Data Cleaning

Everything Data
CompSci 290.01 Spring 2014

DUKE
COMPUTER SCIENCE

Announcements (Mon. Jan. 12)

- Team assignments will be posted before Lab 1

- Office Hours:
  - Jun Yang: M 4:30 – 5:30 PM
  - Ashwin Machanavajjhala: W 4:30 – 5:30 PM

Sources of data …

- Paper records
- Sensors
- Web pages
- Activity Logs
- …

… are noisy

- Data entry errors
- Measurement errors
- Extraction errors

Typical data quality issues

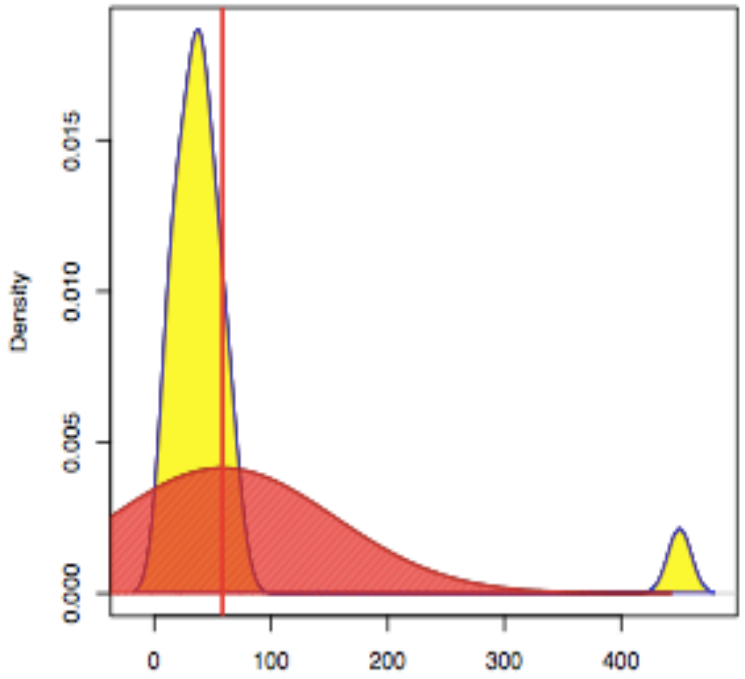
1) parsing text into fields (separator issues)
2) Naming conventions: ER: NYC vs New York
3) Missing required field (e.g. key field)
4) Different representations (2 vs Two)
5) Fields too long (get truncated)
6) Primary key violation (two people with the same social security number)
7) Redundant Records (exact match or other)
8) Formatting issues – especially dates
9) Licensing issues/Privacy/ keep you from using the data as you would like

Adapted from Berkeley CS 194 Lecture notes

Domain knowledge is critical

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

ages of employees (US)

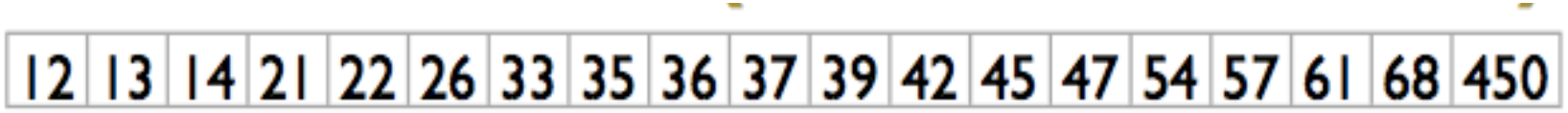


- median 37
- mean 58.52632
- variance 9252.041

Adapted from Berkeley CS 194 Lecture notes

Domain knowledge is critical

- Same data values in a different domain may not have errors …

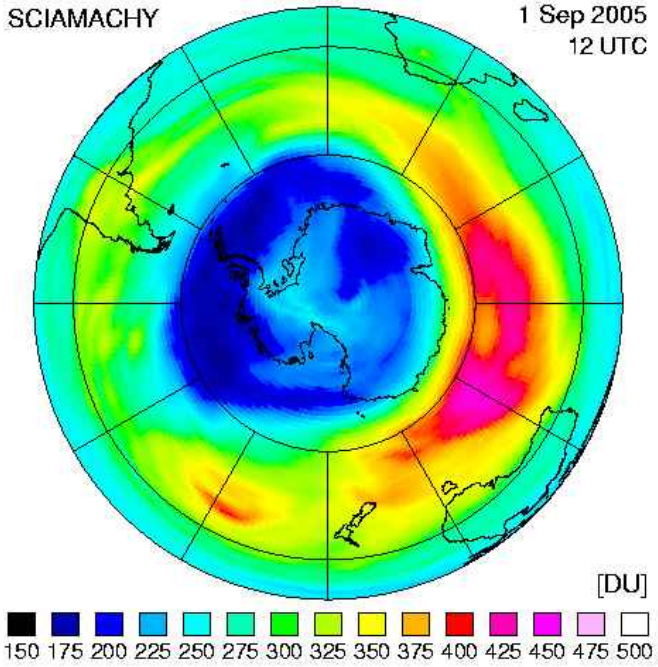


- Number of friends in a social network
  - A few people in Facebook have millions of friends, while the average number of friends is on the order of hundreds

Data Cleaning Makes Everything Okay?

The appearance of a hole in the earth's ozone layer over Antarctica, first detected in 1976, was so unexpected that scientists didn't pay attention to what their instruments were telling them; they thought their instruments were malfunctioning.

National Center for Atmospheric Research



In fact, the data were rejected as unreasonable by data quality control algorithms

Adapted from Berkeley CS 194 Lecture notes

Regular Expressions

Regular expressions

- A formal language to specify sets of strings.
  - Useful for extracting data fields from text
  - Useful for searching through text
  - (think “Find” or “grep”)

Disjunctions

- *Find all occurrences of letter ‘n’ in the following sentence.*

  “I am nobody.
  Nobody is perfect.
  Therefore I am perfect.”

- /[Nn]/

Disjunctions

Expression	Meaning
/[Nn]obody/	Nobody or nobody
/[A-Z]/	All upper case letters
/[0-9]/	All numbers
/[A-Za-z]/	All letters
/[ ]/	All white spaces

- *How do you search for 'nobody' or 'perfect' ?*

nobody | perfect

Negations

- *Find all letters that are not ‘n’ in the following sentence.*

  “I am nobody.
  Nobody is perfect.
  Therefore I am perfect.”

- /[^Nn]/

Negations

Expression	Meaning
/[^]/	Non space characters
/[^A-Z]/	Not capital letters
/[^A-Za-z]/	Not letters

- *How do you search for '^' ?*

/[\^]/

Regular operators: * + ? .

Expression	Meaning
/no?body/	nobody , nbody
/no*body/	nbody, nobody, noobody, nooobody, ….
/no+body/	nobody, noobody, nooobody …
/n.body/	n body, nobody, n3body, npbody …

- ? : 0 or 1 occurrence
- * : 0 or more occurrences
- + : 1 or more occurrences
- . : any one character

Start and End

- ^ : start of a sentence
- $: end of a sentence

() operator

- (no)?body matches body and nobody

- String matched by the expression within () can be captured
  - /(no)body/ when applied to the string "nobody" returns an array of length 1 with the string "no".

Exercise

- *Find all occurrences of the word 'am' in the sentence:*

  "Am I ashamed of who I am?"

Ans: the 2nd group in

/(^|[^A-Za-z])([Aa]m)($|[^A-Za-z])/

Matches start of line or non letter

Matches end of line or non letter

Summary

- Cleaning is an important step before making sense of data.
- Diverse set of techniques are usually employed to fix many types of errors.
- Regular expressions are a useful tool to parse the data into the right format.