



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

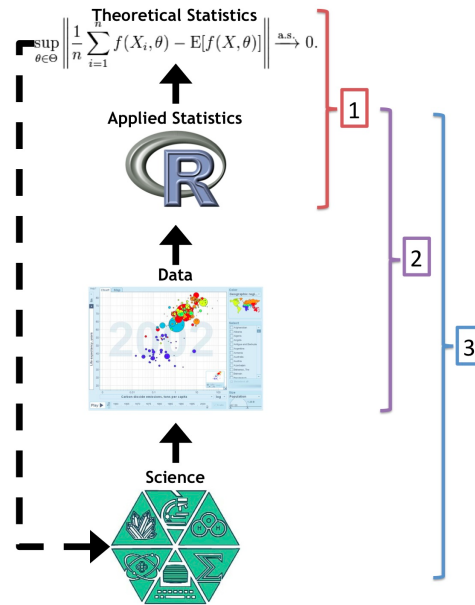
Steps in a data analysis

- 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?”

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?

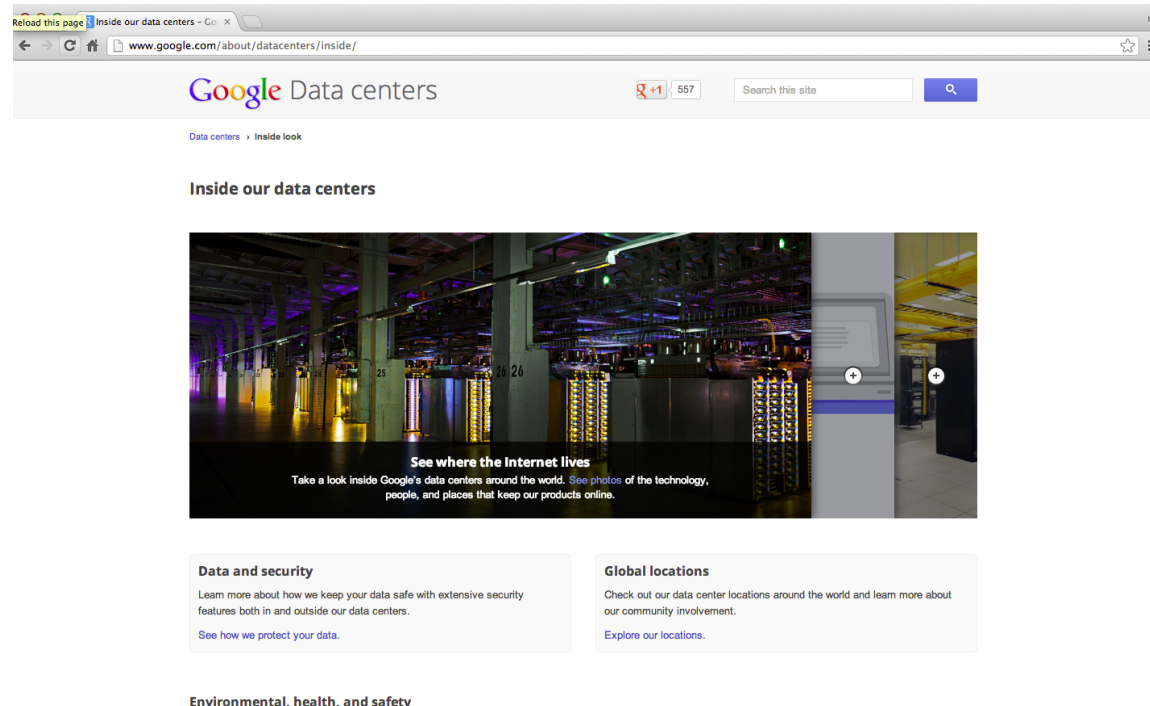
Make it concrete

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

Our example

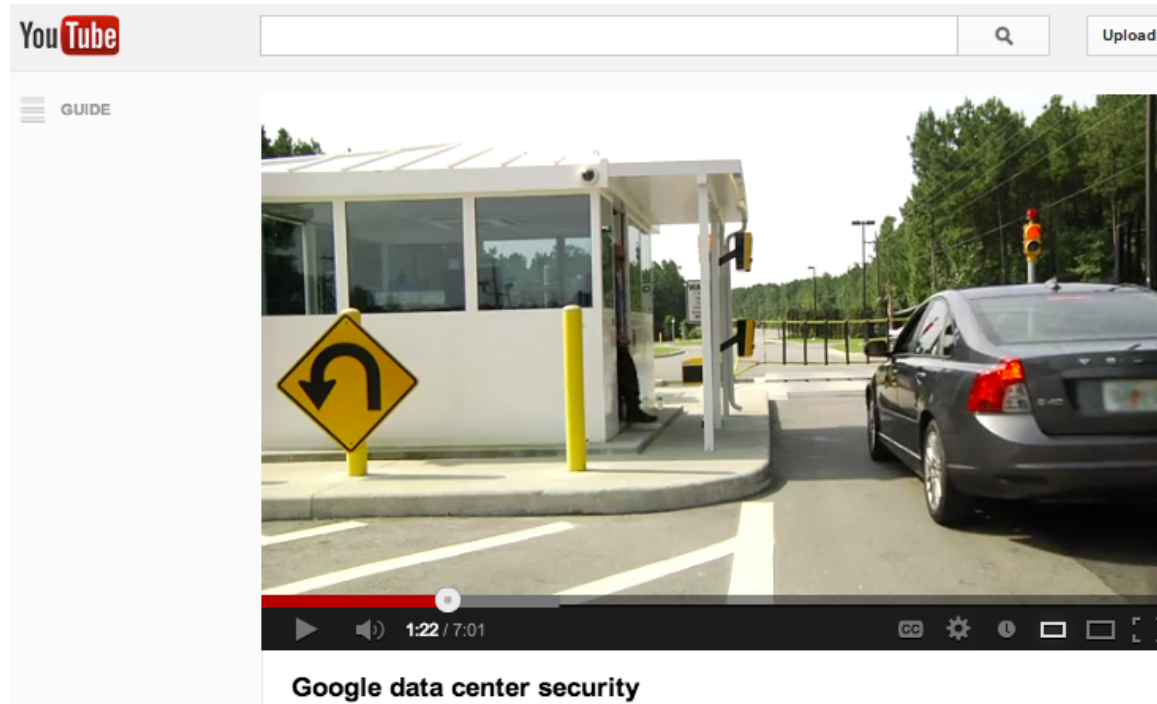


<http://www.google.com/about/datacenters/inside/>

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

The screenshot shows the UCI Machine Learning Repository website. The browser address bar displays `archive.ics.uci.edu/ml/datasets/Spambase`. The page header includes the UCI logo, the text "Machine Learning Repository", and the subtitle "Center for Machine Learning and Intelligent Systems". Navigation links for "About", "Citation Policy", "Donate a Data Set", and "Contact" are present. A search bar and a "Google" logo are also visible. The main content area is titled "Spambase Data Set" and includes links for "Download: Data Folder" and "Data Set Description". An abstract states: "Classifying Email as Spam or Non-Spam". To the right of the abstract is a list of "Related Datasets". Below the abstract is a table with data set characteristics.

Data Set Characteristics:	Multivariate	Number of Instances:	4601	Area:	Computer
Attribute Characteristics:	Integer, Real	Number of Attributes:	57	Date Donated	1999-07-01
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	66346

Source:

Creators:
 Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt
 Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304

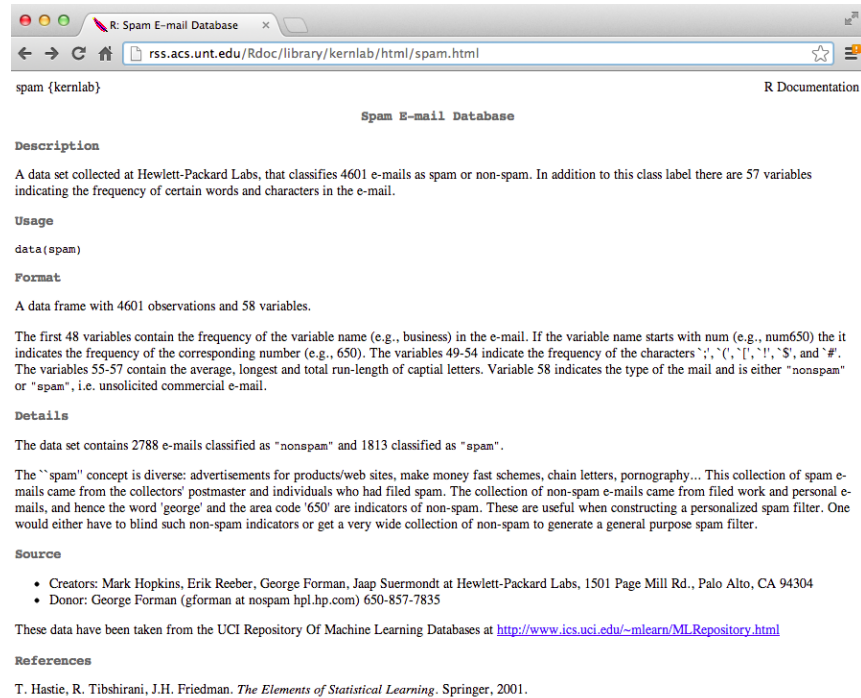
Donor:
 George Forman (gforman at nospam hpl.hp.com) 650-857-7835

<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



The screenshot shows a web browser window with the title "R: Spam E-mail Database". The address bar displays the URL rss.acs.unt.edu/Rdoc/library/kernlab/html/spam.html. The page content is titled "Spam E-mail Database" and includes sections for "Description", "Usage", "Format", "Details", "Source", and "References".

Description

A data set collected at Hewlett-Packard Labs, that classifies 4601 e-mails as spam or non-spam. In addition to this class label there are 57 variables indicating the frequency of certain words and characters in the e-mail.

Usage

```
data(spam)
```

Format

A data frame with 4601 observations and 58 variables.

The first 48 variables contain the frequency of the variable name (e.g., business) in the e-mail. If the variable name starts with num (e.g., num650) the it indicates the frequency of the corresponding number (e.g., 650). The variables 49-54 indicate the frequency of the characters `;`, `!`, `I`, `S`, and `#`. The variables 55-57 contain the average, longest and total run-length of capital letters. Variable 58 indicates the type of the mail and is either "nonspam" or "spam", i.e. unsolicited commercial e-mail.

Details

The data set contains 2788 e-mails classified as "nonspam" and 1813 classified as "spam".

The "spam" concept is diverse: advertisements for products/web sites, make money fast schemes, chain letters, pornography... This collection of spam e-mails came from the collectors' postmaster and individuals who had filed spam. The collection of non-spam e-mails came from filed work and personal e-mails, and hence the word 'george' and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter.

Source

- Creators: Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt at Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304
- Donor: George Forman (gforman at nospam hpl.hp.com) 650-857-7835

These data have been taken from the UCI Repository Of Machine Learning Databases at <http://www.ics.uci.edu/~mllearn/MLRepository.html>

References

T. Hastie, R. Tibshirani, J.H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

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

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 reformatting, subsampling - record these steps
- **Determine if the data are good enough** - if not, quit or change data

Our cleaned data set

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

<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)
```