GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 2: CLEANING DATA

OCTOBER 8, 2014 // DAT10 SF

Francesco Mosconi, PhD

DATA SCIENCE IN THE NEWS

DATA SCIENCE IN THE NEWS

Gyrophone: Recognizing Speech From Gyroscope Signals

Yan Michalevsky Dan Boneh
Computer Science Department
Stanford University

Gabi Nakibly
National Research & Simulation Center
Rafael Ltd.

Abstract

We show that the MEMS gyroscopes found on modern smart phones are sufficiently sensitive to measure acoustic signals in the vicinity of the phone. The resulting signals contain only very low-frequency information (<200Hz). Nevertheless we show, using signal processing and machine learning, that this information is sufficient to identify speaker information and even parse speech. Since iOS and Android require no special permissions to access the gyro, our results show that apps and active web content that cannot access the microphone can nevertheless eavesdrop on speech in the vicinity of the phone.

1 Introduction

Modern smartphones and mobile devices have many sensors that enable rich user experience. Being generally put to good use, they can sometimes unintentionally expose information the user does not want to share. While the privacy risks associated with some sensors like a microphone (eavesdropping), camera or GPS (tracking) are obvious and well understood, some of the risks remained under the radar for users and application developers. In particular, access to motion sensors such as gyroscope and accelerometer is unmitigated by mobile operating systems. Namely, every application installed on a phone and every web page browsed over it can measure and record these sensors without the user being aware of it.

gyroscopes are sufficiently sensitive to measure acoustic vibrations. This leads to the possibility of recovering speech from gyroscope readings, namely using the gyroscope as a crude microphone. We show that the sampling rate of the gyroscope is up to 200 Hz which covers some of the audible range. This raises the possibility of eavesdropping on speech in the vicinity of a phone without access to the real microphone.

As the sampling rate of the gyroscope is limited, one cannot fully reconstruct a comprehensible speech from measurements of a single gyroscope. Therefore, we resort to automatic speech recognition. We extract features from the gyroscope measurements using various signal processing methods and train machine learning algorithms for recognition. We achieve about 50% success rate for speaker identification from a set of 10 speakers. We also show that while limiting ourselves to a small vocabulary consisting solely of digit pronunciations ("one", "two", "three", ...) and achieve speech recognition success rate of 65% for the speaker dependent case and up to 26% recognition rate for the speaker independent case. This capability allows an attacker to substantially leak information about numbers spoken over or next to a phone (i.e. credit card numbers, social security numbers and the like).

We also consider the setting of a conference room where two or more people are carrying smartphones or tablets. This setting allows an attacker to gain simultaneous measurements of speech from several gyroscopes.

Source: https://crypto.stanford.edu/gyropho

Unsupervised joke generation from big data

Saša Petrović
School of Informatics
University of Edinburgh
sasa.petrovic@ed.ac.uk

David Matthews
School of Informatics
University of Edinburgh
dave.matthews@ed.ac.uk

Abstract

Humor generation is a very hard problem. It is difficult to say exactly what makes a joke funny, and solving this problem algorithmically is assumed to require deep semantic understanding, as well as cultural and other contextual cues. We depart from previous work that tries to model this knowledge using ad-hoc manually created

Unlike the previous work in humor generation, we do not rely on labeled training data or hand-coded rules, but instead on large quantities of unannotated data. We present a machine learning model that expresses our assumptions about what makes these types of jokes funny and show that by using this fairly simple model and large quantities of data, we are able to generate jokes that are considered funny by human raters in 16% of cases.

The main contribution of this paper is, to the

LAST TIME

- Where to get Data
- What data we get, json format
- How to get data
- Linear Algebra

QUESTIONS?

CLEANING DATA

DATAIST (HILARY MASON & FRIENDS)

- 1. Obtain - pointing and clicking does not scale (APIs, Python, shell scripting)
- 2. Scrub - “Scrubbing data is the least sexy part of the analysis process, but often one that yields the greatest benefits” (Python, sed, awk, grep)
- 3. Explore - look at the data (visualizing, clustering, dimensionality reduction)
- 4. Model - “All models are wrong, but some are useful” / models are built to predict and interpret!
- 5. Interpret - “The purpose of computing is insight, not numbers”

Source: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

DATAIST (HILARY MASON & FRIENDS)

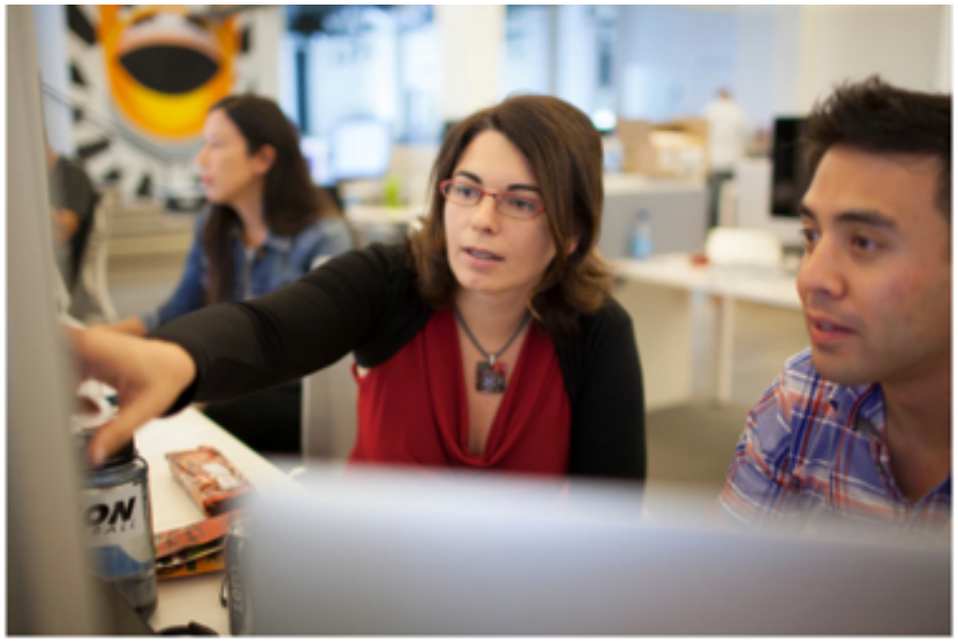
- 1. Obtain - pointing and clicking does not scale (APIs, Python, shell scripting)
- 2. Scrub - “Scrubbing data is the least sexy part of the analysis process, but often one that yields the greatest benefits” (Python, sed, awk, grep)
- 3. Explore - look at the data (visualizing, clustering, dimensionality reduction)
- 4. Model - “All models are wrong, but some are useful” / models are built to predict and interpret!
- 5. Interpret - “The purpose of computing is insight, not numbers”

Source: http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

FOR BIG-DATA SCIENTISTS, 'JANITOR WORK' IS KEY HURDLE TO INSIGHTS

From NYTimes on August 18, 2014:

"Data wrangling is a huge — and surprisingly so — part of the job," said Monica Rogati, vice president for data science at Jawbone, whose sensor-filled wristband and software track activity, sleep and food consumption, and suggest dietary and health tips based on the numbers. "It's something that is not appreciated by data civilians. At times, it feels like everything we do."



http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

DATA MUNGING IS AWESOME

Obtain Data
Scrub Data
Explore
} 80%

Model Algorithms
iNterpret Results
} 20%

Majority of time
is spent data munging

DATA CLEANSING

- Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

- Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

- Remove inconsistencies

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

- Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

- Remove inconsistencies
- Data type harmonization

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

- Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

- Remove inconsistencies
- Data type harmonization
- Standardization, Normalization

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

- Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

- Remove inconsistencies
- Data type harmonization
- Standardization, Normalization
- Typos correction, Formatting (eg. timestamps)

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

- Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

- Remove inconsistencies
- Data type harmonization
- Standardization, Normalization
- Typos correction, Formatting (eg. timestamps)
- Missing data

Source: http://en.wikipedia.org/wiki/Data_cleansing

DATA CLEANSING

- Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

- Remove inconsistencies
- Data type harmonization
- Standardization, Normalization
- Typos correction, Formatting (eg. timestamps)
- Missing data
- Sorting

Source: http://en.wikipedia.org/wiki/Data_cleansing

MISSING DATA

MISSING DATA

- Understand the reasons why data are missing
- Random or not?
- If random, the data sample may still be representative of the population.
- If not random analysis may be harder

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING DATA

- Understand the reasons why data are missing
- Random or not?
- If random, the data sample may still be representative of the population.
- If not random analysis may be harder

- Missing completely at random (MCAR)

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING DATA

- Understand the reasons why data are missing
- Random or not?
- If random, the data sample may still be representative of the population.
- If not random analysis may be harder

- Missing completely at random (MCAR)
- Missing at random (MAR)

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING DATA

- Understand the reasons why data are missing
- Random or not?
- If random, the data sample may still be representative of the population.
- If not random analysis may be harder

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING COMPLETELY AT RANDOM (MCAR)

- Missing value (y) neither depends on x nor y
- Example: some survey questions asked of a simple random sample of original sample

- When data are MCAR, the analyses performed on the data are unbiased; however, data are rarely MCAR.

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING AT RANDOM (MAR)

- Missing value (y) depends on x, but not y
- Example: Respondents in service occupations less likely to report income

Source: http://en.wikipedia.org/wiki/Missing_data

MISSING NOT AT RANDOM (MNAR)

- The probability of a missing value depends on the variable that is missing
- Example: Respondents with high income less likely to report income

Source: http://en.wikipedia.org/wiki/Missing_data

TECHNIQUES TO DEAL WITH MISSING DATA

- Imputation, Partial imputation
- Deletion, Partial deletion
- Analysis
- Interpolation

Source: http://en.wikipedia.org/wiki/Missing_data

TECHNIQUES TO DEAL WITH MISSING DATA

- 1. Identify patterns/reasons for missing and recode
- correctly
- 2. Understand distribution of missing data
- 3. Decide on best method of analysis

Source: https://www.utexas.edu/cola/centers/prc/_files/cs/Missing-Data.pdf

LINKS

- https://www.utexas.edu/cola/centers/prc/_files/cs/Missing-Data.pdf
- http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html
- http://en.wikipedia.org/wiki/Missing_data
- https://www.coursera.org/course/getdata
- Body Level Five

WALK THE WALK OF CLEANING DATA

DATA MUNGING TOOLS AND OPERATIONS

- python, pandas
- sed, awk, bash, perl
- regular expressions
- text editor
- etc. etc. etc.

DATA MUNGING TOOLS AND OPERATIONS

- python, pandas
- sed, awk, bash, perl
- regular expressions
- text editor
- etc. etc. etc.

IN GROUPS

- Choose one tool from the list
- investigate functionality
- find one example
- show use to class

LINKS

- https://www.utexas.edu/cola/centers/prc/_files/cs/Missing-Data.pdf
- http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html
- http://en.wikipedia.org/wiki/Missing_data
- https://www.coursera.org/course/getdata
- Body Level Five