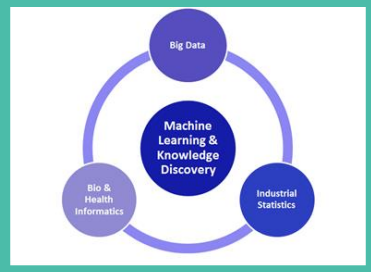
521156S TOWARDS DATA MINING

MATKALLA TIEDONLOUHINTAAN

UNIVERSITY OF OULU

BISG Oulu



Data Analysis and Inference Group

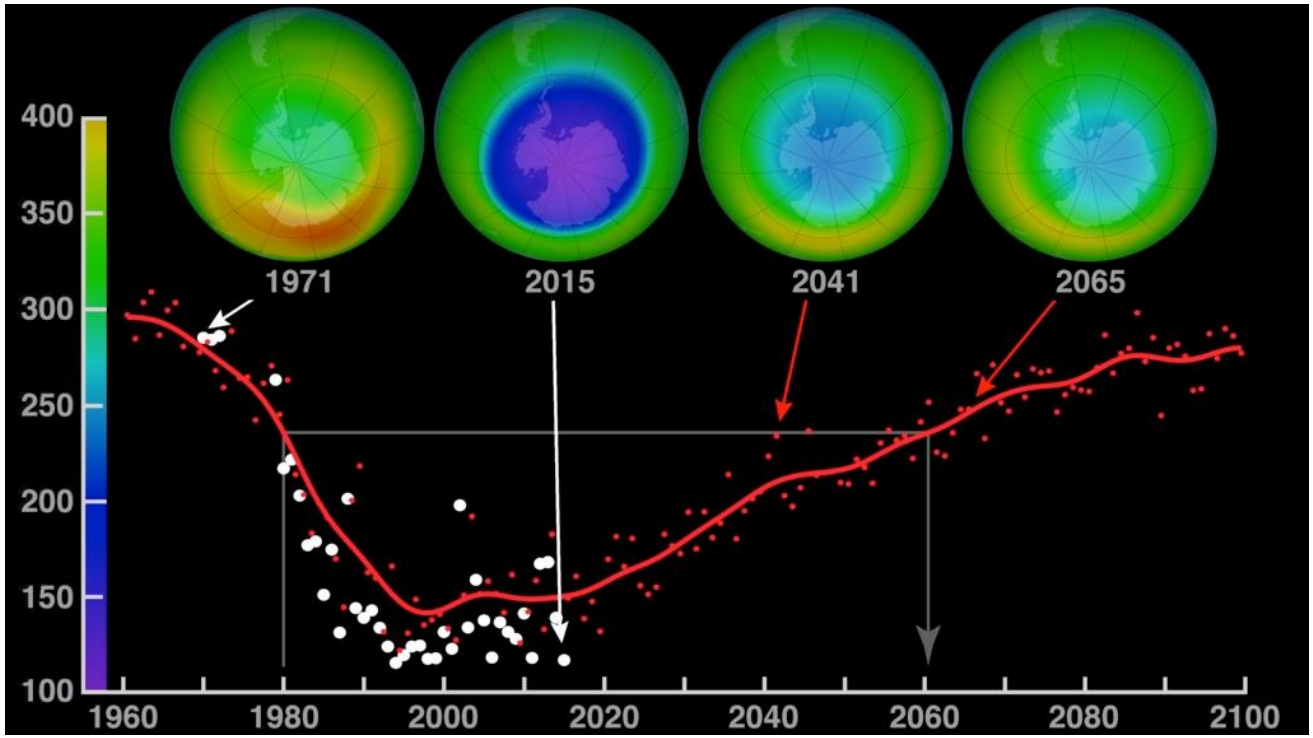
Outline

- Motivation: Data collection gone bad
- Planning, planning, planning
- Differences when collection data from stationary or non-stationary elements
- Differences when collection data from humans
- Data collection example
- Reporting the problems in data collection

Data collection gone bad

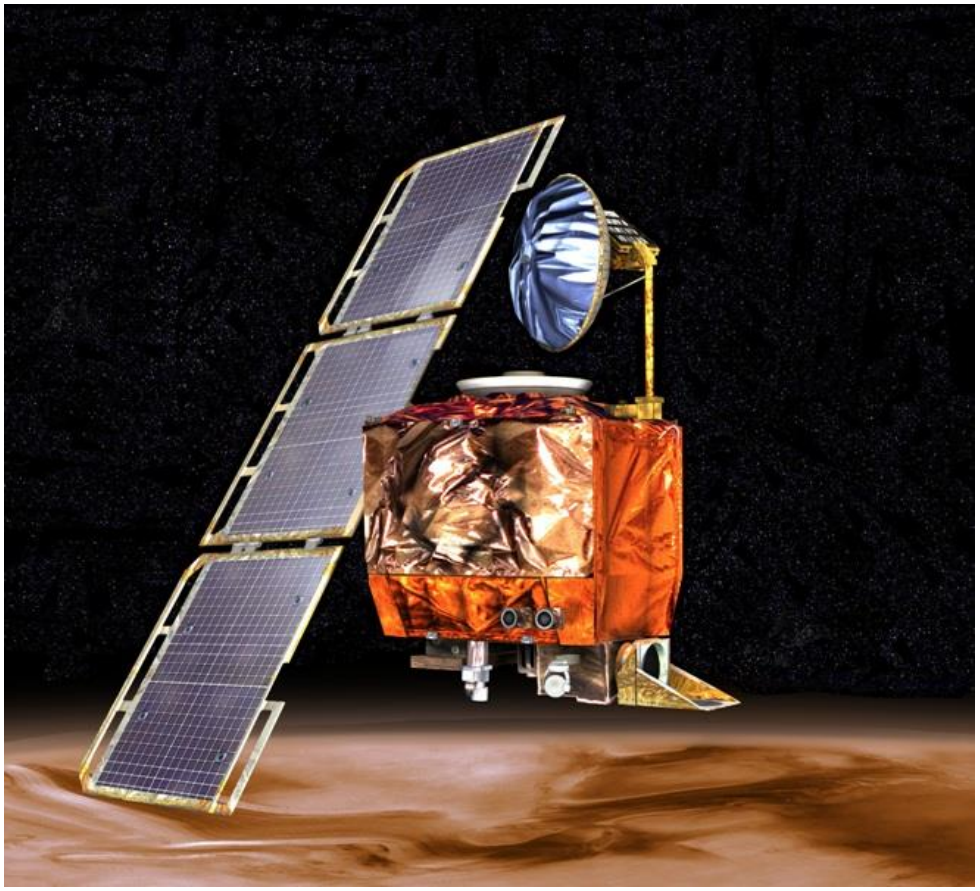
A large shock was needed to motivate the world to get serious about phasing out CFC's and that shock came in a 1985 field study by Farman, Gardinar and Shanklin. Published in Nature, May 1985, the study summarized data that had been collected by the British Antarctic Survey showing the ozone levels had dropped to 10% below normal January levels for Antarctica. The authors has been somewhat hesitant about publishing because Nimbus-7 satellite data has shown no such drop during the Antarctic spring.

But NASA soon discovered that the spring-time "ozone hole" had been covered up by a computer-program designed to discard sudden, large drops in ozone concentrations as "errors". The Nimbus-7 data were rerun without the filter-program and evidence of the Ozone-hole was seen as far back as 1976.



Chlorofluorocarbons (CFCs) and other halogenated ozone depleting substances (ODS)

Nov. 10, 1999: Metric Math Mistake Muffed Mars Meteorology Mission



1999: A disaster investigation board reports that NASA's $125 million Mars Climate Orbiter burned up in the Martian atmosphere because engineers failed to convert units from English to metric.

Europe's Mars Schiaparelli lander crashed due to software glitch

A series of errors resulted in the parachute being released too early

by Andrew Liptak | @AndrewLiptak | May 27, 2017, 12:42pm EDT

SHARE TWEET LINKEDIN

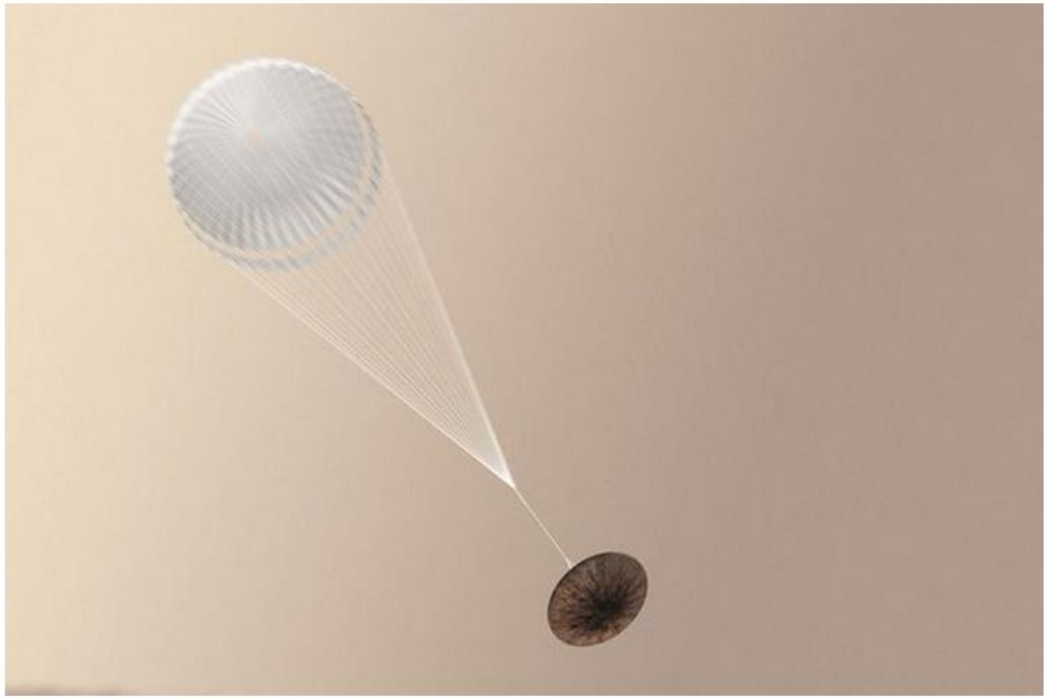
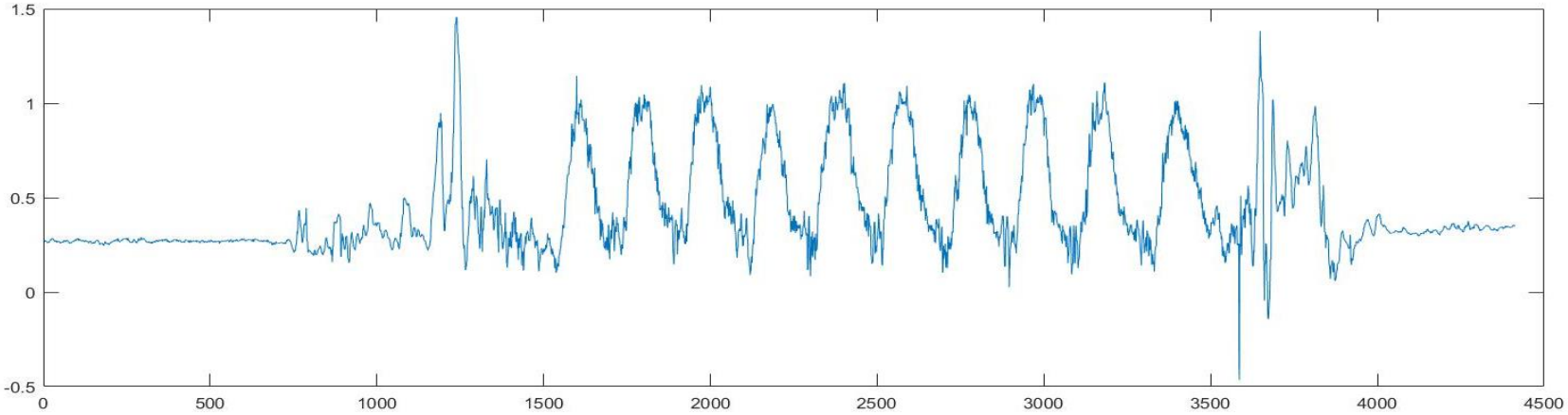
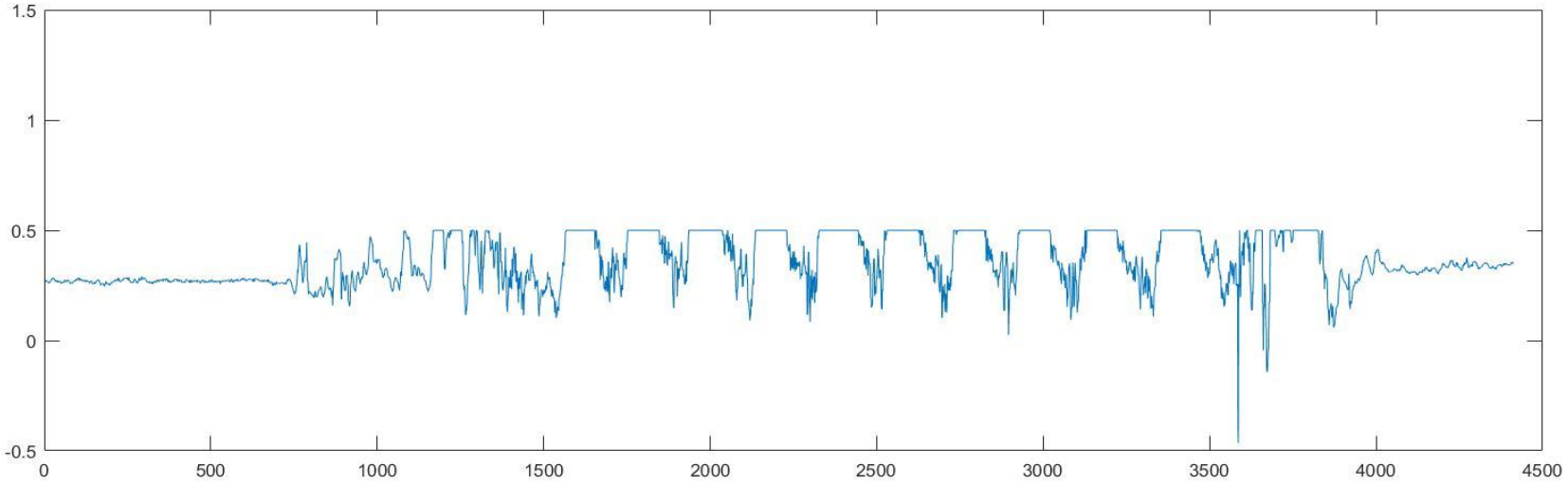


Image: ESA

- **Inertial Measurement Unit angular pitch rate exceeded the systems limit a.k.a. signal saturated**
- **Lander's systems thought the spacecraft was closer to the ground, and released the capsule's back-shell and parachute too early**

Signal saturation





Own data collection problems

- Migraine study: Devices gone missing, rash problems
- IMU studies: Bluetooth connectivity, human blocking the signal
- IMU studies: sensor attachment problems
- Battery running out
- People don’t read the instructions
- With mobile phones the data collection stops when screen is turned off

Know your data

- You have to understand basics of data collection to be critical enough to the trustworthiness of data

Planning, planning, planning

Data collection: planning

- **Define the problem**
  - If the data does not answer to the question there is nothing to do but collect more data
    - Big data problems
  - How do you collect your data
    - Plan the procedure

Data collection: planning

– Define the problem, goals and objectives

- description of the project
- data types, data scales and amounts including time intervals
- methodologies and devices used for data recording
- how to deliver and store the data
- Rationale for collecting the data
- What insight the data might provide
- What will be done with the data later on
- DOCUMENTATION!!!

- Ensuring Repeatability, Reproducibility, Accuracy and Stability

Data collection: procedure

- Plan
- First data collection, one subject, one device
- Plan and iterate
- Final data collection
- Reporting the problems in data collection

Data collection: sample size

- **Defining minimum sample size for statistical significance**
- **What information do you need**
  - Original population size
  - Margin of error (MOE, a.k.a. confidence interval) e.g +/- 5 %
  - Z-score of confidence interval e.g. 90 %, 95 % (http://www.sjsu.edu/faculty/gerstman/StatPrimer/z-two-tails.pdf)
  - Standard of deviation, σ

⇒ https://www.surveysystem.com/sscalc.htm

⇒ μ is $n \geq \left(\frac{z^{*}\sigma}{\mathrm{MOE}}\right)^{2}$;

⇒ $$n = \frac{X^2 * N * P * (1-P)}{(ME^2 * (N-1)) + (X^2 * P * (1-P))}$$

Where:

n = sample size

X^2 = Chi-square for the specified confidence level at 1 degree of freedom

N = Population Size

P = population proportion (.50 in this table)

ME = desired Margin of Error (expressed as a proportion)

Data collection: sample size

- **Drop out rate**
- **Data collection problems**
  - Explain what problems you have found out, this can for instance be a part of ReadMe-file if you publish your dataset as open data

Data collection: time intervals

- How often you need measurement
- From how long scale you need for measurements

Data collection: planning

– Methods of data collection

- Background information collection
- Questionnaires
- Interviews
- Direct observations
- Sensor observations
- Reference information e.g. video

Data collection: planning

- Keep it simple
- Discuss the collection, learn from mistake of others
- Schedule the collection
- Iterate, learn from the data
- Do not be afraid to stop data collection

Differences when collection data from stationary or non-stationary elements

Stationary and non-stationary elements

- **Stationary elements stay similar during time**
- **Non-stationary elements change during time**
  - Environmental objects, living objects
    - Due humans, technology development, pollution, etc
    - Seasonal changes
  - Processes due erosion of parts, drift
  - Humans are non-stationary

- **What is the main difference in data collection?**

Time series data collection

- Best to collect data from distinct time windows (minutes, days, weeks)
- Helps to avoid using “the same data” in modelling and testing

HUMANS

Collecting data from humans

- Every human is different
- Humans change/evolve in during time

Collecting data from humans – ethical aspects

- In US even to be allowed to use some open data sets a basics of human ethics is required
- Web courses are provided for that e.g. https://phrp.nihtraining.com/

Collecting data from humans – ethical aspects

– Why?

– The Syphilis Study at Tuskegee

- A long-term study of black males conducted by the United States Public Health Service in Tuskegee, Alabama initiated in the 1930s and continued until 1972.
- 600 African-American men: about 400 with syphilis (cases) and about 200 without syphilis (controls). These men were led to believe that some of the procedures done in the interest of research (e.g., spinal taps) were actually “special free treatment.”
- By 1946 reports indicated that the death rate among those with syphilis was about twice as high as it was among the controls.
- In the 1940s, penicillin was found to be effective in the treatment of syphilis. The Syphilis Study at Tuskegee continued, however, and the men were neither informed about nor treated with the antibiotic.

Collecting data from humans – ethical aspects

– three principles essential to the ethical conduct of research with humans:

- Respect for persons
  - Individuals should be treated as autonomous agents
  - Persons with diminished autonomy are entitled to additional protections
    - Pregnant Women, Human Fetuses and Neonates
    - Prisoners
    - Children
- Beneficence
  - Do no harm
  - Maximize possible benefits and minimize possible harms
- Justice
  - individuals and groups be treated fairly and equitably in terms of bearing the burdens and receiving the benefits of research

Collecting data from humans – ethical aspects

- **Informed consent:**
- **Potential study participants must:**
  - Give their consent freely and voluntarily
  - Have the decisional capacity to understand the information presented to them
  - Be provided complete information about the study in order to make an informed decision

Collecting data from humans – ethical aspects

– Risks

- Physical
- Psychological
- Social
- Legal
- Economic

– Benefit should be greater than the risk

– Compensation should not be too attractive and blind the risks

– Privacy and Confidentiality

- *Privacy* means being “free from unsanctioned intrusion.”
- *Confidentiality* means holding secret all information relating to an individual, unless the individual gives consent permitting disclosure.

Randomized controlled trial (RCT)

- **a type of scientific (often medical) experiment which aims to reduce bias when testing a new treatment**
- **Participants are randomly allocated to either the group receiving the treatment or to a control group receiving standard treatment (or placebo treatment)**
- **The major categories of RCT study designs are:[32]**

- Parallel-group – each participant is randomly assigned to a group, and all the participants in the group receive (or do not receive) an intervention.
- Crossover – over time, each participant receives (or does not receive) an intervention in a random sequence.
- Cluster – pre-existing groups of participants (e.g., villages, schools) are randomly selected to receive (or not receive) an intervention.
- Factorial – each participant is randomly assigned to a group that receives a particular combination of interventions or non-interventions (e.g., group 1 receives vitamin X and vitamin Y, group 2 receives vitamin X and placebo Y, group 3 receives placebo X and vitamin Y, and group 4 receives placebo X and placebo Y).

BIAS

- **RCT is considered a golden standard to avoid bias in medical treatment tests but …**
- **Researcher bias**

- Giving own opinion through questions
- Don't want to find negative things about the research financier

- **Respondent bias**

- survey of who you will vote?
- Salary surveys?
- Test of new technology with volunteers?
- Human perspective of what they consider big, small e.g. how much they sit daily, how big portions they eat?

- **Human data bias**

- Running speed estimation model based on heart rate?
  - 1 person data set
  - All athletes data set
  - All men data set

Data collection example

An example:

- **Collecting gym data using Myo (wearable sensor worn in arm)**

Description of the project

To recognize automatically and person independently gym activities and make a activity diary using wearable sensor collected data

Data types, data scales and amounts including time intervals

- Streaming data, 50Hz, 3D acceleration, 3D angular velocity, electromyogram of 8 channels
- 10 individuals
- 30 gym exercises (defined before hand, do the exercise order matter?), 1 set of 10 repetitions of every exercise
- Normal movements between exercises, changing locations, changing weight, stretching
- User-defined weights

Methodologies and devices used for data recording

- Myo device
- Define the location: right arm, icon upwards
- Sends data using Bluetooth to laptop
- To spot Bluetooth problems run data on screen real time
- Paper of needed exercises, their instructions and their order
- Informed consent

An example:

- **Collecting gym data using Myo (wearable sensor worn in arm)**

How to deliver and store the data

- makes a csv file
- Store it on network drive on University of Oulu (automatically backups the data)
- Label the data manually by plotting the signals with Matlab (due the periodic movements can be seen from the signals)

Rationale for collecting the data and what insight the data might provide, what will be done with the data later on

Versatile enough to be used in research. Activity diary, unseen activity detection, novel sensor combination, methodology development.
Will be made open for other researchers

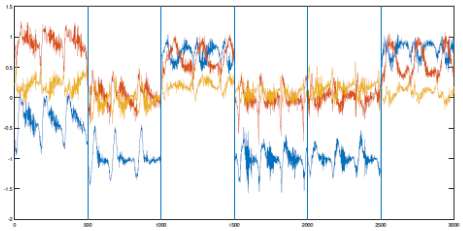
An example:

- **Collecting gym data using Myo (wearable sensor worn in arm)**

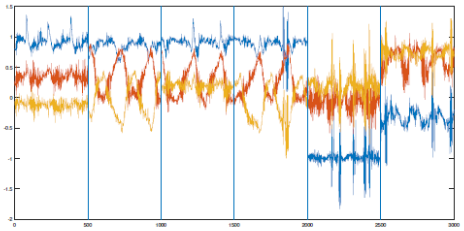
DOCUMENTATION

Example of data (acceleration signals of 30 activities (a, b, c, d, e) + null data (f)

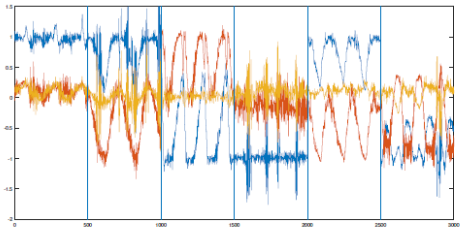
(a)



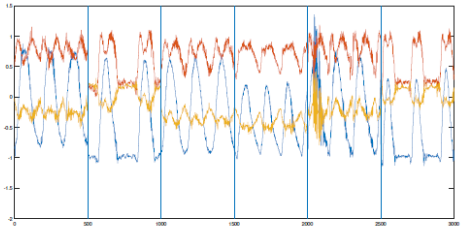
(b)



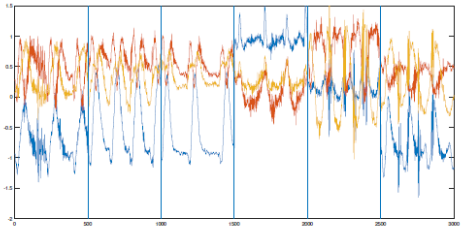
(c)



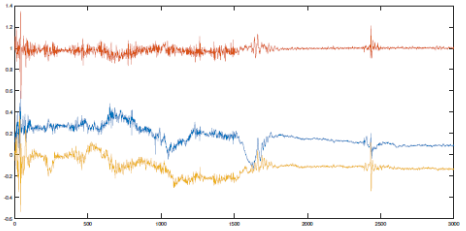
(d)



(e)



(f)



An example:

- Collecting gym data using Myo (wearable sensor worn in arm)

DOCUMENTATION

http://www.oulu.fi/bisg/node/40364

Data collection: lessons learned

- Volunteer recruiting challenging
- Bluetooth connectivity issues
- Change of exercise order due the other people in the gym
- No video, so you have to be “available” all time
- Some exercises selected were novel so instruction were needed
- Volunteer with physical problems like back problem
- Some volunteers did not took easily instructions

Data collection: planning

- **Methods of data collection in previous example?**
  - Background information collection
  - Questionnaires
  - Interviews
  - Direct observations
  - Sensor observations
  - Reference information e.g. video

Data collection: planning

– How to improve the data collection

Reporting the problems

Data collection: problems

- There will always be some problems
- Be open with those problems
- Analyze if the problems will bias your study
- Instructions to others

Data collection: problems

– **Reality always amazes me**

- Summer versus winter in steel mill
- How many swimming styles there are
- What human do when nobody is watching

For exercise

Exercise preparation

– Upload an app for raw acceleration data collection

For example:

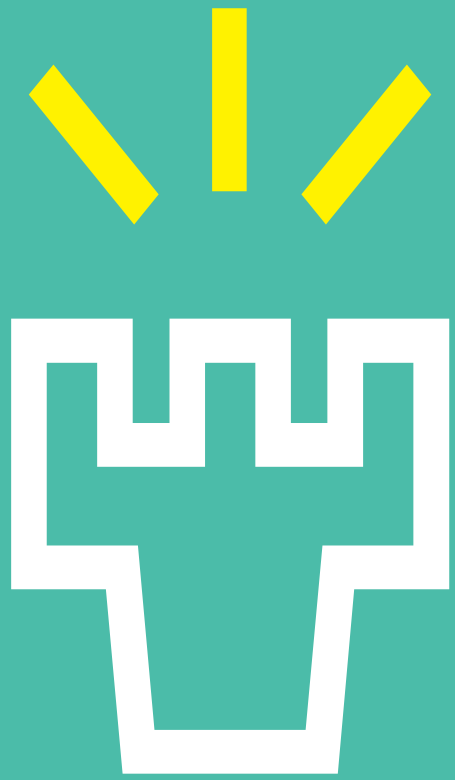
Android: Apps Medion Accelerometer

Ios: Mobile Science - Acceleration

Summary

Summary

- Planning, planning, planning
- Know your data
- Define what you want to do with the data
- Be extra cautious when dealing with humans



UNIVERSITY
OF OULU