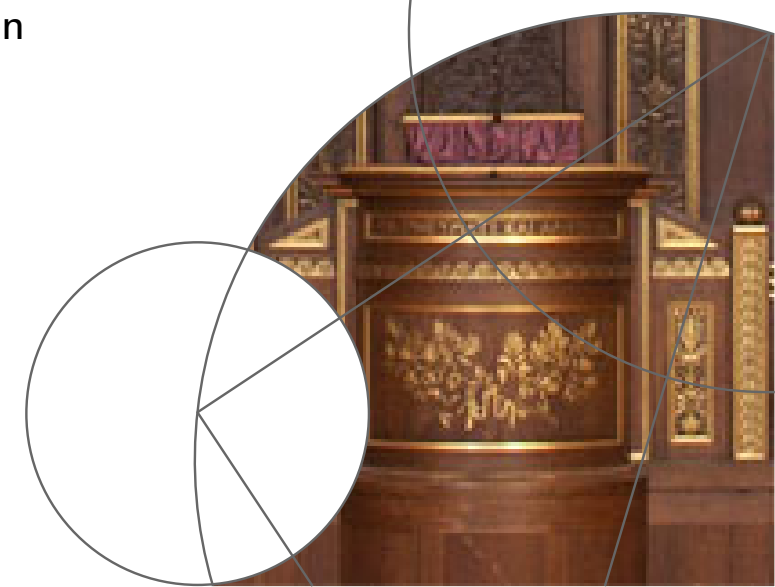
UNIVERSITY OF COPENHAGEN

Cleaning up the data cleaning process: Reproducible data cleaning in R

Anne Helby Petersen & Claus Thorn Ekstrøm
Department of Public Health, University of Copenhagen



A note on terminology

Reproducibility

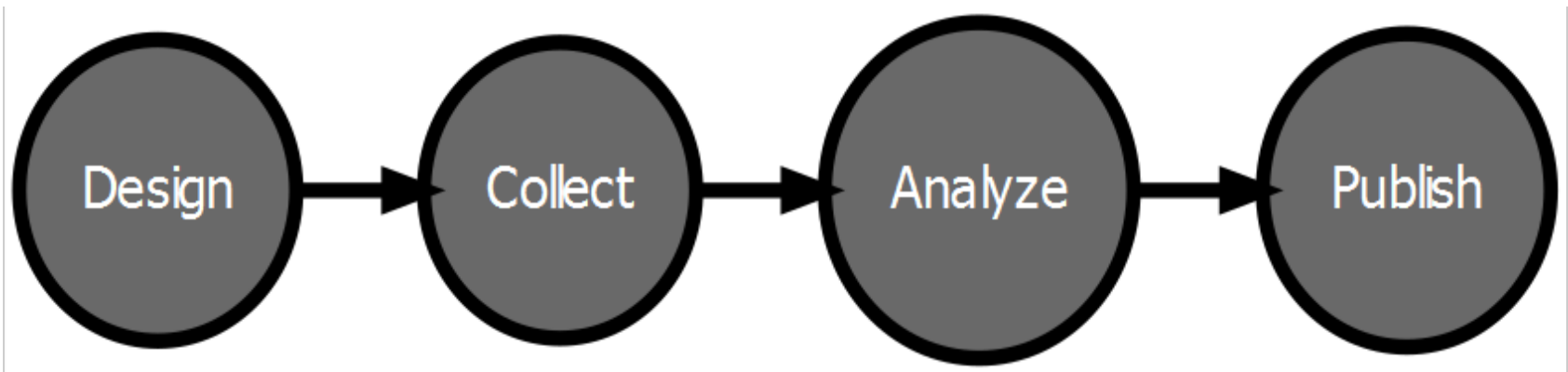
Given code/data/materials, can I get *the same* (=identical) numbers that you did?

Replicability

Given scientific protocol, can I get *the same* (=in agreement) result that you did in my own study?

Stylized research process

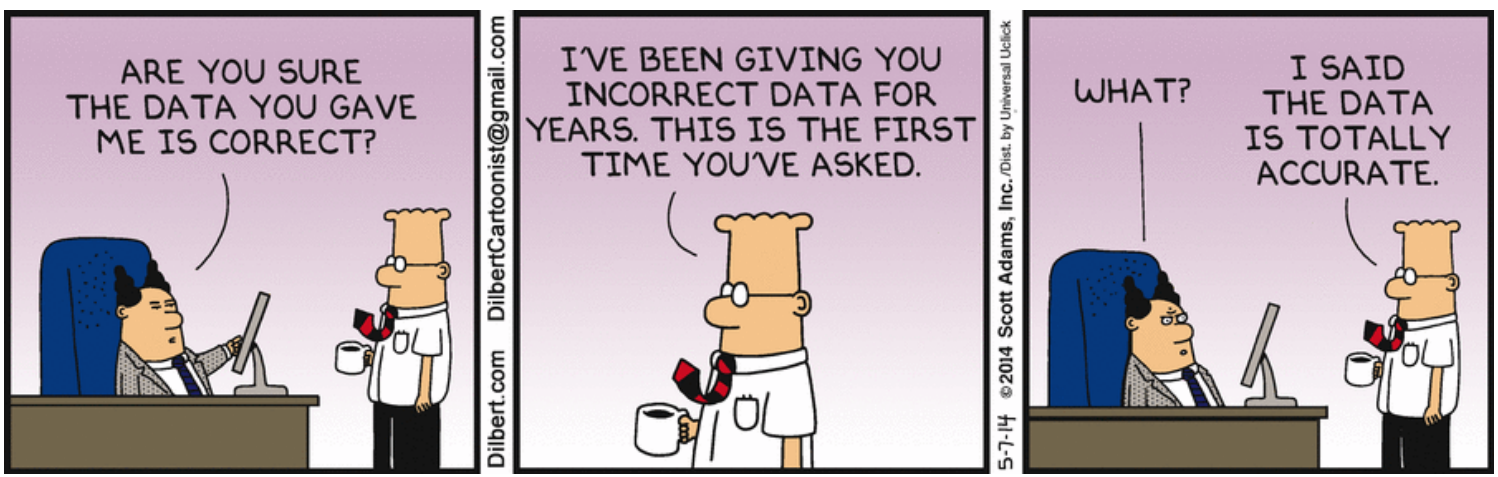
From idea . . .



. . . To publication

Reproducible computational social science

- Requires reproducible data screening/cleaning/wrangling
- Often collaborative effort across disciplines
- Discussions about data should include both the R-savvy and the more "classical" social scientist!



Data screening in dataMaid

Example: Cleaning dirty US president data

```
> library(dataMaid)
```

```
Loading required package: ggplot2
```

```
> data(bigPresidentData)
> bpd <- bigPresidentData
> dim(bpd)
```

```
[1] 47 15
```

lastName	firstName	orderOfPresidency	birthday	dateOfDeath	stateOfBirth	party
Ford	Gerald	38	1913-07-14	2006-12-26	Nebraska	Republican
Adams	John	2	1735-10-30	1826-07-04	Massachusetts	Federalist
Hoover	Herbert	31	1874-08-10	1964-10-20	Iowa	Republican
Washington	George	1	1732-02-22	1799-12-14	Virginia	Independent
Fillmore	Millard	13	1800-01-07	1874-03-08	New York	Whig
Clinton	William	42	1946-08-19	*NA*	Arkansas	Democratic
Harrison	Benjamin	23	1833-08-20	1901-03-13	Ohio	Republican
Monroe	James	5	1758-04-28	1831-07-04	Virginia	Democratic-Republican
Reagan	Ronald	40	1911-02-06	2004-06-05	Illinois	Republican
Kennedy	John	35	1917-05-29	1963-11-22	Massachusetts	Democratic
Coolidge	Hobbes	30	1872-07-04	1933-01-05	Vermont	Republican
Polk	James	11	1795-11-02	1849-06-15	North Carolina	Democratic
Nixon	Richard	37	1913-01-09	1994-04-22	California	Republican
Lincoln	Abraham	16	1809-02-12	1865-04-15	Kentucky	Republican/National Unior
Roosevelt	Theodore	26	1858-10-27	1919-01-06	New York	Republican
Madison	James	4	1751-03-16	1836-06-28	Virginia	Democratic-Republican
Bush	George	41	1924-06-12	*NA*	Massachusetts	Republican
Bush	George	43	1946-07-06	*NA*	Connecticut	Republican
Harrison	William	9	1773-02-09	1841-04-04	Ohio	Whig
Arthur	Chester	21	1830-10-05	1886-11-18	Vermont	Republican
Taylor	Zachary	12	1784-11-24	1850-07-09	Virginia	Whig

Creating a data report

One-liner:

```
> makeDataReport(bpd)
```

Creating a data report

Or multi-liner:

```
> makeDataReport(bpd,
+    output = "html",
+    useVar = "onlyProblematic",
+    mode = c("check", "summarize"),
+    replace = TRUE,
+    visuals = setVisuals(all = "basicVisual"),
+    checks = setChecks(numeric = "identifyOutliers"),
+    maxDecimals = 6,
+    reportTitle = "Very nice report",
+    maxProbVals = Inf,
+    treatXasY = list(complex = "numeric"))
```

dataMaid beyond the data report

- All summary/visual/check functions can be used interactively too (e.g. `identifyMissing(data$var)`)
- Easy to build custom summary/visual/check functions that can be added to the `dataMaid` toolbox
- Codebook generator function that makes a data report more suited for documenting clean data: `makeCodebook(data)`

Row-wise or column-wise checks?

bigPresidentData

Filter

	lastName	firstName	orderOfPresidency	birthday	dateOfDeath	stateOfBirth	party
38	Ford	Gerald	38	1913-07-14	2006-12-26	Nebraska	Republican
2	Adams	John	2	1735-10-30	1826-07-04	Massachusetts	Federalist
31	Hoover	Herbert	31	1874-08-10	1964-10-20	Iowa	Republican
1	Washington	George	1	1732-02-22	1799-12-14	Virginia	Independent
13	Fillmore	Millard	13	1800-01-07	1874-03-08	New York	Whig
42	Clinton	William	42	1946-08-19	NA	Arkansas	Democratic
23	Harrison	Benjamin	23	1833-08-20	1901-03-13	Ohio	Republican
5	Monroe	James	5	1758-04-28	1831-07-04	Virginia	Democratic-Republican
40	Reagan	Ronald	40	1911-02-06	2004-06-05	Illinois	Republican
35	Kennedy	John	35	1917-05-29	1963-11-22	Massachusetts	Democratic
30	Coolidge	Hobbes	30	1872-07-04	1933-01-05	Vermont	Republican
11	Polk	James	11	1795-11-02	1849-06-15	North Carolina	Democratic
37	Nixon	Richard	37	1913-01-09	1994-04-22	California	Republican
16	Lincoln	Abraham	16	1809-02-12	1865-04-15	Kentucky	Republican/National Union
26	Roosevelt	Theodore	26	1858-10-27	1919-01-06	New York	Republican
4	Madison	James	4	1751-03-16	1836-06-28	Virginia	Democratic-Republican
41	Bush	George	41	1924-06-12	NA	Massachusetts	Republican
43	Bush	George	43	1946-07-06	NA	Connecticut	Republican
9	Harrison	William	9	1773-02-09	1841-04-04	Ohio	Whig
21	Arthur	Chester	21	1830-10-05	1886-11-18	Vermont	Republican
12	Taylor	Zachary	12	1784-11-24	1850-07-09	Virginia	Whig

Row-wise or column-wise checks?

bigPresidentData

Filter

	lastName	firstName	orderOfPresidency	birthday	dateOfDeath	stateOfBirth	party
38	Ford	Gerald	38	1913-07-14	2006-12-26	Nebraska	Republican
2	Adams	John	2	1735-10-30	1826-07-04	Massachusetts	Federalist
31	Hoover	Herbert	31	1874-08-10	1964-10-20	Iowa	Republican
1	Washington	George	1	1732-02-22	1799-12-14	Virginia	Independent
13	Fillmore	Millard	13	1800-01-07	1874-03-08	New York	Whig
42	Clinton	William	42	1946-08-19	NA	Arkansas	Democratic
23	Harrison	Benjamin	23	1833-08-20	1901-03-13	Ohio	Republican
5	Monroe	James	5	1758-04-28	1831-07-04	Virginia	Democratic-Republican
40	Reagan	Ronald	40	1911-02-06	2004-06-05	Illinois	Republican
35	Kennedy	John	35	1917-05-29	1963-11-22	Massachusetts	Democratic
30	Coolidge	Hobbes	30	1872-07-04	1933-01-05	Vermont	Republican
11	Polk	James	11	1795-11-02	1849-06-15	North Carolina	Democratic
37	Nixon	Richard	37	1913-01-09	1994-04-22	California	Republican
16	Lincoln	Abraham	16	1809-02-12	1865-04-15	Kentucky	Republican/National Union
26	Roosevelt	Theodore	26	1858-10-27	1919-01-06	New York	Republican
4	Madison	James	4	1751-03-16	1836-06-28	Virginia	Democratic-Republican
41	Bush	George	41	1924-06-12	NA	Massachusetts	Republican
43	Bush	George	43	1946-07-06	NA	Connecticut	Republican
9	Harrison	William	9	1773-02-09	1841-04-04	Ohio	Whig
21	Arthur	Chester	21	1830-10-05	1886-11-18	Vermont	Republican
12	Taylor	Zachary	12	1784-11-24	1850-07-09	Virginia	Whig

Row-wise or column-wise checks?

bigPresidentData

Filter

	lastName	firstName	orderOfPresidency	birthday	dateOfDeath	stateOfBirth	party
38	Ford	Gerald	38	1913-07-14	2006-12-26	Nebraska	Republican
2	Adams	John	2	1735-10-30	1826-07-04	Massachusetts	Federalist
31	Hoover	Herbert	31	1874-08-10	1964-10-20	Iowa	Republican
1	Washington	George	1	1732-02-22	1799-12-14	Virginia	Independent
13	Fillmore	Millard	13	1800-01-07	1874-03-08	New York	Whig
42	Clinton	William	42	1946-08-19	NA	Arkansas	Democratic
23	Harrison	Benjamin	23	1833-08-20	1901-03-13	Ohio	Republican
5	Monroe	James	5	1758-04-28	1831-07-04	Virginia	Democratic-Republican
40	Reagan	Ronald	40	1911-02-06	2004-06-05	Illinois	Republican
35	Kennedy	John	35	1917-05-29	1963-11-22	Massachusetts	Democratic
30	Coolidge	Hobbes	30	1872-07-04	1933-01-05	Vermont	Republican
11	Polk	James	11	1795-11-02	1849-06-15	North Carolina	Democratic
37	Nixon	Richard	37	1913-01-09	1994-04-22	California	Republican
16	Lincoln	Abraham	16	1809-02-12	1865-04-15	Kentucky	Republican/National Union
26	Roosevelt	Theodore	26	1858-10-27	1919-01-06	New York	Republican
4	Madison	James	4	1751-03-16	1836-06-28	Virginia	Democratic-Republican
41	Bush	George	41	1924-06-12	NA	Massachusetts	Republican
43	Bush	George	43	1946-07-06	NA	Connecticut	Republican
9	Harrison	William	9	1773-02-09	1841-04-04	Ohio	Whig
21	Arthur	Chester	21	1830-10-05	1886-11-18	Vermont	Republican
12	Taylor	Zachary	12	1784-11-24	1850-07-09	Virginia	Whig

Row-wise inconsistency: Undead Lincoln?

```
> t(bpd[bpd$firstName == "Theodore" &
+       bpd$lastName == "Roosevelt",
+       c("birthday", "dateOfDeath",
+         "presidencyBeginDate", "presidencyEndDate")])
```

```
                      26
birthday              "1858-10-27"
dateOfDeath           "1919-01-06"
presidencyBeginDate   "1933-03-04"
presidencyEndDate     "1945-04-12"
```

Row-wise *and* column-wise constraints!

An R-packages that performs row-wise checks: `validate` (van der Loo & de Jonge)

```
> library(validate)
> val <- validator(`No zombies` = dateOfDeath >=
+                    presidencyBeginDate)
> res <- confront(bpd, val)
> summary(res)[, 1:6]
```

```
        name items passes fails nNA error
1 No.zombies    47     39     1   7 FALSE
```

Want to try out dataMaid?

- Package available on CRAN, development version on Github: github.com/ekstroem/dataMaid
- Article accepted for publication in the Journal of Statistical Software (manuscript on Github)
- Vignette about how to make custom extensions: `vignette("extending_dataMaid")`
- Alpha version of Shiny app:

```
> library(shiny)
> runGitHub("dataMaid", "ekstroem", subdir = "app")
```