LAB WEEK 2 – PREPROCESSING

DATA MINING SPRING 2013 | ANDERS HARTZEN (ANDERSHH@ITU.DK) & JENS ANDERSSON GRØN (JANG@ITU.DK)

WELCOME TO THE DATA MINING LABS!

HELLO FROM YOUR TWO TEACHING ASSISTANTS

Anders Hartzen

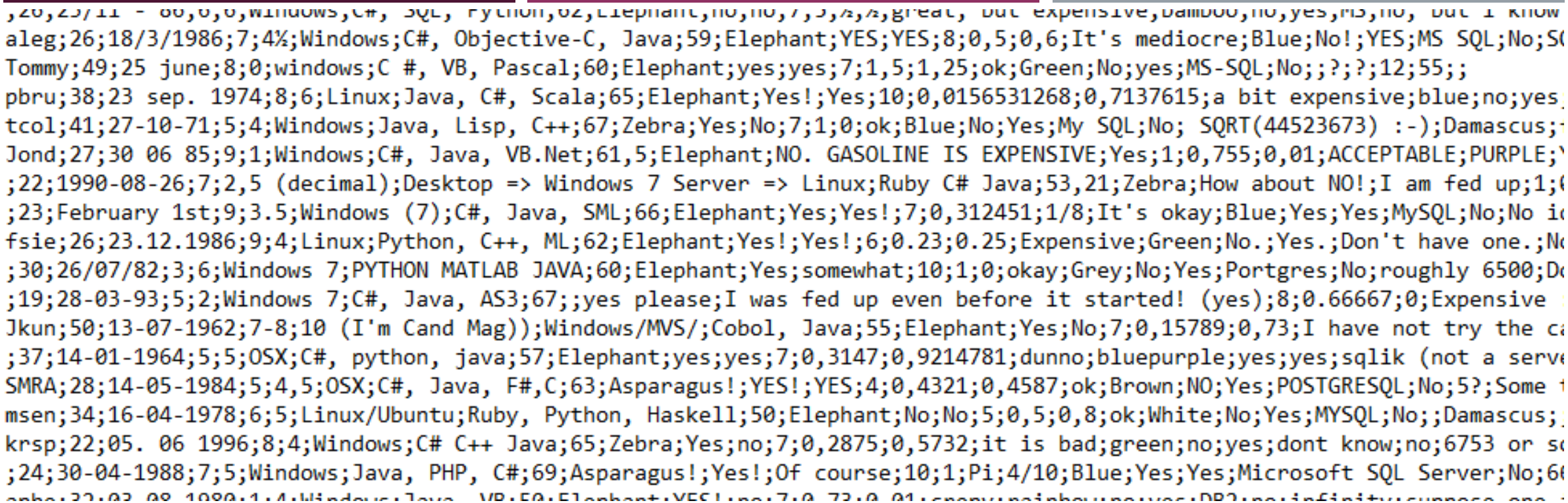
- B.Sc. Software Development from ITU, 2010
- M.Sc. Media Technology and Games, Technology from ITU, 2012
- Working as Teaching Assistant at ITU since then
- Trying to get Ph.D. project funded
- Email: andershh@itu.dk

Jens Andersson Grøn

- B.Sc. Software Development from Copenhagen Business, Lyngby, 2011.
- Currently studying for a M.Sc. in Software Development and Software Engineering at ITU.
- Email: jang@itu.dk

DATA MINING LABS

- Most exercises during the course will focus on you implementing algorithms in a programming language of your choice.
  - We recommend Java, as you will be able to use code frameworks we will provide you in the different labs.
- We will be here to help you and Julian will usually be around as well to help you if needed.
- Doing the labs will help you do the groundwork for the individual mandatory assignment you have to hand in on Friday March 22.
- Email us (andershh@itu.dk and jang@itu.dk) anytime with your questions and/or feedback
  - Send emails to both of us.
  - Will reply ASAP.



TODAY'S LAB

Preprocess this!

PREPROCESSING – AKA CLEANING DATA

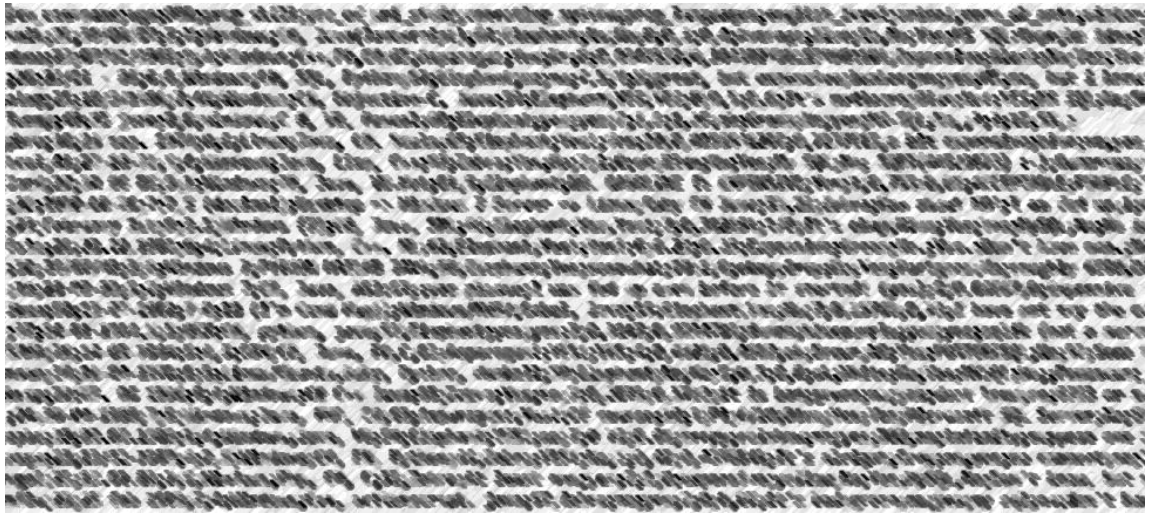
- Today you will be working with data from an ITU questionnaire from last years course.
- You will be cleaning up the data by using pre-processing techniques you get to implement yourself.
- The data is the literal transcription of the freetext (i.e. comments) participants wrote. No assumptions or corrections were made.
- Therefore the data needs heavy cleaning and preprocessing to be more useful for further experiments.

OVERVIEW OF TODAY'S LAB

- Part 1 – Load the data set using code.
- Part 2 – Clean the data set using code.
- Part 3 – Normalize attributes using code.
- Part 4 – Use descriptive statistical methods to describe the data set using code.

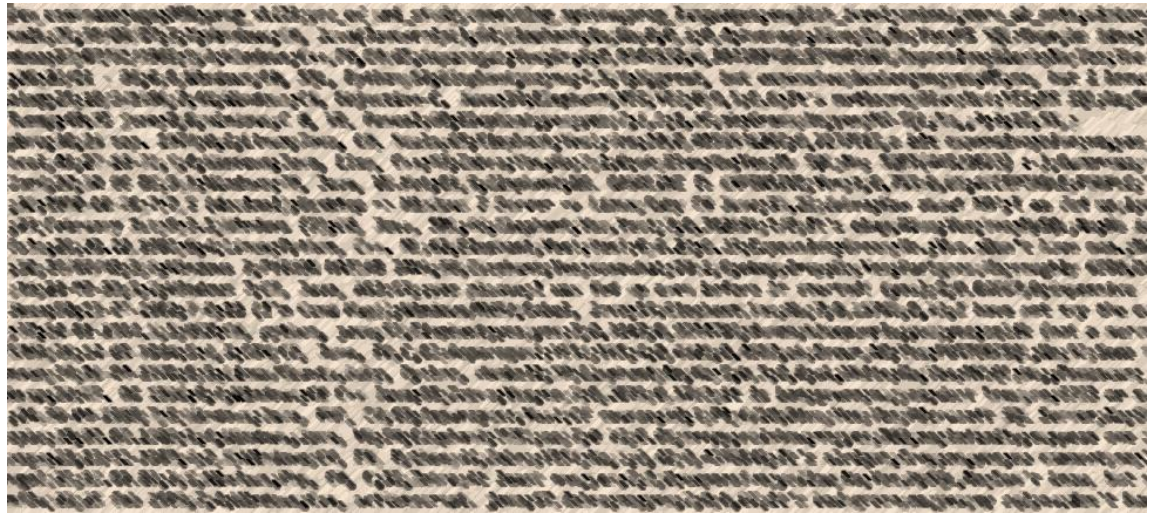
LAB PART 1 - LOAD IN DATA

- Java code is available from the course page on learnIT to help you get started loading in the data.
  - Is pretty basic, but works.
- Feel free to write your own code and/or use another programming language.
  - Though we are most able to help with Java/C# questions.



LAB PART 2 – CLEAN DATA

- Using code!
- Issues worth considering:
  - Missing values?
  - Different formats?
  - Noise? Outliers?
  - Data transformation?



LAB PART 3 – NORMALIZATION

- In your own code normalize the numerical values
  - Min-max
  - Z-score
  - Decimal scaling

```
itu_username;age;DOB;prog_skill;yrs_of_uni_study;operating_system;prog_languages;english_level;animal;more_dk_mnts;winter_
seik;25;30/08/87;7;½;Windows;C#, Java(Javascript);64;Elephant;No;No;8;0,3;0,4;Average;Blue;No,No;Yes;MS SQL;Nope;No idea;
mbma;23;07/01/1989;9;4;Windows;C++,F#,C#;65;Asparagus;Yes;No;8;0,0347532;0,1043027301879981101;Expensive;Black;No;Yes;The
;26;23/11 - 86;6;6;Windows;C#, SQL, Python;62;Elephant;no;no;7,5;½;½;great, but expensive;bamboo;no;yes;MS;no, but i know
aleg;26;18/3/1986;7;4½;Windows;C#, Objective-C, Java;59;Elephant;YES;YES;8;0,5;0,6;It's mediocre;Blue;No!;YES;MS SQL;No;S
Tommy;49;25 june;8;0;windows;C #, VB, Pascal;60;Elephant;yes;yes;7;1,5;1,25;ok;Green;No;yes;MS-SQL;No;;?;?;12;55;;
pbru;38;23 sep. 1974;8;6;Linux;Java, C#, Scala;65;Elephant;Yes!;Yes;10;0,0156531268;0,7137615;a bit expensive;blue;no;yes
tcol;41;27-10-71;5;4;Windows;Java, Lisp, C++;67;Zebra;Yes;No;7;1;0;ok;Blue;No;Yes;My SQL;No; SQRT(44523673) :-);Damascus;
Jond;27;30 06 85;9;1;Windows;C#, Java, VB.Net;61,5;Elephant;NO. GASOLINE IS EXPENSIVE;Yes;1;0,755;0,01;ACCEPTABLE;PURPLE;
;22;1990-08-26;7;2,5 (decimal);Desktop => Windows 7 Server => Linux;Ruby C# Java;53,21;Zebra;How about NO!;I am fed up;1;
;23;February 1st;9;3.5;Windows (7);C#, Java, SML;66;Elephant;Yes;Yes!;7;0,312451;1/8;It's okay;Blue;Yes;Yes;MySQL;No;No i
fsie;26;23.12.1986;9;4;Linux;Python, C++, ML;62;Elephant;Yes!;Yes!;6;0.23;0.25;Expensive;Green;No.;Yes.;Don't have one.;N
;30;26/07/82;3;6;Windows 7;PYTHON MATLAB JAVA;60;Elephant;Yes;somewhat;10;1;0;okay;Grey;No;Yes;Portgres;No;roughly 6500;D
;19;28-03-93;5;2;Windows 7;C#, Java, AS3;67;;yes please;I was fed up even before it started! (yes);8;0.66667;0;Expensive
Jkun;50;13-07-1962;7-8;10 (I'm Cand Mag));Windows/MVS/;Cobol, Java;55;Elephant;Yes;No;7;0,15789;0,73;I have not try the c
;37;14-01-1964;5;5;OSX;C#, python, java;57;Elephant;yes;yes;7;0,3147;0,9214781;dunno;bluepurple;yes;yes;sqlik (not a serv
SMRA;28;14-05-1984;5;4,5;OSX;C#, Java, F#,C;63;Asparagus!;YES!;YES;4;0,4321;0,4587;ok;Brown;NO;Yes;POSTGRESQL;No;5?;Some
msen;34;16-04-1978;6;5;Linux/Ubuntu;Ruby, Python, Haskell;50;Elephant;No;No;5;0,5;0,8;ok;White;No;Yes;MYSQL;No;;Damascus;
krsp;22;05. 06 1996;8;4;Windows;C# C++ Java;65;Zebra;Yes;no;7;0,2875;0,5732;it is bad;green;no;yes;dont know;no;6753 or s
;24;30-04-1988;7;5;Windows;Java, PHP, C#;69;Asparagus!;Yes!;Of course;10;1;Pi;4/10;Blue;Yes;Yes;Microsoft SQL Server;No;6
apbe;32;03.08.1980;1;4;Windows;Java, VB;50;Elephant;YES!;no;7;0.73;0.01;crepy;rainbow;no;yes;DB2;no;infinity;suppose one
asgs;23;13/12-1989;4;2,5;Windows;C#, F#, Python;65;Zebra;YES!;No;1;0,42;0,4376842;It's alright;Grey;no;yes;no comment;no;
vgol;25;13.02.87;7;5;Windows;Php, C#, C;55;Zebra;At least one;yes;7;1;0;expensive;yellow;No;Yes;MySQL;No;Don't know, but
mfio;23;25 feb 1989;7;4;Windows;C++, actionscript, java;57;Zebra;no;yes;7;0;0,07;it's bad;green;no;yes;my sql;no;don't kn
```

LAB PART 4 – DESCRIPTIVE STATISTICS

- In your own code try to describe the data using descriptive statistics.
- Central tendency of the data
  - Mean
  - Median
  - Mode
  - Etc. (See pg. 45 in book for overview)
- Dispersion of the data
  - Standard deviation
  - Five-number summary
    - Min
    - Quartiles
    - Median
    - Max
  - Etc. (see pg. 48 in book)

HIDDEN TRUTHS? LARGE DATASETS.

At the end of the lab think about the following:

- Were there any meaningful correlations between parts of the data?
- Are there other methods I could have used to detect possible correlations?
- Would my preprocessing code work well if applied to a very large dataset?
  - Any changes I would make in my code?
  - Any new constructs I could add to my code to make it work with a large dataset?
    - Mixed-initiative via a GUI?
    - Output troublesome data via File-output?
    - ???

LET'S GO!

THANK YOU FOR LISTENING!