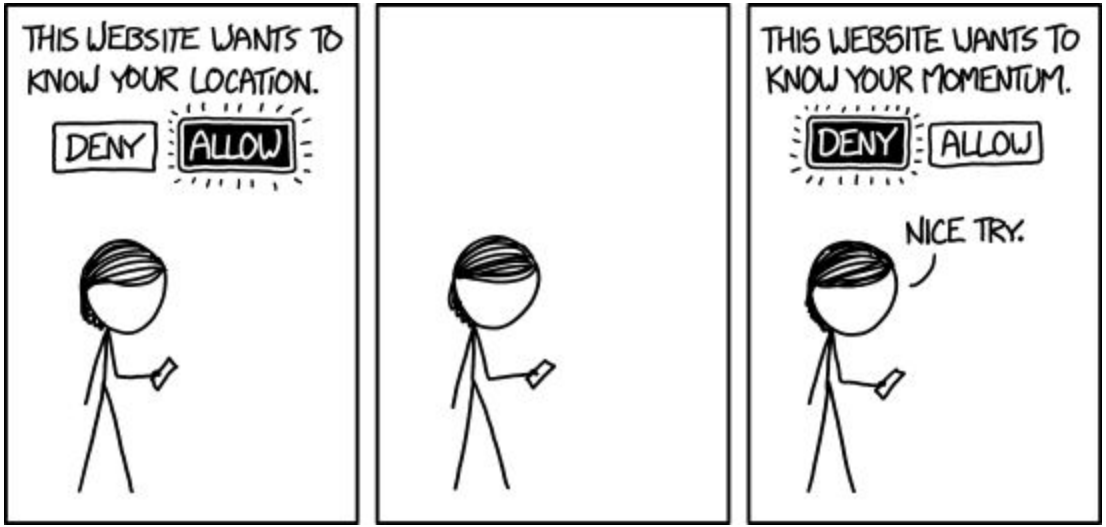
Sensor Data Streams

Stephen Voida , Donald J. Patterson , and Shwetak N. Patel

Presented by Aravind Sagar

Overview

- Introduction
- Questions it can answer and units of analysis
- Data captured by sensors
- How to do it?
- Limitations



https://xkcd.com/1473/

Why Use Sensors?

- Detailed low level data
- Samples as frequent as we need
- Automated data collection
- With or without intervention
- No recall/recollection limitations

Questions that can be Answered

Sample Questions

- Where do people travel over the course of a day?
- With whom do they normally communicate or collaborate?
- What tools or information resources do they use at various points during the day? When, where, and with whom?
- What routines help to define a “typical” or “atypical” day?
- How healthy are a person’s daily behaviors? Is he or she making good health choices?

More about what can sensors answer

- Assess current behavior, as well as change over time
- Fine-grained detailed events
- Coupling of action to specific sensor
- Determine high-level behaviour such as activity and routing
- Triangulation with multiple sensors

Units of Analysis

Egocentric

Group-centric

Space-centric

Egocentric Sensor Data Streams

"Sensors focused on monitoring the movements, activities, and interactions of a **single individual** can answer questions at an egocentric unit of analysis."

Data collected:

- From sensors on smartphones, wearables and other technology
- Movements, location or interaction
- Activity in a private or semi-private space

Group-centric Sensor Data Streams

"Group-centric approach can involve simply capturing the same signals as for a single person, but across a **group over the same window of time**, or it might involve deploying a **broader set of environmental or infrastructural sensors** in a shared/community space or collecting data about more **interpersonal types of interactions**"

Space-centric Sensor Data Streams

"... to answer questions about how spaces are used, irrespective of their particular inhabitants"

- Often in public spaces
- Sensors include cameras, microphones, pressure sensors, motion detectors
- Special networking support may be required
- Focus on maximizing area of coverage
- "Infrastructure Mediated Sensing"

Data Captured by Sensors

Sensors in the Physical World

- Virtually limitless number of available sensors
- New ones developed all the time
- Stream of time-stamped fine-grained data
- Examples: electrical switches, motion sensors, pressure sensors, voltmeters, photometers, thermometers, moisture sensors, proximity sensors

“Virtual” sensors

- Some examples
  - Log files
  - Non-automated human observations - for example a human analyzing a video stream
- Quantity typically lower
- Richer data

Enriching collected data

- Combine multiple sensor streams including physical and virtual sensors
- Augment with other data collection techniques
    - For example, reduce the burden of recall during an interview

How to Do It?

Issues related to sensor data streams

- Generating the research questions and planning how to analyze the data streams
- Building, acquiring, or provisioning the sensors
- Determining how frequently and at what level of fidelity to collect data samples
- Installing the sensors
- Storing the collected data
- Making sense of the collected corpus of data

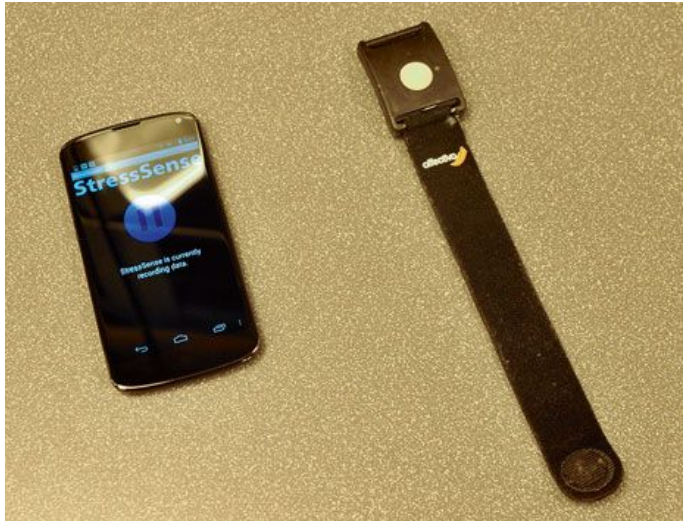
Generating the Research Questions and Planning How to Analyze the Data Streams

Privacy and ethical concerns

- Capturing data without knowledge of participants or passers by
- Ability of participants to revoke consent
- Ability to suspend data collection/delete collected data

Building, Acquiring, or Provisioning the Sensors

- A function of cost, availability, technological capability, intrusiveness, or methodological needs
- Instrumenting a space vs instrumenting a participant
- Effectiveness of sensor in different environments
  - For instance, GPS is ineffective indoors
- Intrusiveness of the sensors
  - Can affect ecological validity of the study



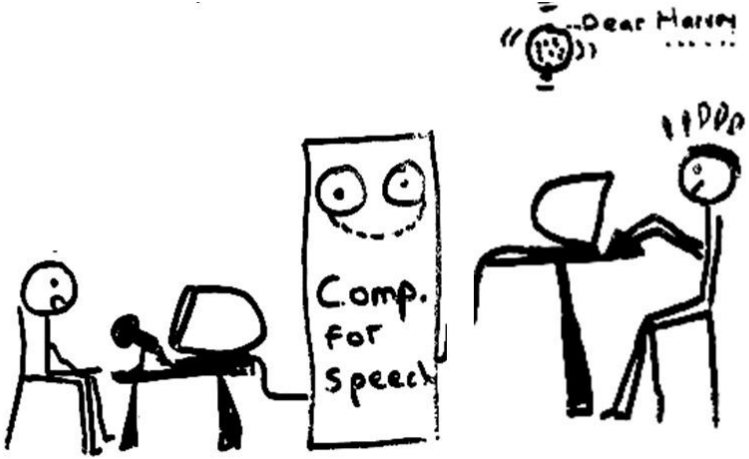
Poh, M.-Z., Swenson, N. C., & Picard, R. W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity.

Determining How Frequently and at What Level of Fidelity to Collect Data Samples

- Balance between sampling rate and storage/bandwidth, processing, and power requirements
  - Consider case of mobile phones collecting data
- Privacy improvements by blurring data
- Still privacy concerns exist:
  - Sensors *can* collect higher fidelity data
  - Choice of interpretation of data impacts perception
- One method used is to vary the fidelity of data collection based on input from other sensors
  - For example, record location only when accelerometer readings indicate participant motion

Installing the Sensors

- Coverage of large spaces
  - High labor costs of installation
  - Ongoing sensor network management
  - Power source - additional maintenance
- Wizard-of-Oz to mitigate deployment risk



What the user sees

The Wizard

Storing the Collected Data

Two broad options

- Device connected to (or nearby) the sensor data source
  - Collect a copy at checkpoints or end of study
  - Unsuitable for large volumes of data
- Continually transmit collected data back to a central server
  - Power and connectivity needs to be considered

Hybrid approach when using a large number of sensors

- Aggregating data from multiple sources into a single “stream” at the time of collection
- Reduces transmission and sync issues
- Simplifies further analysis

Making Sense of the Collected Corpus of Data

- Quantitative analysis should into account noisy sensor sources, ambiguities in sensed data, or technical problems with the sensors
- Usually using standard statistical software packages
- Example, comparing frequency or distribution of sensed events under different conditions
- Machine-learning techniques to classify sensor data into higher-level representations of participants' behaviors or to perform hypothesis testing
    - Helps with noise in sensor data

What to Report in a Study Using Sensors

Sufficient detail to reproduce the study, along with reasons

- **Hardware**: Type and quantity of hardware, configuration
- **Experimental setup**: Placement of sensors, installation details, environment details
- **Participant knowledge**: About the hardware and data collection
- **Experimental execution**: Anything during the experiment like maintenance and outages
- **Software infrastructure**: Used to collect and transmit the data
- **Analysis**: Details of analysis on data, software used, parameters of the algorithms

Limitations

- Privacy and ethics concerns are sometimes hard to avoid, or they slip under the radar
- Sensor data generally does a poor job of answering questions of **why**
  - Provide little insight about the intention behind the actions, or mental states of a participant
  - Data fusion/triangulation can help
- Phenomena to be measured or observed must be well understood at the onset of the study
- Sensors may have limitations in the quality of data they can collect
  - Cost and intrusiveness can play a part here
- Large streams of data can get challenging
- Additional technical complexity compared to other methods

Thank You!