Structure of a Data Analysis

Part 1

Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Steps in a data analysis

- Define the question

- Define the ideal data set

- Determine what data you can access

- Obtain the data

- Clean the data

- Exploratory data analysis

- Statistical prediction/modeling

- Interpret results

- Challenge results

- Synthesize/write up results

- Create reproducible code

Steps in a data analysis

- Define the question

- Define the ideal data set

- Determine what data you can access

- Obtain the data

- Clean the data

- Exploratory data analysis

- Statistical prediction/modeling

- Interpret results

- Challenge results

- Synthesize/write up results

- Create reproducible code

The key challenge in data analysis

"Ask yourselves, what problem have you solved, ever, that was worth solving, where you knew knew all of the given information in advance? Where you didn't have a surplus of information and have to filter it out, or you didn't have insufficient information and have to go find some?"

Dan Myer, Mathematics Educator

4/16

Defining a question
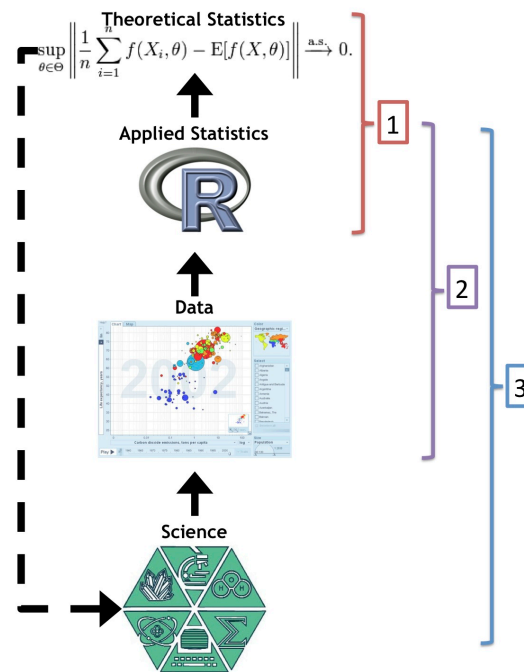
1. Statistical methods development

2. Danger zone!!!

3. Proper data analysis

An example

Start with a general question

Can I automatically detect emails that are SPAM that are not?

Make it concrete

Can I use quantitative characteristics of the emails to classify them as SPAM/HAM?

Define the ideal data set

- The data set may depend on your goal

    - Descriptive - a whole population

    - Exploratory - a random sample with many variables measured

    - Inferential - the right population, randomly sampled

    - Predictive - a training and test data set from the same population

    - Causal - data from a randomized study

    - Mechanistic - data about all components of the system

7/16

Our example

http://www.google.com/about/datacenters/inside/
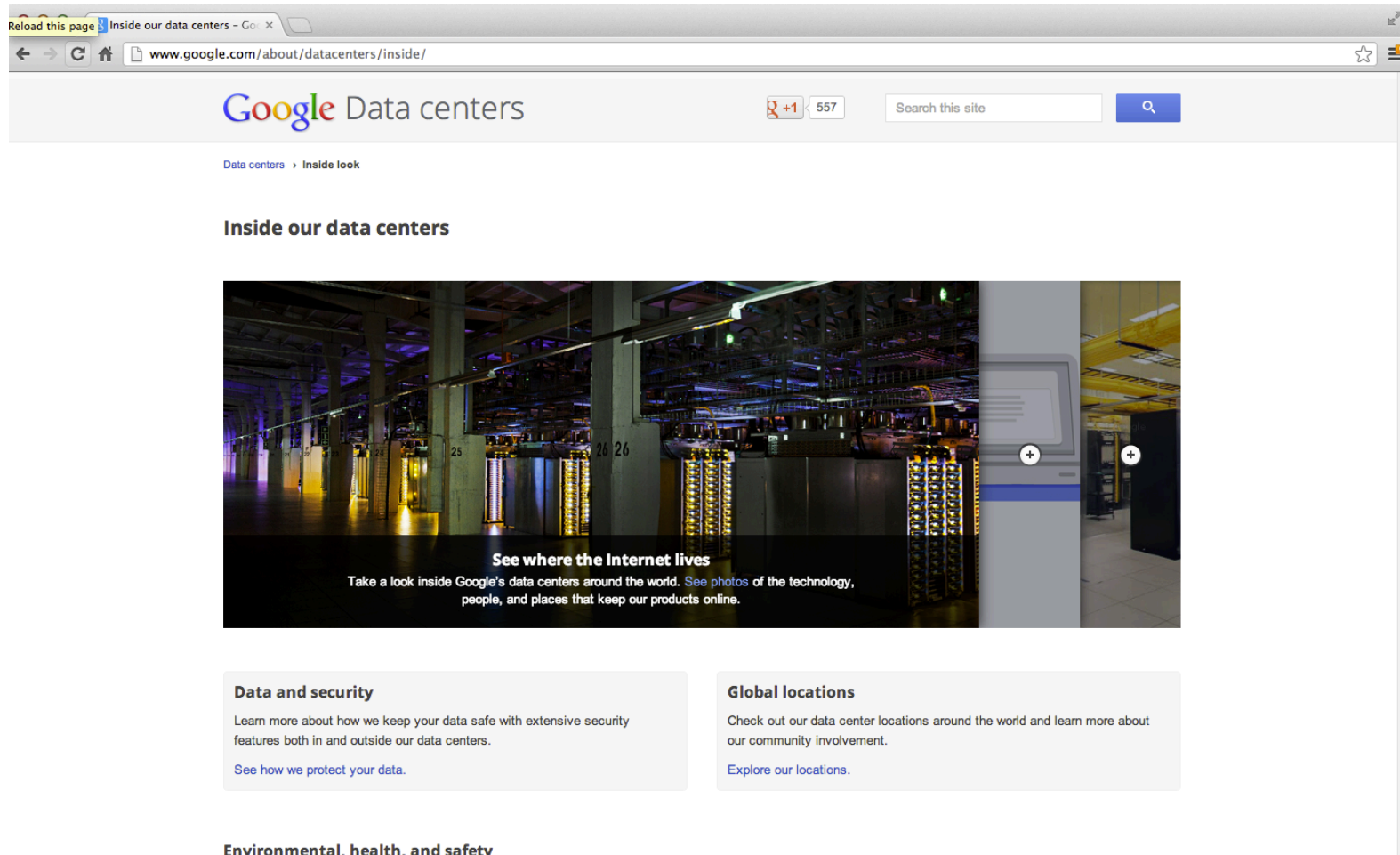
Determine what data you can access

- Sometimes you can find data free on the web

- Other times you may need to buy the data

- Be sure to respect the terms of use

- If the data don't exist, you may need to generate it yourself
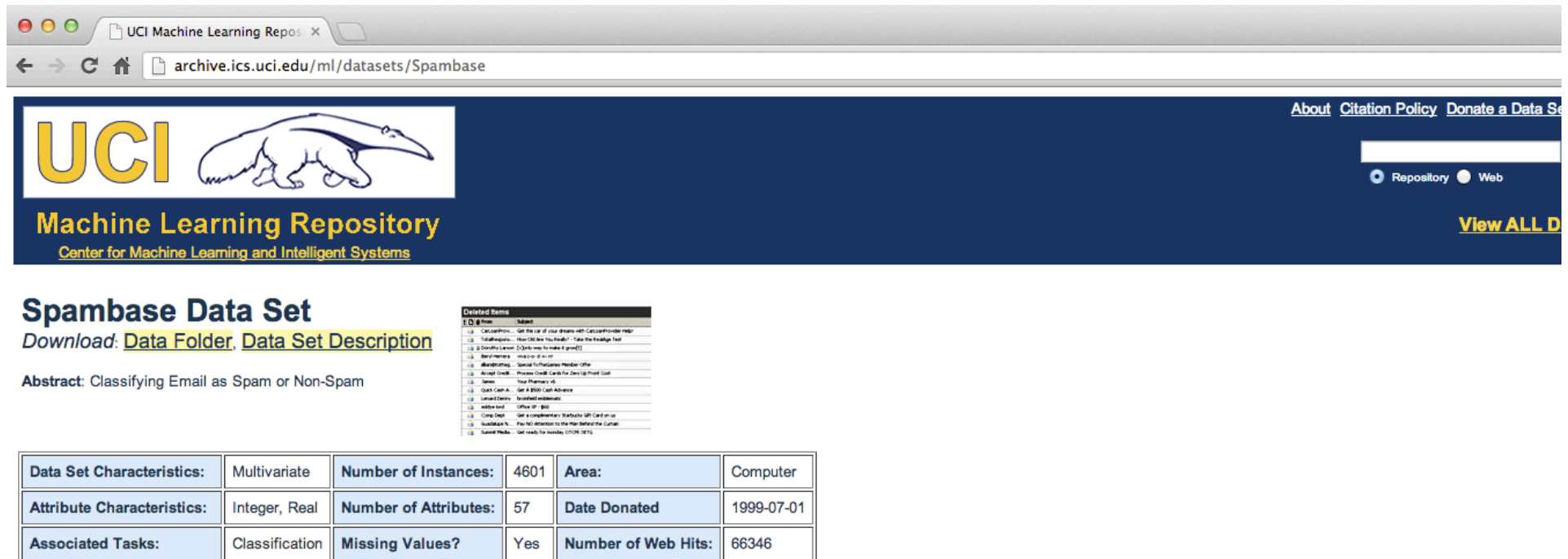
Back to our example
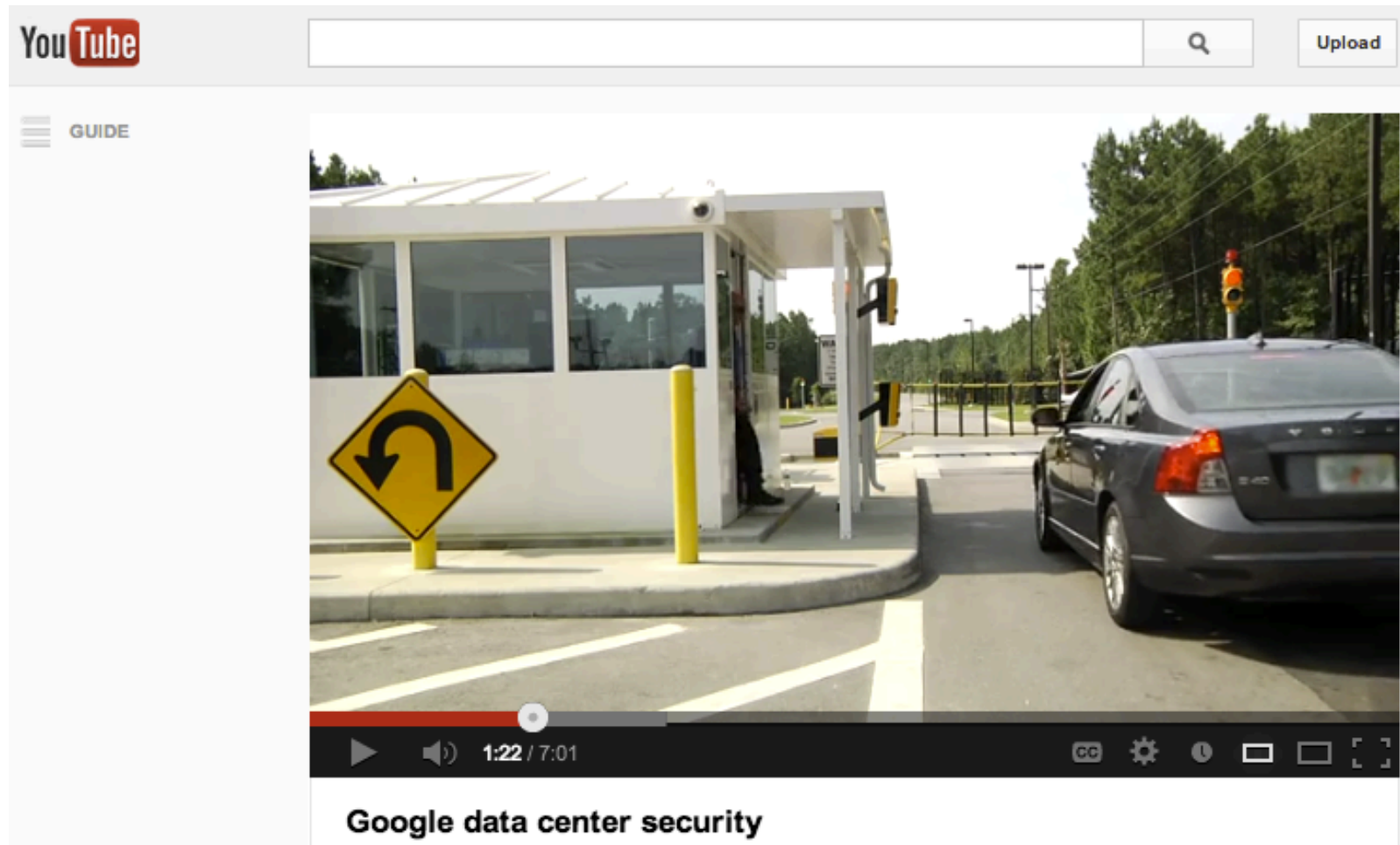
Google data center security

A possible solution

http://archive.ics.uci.edu/ml/datasets/Spambase

Obtain the data

- Try to obtain the raw data

- Be sure to reference the source

- Polite emails go a long way

- If you will load the data from an internet source, record the url and time accessed

Our data set

13/16

Clean the data

- Raw data often needs to be processed

- If it is pre-processed, make sure you understand how

- Understand the source of the data (census, sample, convenience sample, etc.)

- May need reformating, subsampling - record these steps

- **Determine if the data are good enough** - if not, quit or change data

14/16

Our cleaned data set

```
# If it isn't installed, install the kernlab package
library(kernlab)
data(spam)
dim(spam)
```

```
[1] 4601   58
```

http://rss.acs.unt.edu/Rdoc/library/kernlab/html/spam.html

Subsampling our data set

We need to generate a test and training set (prediction)

```
set.seed(3435)
trainIndicator = rbinom(4601,size=1,prob=0.5)
table(trainIndicator)
```

```
trainIndicator
   0    1
2314 2287
```

```
trainSpam = spam[trainIndicator==1,]
testSpam = spam[trainIndicator==0,]
dim(trainSpam)
```

```
[1] 2287   58
```