Getting to Know Your Data


What's in the data?

Tukey

Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis.

Tukey (cont.)

Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure;

Tukey (cont.)

Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the **informality** and **flexibility** appropriate to the **exploratory character of exposure** can be fitted into any of the structures of formal statistics so far proposed.

Tukey (cont.)

Nothing—not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers—nothing can substitute here for the **flexibility of the informed human mind**.

Tukey (cont.)

Nothing—not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers—nothing can substitute here for the **flexibility of the informed human mind**.

Accordingly, both approaches and techniques need to be structured so as to **facilitate human involvement and intervention**.

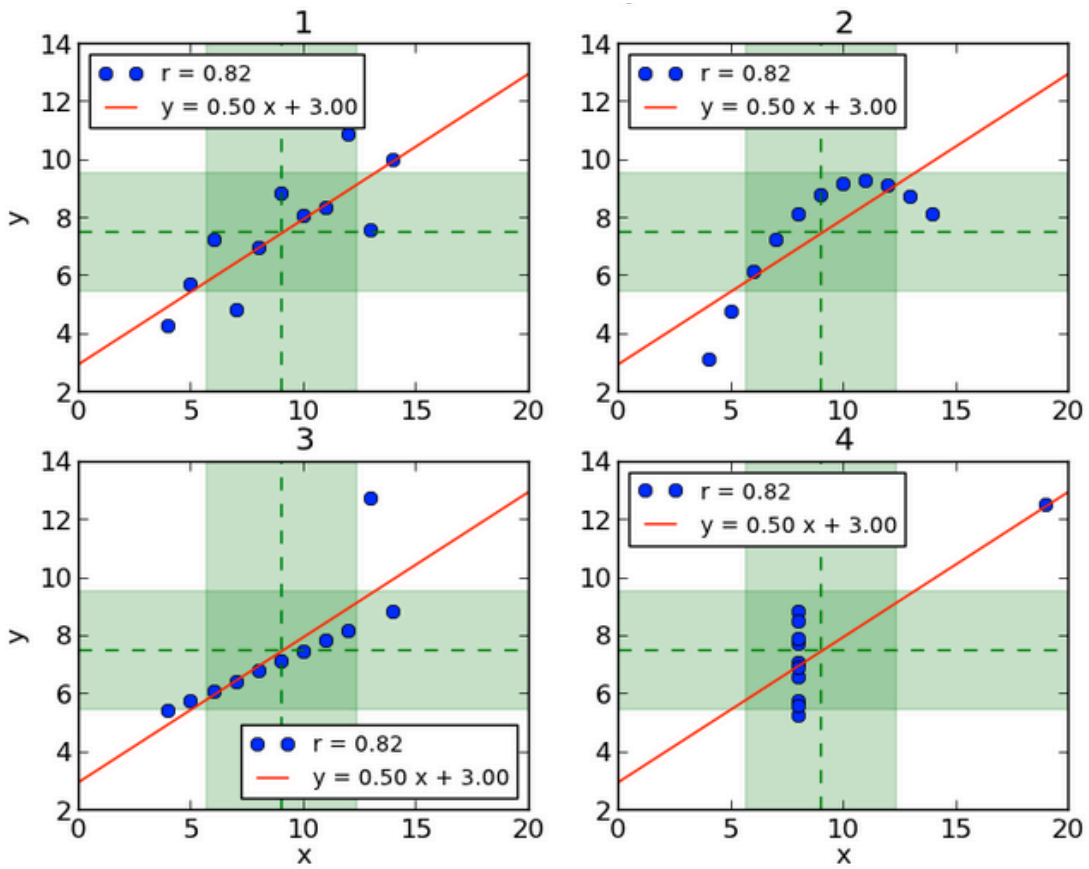
Summary Statistics

Useful to look at clean data, that you understand and trust

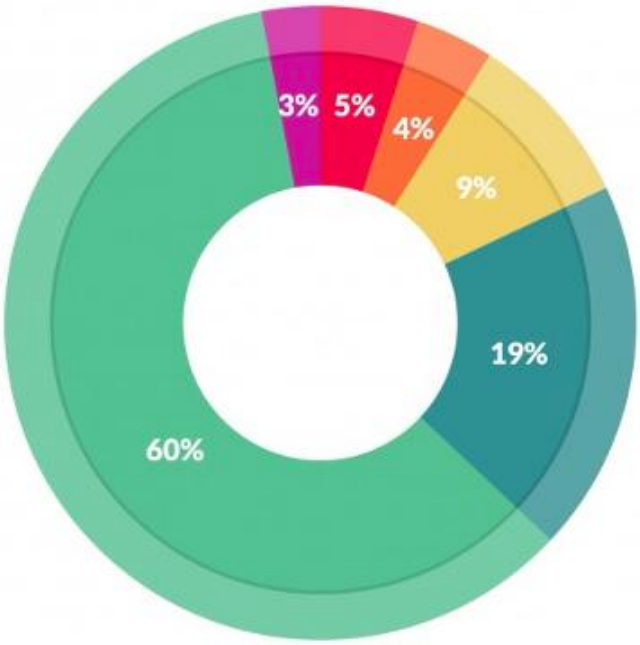
Can be misleading

Summary Statistics

Can be misleading



Data Munging



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Data Munging

60%

Data Munging

data_file.tsv:

```
# userID  gender  is_pregnant
  9       M       0
  10      F       0
  88      N/A     1
  11      .       1
  109     M       1
```

Data Munging

data_file_cleaned_2015-09-21.tsv:

```
# userID  gender  is_pregnant
  9       M       0
  10      F       0
  11      N/A     1
  88      N/A     1
```

Data Quality Hurdles

Missing Data

Erroneous Values

Type Conversion

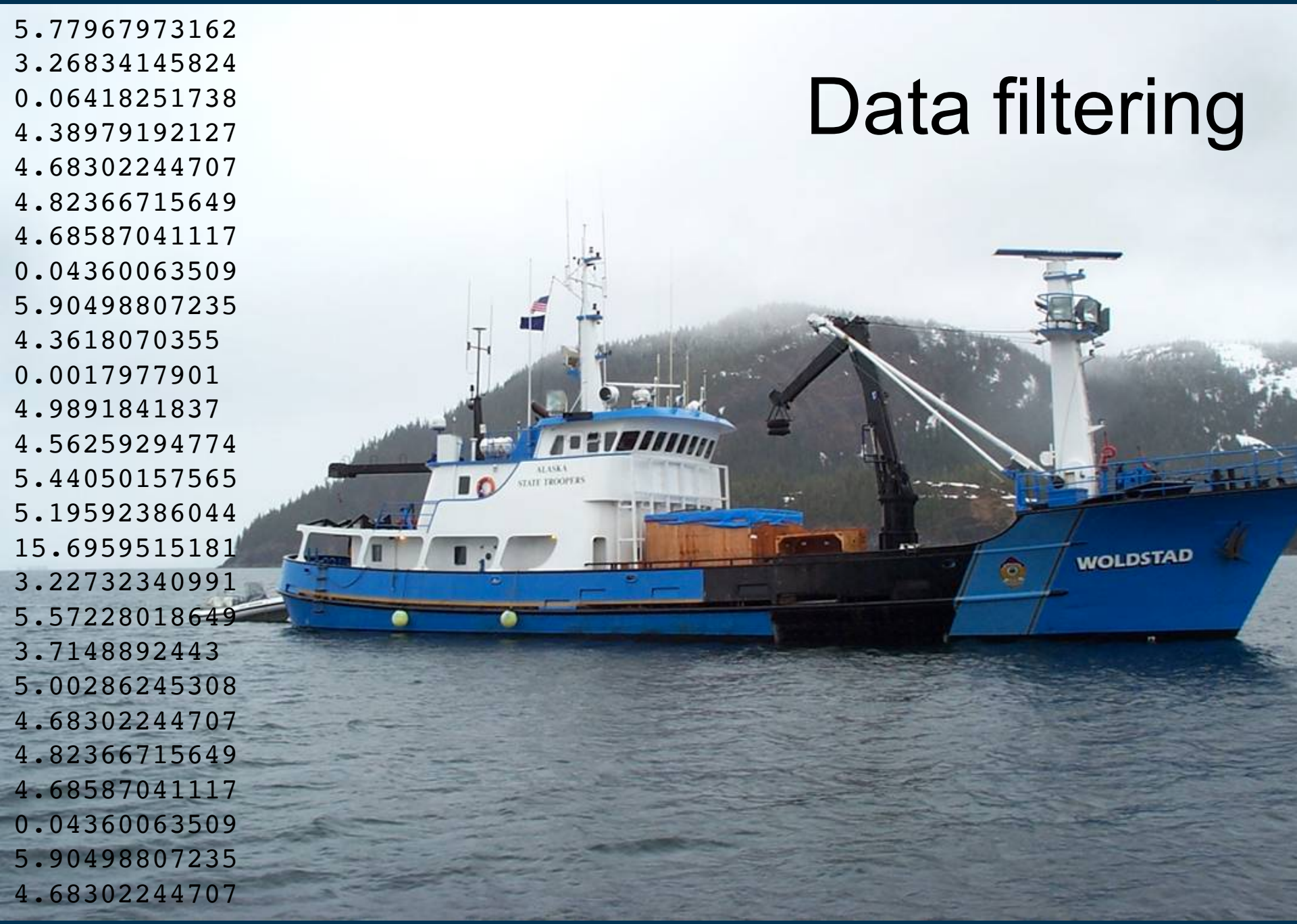
Entity Resolution

Data Integration

More Data Munging

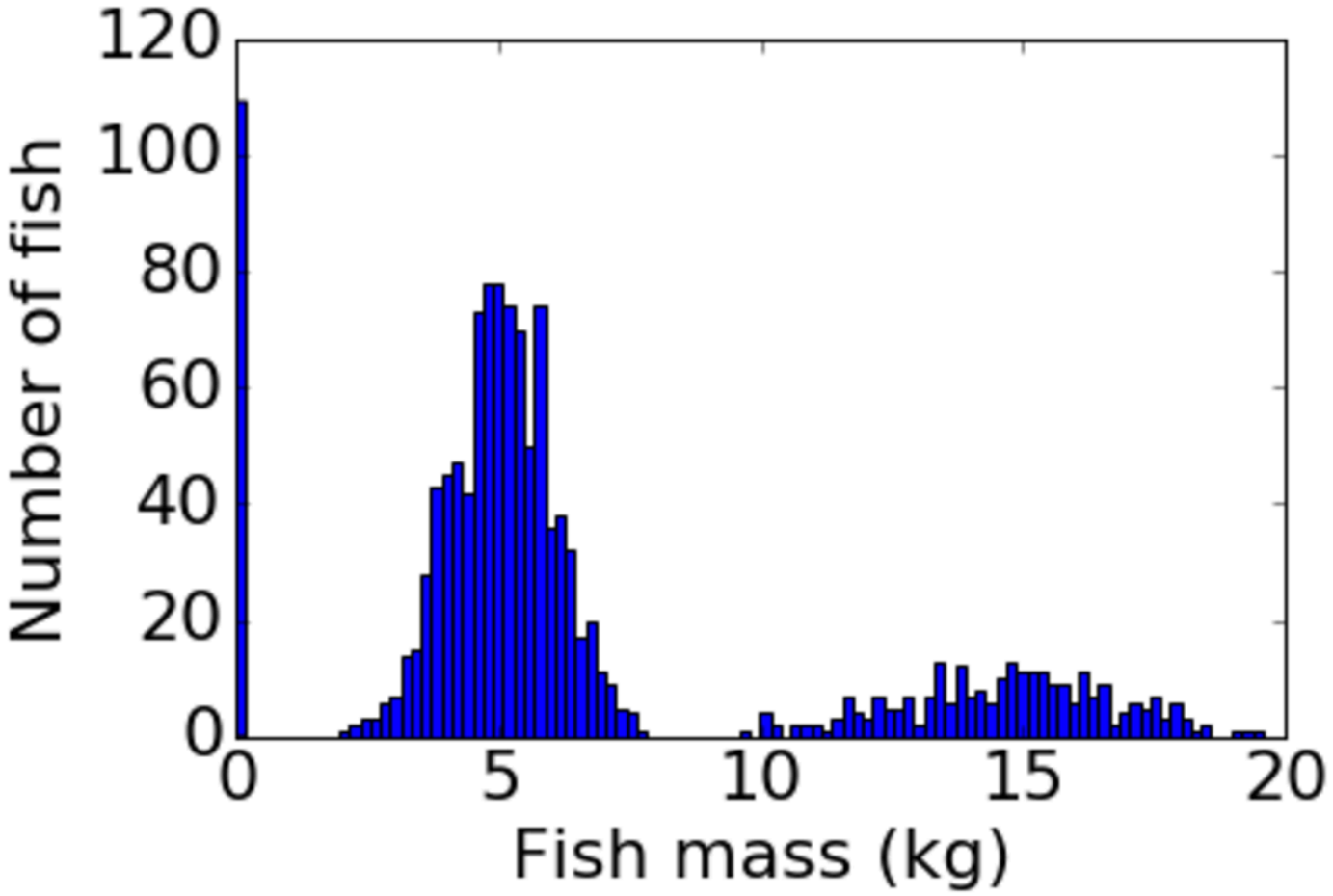
```
'1911 - 1961'
'1958\u201386'
'1921\u201376'
'427 BC \u2013 386 BC'
'1983 \u2013 present'
'1983\u2013present'
'1991\u20132001'
'1983{{spaced ndash}}present'
'<!-- YYYY\u2013YYYY (or \u2013present) -->'
'1989\u2013present'
'1984\u20132001<br /> 2005\u2013present'
'c. 1914\u20131971'
'1960&ndash;present'
'1888---c.1920'
```

Data filtering



5.77967973162
3.26834145824
0.06418251738
4.38979192127
4.68302244707
4.82366715649
4.68587041117
0.04360063509
5.90498807235
4.3618070355
0.0017977901
4.9891841837
4.56259294774
5.44050157565
5.19592386044
15.6959515181
3.22732340991
5.57228018649
3.7148892443
5.00286245308
4.68302244707
4.82366715649
4.68587041117
0.04360063509
5.90498807235
4.68302244707

5.77967973162
3.26834145824
0.06418251738
4.38979192127
4.68302244707
4.82366715649
4.68587041117
0.04360063509
5.90498807235
4.3618070355
0.0017977901
4.9891841837
4.56259294774
5.44050157565
5.19592386044
15.6959515181
3.22732340991
5.57228018649
3.7148892443
5.00286245308


120
100
80
60
40
20
0
Number of fish
0
5
10
15
20
Fish mass (kg)

“The first sign that a visualization is good is that it shows you a problem in your data.”

—Wattenberg

Berkeley SCHOOL OF INFORMATION