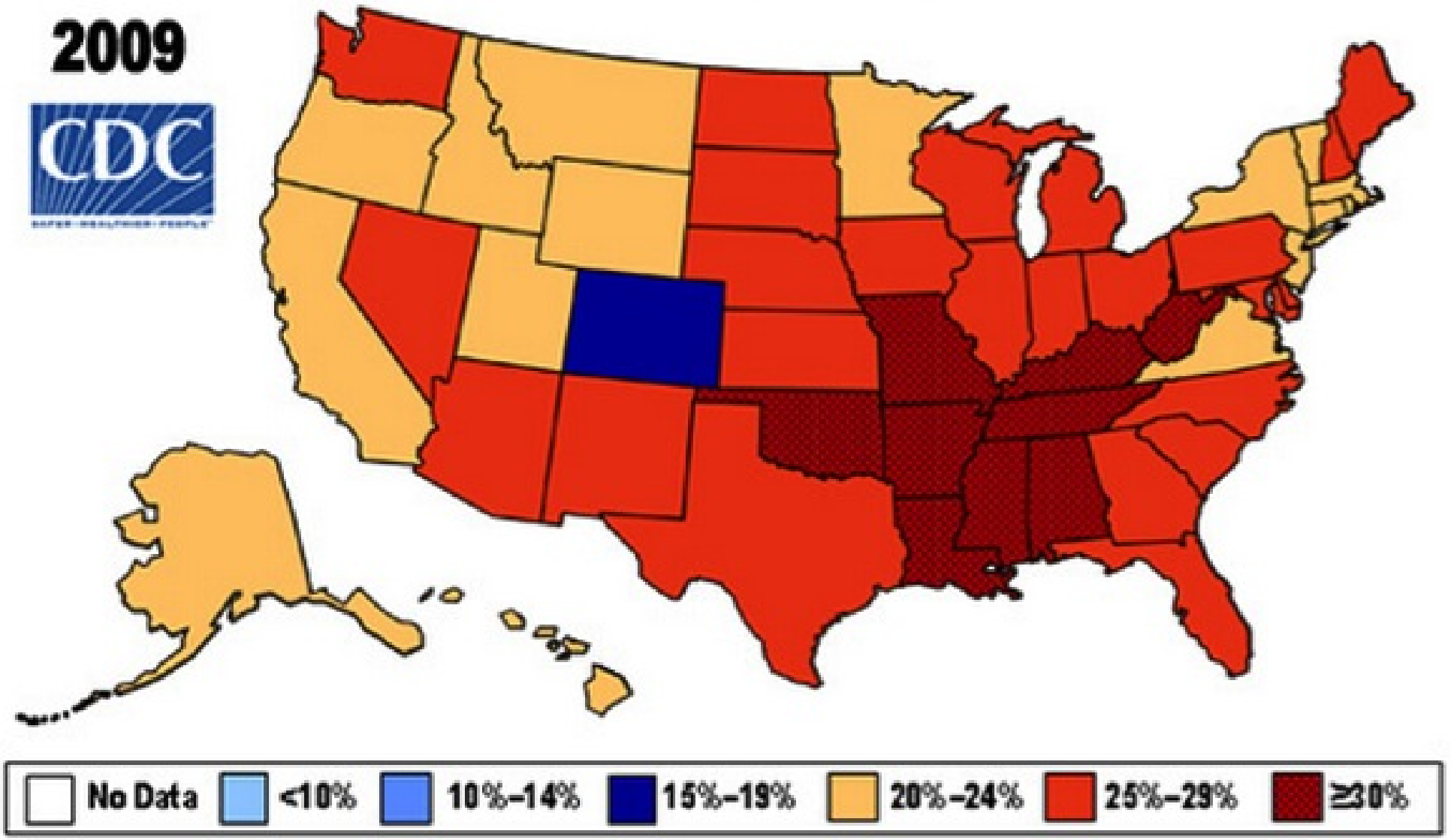
Percent of Obese (BMI ≥ 30) in U.S. Adults

<previous next> play stop

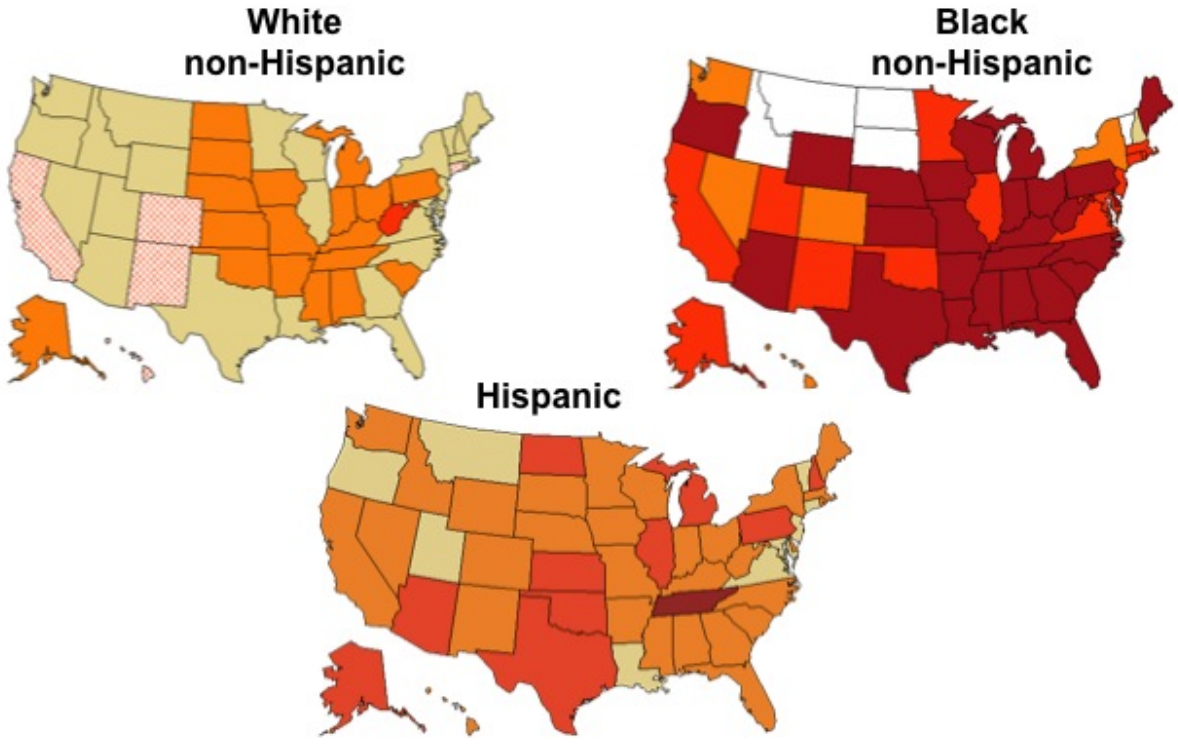


Percent of Obese (BMI ≥ 30) in U.S. Adults

<previous next> play stop



Another dimension of the issue



State-specific prevalence of obesity (BMI $\geq$ 30) among U.S. adults, by race/ethnicity, 2006–2008. Source: CDC.

Practice session

Class

```
* Get the demo code.
srqm fetch week2.do
* Open the do-file for this week.
doedit code/week2
```

Coursework

- Finish the do-file and read all comments at home.
- Start writing some draft code on your own interests.
- Start thinking about which course dataset to analyse.
- Read from the codebooks in the data/ folder.

Exercises

Ex 2.1. European Social Survey 2008

1. Load the data.
2. Find all variables on discrimination.
3. How many countries are there in the dataset?

Ex. 2.2. Quality of Government 2011

1. Load the data.
2. Find all variables on corruption.
3. Which one has the most observations?

Other sources

Personal selections

- `https://github.com/briatte/srqm/wiki/data`
- `http://srqm.tumblr.com/tagged/data`
- `https://pinboard.in/u:phnk/t:data:social-science`

Important

For this course, use one of the datasets from the `data/` folder. Those are high-quality, pre-processed datasets with minimal data preparation left for you to code.

Using an external data source would require several hours of additional work. Unfortunately, you do not have that time.

Opening and animating data online