VISUAL DATA ANALYSIS

LECTURE 1
INTRODUCTION

UNIVERSITY OF SKÖVDE

AGENDA

- Teachers involved in the course

- What is visual data analysis and why is it needed?

- Course contents: teaching goals and examinations
  - lectures
  - seminars
  - project work + supervisions

- Canvas page

- Schedule

TOVE HELLDIN

Tove Helldin
Lecturer and course coordinator
http://his.se/helo/

tove.helldin@his.se
+46 (0) 500 448383
Signature: helo
Room: PA520M (Portalen, 5th floor)

PhD in computer science, Örebro University, Sweden
Main research areas: human-computer interaction, usability, explainable AI, visual analytics

JUHEE BAE

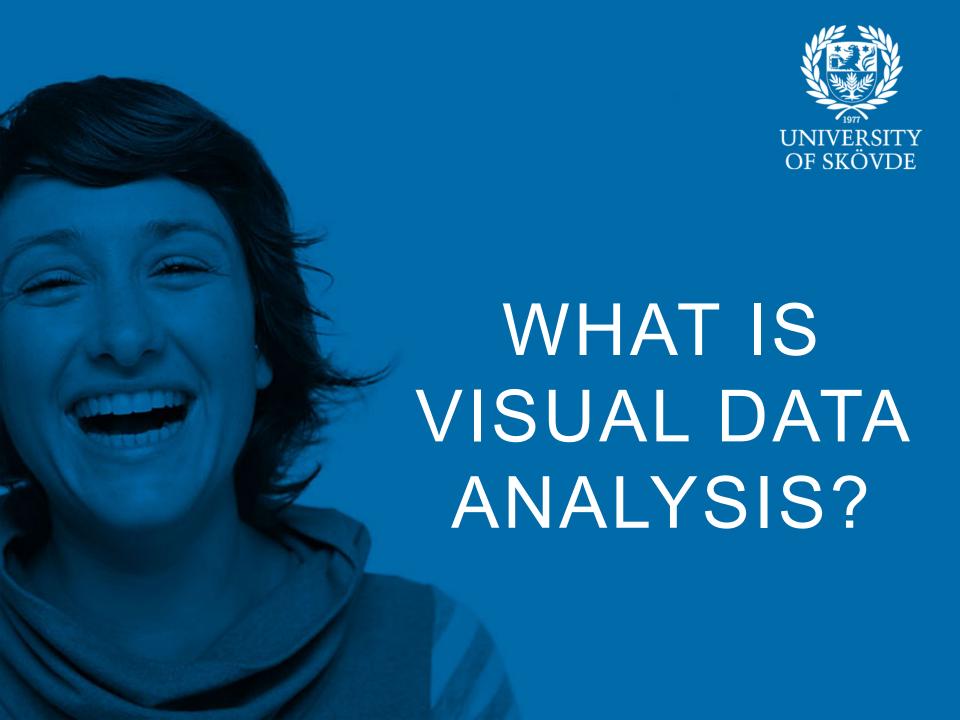Lecturer and course examiner
http://his.se/baej/

juhee.bae@his.se
Phone: +46(0) 500 448910
Signature: baej
Room: PA520J (Portalen, 5th floor)

Ph.D. in computer science from North Carolina State University
Main research areas: data visualization, human-computer interaction, visual analytics, machine learning

WHAT IS VISUAL DATA ANALYSIS?

WHY VISUAL DATA ANALYSIS?

The information age
- Increasing amount of data to be dealt with
- Steady improvement of data storage devices
- Easier and cheaper to collect and store data
- Data is often stored without filtering and
   refinement for later use

The possibilities to collect and store data increase at a faster rate than our ability to use it for making decisions

However, in most applications, raw data has no value in itself, instead we want to extract the information contained in it
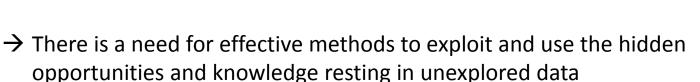
WHY VISUAL DATA ANALYSIS?

The information overload problem
We are often confronted with large amounts of disparate, conflicting and dynamic information, collected from heterogeneous sources

The information overload problem refers to the danger of getting lost in the data which may be
- Irrelevant to the current task at hand
- Processed in an inappropriate way
- Presented in an inappropriate way

→ There is a need for effective methods to exploit and use the hidden opportunities and knowledge resting in unexplored data
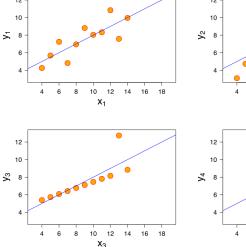
WHY REPRESENT ALL THE DATA?
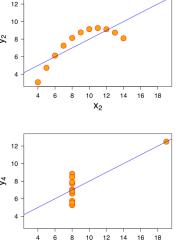
Summaries lose information, details matter
- Confirm expected and find unexpected patterns
- Assess validity of statistical model, is the model appropriate for our data?

- The datasets included in the Anscombe's quartet are only 20 static data points! It gets worse…
- We want the human to be able to review the complexities of a dataset and try to make a judgement call
- We will use statistics along the way as a help, but that cannot give us everything, we want to dig deeper

Anscombe's quartet

	I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe's quartet

Property	Value
Mean of X	9
Variance of X	11
Mean of Y	7.5
Variance of Y	4.1
Correlation	0.816
Linear Regression	y = 3.0 + 0.5x

WHY VISUAL DATA ANALYSIS?

We don't need vis when fully automatic solutions exist and are trusted
- But many analysis problems are ill-specified!
    - We might not know exactly what questions to ask in advance

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively
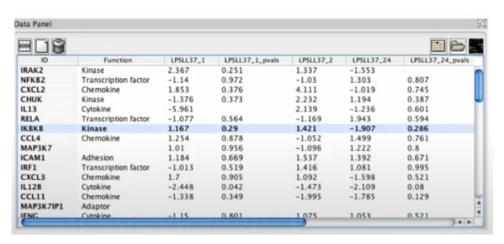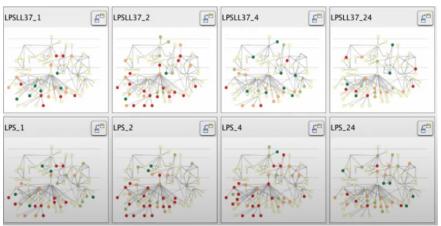(Munzner, T., 2020)
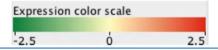
WHY USE AN EXTERNAL REPRESENTATION?

Why visual representations?

- Getting something out of your head to an external representation helps to replace cognition with perception
  - Freeing up our cognitive resources for higher level questions (not just comparing numbers but enabling us to focus on the underlying question)

WHY HAVE A HUMAN IN THE LOOP?

Possibilities

- Presentation of known results
- Stepping stone to assess requirements before developing models
- Help automatic solution developers refine and determine parameters
- Help end users of automatic solutions verify, build trust
- Long-term use for end-users (exploratory analysis of scientific data)

ANALYSIS FRAMEWORK
4 LEVELS, 3 QUESTIONS

Domain situation

- Who are the target users?
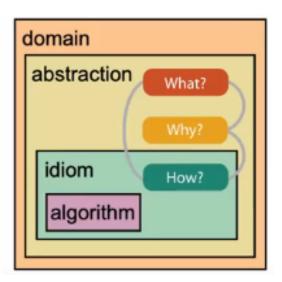
Abstraction

- Translate from specifics of domain to vocabulary of vis
  - What is shown? Data abstraction
  - Why is the user looking at it? Task abstraction

Idiom

- How is it shown?
- Visual encoding idiom: how to draw
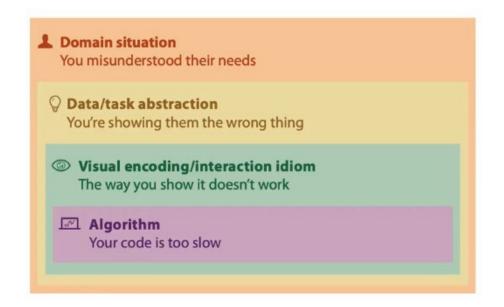- Interaction idiom: how to manipulate

Algorithm

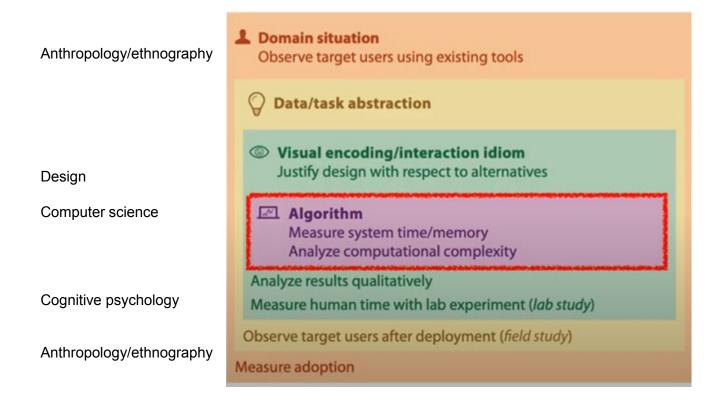- Efficient compuation

DIFFICULTIES OF VALIDATION

Different ways to get it wrong at each level…
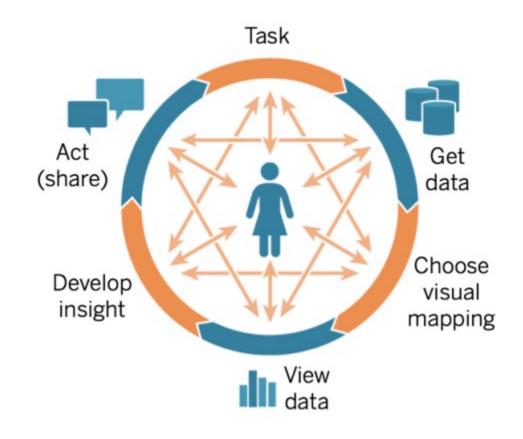
DIFFICULTIES OF VALIDATION

Different ways to get it wrong at each level…

Anthropology/ethnography

Design

Computer science

Cognitive psychology

Anthropology/ethnography

👤 **Domain situation**
Observe target users using existing tools

💡 **Data/task abstraction**

👁 **Visual encoding/interaction idiom**
Justify design with respect to alternatives

🖥 **Algorithm**
Measure system time/memory
Analyze computational complexity

Analyze results qualitatively
Measure human time with lab experiment (*lab study*)
Observe target users after deployment (*field study*)
Measure adoption

TURN THE INFORMATION OVERLOAD INTO AN OPPORTUNITY

MYRIAD OF VDA TOOLS

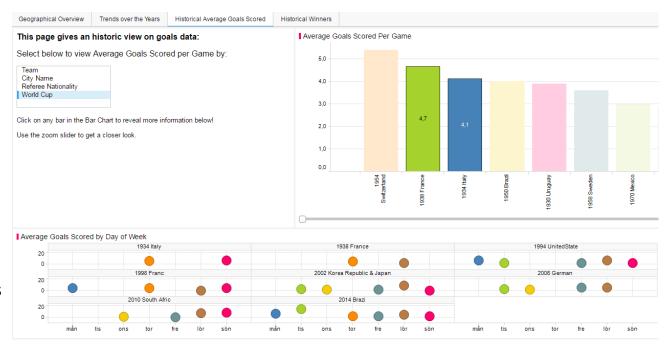A myriad of tools have been developed/are being developed
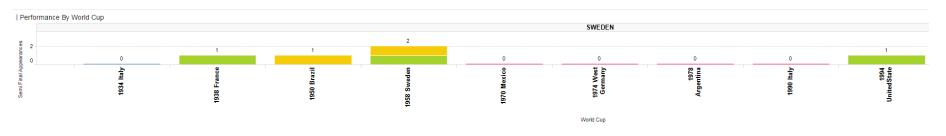
SPOTFIRE EXAMPLE

Statistics from the soccer world cup 1930-2014

- Average goals/game in different world cups

- which days of the week these goals were scored

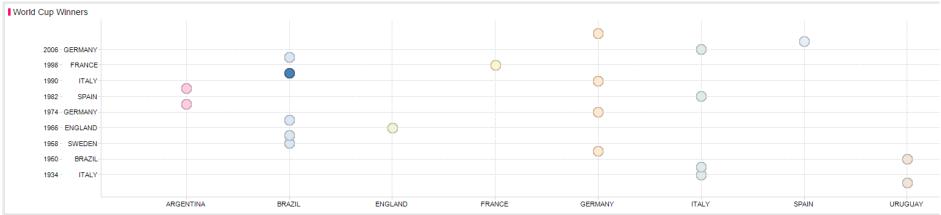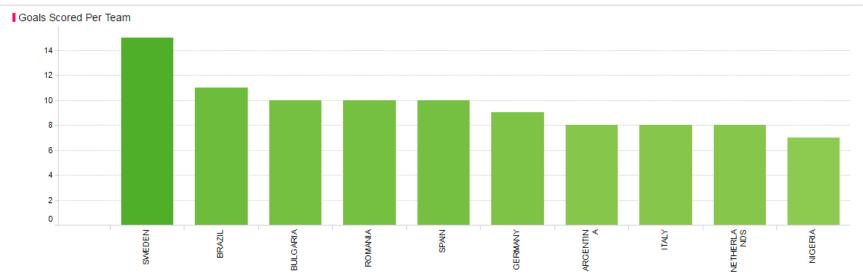- How many semi-final apperances have Sweden made during this time period?

SPOTFIRE EXAMPLE

Statistics from the soccer world cup 1930-2014

- World cup winners

- In the world cup in 1994, Brazil won
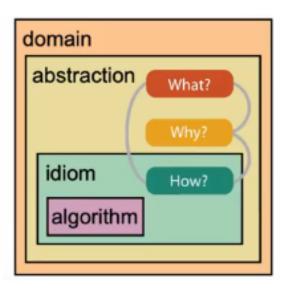 - But it was actually Sweden that scored the most goals

WHAT IS THE COURSE ABOUT?

Evaluation

Human-computer interaction

Human perception and cognition

Information visualization

TEACHING ACTIVITIES

- We will not be able to cover it all in 7,5 hp…

- 6 lectures
  - L1: Introduction (today)
  - L2: Information visualization 1
  - L3: Human perception and cognition
  - L4: Information visualization 2
  - L5: Guest lecture – Spotfire
  - L6: Human-computer interaction and evaluation of VDA applications

- 3 research article seminars (1 reserve)

- 2 project work seminars
  - Mid-check and final seminar

- 3 supervisions using Spotfire
  - 1 general
  - 2 for your project work

EXAMINATION

- Seminar assignment   (3 hp , 80 hours)   Pass/Fail

- Project work   (4,5 hp, 120 hours)   A-F

                                                      Will determine your grade

SEMINAR ASSIGNMENT

SEMINAR ASSIGNMENT
SELECT SEMINAR PAPERS

Seminars with a focus on research within Visual Data Analysis (VDA)
In groups of 3 students, select one research paper/student related to VDA

- The following venues can be of interest (others of course as well as previous of these)

    - IVAPP 2020
    - VIS2020
    - CHI2020
    - IUI 2020
    - IEEE Transactions on Visualization and Computer Graphics
    - Computer Graphics Forum

- Example papers can also be found on Canvas → Files → Seminar articles

- Preferably, choose papers that support your project work (optional though)

- Since you will present the three papers in a group, make sure that there is a common theme in the papers that you can convey

    - Methods, visualization techniques chosen, evaluation, problem domain etc.

SEMINAR ASSIGNMENT
BOOK SEMINAR PAPERS

Each group books their papers by adding it to the Google docs file
- Write the title of the papers + URL + your names on the preferred seminar slot
- First-come, first-served basis
- You may not choose the same paper as anyone else

- I.e. the booking decides your papers to present as well as the time for the presentation

- **Deadline for the selection of papers is 2021-01-28, 23:59**

SEMINAR ASSIGNMENT
SEMINAR PRESENTATION

- Each group of students presents their chosen papers during a seminar using slides

  - Think of the common theme in the papers to present something coherent!

- The presentation is expected to last about 15 minutes, followed by a 5-10 minutes Q&A session. All students in the group must present one/paper each as well as discuss their common issues

- If you want to use the teacher's computer, you have to email your slides to Tove (tove.helldin@his.se) one day before your seminar. The slides should be in .pdf or .ppt format.

- Active participation during all seminars is vital!

  - The presenters should **prepare 2-3 questions** related to their selected papers

    - Can range from collecting the audience's experiences of the issues being handled in the papers, what you believe are the strengths/limitations of the papers from a visual data analysis perspective, the methodologies used etc.

  - The audience should have read the papers beforehand and should keep the discussions active!

See more detailed criteria for the seminar presentations on Canvas→ Assignments→ Book seminar papers and information about the seminar presentations

SEMINAR ASSIGNMENT
SEMINAR REPORT

The students are to write a report of the 3 papers.
The report is to be written individually!

Max 6 pages long, min 4 pages long (11 point font, no figures , tables or references included)
- The report should inform the reader of the purpose of the papers, their main content and conclusions in relation to visual data analysis.
- Not just a summary of their contents → your reflections are vital
- A template is provided on Canvas→ Files→ Templates→ Seminar report template in both .pdf and .docx.
- The text must be your own! All sources need to be cited and referred to correctly. If you don't know how, please refer to this library guide.
- **Deadline for the seminar report is 2021-03-04, 23:59**
- Submit the report through Canvas, as well as send it to Tove's urkund address: tove.helldin.his@analys.urkund.se

See more detailed criteria for the seminar report writing on Canvas→ Assignments→ Seminar report

SEMINAR ASSIGNMENT
FAILURE TO ATTEND A SEMINAR + RE-EXAMINATION

Participation and attendance to the seminars is obligatory.

If you fail to attend to a seminar, you have to write a report of the seminar papers presented at that particular seminar
- See the Google docs booking page for the list of papers/seminar

Also, if you fail to hand-in or do not fulfil the criteria for the seminar report, you will need to hand in a revised seminar report.

The summary should be 4-6 pages long (without figures, tables, references)
→ The same criteria as for the Seminar report writing
→ **Deadline for the missed seminar reports/the seminar report re-examination is 2021-04-05, 23:59**

PROJECT
WORK

PROJECT WORK

Choose a topic/problem related to VDA and work on this in groups (the same groups as for the seminar assignment)

The idea is that all groups of students

- Set up research question(s)/hypotheses or problem
- Choose a dataset(s) that can help to solve the problem
- Prepare dataset for analysis (i.e. data management)
- Analyse the data
- Appropriately visualize and interact with the data
  - Solve/investigate the problem selected
  - Validate and evaluate (user study)
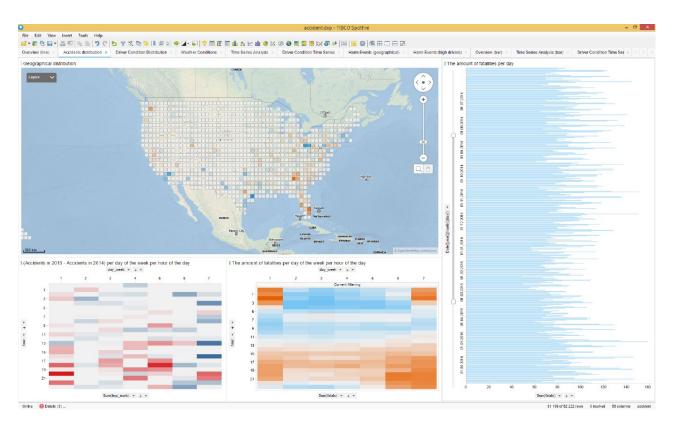 - Discuss your work!

PROJECT WORK

You are free to choose which tool(s) you would like or make your own implementation, however, it is recommended to use some ready to use tool

- We will have supervisions using Spotfire (3 occasions)

    - An email with a Spotfire licence has been sent to all registered students' emails

    - All Spotfire occasions will take place here: https://tibco.zoom.us/j/2468558879

    - Tableau licences are provided as well (see Canvas → Files → Project→ Tableau license keys.pdf)

- This is a master's course → you will need to spend time to learn how to use the tool by yourself and use the 3 supervisions with Spotfire carefully!

    - There is a lot of good tutorials online!

- General supervision during the whole course

    – Contact Juhee, easiest, fastest through by email – juhee.bae@his.se

    – Previous project examples are available on Canvas

    – Spotfire supervision by the people from Spotfire

PREVIOUS EXAMPLES
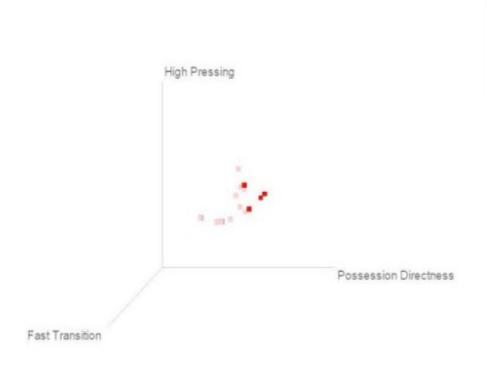
Cause of the increase on US Traffic Fatalities 2015

Fatality analysis reporting system,
https://www.transportation.gov/fastlane/2015-traffic-fatalities-data-has-just-been-released-call-action-download-and-analyze

PREVIOUS EXAMPLES

Soccer team analysis

PREVIOUS EXAMPLES
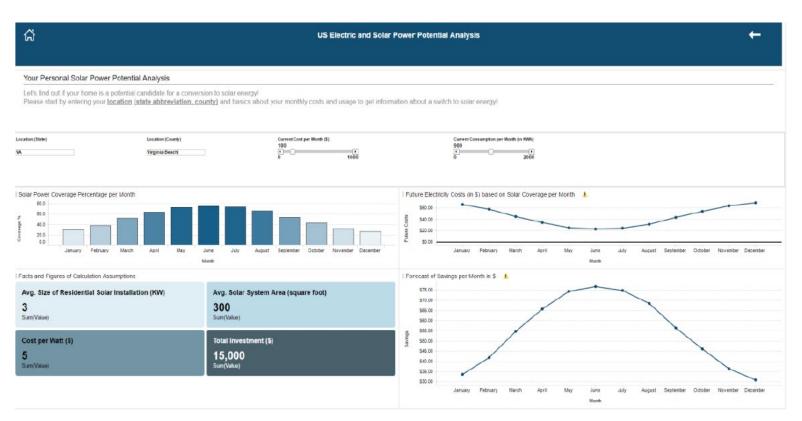
Solar Power analysis for Virginia Beach area

U.S. Electric Utility Companies and Rates: Look-up by Zipcode (2015)
https://openei.org/datasets/dataset/u-s-electric-utility-companies-and-rates-look-up-by-zipcode-2015

PREVIOUS EXAMPLES
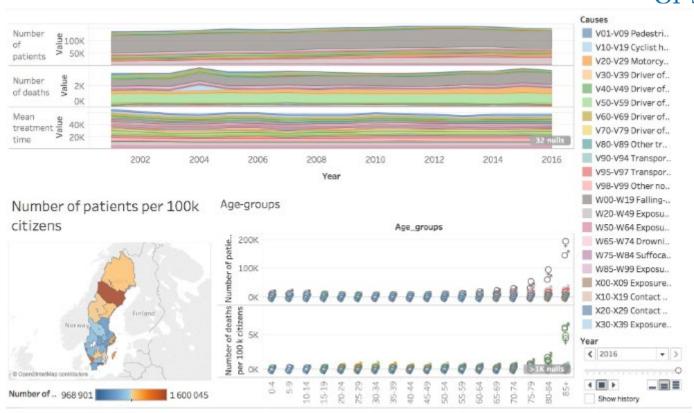
Common types of injuries in Sweden

Figure 1. Injuries in the regions over the years

DATA

Amazon public data sets https://aws.amazon.com/public-datasets/

Kaggle https://www.kaggle.com/datasets

Stanford Large Network Dataset Collection https://snap.stanford.edu/data/

Social Computing Data Repository http://socialcomputing.asu.edu/pages/datasets

UCI https://archive.ics.uci.edu/ml/datasets.html

Yahoo https://webscope.sandbox.yahoo.com/

InfoVis http://www.infovis-wiki.net/index.php?title=Data_Libraries#InfoVis_Contest_Datasets

and many, many more …..

PROJECT TEMPLATES & DEADLINES

Project report template on Canvas:

- final_project_template.docx/.pdf

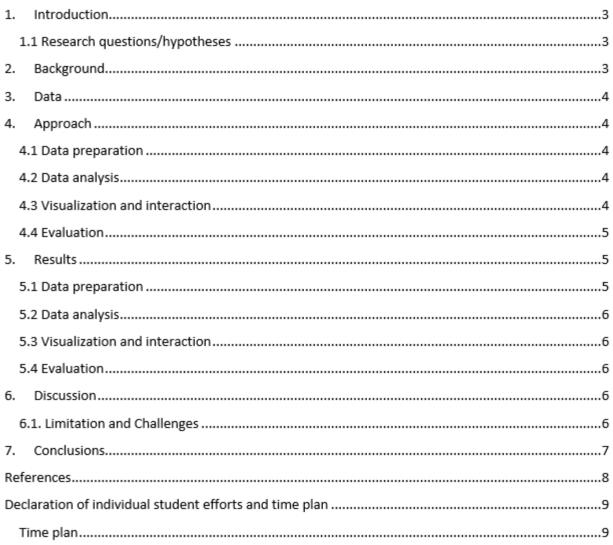Start populating this report template already from the project proposal deadline.

Due dates to submit on Canvas (and Urkund – juhee.bae.his@analys.urkund.se)

- Project proposal (10%): **Feb 8**. → Project proposal folder

- Mid-check (report and presentation 20%):
  - Mid-check report: **Feb 22** → Mid check seminar – for peer-review and teacher comments folder + Urkund
  - Peer-review comments on mid-check: **Feb 26** → Mid-check comments on another student group's work folder

- Final (report and presentation 70%):
  - Draft of final report: **March 15** → Final project work seminar folder + Urkund
  - Final project work report: **March 25** → Final project work report folder + Urkund

FINAL REPORT TEMPLATE

Contents

RECAP
HOW TO SUBMIT ASSIGNMENTS?

Have a look at the Canvas site for the different assignments.

Remember to upload your seminar report, project mid-check report and final project report on Canvas **AND** send it to our Urkund addresses:
Seminar report: tove.helldin.his@analys.urkund.se
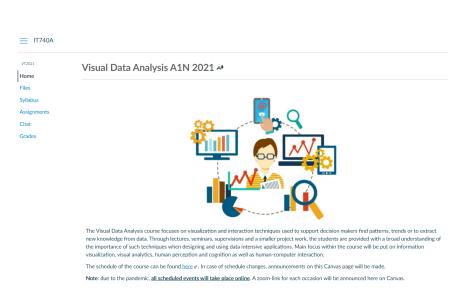Project report: juhee.bae.his@analys.urkund.se

The VDA course is a prerequisite for the DDDM course (starts autumn 2021) for those taking the 2 year master

COURSE HOMEPAGE (IT740A)

Keep an eye on the Canvas page!

- Announcements
- Assignment descriptions
- Submission of assignments
- Seminar papers
- Templates
- Lecture slides
- Course literature & reference material
- Schedule
- Contact information

SCHEDULE

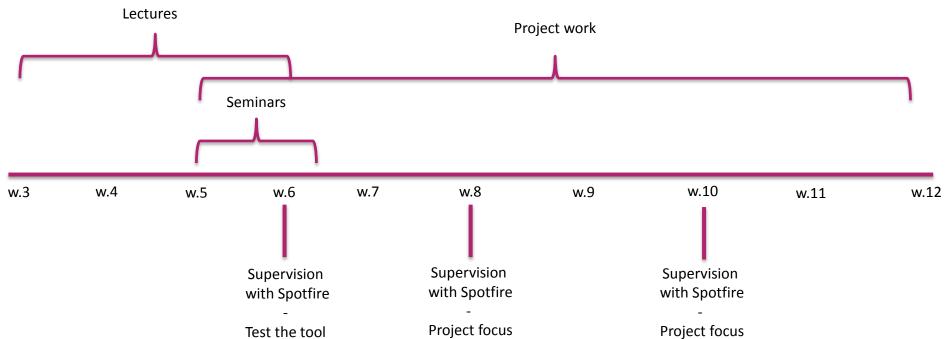TimeEdit – search for course IT740A

Due to the pandemic, all occasions will take place online through Zoom.
Zoom links are available through Canvas (course start page). Keep updated here!

Announcements on Canvas will be made for changes
- For example, the seminar occasions might change depending on your bookings + if we need a reserve

COURSE OUTLINE
LECTURES, SEMINARS AND PROJECT WORK

Lectures

Project work

Seminars

| w.3 | w.4 | w.5 | w.6 | w.7 | w.8 | w.9 | w.10 | w.11 | w.12 |

Supervision with Spotfire

-

Test the tool

Supervision with Spotfire

-

Project focus

Supervision with Spotfire

-

Project focus

COURSE TIMELINE
DEADLINES AND PRESENTATIONS

Submit
mid-check for
peer-review

Book
seminar
papers

Submit
project
proposal

Submit
mid-check
peer-review

Submit
seminar
report

Submit
draft of
final report

Submit
final
project
report

w.3   w.4   w.5   w.6   w.7   w.8   w.9   w.10   w.11   w.12

Supervision
with Spotfire
-
Test the tool

Mid-check
presentation

Supervision
with Spotfire
-
Project focus

Supervision
with Spotfire
-
Project focus

Final project
presentation