UNIVERSITEIT TWENTE.

Data Science [201400174]

Course year 2022/2023, Quarter 1B

DATE
December 14, 2022

EXCERPT

Projects

TEACHERS
Faizan Ahmed
Ellen-Wien Augustijn
Nacir Bouali
Faiza Bukhsh
Rolf de By
Karin Groothuis-Oudshoorn
Maurice van Keulen
Mahdi Khodadadzadeh
Elena Mocanu
Estefania Talavera
Brenda Voorthuis
Shenghui Wang

COURSE COORDINATOR
Karin Groothuis-Oudshoorn (quartile 1A)
Maurice van Keulen (quartile 1B)
Faizan Ahmed (quartile 2A)

PROJECT OWNERS
Faiza Bukhsh
Karin Groothuis-Oudshoorn
Maurice van Keulen
Elena Mocanu
Mannes Poel
Michel van Putten
Mohsen Jafari Songhori
Luc Wismans

Contents

Part II

Projects

Project 1

Project 1: Predicting surgical case durations for a Thorax centre [PSCD]

1.1 Introduction

Project owner: Karin Groothuis-Oudshoorn

Primary topic: DPV and/or DM

In modern healthcare, organizations face the challenge of delivering more and better quality care with less human and financial resources. This is mainly due to rising demand for healthcare and increasing expenditures. Efficiency is directly linked with quality, as inefficient care processes use up valuable resources and displace more useful care. Efficiency improvements are therefore very valuable for hospitals. MST is a top-clinical medical center located in the region of Twente and is one of the biggest non-academic hospitals in the Netherlands. The medical center compromises of two inpatient clinics in Enschede and Oldenzaal and two outpatient clinics in Haaksbergen and Losser. The inpatient clinic in Enschede has moved her patients as of 2016 to the newly built location Koningsplein. This new location provides a capacity of 739 beds. Thorax Centrum Twente (TCT) is a center within MST, specializing in diagnosis and treatment of cardiothoracic diseases. Multidisciplinary medical care is delivered through several cardiothoracic-related specialties such as cardiology and cardiac surgery. TCT is one of the 16 thorax centers in the Netherlands and has grown rapidly after its establishment in September 2004. One reason for this is their short waiting list for open heart surgery, making TCT an interesting medical center for patients. TCT performs approximately 1,100 to 1,200 open-heart surgeries per year, mainly coronary and heart valve surgeries. TCT experiences a high rate of operating rooms working beyond regular operating time. High amounts of overtime result in unnecessary costs and low staff satisfaction. A recent study among Dutch hospitals suggests that more accurate predictions of surgical case duration and altering the sequencing of surgical cases on an OR-schedule can improve efficiency[1].

References

[1] van Veen-Berkx E, Elkhuizen SG, van Logten S, et al. Enhancement opportunities in operating room utilization; with a statistical appendix. J. Surg Res. 2015;194(1):43-51.

Deliverables

- Report (according to the DS template)
- Presentation

1.2 Description of data set

The dataset comprised of 4087 surgical cases performed from January 2013 to January 2016 at TCT. The surgical case duration is given in minutes. The hospital stay time and IC stay time is given in days. Unknown data is indicated with 'NULL' or 'onbekend'. The data is in the file 'surgical_case_duration.csv'. In tables 1.1-1.4 you can find descriptions of the variables in the dataset. The different levels are given. In case of type of surgery not all labels are translated (we left out the less frequent ones).

1.3 Description of challenge

The challenge of this project is to identify patterns in surgical case durations and to derive prediction models and / or classification models for surgical case duration to support OR-planners at TCT in making the most efficient OR-schedules with the available patient level data in order to decrease the overtime at TCT, while maintaining the current OR-utilisation rate.

1.4 Tips and suggestions

- Some variables (features) have a lot of categories and some of those categories have few observations. In that case it can be wise to recode these categories to e.g. 'other types' or leave those observations out. But please explain and give arguments if you do so!

Table 1.1: Description of Surgery-related variables

Variablename	Variable (English)	Categories (definition)
Operatietype	Surgery type	Aortic Valve Replacement (AVR) AVR + MVP Bentall Procedure Coronary Artery Bypass Graft (CABG) CABG + AVR CABG + MVP Epicardial LV-lead (Epicardiale LV-lead) Lobectomy or segment resection (Lobectomie of segmentresectie) Mediastinoscopy Mitral Valve Plasty (MVP) MVP + Tricuspid Valve Plasty (TVP) Mitral Valve Replacement (MVR) Nuss bar removal Nuss-procedure Refixation of the sternum (Refixatie sternum) Rethoracotomy (Rethoractomie) Removal of steel wires (Staaldraden verwijderen) VATS Boxlaesie (video assisted thoracic surgery) Wound debridement (wondtoilet) Other types
Benadering	Surgical approach	Full sternotomy (Volledige sternotomie) Left antero lateral (Antero lateraal links) Right antero lateral (Antero lateraal rechts) Left postero lateral (Postero lateraal links) Right postero lateral (Postero lateraal rechts) Partial sternotomy (Partiele sternotomie) Other approaches: Parasternaal links, Parasternaal rechts, Dwarse sternotomie, Xiphoidaal, NULL)
Chirurg	Surgeon	Surgeon 1 – 15 Other specialism (Ander specialisme)
Anesthesioloog	Anesthesiologist	3 – 19, Unknown (onbekend)
OK	Operation room	HCK1, HCK3, HCK4, OK 1, OK 10, OK 11, OK 3, OK 4, OK 5, OK 9, TOK1, TOK2, TOK3, TOK4, else (Elders)
Casustype	Urgency	Elective (planned on the elective program) (Electief) Emergency ($<$ 24 hours) (Spoed $<$ 24 uur) Acute ($<$ 30 minutes) (Acuut $<$ 30 minuten) Acute (Spoed) Acute ($<$ 5 hours) (Spoed $<$ 5 uur) Unknown (NULL)
Dagdeel	Time of day	Morning (7:00 – 12:00) Afternoon (12:00 – 18:00) Evening and night (18:00 – 7:00)
Aantal anastomosen	Amount of bypasses	Continuous variable
HLM	Cardiopulmonary bypass use	Yes (heart-lungmachine usage planned for surgery) No

Table 1.2: Description of Surgery-related variables

Variablename	Variable (English)	Categories (definition)
Leeftijd	Patient age	Continuous
Geslacht	Patient gender	Male Female
AF	Presence of atrial fibrillation	Yes (AF rhythm present) No
Chronische longziekte	Presence of chronic lung disease	Yes (long term use of bronchodilators or steriods for lung disease) No
Extracardiale vaatpathie	Presence of extracardial arteriopathy	Yes (claudication, carotid occlusion or 50% stenosis, amputation for arterial disease or previous or planned intervention on the abdominal aorta, limb arteries or carotids) No
Active endocarditis	Presence of active endocarditis	Yes (patient still on antibiotic treatment for endocarditis) No
Hypertensie	Presence of hypertension	Yes No
Pulmonale hypertensie	Presence of pulmonary hypertension	Normal (no increased pulmonary artery pressure) Moderate (pulmonary artery systolic pressure 31-55 mmHg) Severe (pulmonary artery systolic pressure $>$ 60mmHg)
Slechte mobiliteit	Presence of poor mobility	Yes (severe impairment of mobility secondary to musculoskeletal or neurological dysfunction) No
Hypercholesterolemie	Presence of hypercholesterolemia	Yes No
Perifeer vaatlijden	Presence of peripherial vascular disease	Yes No
Linker ventrikel	Left ventricle	Good (LV ejection fraction $>$50%) (Goed) Moderate (LV ejection fraction 31-50% (Matig) Poor (LV ejection fraction $\leq$ 30% (Slecht) Very poor (Heel slecht)
Nierfunctie	Renal function	Normal (creatine clearance $>$ 85 ml/min Moderate (creatine clearance 50-85 ml/min) (Matig) Poor (creatine clearance $<$ 50 ml/min or dialysis) (Slecht) Dialyse
DM	Presence of diabetes mellitus requiring insulin	Yes (diagnosis DM requiring insulin) No
Eerdere hartchirurgie	Previous heart surgery	Yes (heart surgery in the patient's history) No
Kritische preoperatieve status	Critical pre-OR state	Yes (ventricular tachycardia or ventricular fibrillation or aborted sudden death, preoperative cardiac massage, preoperative ventilation before anesthetic room, preoperative inotropes or IABP, preoperative acute renal failure (anuria or oliguria $<$10ml/hr)) No
Myocard infact $<$90 dagen	Mycordial infarction before surgery	Yes (MI within 90 days before surgery) No
Aorta chirurgie	Aortic surgery	Yes (planned surgery on the aorta) No
Euroscore1	Euroscore1	Continuous variable
Euroscore2	Euroscore II	Continuous variable

Table 1.3: Description of Surgery-related variables (Continuation)

Variablename	Variable (English)	Categories (definition)
CCS	Canadian Cardiovascular Society (CCS) score for angina	0 (no symptoms) 1 (angina only during strenuous or prolonged physical activity) 2 (slight limitation, with angina only during vigorous physical activity) 3 (symptoms with everyday living activities, i.e. moderate limitation) 4 (inability to perform any activity without angina or angina at rest, i.e. severe limitation)
NYHA	New York Heart Association (NYHA) score - dyspnea	1 (cardiac disease, but no symptoms and no limitation in ordinary physical activity, e.g. no shortness of breath when walking, climbing stairs etc.) 2 (mild symptoms (mild shortness of breath and/or angina) and slight limitation during ordinary activity) 3 (marked limitation in activity due to symptoms, even during less-than-ordinary activity, e.g., walking short distances (20-100 meters). Comfortable only at rest 4 (severe limitations, experiences symptoms even while at rest, Mostly bedbound patients).

Table 1.4: Description of Outcome variables

Variablename	Variable (English)	Categories (definition)
Geplande operatieduur	Planned surgery duration	Continuous outcome
Operatieduur	Surgery duration	Continuous outcome
Ziekenhuis ligduur	Hospital days	Continuous outcome
IC ligduur	Intensive care days	Continuous outcome

Project 2

Project 2: Public Priorities for Primary Child Health Care for Children [MOCHA]

2.1 Introduction

Project owner: Karin Groothuis-Oudshoorn & Brenda Voorthuis

Primary topic: DPV

The health of our children is of utmost important not only for themselves and their families, but for the whole society. As future workers, parents and carers, these children are the ones that will build the world of the future. Health services for children are structured differently throughout the European Union, and there is little research into what works best. Therefore, the MOCHA project (Models of Child Health Appraised) performed a systematic, scientific evaluation of the types of primary care for children that exist in Europe (`http://www.childhealthservicemodels.eu/`). The aim of this project was to elicit formative values from the general public in five different countries and to determine public priorities in the assessment of the quality of a child-oriented primary health care system. The following research question and sub-questions were formulated. What are the priorities of European citizens in assessing the quality of primary care for children in Europe?

1. What are the experiences and/or perceptions of European citizens with the quality of currently provided primary care for children?
2. What are the preferences of European citizens with respect to the quality attributes of primary care for children?

Based on the diversity of the primary health care systems (general-practitioner-led, pediatrician-led or mixed), the United Kingdom, the Netherlands, Germany, Spain and Poland were chosen for studying these research questions. In a descriptive, cross-sectional, quantitative design, a uestionnaire was used to elicit preferences of a representative sample of the general public with respect to children's primary care and to measure experiences with the quality of currently provided care. Based on the review of literature, the child and carer centred outputs of each of the attributes of the primary health care system were efined. Through an iterative process within the MOCHA project team, the outputs were operationalized in 40 attributes-items. ttribute-items were operationalized in plain language and technical jargon was avoided as much as possible. Between one and nine attribute-items were operationalized for each of the nine attributes (outputs) of a child-oriented health care system (see Table 2.1).

Table 2.1: Description of the 9 attributes of the primary health care system

Attributes	Definition	Attribute-items
ACCESSIBLE	Accessible primary care is available within reasonable reach of parents and children, with ample opening hours, good appointment systems and other aspects of service organization and delivery that allow children to obtain the services when they need them"	1, 2, 3, 4, 5, 6, 7, 8, 9
AFFORDABLE	Affordable primary care can be accessed without inordinate financial barriers, such as high co-payments or cost-sharing arrangements	10, 11
APPROPRIATE	Appropriate primary care is effective in meeting the child's needs, timely and of high technical quality	12, 13, 14, 15, 16, 17
CONFIDENTIAL	Confidentiality in primary care is the right of a child to have personal, identifiable medical information kept private if they choose to, from medical professionals as well as parents	18, 19, 20
CONTINUOUS	Continuous primary care is the experience of a continuous caring relationship with the health care professional(s) by a single child and its parents over time, that is responsive of the child's changing needs.	21, 22, 23, 24, 25, 26
COORDINATED	Coordinated primary care is deliberately organizing child care activities and sharing of information among all of the participants concerned with a child's care with the aim to achieve safer and more effective care.	27, 28, 29, 30, 31
EMPOWERING	Empowerment in primary care is a process through which children and parents gain greater control over decisions and actions affecting a child's health	32, 33, 34, 35, 36, 37
EQUABLE	Equable primary care is the absence of systematic and potentially remediable differences in access to primary care and health status across population groups"	38, 39
TRANSPARENT	Transparent primary care is the degree to which a healthcare service or provider is open to children and parents about their quality, cost structure, services and work method	40

The description of the attribute items can be found in the file **mocha_items.csv**.

Deliverables

- Report (according to the DS template)
- Presentation

2.2 Description of data set

Two data files are provided containing the same data. In the file **POCHA_FULL_DATASET_names.xlsx** the column names are the variable names and the labels are the values. In the file **POCHA_FULL_DATASET_labels.xlsx** the column names are the labels of the variable names, i.e. the question itself. The values in the cells are not the labels but the values. The file **mocha_items.csv** contains the description of the attribute items.

2.2.1 The POCHA questionnaire

The POCHA questionnaire consisted of five parts (see : `https://www.childhealthservicemodels.eu/wp-content/uploads/Final-Report-POCHA.pdf`).:

1 background characteristics

2 the health status of the child(ren) and current health care consumption of any child(ren) below the age of 18

3 child independence, i.e. the age at which a child could be or should be able to make decisions independent from its parents

4 the quality of the primary health care system

5 the prioritization of the 40 attribute-items of a child-oriented primary health care system

The following background characteristics of respondents were measured:

- age: 19 years or younger, between 20 and 24, . . . , between 65 and 69, 70 years or older;
- gender: female, male;
- country: United Kingdom, Netherlands, Germany, Spain, Poland; and the countries' region;
- number of children: 1, 2, 3 other;
- number of children ¡ 18: yes/no;
- highest level of completed education: with the following categories for the UK: Entry level, GCSE (grades D-G), GCSE (grades A*-C), A-Level, Higher National Certificate/National Diploma, Bachelor Degree, Master Degree, Doctoral Degree
- size of the city, where the respondent lives: ¡ 100, 100 – 1.000, 1.000 – 10.000, 10.000 – 20.000, 20.000 – 100.000, 100.000 – 200.000, 200.000 – 1.000.000, ¿ 1.000.000.

Health Status and Health Care Consumption

Health status was measured was asking: "In the past 12 months, has (one of) your child(ren) had a medical condition that lasted longer than 6 weeks?". "What condition was this/were these? Please indicate all conditions that apply for any child(ren) below the age of 18." Conditions to choose were: Eczema, Asthma, Hay fever, Allergy, Stomach ache, Headache, Back problems, Fatigue, Sleep problems, Depressive complaints, Hyperactivity and ADHD, Constipation, Overweight and obesity, Other: . . . .

Following the question on health status, health care consumption for children ¡ 18 with long-term disease (¿ 6 weeks) in the previous 12 months was measured with the frequency of contact with primary and secondary health care providers.

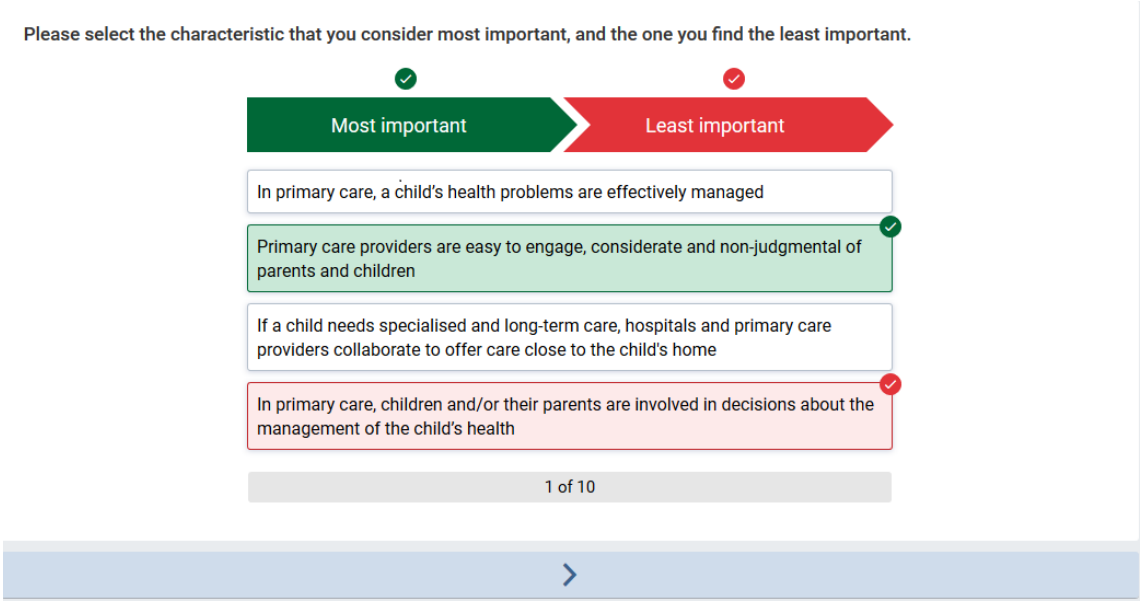

Figure 2.1: Example of a prioritization question from the English Survey.

Quality of the Primary Health Care System

In order to measure experiences with and/or perceptions of primary care for children, ten questions with statements were presented about the quality of primary health care system for children in place in the respondent's country. The respondent was asked to indicate to what extent he agreed with each statement, on a 5-point rating scale (strongly disagree to strongly agree). If the respondent had no direct experience of primary care for children, he was asked to react on the statement based on his perceptions of primary care for children in his country. These perceptions could be based on media coverage and/or on stories from friends and family. For each respondent on the questionnaire, the ten statements were randomly selected out of 40 statements on the quality of the primary health care system for children.

Prioritization of Attribute-items of a Child Oriented Primary Health Care System

To determine what people find important characteristics (= attribute-items) of the quality of a primary health care system, we asked respondents in 20 questions to indicate what they think is important in the assessment of primary care for children. To this respect, best-worst scaling questions were used; for an example of this type of question used, see Figure 1. In each question, a list of four potential characteristics of primary care was presented. The respondent is asked to read these characteristics carefully and to choose which of these he finds most important in primary care for children and which of these he finds least important.

Respondent priorities in each question are likely influenced by the combination of attribute-items in that question. Thus, combinations of attribute-items within and over respondents was varied using an experimental design. Each attribute-item was presented to each respondent twice, in a combination different from each other. Eight different sets of ten questions were made, in which each attribute-item was presented once. Two different sets of ten questions were presented to each respondent.

2.3 Description of challenge

The challenge in this project is to provide an appropriate star schema for the data that is obtained from the questionnaire and to summarize and visualize the data such that the research questions are answered. Look for relevant differences between countries and based on other background characteristics.

2.4 Tips and suggestions

- The best-worst scaling questions can be analyzed with different methods, where the most easy methods is the best minus worst counts analysis. As an example you can look at the article from Louviere & Flynn "Using Best-Worst Scaling Choice Experiments to Measure Public Perceptions and Preferences for Healthcare Reform in Australia, the Patient 2010". Priorities for attribute-items of quality of care can be calculated using counts analysis on an individual and a group level. First, the number of times an attribute-item of quality was selected as most (max 2) and least important (max 2) was counted per respondent. Then, best-worst scores were calculated by subtracting the number of times each attribute-item was selected as least important from the number of times it was selected as most important. Individual best worst counts ranged from – 2 (not important) to +2 (important). Group priorities of attribute-items can then be defined as the best-worst score of that attribute-item for that group (defined e.g. by country). Best worst scores should then be normalized over groups by dividing the best-worst count by the number of times each attribute-item was presented to the group (2 times for each respondent in the group) and multiplying this ratio with 100. Group best worst scores can be calculated for respondents from each country, male and female respondents and respondents with and without children below the age of 18.

Project 3

Project 3: Predicting mortality for COVID-19 patients [COVID]

3.1 Introduction

Project owner: Brenda Voorthuis & Karin Groothuis-Oudshoorn

Primary topic: DM

The COVID-19 outbreak was first identified in Wuhan, China, in December 2019 and was declared a global pandemic in March 2020. By the end of July 2020, there were over 17 million confirmed cases worldwide and close to 700,000 deaths. The risk of death increases with age and with underlying conditions such as obesity and diabetes. The worldwide increase in COVID-19 cases is putting high pressure on healthcare services. Early assessment of the severity of COVID-19 cases is essential for logistic planning. Studies by Yan et al. (2020) and Zhou, Chen, and Lei (2020) suggest that the mortality of an individual COVID-19 case can be predicted with more than 90% accuracy and over 10 days in advance based on just three biomarkers: lactic dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (hs-CRP).

References

Yan, L., Zhang, H-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., ... Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence, 2*, 283-288. doi:10.1038/s42256-020-0180-7

Zhou, F., Chen, T., & Lei, B. (2020). *Do not forget interaction: Predicting fatality of COVID-19 patients using logistic regression*. Retrieved from `https://arxiv.org/abs/2006.16942`

Deliverables

- Report (according to the DS template)
- Presentation

3.2 Description of data set

Two datasets are available. The first dataset was used as training data in Yan et al. (2020) and was collected between January 10, 2020 and February 18, 2020, at the Tongji Hospital in Wuhan, China. It consists of 375 patients, of which multiple biomarker measurements were taken and 174 died. The second dataset of another 110 patients from the same hospital was used as test data and was collected between February 19, 2020, and February 24, 2020. Of these 110 patients, 13 died.

The training data file **time_series_375_preprocess_en.xlsx** contains data on the age and gender of each patient, along with admission time, the outcome (death/discharge) and time of discharge/death. The other 74 columns contain information on different biomarkers. The test dataset, **time_series_test_110_preprocess_en.xlsx**, contains the same patient information but only the biomarkers LDH, lymphocytes, and hs-CRP.

3.3 Description of challenge

The challenge of this project is to develop a prediction model for the mortality of COVID-19 patients based on the biomarkers that are available in the data. Create multiple machine learning models (i.e. different DM methods) and compare them based on the applicable performance measures to identify the best fitting model.

3.4 Tips and suggestions

- Multiple measurements per patient are available in the datasets. In the development and testing of the models by Yang et al. (2020) and Zhou, Chen, and Lei (2020), the last known sample of each patient was used. Other methods are allowed, but explain and substantiate your choice if you do so!
- The model output corresponds to mortality. The variable `outcome` is coded with 0 (discharge) and 1 (death).
- The test dataset only contains information on the three biomarkers LDH, lymphocytes, and hs-CRP. If you want to test a model containing different biomarkers, split up the training data into another training dataset and testing dataset or use cross-validation.

Project 4

Project 4: Predicting neurological outcome in patients with a severe postanoxic encephalopathy [EEG]

4.1 Introduction

Project owner: Michel J.A.M. van Putten
Primary topic: DM

Each year, about 7000 patients with a postanoxic coma after a cardiac arrest are admitted to the Intensive Care Unit. Early prediction of neurological outcome is highly relevant, not only for the treating physicians, but also for family members. This can prevent futile treatment, but will also assist in providing care for those with a high probability of good recovery.

We and others have shown that early recording of the electroencephalogram (EEG) allows reliable prediction of both poor and good outcome in a significant percentage of patients (about 50-60%). While these recordings are typically assessed by visual analysis, machine learning may assist or even outperform human visual assessment.

We provide you with a dataset with various quantitative EEG features and the neurological outcome, the Cerebral Performance Category Score (CPC).

Deliverables

- key result: ROC curves for poor and good outcome, including confidence intervals
- report
- presentation

References

The description of the data set contains references to several papers where more details are given on the specific features. Furthermore, these two papers are recommended:

- Marleen C Tjepkema-cloostermans, Jeannette Hofmeijer, Albertus Beishuizen, Harold W Hom, Michiel J Blans, Frank H Bosch, and Michel J A M Van Putten. Cerebral Recovery Index: Reliable Help for Prediction of Neurologic Outcome After Cardiac Arrest. Critical Care Medicine, pages 1–9, 2017.
- MC Tjepkema-Cloostermans, FB van Meulen, G Meinsma, and MJAM van Putten. A Cerebral Recovery Index (CRI) for early prognosis in patients after cardiac arrest. Critical Care, 17(5):R252, 2013.

As a general text on machine learning, the project owner recommended

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. http://doi.org/10.1023/A:1010933404324

4.2 Description of data set

The data is contained in two excel files, featuresNEW_12hrs.xls (corresponding to 12 hours after CA) and featuresNEa_24hrs.xls (corresponding to 24 hours after CA) from several patients. Some patients can have both 12hrs and 24hrs EEG.

In excel files, column 1 corresponds to Neurological outcome (or `$Patient Outcome`) which is the CPC score grouped into two categories (binary): good (CPC scores [1-2] denoted as "1") and poor (CPC scores [3-5] denoted as "0").

Columns 2 to 45 corresponds to different features extracted from the EEG. In addition to standard features, we also extract amplitude and frequency modulation features from EEG. The HT is a popularly used tool in the field of neuroscience that provides an automatic method for separating the signal spectrum into amplitude modulation (AM) and frequency modulation (FM) components [**?**]. Given an input EEG signal $x[n]$, the AM and FM can be obtained as,

$$\begin{aligned} AM &= w[n,n]\,|z[n]| \\ FM &= \frac{1}{2\pi} w[n,n] \frac{d\angle z[n]}{dn} \end{aligned}$$

where $z[n]$ is the analytic associate of the signal $x[n]$, and $w[n,n]$ is 2D Hamming window of duration H_t (4s) and bandwidth H_f seconds (4s).

Below is the description of features:

- **Time domain features** [**?**]:
  'Nonlinear energy', 'Activity' ,'Mobility', 'Complexity', 'RMS Amplitude', 'kurtosis', 'skewness'
  AM - 'meanAM' (mean AM), 'stdAM' (standard deviation of AM), 'SkewAM' (Skewness of AM), 'KurtAM' (Kurtosis of AM)
  'BSR' - Burst suppression ratio defined as

$$BSR = \frac{\text{duration of EEG in suppression state (amplitude} \leq 5\mu V)}{\text{total duration of EEG}} \times 100$$

- **Frequency domain features** (obtained using standard Fourier transform):
  Power in subband: 'delta'(0.5-4 Hz), 'theta'(4-8 Hz), 'alpha'(8-12 Hz), 'spindle' (12-16 Hz), 'beta'(16-32 Hz), 'total' (0.5-32Hz)
  Corresponding normalized power (normalized with total spectral power): 'delta_tot' (delta/total), 'theta_tot' (theta/total), 'alpha_tot' (alpha/total), 'spindle_tot' (spindle/total), 'beta_tot' (beta/total)
  Corresponding normalized power (normalized with delta spectral power): 'alpha_delta' (alpha/delta), 'theta_delta' (theta/delta), 'spindle_delta' (spindle/delta), 'beta_delta' (bata/delta)
  Corresponding normalized power (normalized with theta spectral power): 'alpha_theta' (alpha/theta), 'spindle_theta' (spindle/theta), 'beta_theta' (beta/theta)
  FM: 'fhtife1' (mean FM), 'fhtife2' (standard deviation of FM), 'fhtife3' (skewness of FM), 'fhtife4'

(Kurtosis of FM)
'sef' (spectral edge frequency) [**?**, **?**], 'df' (peak frequency)

- **Entropy domain features** [**?**, **?**]:
'svd_ent' (Singular value decomposition entropy), 'H_spec' (spectral entropy) [**?**], 'SE' (State entropy) [**?**], 'saen' (sample entropy) [**?**], 'abs(renyi)' (Renyi entropy) [**?**], 'abs(shan)' (Shannon entropy) [**?**], 'perm_entr' (permutation entropy) [**?**], 'FD' (fractal dimension) [**?**]

4.3 Description of challenge

We would like to improve on our earlier estimates, about 50%, in prediction accuracy, both for good outcome (CPC=1 or 2) and poor outcome (CPC =3,4 or 5). For poor outcome, this should be reached at a specificity of 100% and for a good outcome for specificity of 95% or better.

4.4 Tips and suggestions

You can use various classifiers, e.g. SVM, decision trees, random forests. Results must be shown as ROC curves, including confidence intervals. Apply techniques that provide information about which features are most relevant for the prediction.

Note that for most patients, data from various hours after arrest is available. Start with using data for the same hours after arrest, e.g 12h or 24h.

How does the prediction change if you use different hours after arrest? try to explain from a neurophysiological and biological perspective.

Project 5

Project 5: Text classification or Named Entity Recognition [TCNER]

5.1 Introduction

Project owner: Nacir Bouali and Maurice van Keulen
Primary topic: IENLP
Well combinable with DM or SEMI.

In the realm of information extraction and natural language processing, there are two suggested projects to select from

- **Text Classification**
  Recommend a conference to a researcher given the title of his new article
- **Named Entity Recognition**
  Extract and classify named entities from tweets

In case that a group wants to suggest a different project, the group should submit an initial project proposal to be reviewed. In this case the group is encouraged to meet with the project owner / topic teacher to discuss project ideas.

The suggested projects represent two challenges. For each challenge, you will be given a training set to train and tune your system on. The models should only be optmized using the training data. The test set then gives you a good estimate how your method behaves on unseen data.

Note that the project grading is not related to the achieved results, but rather on the methodology and the soundness of the evaluation.

5.2 Description of data set

The data for *Text Classification* is Conference Proceedings training data from the paper below.

> **Reference**: Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1 (March 2002), 1-47.`http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf`.

The data for *Named Entity Recognition* is NER Twitter training data available on Canvas.

References: (1) Nadeau, David & Sekine, Satoshi. (2007). A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes. 30. 10.1075/li.30.1.03nad. (2) Sharnagat, R. (2014). Named entity recognition: A literature survey. Center For Indian Language Technology. http://www.cfilt.iitb.ac.in/resources/surveys/rahul-ner-survey.pdf.

5.3 Description of challenge

5.3.1 Text Classification

- Design and implement a system to recommend a conference to a researcher given the title of his new article.
- The system should use the provided Conference Proceedings training data. You should implement the sub tasks (feature extraction, dimensionality reduction, and classifier) by yourself.
- You are free to select the algorithms you prefer for each sub task. However it is recommended that you test and compare multiple methods.
- Evaluate you system on the training set by using the cross-validation approach. Provide the confusion matrix of your system output.
- Evaluation should be done in terms of Micro-average precision, recall and F1 measures.
- Once you fount the best model on the training set, evaluate your model on the test set and report the results.

5.3.2 Named Entity Recognition

- Design and implement a system to extract and classify named entities in tweets.
- The system should use the provided NER Twitter training data. You should implement the sub tasks (feature extraction, and classifier) by yourself.
- You are free to select the algorithms you prefer for each sub task. However it is recommended that you test and compare multiple methods.
- Evaluate you system on the training set by using the cross-validation approach.
- Evaluation should be done in terms of micro-average of precision, recall and F1 measures.
- Evaluate your best model you found on the training set on the test set. Report your results.

5.4 Tips and suggestions

Project 6

Project 6: Automatic detection of Atrial fibrillation (AF) episodes [AF]

6.1 Introduction

Project owner: Mannes Poel
Primary topic: DM or TS

Atrial fibrillation (AF) occurs as a complication postoperatively form cardiac surgery. AF results in stasis of the blood. In the postoperative period AF can induce delirium and neurocognitive decline, thereby prolonging the hospital stay. [1] On the long term serious complications like thromboembolic diseases, stroke and heart failure can be induced by AF. These complications result in increased morbidity and mortality and prolonged hospital stays. [2-7] Precise ECG monitoring is important to detect AF as soon as possible. Then complications caused by AF can be obviated due to a fast intervention. The challenge of this project it is to develop an algorithm/method that can detect automatically episodes of AF (minimum of 30 seconds) from (preprocessed) ECG data. Framing it differently, the research questions is: "To what extent can one automatically detect episodes of AF?"

6.1.1 Background

Manual detection of AF in ECG record is time-consuming, especially in the case of large datasets consisting of 24-hour ECGs. When automating the detection, the physician can be deprived of work and research can be accelerated. Also, such an algorithm may result in the direct detection of AF during ECG monitoring, thereby creating the possibility for a fast treatment of AF. This underlines the need for an algorithm to automatically detect AF for analysis purposes. Automatic AF detection provides a faster analysis of long term ECGs. Hereby opportunities arise for better diagnostics and for gaining more insight into postoperative AF on a larger scale. Automatic quantification of AF may help to get insight in the yet unsolved underlying problem of AF.

AF is defined as a period of at least 30 seconds in which an irregular ventricular rate and P peaks are absent. [8] These two ECG characteristics indicate the rapid abnormal atrial activity seen in AF. An AF detection algorithm based on R-R interval irregularity is preferred, due to the prominence of QRS complexes, making it more robust to noise. [9] Therefore this algorithm is also based on the R-R interval irregularity.

6.2 Description of data set

In this project you will work with preprocessed ECG data from the Erasmus Medical Centre in Rotterdam of the department electrophysiology. Data was obtained within 10 days post-operatively of CABG surgery. Atrial fibrillation (an arrhythmia) occurs frequently after cardiac surgery. Atrial fibrillation is an arrhythmia that does not have to sustain constantly, it comes and goes, and therefore often passes by unnoticed in the patient's hospital stay.

ECG

From the patient a 12 lead ECG is recorded. A semi-automatic program analyzed the ECGs for annotation of the R peaks (see green dots in Figure 6.1). R peak detection was manually audited by a physician and atrial fibrillation or other arrhythmias were labeled.

Figure 6.1: Example ECG. The green dots annotate the R peaks

From this a text file was formed (Figure 6.2) with the time, R-R intervals in milliseconds, and more details on the observed rhythm.

Preprocessing

The algorithm you will build will be based on the R-R intervals, and more specifically on the irregularity of these intervals. This data has been preprocessed for you to make it easier to work with.

1. First of all the array of R-R intervals was split in arrays covering periods of 30 seconds. Each of this we will call a sample.
2. All samples for which no control could be made (due to artifacts, loss of server contact, limited data, etc.) were excluded.
3. All samples that included unphysiological high R-R intervals were excluded.
4. 30 bins of 50 milliseconds were created covering R-R intervals of 200 ms up to 1700 ms. For each sample the frequency of an R-R interval occurring in a certain bin was counted. See Figure 6.3.
5. These frequencies of occurring were then normalized.
6. All samples used were shuffled.

Excel file

In the excel file placed on blackboard is the data of 150.000 periods of half a minute of different patients. Around 36.000 of these periods are labeled as AF. Each row represents a sample. The first 30 columns show the values as explained above (normalized frequency of R-R interval in 30 bins). The 31st column is the label. Zero (0) means no AF was reported in this sample by the physician. One (1) means AF was reported by the physician. For students who want to investigate the raw data, this data can be made available on request.

[1] Alqahtani AA. Atrial fibrillation post cardiac surgery trends toward management. Heart Views. 2010 Apr 1;11(2):57.

[2] W.H. Maisel, J.D. Rawn, and W.G.Stevenson. Atrial Fibrillation after Cardiac Surgery; Review. Annals of Internal Medicine, 135(12):1061–1073, 2001.

23:56:21	845	N
23:56:22	845	N
23:56:23	855	N
23:56:24	845	N
23:56:25	840	N
23:56:26	840	N
23:56:26	845	N
23:56:27	830	N
23:56:28	825	N
23:56:29	825	N
23:56:30	835	N
23:56:31	815	N
23:56:31	815	N
23:56:32	820	N
23:56:33	830	N
23:56:34	830	N
23:56:35	835	N
23:56:35	835	N
23:56:36	840	N
23:56:37	850	N
23:56:38	850	N
23:56:39	835	N
23:56:40	830	N

Figure 6.2: Example of text file that is the result of the processed ECG. In the first column record time, second column R-R interval in milliseconds and in the last columns annotations regarding the rhythm have been made.

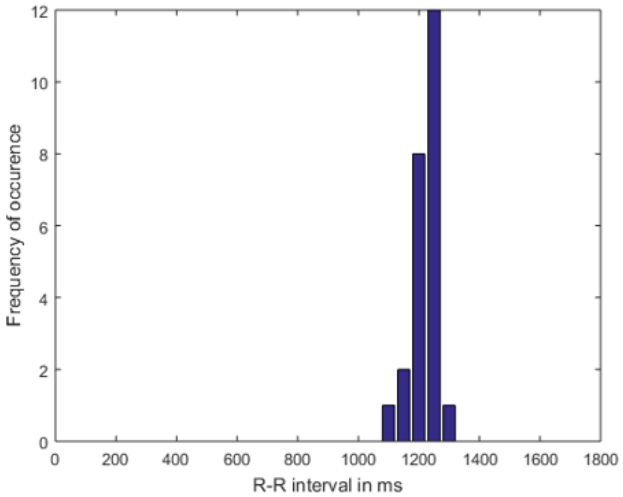

Figure 6.3: This is an example of how the data would look like visualized after step 4. Here you see that within this 30 second episode 24 R-peaks were detected mostly with an R-R interval around 1200 ms

[3] N. Echahidi, P. Pibarot, G. O'Hara, and P. Mathieu. Mechanisms, prevention, and treatment of atrial fibrillation after cardiac surgery. Journal of the American College of Cardiology, 51(8):793–801, February 2008.

[4] N.S. Peters, R.J. Schilling, P. Kanagaratnam, and V. Markides. Atrial fibrillation: strategies to control, combat, and cure. Lancet, 359(9306):593–603, February 2002.

[5] S.M. Narayan, M.E. Cain, and J.M. Smith. Atrial fibrillation. Lancet, 350(9082):943–50, September 1997.

[6] S.S. Chugh, R. Havmoeller, K. Narayanan, D. Singh, M. Rienstra, E.J. Benjamin, R.F. Gillum, Y.H. Kim, J.H. McAnulty, Z.J. Zheng, M.H. Forouzanfar, M. Naghavi, G. Mensah, M. Ezzati, and C.J.L. Murray. Worldwide epidemiology of atrial fibrillation: A global burden of disease 2010 study. Circulation, 129(8):837–847, 2014.

[7] M.P. Turakhia, M.D. Solomon, M. Jhaveri, P. Davis, M.R. Eber, R. Conrad, N. Summers, and D. Lakdawalla. Burden, timing, and relationship of cardiovascular hospitalization to mortality among Medicare beneficiaries with newly diagnosed atrial fibrillation. American Heart Journal, 166(3):573–580, 2013.

[8] V. Fuster, L.E. Ryden, R.W. Asinger, D.S. Cannom, H.J. Crijns, R.L. Frye, J.L. Halperin, and et al. ACC/AHA/ESC Guidelines for the Management of Patients With Atrial Fibrillation: Executive Summary A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the European Society of Cardiology Committee, volume 104. 2001, ISBN: 8006116083.

[9] N. Larburu, T. Lopetegi, and I. Romero. Comparative study of algorithms for atrial fibrillation detection. In 2011 Computing in Cardiology, pages 265–268. IEEE, 2011

6.3 Description of challenge

The goal of the project is to answer the research question "To what extent can one automatically detect episodes of AF?"

Method.

Of course one should apply DM techniques to answer this question based on the given data. This implies that you should:

1. Design/select different DM models, such as Decision Trees, which one you would like to validate. The selection of promising models can be based on related work. Which models did others use and what was their performance. Select 3 or 4 models which you want to compare on the given data set.
2. Validate (measure the performance) of the selected models on the given data set. Of course one is interested in the performance of the models on new, unseen, data. How do you estimate such a performance, and which performance measure to use (accuracy, precision, recall). Are the results sound, i.e. what you expect?
3. One problem with the data set is that it is imbalanced, much more non AF then AF samples. Look up in the literature how to deal which such imbalanced data sets.
4. Compare the different models and select the best one. Explain the selection procedure.
5. Finally answer the research question and discuss the practical applicability of an automatic AF detector.

6.4 Tips and suggestions

Here you can find some practical hints for dealing with the data.

- Determine how you will deal with the unbalanced data, much more examples of non-AF than AF.

- Visualize the difference in features between AF and non-AF.
- Inspect your results, for instance the decision tree. Does it contain sound features, does it make sense?
- It training takes too long you can sample the data and afterwards validate the most promising models on the whole dataset. Be aware that you do not filter out all AF instances.
- Determine if feasible which features are most predictive and does this make sense.
- If time allows determine new features (based on related work) and validate if this increases the performance.
- If feasible compare with human performance. Maybe some information on this can be found in the literature or on the web.

Project 7

Project 7: Linked Open Data [LOD]

7.1 Introduction

Project owner: Maurice van Keulen
Primary topic: SEMI

The project is about enriching data about cultural events:[1]

> Pick a theatre of your own choosing and demonstrate how you can enrich the data of that theater.

7.2 Description of data set

Data is not available directly, but based on information from the theater's website and established data sets and ontologies:

- Construct an RDF data set containing the information from the theatre's website
- Linking the data to established Linked Open Data sets, such as DBpedia.

7.3 Description of challenge

The added value of Semantic Web technology can be *demonstrated* by, among other things,

- Running example SPAQRL queries finding data in a way not possible with the original website's or Google's search facilities.
- Using the remote querying facilities of SPARQL to query Linked Open Data on the web.

Note that these are possibilities and suggestions. There actually is much freedom in this project to go your own way and to focus your attention. You need to, however, stay within the domain of cultural performances and the global goal of semantic enrichment. Minimal requirements for the project are

[1] You may choose to do something else than cultural events. This has been chosen because it is a semantically rich domain. If you want to focus on another domain, please consult the topic teacher to determine if the domain is suitable enough.

- It should enrich the original data with Linked Open Data on the web.
- It should demonstrate that now something non-trivial can be done.

Deliverables for the project are

- Presentation slides
- Report (PDF) explaining
  1. the specific goal of your project,
  2. the applied Semantic Web technogies,
  3. the resulting RDF data set and how it was constructed,
  4. the design of the software / system / website you developed, and
  5. examples that demonstrate non-trivial queries.
- Source files of all artifacts used and produced in the project such as data sets, web pages, software code, example queries, etc.

Please combine the above files in a PDF file. Source code can be put in appendices. Submit the PDF or PDFs to Canvas.

The grading for the project is based on the following principles

- How much *understanding* is shown
- The depth and quality of the developed components
- The power and potential of the enrichment

7.4 Tips and suggestions

Project 8

Project 8: Music album deduplication [ALBUM]

8.1 Introduction

Project owner: Maurice van Keulen
Primary topic: PDBDQ or DINT

This project is about (probabilistically) deduplicating semantic duplicates in a given data set about music albums.

A semantic duplicate is a set of different data items that actually refer to the same entity/object in the real world. In this case, one can have two or more data items describing a music album with slightly different strings for the album, artist and song names, or there may be songs missing, typos, language differences, and many more possible problems.

The approach suggested here is:

- *Match* the data from the discs to determine a *matching score* for each pair of discs.
- Set *two thresholds* τ_b and τ_t that define three classes of matches:
  - **Non-match** any match below τ_b is for sure **not a match**
  - **Match** any match above τ_t is for sure **a match**
  - **Uncertain** for any match between τ_b and τ_t it is uncertain whether or not it refers to the same disc or not.
- *Merge* the matching and uncertain matching discs.
- *Evaluate* the result using the "ground truth" data provided.
- Experiment with and evaluate any other aspects that interests you.

8.2 Description of data set

The data set about music albums is `cddb_discs.xml`. The data is a list of discs representing music albums. The ground truth is `cddb_9763_dups.xml` where the duplicates are already indicated in the data as it is a list of pairs of music albums. Source of this data can be found at `http://hpi.de/naumann/projects/repeatability/datasets/cd_datasets.html`

It is also possible to bring your own data. In that case, you need to pay attention to the following

- You need approval from the project owner / topic teacher.
- For topic DINT, it needs to have an element of content-based matching.
- For topic PDBDQ, it needs to have an element of uncertain matching as well as other imperfections such as inconsistency or untrustworthiness (possibly artificallyl created).

8.3 Description of challenge

The goal of the project is

> Write a program in a language of your own choosing that reads the discs file and integrates the duplicates.

If you have chosen topic Data Integration [DINT] , then use its knowledge and skills for determining a best matching step. A ground truth is provided for evaluating your matching. You can set $\tau_b = \tau_t$ to leave no uncertain matches.

If you have chosen topic Probabilistic Databases and Data Quality [PDBDQ] , then use its knowledge and skills for establishing a merging step that faithfully captures all uncertainty due to matching and inconsistency in the merging.

If you have chosen both DINT and PDBDQ topics, then you are of course well equipped to do both: determining a best matching *and* a best merging step.

The file `cddb_9763_dups.xml` containing the correct duplicates (the "ground truth") is provided for evaluation.

Deliverables for the project are

- Presentation slides
- Report (PDF) explaining
  - The specific analytical goal for the integration.
  - The identified data quality problems.
  - The matching algorithm(s) used (and why).
  - The merging algorithm(s) used (and why).
  - The results and interpretation of an evaluation.
  - (PDBDQ) How you demonstrate the strength of probabilistic representation of data quality problems in your project.
  - Conclusions and recommendations.
- Provide as appendices source files of all artifacts used and produced in the project such as datalog programs, raw data files, data conversion scripts, example queries, etc.

It doesn't matter what your main focus is or which other tools and programming languages you use; it does matter how well well you have accomplished the task you have set out to do and how much understanding you show.

8.4 Tips and suggestions

It is possible to 'skip' the matching step by (mis)using the ground truth. So, if you have only chosen the PDBDQ topic and want to (first) focus on the merging step, it will provide you with the actual duplicates in the data set.

Note also that for applying the knowledge and skills of PDBDQ, a perfectly functioning matching step is not necessary. If the matching is rough, identifying many possible matches, the approach still works provided your lower threshold τ_b is low enough to include the correct matches.

Project 9

Project 9: Referral Advice [RA]

9.1 Introduction

Project owner: Mannes Poel
Primary topic: DM

Low back pain (LBP) is the most common cause for activity limitation and has a tremendous socioeconomic impact in Western society. When people get low back pain (LBP), it is not always evident what the correct referral advice is. A direct correct referral is essential for effective treatment to prevent the development of chronic LBP the utmost. For a data driven approach to getting more insight in the process of referral advice referral advice was stored in a database provided by the Groningen Spine Center (GSC) containing patient reported answers to questionnaires and the advised treatment, see the corresponding dataset (CSV file) for more information on the features and possible treatments.

We would like to know to what extend one can automatize the referral advice by constructing a data driven Decision Support System, using Data Mining, and we also want to know which features are relevant in this referral advice, since this could shorten the questionnaire.

9.2 Description of data set

The dataset contains data of 1547 real patient cases on low back pain that were judged by healthcare professionals on referral advices: 1. advice, 2. rehabilitation, 3. surgery, 4. injection/medication, 5. combination of 1-4. Explanation of the features can be found in the dataset (CSV file) accompanying this project. Be aware that the dataset has missing values.

9.3 Description of challenge

The research questions of this assignment are:

1. To what extend can one, based on the answers given in the questionnaire, automatically determine the advised treatment for a patient with back pain.

2. To what extend can one reduce the number of questions, i.e. which questions can be left out without major decrease in performance?

9.4 Tips and suggestions

Some tips for this project.

- Most features have (a lot of) missing values. Design and implement methods to deal with this.
- Evaluate techniques beyond Decision Trees and Regression.

Project 10

Project 10: Transport [TRANSPORT]

10.1 Introduction

Project owner: Luc Wismans
Suitable primary topic: DPV, DM
Possibly combinable with TS as the data can be viewed as a time series.
Furthermore, it is useful to have some knowledge on GIS software packages like the open software QGis.

Transportation is about moving of people and goods from A to B. Being able to transport people and goods is a prerequisite for economic growth and the consequence of the separation of production and consumption. There are various means of transportation (e.g. car, train or bicycle) being serviced by private companies or public authorities. Although transportation brings utility it also comes with a cost. Unwanted side effects (i.e. externalities) are caused by transportation affecting for instance the air quality, climate, safety and noise. Furthermore, there is a difference between the user needs and resulting behaviour and the societal needs and desired behaviour. Simple examples show that individuals pursuing their own objectives (e.g. shortest travel time from A to B) does not result in the optimal situation for society as a whole (e.g. minimal total delay in the system).

Road authorities are always working on improving the transport system balancing the societal objectives related to economic growth, minimizing externalities and user needs (i.e. sustainability). For this purpose they can adapt the system taking hard measures (infrastructural changes including deployment of intelligent transport systems) or influence the system providing services like traveller information. In most cases the infrastructure is owned by and a responsibility for governmental authorities as were the services provided on these networks. In the case of services these were at least controlled by the government (e.g. public transportation and provision of information). However, the past few years there is a shift of services provided by private parties not only because governmental authorities allow them to do so (providing data to such parties), but also as a result of an increase in data availability as well as ICT technologies not necessarily for transportation purposes deployed by private companies. Loop detector data, GPS, GSM, Bluetooth, WiFi, camera, smart card data, AVL and dedicated smartphone apps are examples of the sources capable of providing data of interest for transportation. Accurate maps/ topology of networks are needed to be able to map these data sources, offering the opportunity to connect and interpret this data for transportation purposes. Other spatial and temporal types of data like, socio economic data, points of interests, weather, deployment of measures, time tables and lines of PT might be of interest because of correlations with traffic conditions.

This data is obviously of interest for governmental authorities and private parties, because it allows to

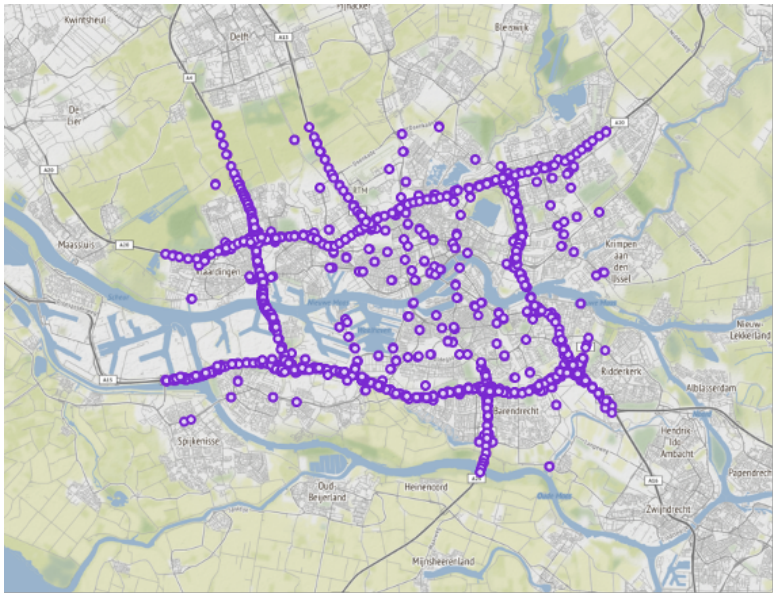

Figure 10.1: Rotterdam area.

improve the decision support information for themselves or their customers. The enormous increase of data availability opens opportunities to better understand the current transport system (e.g. what are the traffic conditions, where are problems and when do they occur, and how do traffic conditions change as a result of construction works), to monitor the transport system (e.g. route choice effects of measures taken), as well as to improve predictions of the future (e.g. what will be the traffic conditions in the coming hour, what will be the travel time from A to B tomorrow during rush hour, etc.). The recent COVID-19 outbreak drastically influenced society as well as mobility of people, not necessarily in a negative way (intelligent lockdown started March 12th). Number of trips, use of modes and resulting traffic conditions changed and can be monitored using data.

Advanced Domain topic There are also more advanced challenges mentioned that you can do in this project. In this advanced transport domain topic we focus on a specific challenge related to state estimation and data driven predictions of traffic conditions or deeper analysis of COVID-19 implications on mobility. Historic data provides the opportunity to recognize and cluster traffic condition patterns which can be used for predictions (tomorrow or next week, or given roadworks), or to train a model to predict future traffic states given the current measurements (next minute or next hour) using machine learning techniques or time series approaches, but also to complete/estimate data in case measurements are missing providing the current traffic conditions. There are several methods which can be used for this purpose. Furthermore, the data available also provides the opportunity to get a deeper understanding of changes in mobility during COVID-19 (just after the intelligent lockdown versus weeks later) using the analysis of the evolution in traffic states. Knowledge of traffic flow dynamics is advised, which is extensively explained in the book of Treiber and Kesting (Treiber, M. & A. Kesting (2013). Traffic flow dynamics. Springer-Verlag Berlin Heidelberg. ISBN 978-3-642-32459-8, DOI 10.1007/978-3-642-32460-4. Especially Part I and Part III will be of interest, but it doesn't hurt to have a look at Part II as well ;-).

10.2 Description of data set

For this domain there are several data sources available. All data has been provided by NDW (`https://ndw.nu/`) for two time periods (i.e., March 1st – April 5th and April 27th – May 31st) for the Rotterdam area (see Figure 10.1). See also `http://opendata.ndw.nu/` for more information.

1. Data delivered by NDW contains

- Flows and speeds from loop detectors, 1 minute aggregates (CSV-files)
- Travel times of predefined routes (CSV-files)
- Status information on occurrence, whole of Netherlands (CSV-files)
  - opening bridges,
  - road works,
  - traffic jams
  - incidents
- Cycling data providing counts, 1 minute and 1 hour aggregates (CSV-files)

2. Shapefiles with measurement locations and trajectories
3. Optionally: access to Floating Car Data (FCD), i.e. speed data derived from FCD can be arranged. This needs signing of contract and provides access to the data lake Dexter of NDW.

10.2.1 Description NDW data speed, flow and traveltime

More information can be found via `http://opendata.ndw.nu/`, however most of the information is in Dutch

Datasets:

1. Datasets containing all measurements:
   - Filenames speeds and flows:
     - intensiteit-snelheid-rotterdam-1-7-Maart.zip
     - intensiteit-snelheid-rotterdam-8-14-Maart.zip
     - intensiteit-snelheid-rotterdam-15-21-Maart.zip
     - intensiteit-snelheid-rotterdam-22-29-Maart.zip
     - intensiteit-snelheid-rotterdam-29-Maart-4-april.zip
     - intensiteit-snelheid-rotterdam-27-april-3-mei.zip
     - intensiteit-snelheid-rotterdam-4-10-mei.zip
     - intensiteit-snelheid-rotterdam-11-17-mei.zip
     - intensiteit-snelheid-rotterdam-18-24-mei.zip
     - intensiteit-snelheid-rotterdam-25-31-mei.zip
   - Filename travel times:
     - reistijd-export_rotterdam_1 - 22 maart.zip
     - reistijd-export_rotterdam_23 maart - 5 april.zip
     - reistijd-export_rotterdam_27 april - 17 mei.zip
     - reistijd-export_rotterdam_18 mei - 31 mei.zip
   - Filename cycling data:
     - fiets-data-export_rotterdam_1 maart - 5 april_minuutniveau.zip
     - fiets-data-export_rotterdam_27 april - 31 mei_minuutniveau.zip
     - fiets-data-export_rotterdam_1 maart - 5 april_uursniveau.zip
     - fiets-data-export_rotterdam_27 april - 31 mei_uursniveau.zip
   - Filename status data:
     - sb-files-export_Rotterdam 1 maart - 31 augustus.zip
     - sb-wegwerkzaamheden-export _ Rotterdam 1 maart - 31 augustus.zip
     - sb-incidenten-export_rotterdam 1 maart - 31 augustus.zip
     - sb-bruggen-export_rotterdam 1 maart - 5 april.zip
     - sb-bruggen-export_rotterdam 27 april - 31 mei.zip
2. Datasets containing metadata:
   - Filename speeds and flows:
     - intensiteit-snelheid-export 1 maart_metadata.zip

  - intensiteit-snelheid-rotterdam-1 mei_metadata.zip

- Filename travel times:

  - reistijd-export_rotterdam_1 maart metadata.zip
  - reistijd-export_rotterdam_1 mei metadata.zip

3. Additional datasets:

   - Fietstellingen_NDW-MRDH.Rapport.2019-06.xlsx
   - 062_Levering_NDW_Shapefiles_20200828.zip

Files within zip-files stated under 1 and 2: CSV-files: "," separates fields.

The raw data used by NDW are 1-minute aggregates. This means that the available datasets contain the raw data of NDW (only for cycling data there are also datasets on 1 hour aggregates available). As a result several fields are not filled, because these are used when higher aggregates would be delivered by NDW, furthermore not all fields are filled for all measurements.

For some of the datatypes, datasets containing all measurements need to be combined with the datasets containing metadata to add e.g. locational data (i.e. to be able to connect the measurements to the location of measurement). The metadata contains 1 day of data containing all available data and the measurements for one day. Because the available locations and information can change over time, two days of metadata are provided. Furthermore, if there is data available per lane and/or data per vehicle class (e.g. cars and trucks), these measurements for the same location and same minute will be presented in a separate line. For this purpose you also need to combine the complete dataset with the metadata file. Cycling data does not yet contain latitude or longitude information. However most of the locations can be found using Fietstellingen_NDW-MRDH.Rapport.2019-06.xlsx

Important attribute fields for example are:

- Id_meetlocatie and index: to connect the metadata information to all measurements
- start_meetperiode and eind_meetperiode indicating the timeinterval of measurement
- gem_intensiteit: measurement of flow
- gem_snelheid: measurement of speed
- rijrichting: side of road
- rijstrook_rijbaan: lane nr, number starts left
- voertuigcategorie: vehicle class
- start_locatie_latitude and start_locatie_longitude: lat-lon location of measurement (or for travel time starting point) based on ETRS89 system, which is the same as WSG84

Data can be connected with a network (e.g. NWB, downloadable via: https://nationaalwegenbestand.nl/index.php/aanbieders). For this you need at least the location data. For this purpose additional use can be made for at least speed, flow and travel time data combining the location data presented in the NDW shapefiles: 062_Levering_NDW_Shapefiles_20200828.zip, using the field "dgl_loc".

10.3 Description of challenge

We provide you with four example challenges you could work on in your project.

1. State estimation:
   The data provided for Rotterdam for the same time period contain different sources which can be used to estimate the traffic conditions (speeds and flows) on the entire network as well as derived performances on level of congestion or total delay in the network. NDW data provides speed and flow measurements on locations based on all passing vehicles. NDW also provides travel times on routes and optionally FCD data. The challenge is to use this data to provide and visualize the spatio-temporal state estimates as complete as possible (e.g. providing space-time diagrams for roadstretches; see Figure 10.2).

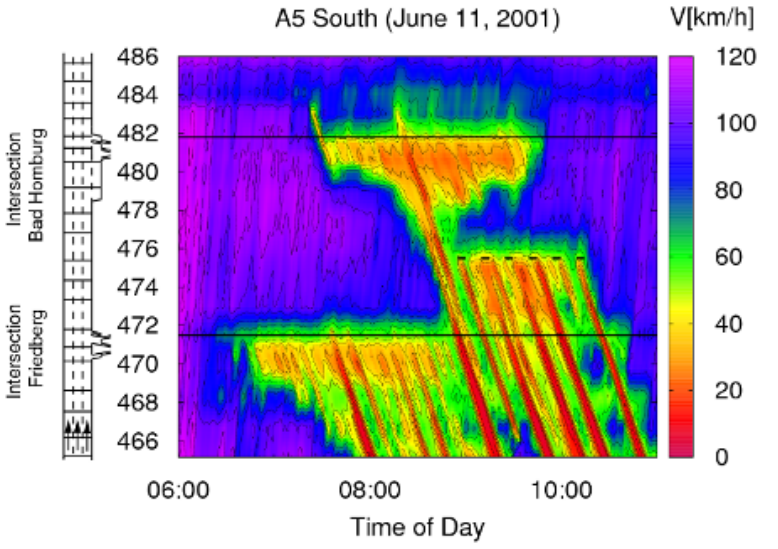

Figure 10.2: Example of space time diagram illustrating spatiotemporal dynamics

Advanced challege: Is it possible to estimate missing values to complete data or to reduce measurements (e.g. what if a loop detector is removed or fails to deliver information, is it possible to estimate the state on this location using the other measurements).

2. **Advanced challenge**: State estimation, one area indicator
   Can you develop an indicator which can summarize the available spatiotemporal data into one or two indicators, which directly shows the state (for inspiration, look for example at the Macroscopic Fundamental Diagram. For more information, see: `http://faculty.ce.berkeley.edu/daganzo/Research/MacroFD/index.html`) of an area and/or the changes over time.

3. Influence of COVID-19:
   The data provided (provided and state estimations (see challenge 1 and 2) can be used to analyse the differences between the period before the intelligent lockdown and directly after as well as several weeks after. The challenge is to analyze the impact, by e.g. showing to what extent there was an increase or decrease of cars using the network compared with for example cyclists, or (**advanced challenge**) visualising this evolution or differences, or analyzing the changes using for instance challenge 2.

4. Correlations:
   The provided NDW data also contains information on for example traffic measures, road works, incidents and bridge openings, but there are also several other sources (open) available. This challenge is about analysing the impact of these aspects on traffic conditions (e.g. classification of incidents based on their impact or the use of bicycles depending on weather conditions; or **advanced challenge**: conditions connecting the status information with the traffic operation data (flows, speeds and traveltimes)). Are there correlations and can we derive knowledge / estimate models which we can use for future decisions.

5. **Advanced challenge**: Prediction of traffic conditions
   Historic data on traffic conditions provides the opportunity to predict future conditions. Using artificial intelligence techniques like learning algorithms or pattern recognition it might be possible to predict traffic conditions for the next minute, next hour or even the next day or next week. The challenge is to build such predictor and test its ability to predict the future. Special attention should be paid on the ability to predict non-regular traffic conditions and the ability of predicting correctly and timely if traffic conditions change from free flow towards congestion and vice versa.

For all challenges it is required to connect the available data with a network and provide visualizations. First steps could be:

- Analysing and understanding of data set, e.g. by making figures of measurements for a specific location or locations for a day or multiple days, computing averages, checking plausibility, determine whether there is data missing, etc
- Visualize data geographically
- Select suitable part of network for case study and determine what data is available for this part and which is not
- Determine desired outcome and possible ways to compute this
- ...

10.4 Tips and suggestions

10.4.1 Creating new geographic maps for use in Tableau

If you'd want to show data using a geographic map and the available maps in Tableau do not suffice, then you can also import your own from *shapefiles* using a Geographic Information System (GIS) such as QGIS or ArcGIS. A shapefile is a special file that stores the polygones that make up a map.

See `http://onlinehelp.tableau.com/v10.1/pro/desktop/en-us/help.htm#maps_shapefiles.html%3FTocPath%3DDesign%2520Views%2520and%2520Analyze%2520Data%7CBuild%2520and%2520Use%2520Maps%7C_____1` for details.

Project 11

Project 11: Decision support for University timetables [TIMETABLES]

11.1 Introduction

Project owner: Maurice van Keulen (Original source of the data: Rudy Oude Vrielink)
Primary topic: DPV

Every year students in higher education grade the institution they are studying at. The National Students Enquiry (NSE) shows that many higher education institutions have work to do concerning their timetabling. The combined results give us information about the opinion on many topics. Unfortunately, the timetabling is one of the processes that are hard to improve but is considered below standard in Twente, being at both the UT and Saxion. You may consult the Elsevier's survey, (in Dutch), at: `http://onderzoek.elsevier.nl/onderzoek/beste-studies-2015/17/universiteit-twente-enschede/1152`. Although the results are not unsatisfactory, we think that we should spend effort to obtain improvement. You will dive into the results of timetabling and advise us how to do that.

The project has the following deliverables to be submitted on Blackboard.

1. Slides of your presentation
2. Report
3. Any source files or intermediary data files

11.2 Description of data set

You have the data, consisting of all timetables from all educational programmes of several years, from both organisations.

Note that the data may be considered incomplete in some ways. This is not something that we 'fix', because it is a real life situation. In daily life you almost never find complete datasets, all cleaned and ready to be explored. Nevertheless, should you need to find more data or other views on this data, please consult the following websites: roosters.saxion.nl or rooster.utwente.nl, where you can find the source data of the given timetables.

Please note that a lot of this data contains names, mostly of teachers. Since the data is freely available on the internet, there is no issue with privacy. Nevertheless, it is considered wise and decent to not spread this data outside our institutions. Please do not share, forward, mail, copy or in any way distribute this data, because it is meant solely for the purpose of learning and advising.

Blackboard contains a lot of files concerning the timetables of Saxion and UT. The data files are called "ActivitiesUT_yyyy1-yyyy2_v2.xlsx", with yyyy1 being the first, and yyyy2 the last year of the file.

11.3 Description of challenge

The UT and Saxion are both planning to start a programme in the field of education logistics. This programme is set up as a series of interrelated projects in an action design research method. This means: small steps, starting with scientific research with pilots, giving the results as advice to the organisation before they make decisions, then setting up the project based on the advice followed by the implementation. Results after implementation go back to being researched, and the next step can be taken. The research you are about to be doing, is one of those steps.

You have the data, consisting of all timetables from all educational programmes of several years, from both organisations. These data can be researched in order to advise the organisation. As explained below, there are three distinct strategies that one can follow. Formulate a clear objective for your project that is based on a combination of strategies and an concrete focus that determines the relevant KPIs, questions, and/or trends.

Please remember that the goal of this project is to give your advise to the management. What is your interpretation and what should UT and/or Saxion do, after your study of the facts and figures?

11.3.1 Strategy 1: Compliance

Given the set of Key Performance Indicators (KPIs) per institution, check to what extent the timetables comply with their own set. In case you interpret the given sets of KPIs as a 'wish list' more than as a list of measurable indicators, you may feel free to translate the set of KPIs to more measurable ones. See to what extent the timetables comply with the KPIs, but also look at the other rules and preferences. To what extent do the timetables follow those? What can you tell?
(Congratulations! You have just done what almost no higher education institution has ever done before!)

11.3.2 Strategy 2: Exploration

This task is about exploring the data and looking for pattern that nobody has seen before. Examples of questions are:

- How far do students and teachers have to walk on campus each day?
- How much waste (free hours in between classes) do teachers have?
- How many free hours do teachers and students have, on average? Who has the most?
- How many contact hours do teachers and students have? Which course has the most?
- What is the maximum used capacity for each building? For each course/programme? For each faculty? What is the minimum?
- How many inconsistencies do you see?
- How much do teachers or students have to walk around campus to follow courses? Who is leading with the most kilometres per day/week/module?
- How many hours does each course schedule, on average, at minimum, at maximum?
- How many students are in every course? Per week/module?

- Can you compare the contact hours of the 2nd year in 2013/14 (non-TOM education) to 2014/15 (TOM education)? How much is the increase in contact hours due to the introduction of TOM?
- Do the same for 1st years.
- What teacher teaches with most other teachers?

Write down your findings in a report directed to the management.

11.3.3 Strategy 3: Trend analysis

Both Strategy 1: Compliance, and Strategy 2: Exploration, indicate which persons, courses, are important *now*. In Strategy 3: Trend analysis, we will analyse the data as time series data. Divide the data in sections of a module, a quarter, a semester, or a year, ranging from September 2013 to July 2016. What can you tell about the trends that you see? Also compare your data between both institutions, UT and Saxion. Examples of analyses that you might do are:

- Plot how many courses are given in each time span;
- Determine the correlation using regression analysis or Pearson's correlation;
- What lecture halls gained importance over the years? (i.e. what are trending halls?)
- What lecture halls show decreased importance over the years? (i.e. what are nostalgic halls?)
- What teachers / faculties / gained importance over the years?
- etc.

11.4 Tips and suggestions

11.4.1 Creating new geographic maps for use in Tableau

If you'd want to show data using a geographic map and the available maps in Tableau do not suffice, then you can also import your own from *shapefiles* using a Geographic Information System (GIS) such as QGIS or ArcGIS. A shapefile is a special file that stores the polygones that make up a map.

See `http://onlinehelp.tableau.com/v10.1/pro/desktop/en-us/help.htm#maps_shapefiles.html%3FTocPath%3DDesign%2520Views%2520and%2520Analyze%2520Data%7CBuild%2520and%2520Use%2520Maps%7C_____1` for details.

Project 12

Project 12: Web Harvesting for Smart Applications [SDSI]

12.1 Introduction

Project owner: Maurice van Keulen
Primary topic: SEMI, DINT, or IENLP
Good to combine with PDBDQ, DPV, or DM

The High Tech Systems Park of Thales in Hengelo[1] is also used as a kind of laboratory, called "Fieldlab The Garden".[2] One of the projects making use of this lab is "Secure Data Sharing Innovation" (SDSI). Part of this project focuses on the development of and experimenting with *Smart Applications*.

This project is related to this activity. The idea is that with technology that can autonomously and robustly harvest data from the web, one can develop smart applications. For example, finding indications of possible unknown side effects of medicines. One could harvest all messages from a web forum for a certain disease, extract information about (a) medicines people report using and (b) which side effects they report having, and compare that with the leaflets of these medicines to determine if some reported side effects are unknown (i.e., not mentioned in the leaflet).

This project doesn't have one specific challenge, but we describe two example challenges with the option to define your own using the examples as inspiration for what you could do. The main goal of this project is to demonstrate the potential of data harvested from the web for developing smart applications.

Topic SEMI is related to the processing of the harvested web pages (typically HTML, which is a dialect of XML). Extraction of the medicines and side effects can be done using the knowledge and skills of topic IENLP (as a classification task) or DINT (as a matching task) or both (evaluating which performs best). PDBDQ is related to the faithful representation of the uncertainty in the extracted information. For an analytical purpose of the smart application, one can develop a predictive model (DM) or a dashboard with visualizations (DPV).

[1] http://hightechsystemspark.com/about-high-tech-systems-park/
[2] http://hightechsystemspark.com/smart-industry/the-garden/

12.2 Description of data set

There is no given data set, but you are expected to choose a website yourself and harvest data from it. There is a requirement to use data from at least two sources, where at least one is harvested from the web. The other could, for example, also be a linked open data set (see topic SEMI).

There are web harvesters / crawlers available on the web that you can use as a service. You can also write your own program that fetches pages (websites that use JavaScript to dynamically load data and construct a web page in the client, can be harvested by using the FireFox browser in *headless* mode, see `https://developer.mozilla.org/en-US/Firefox/Headless_mode`. In certain circumstances, you could resort to manually saving the individual web pages, but that obviously has its limitations.

12.3 Description of challenge

The overall challege is to

> Demonstrate potential of data harvested from the web for developing smart applications by
> 1. integrating data from at least 2 sources (at least one is harvested from the web), and
> 2. demonstrate the potential by providing analytical results (visualizations or a predictive model) or by designing and implementing a proof of concept of a smart web/mobile app

Please find below two example challenges. You could focus on one of these or define your own web harvesting-based smart application that complies with the above conditions.

12.3.1 Example challenge: Online Healthcare Communities

Nowadays, ubiquity of online communities gives us invaluable dataset in the form of electronic peer-to-peer communication, which may help us understand social influence and collective behavior dynamics. Especially, to some extent, messages exchanged in health-related online communities can reflect the intricacies of human health behavior as experienced in real time at individual, community, and societal levels [**?**]. Relevantly, in some online platforms, the online community is not just a place for the public to share physician reviews or medical knowledge, but also a physician-patient communication platform [**?**]. Those platforms can be seen as a solution for lack of medical resources in developing countries (ibid). That is, the online health care community is a potential solution to alleviate the phenomenon of long hospital queues and the lack of medical resources in rural areas. However, except a few papers [**?**, **?**], online healthcare communities have received little attention and investigation efforts.

This project is related to one of the online healthcare communities, called *CancerConnect* (`http://cancerconnect.com/`). The idea is that with technology that can autonomously and robustly harvest data from this website, one can develop smart applications and also data analytics-based implications. For example, one could harvest all posts and comments of active users in one of the communities (like Breast Cancer community on this platform), and then, finding indications of possible unknown side effects of medicines. In more detail, one could harvest all messages from a web forum for a certain disease, extract information about (a) medicines people report using and (b) which side effects they report having (topic IENLP!), and compare that with the leaflets of these medicines to determine if some reported side effects are unknown (i.e., not mentioned in the leaflet).

The challenge is to analyse the data according to interesting questions you come up with, potentially leading to new insights. You could use different approaches including visualizations, natural language processing, semantic analysis or data mining. Additionally, it is possible to build network of interactions among users (who has commented/liked/disliked the posts of whom) across times and gain some insights about the community.

12.3.2 Example challenge: Mining Crowd-based Inventions

Companies are increasingly turning to distributed groups of volunteers to help them design their new products and services [**?**]. This trend, sometimes referred to as Open Innovation or Crowd Design, takes place

across industries and scales. These communities leverage the creative power of thousands of volunteers, often creating novel design solutions at unprecedented speeds. However, building and maintaining an effective design community is no short order: we need to understand how they develop and evolve over time if we wish to create mechanisms that support their growth and effectiveness.

While researchers have studied collaboration networks, few have explored the growth of these recent design networks. This project aims to analyze the growth and evolution of OpenIDEO (`https://www.openideo.com/`), a successful online open innovation community centered around designing products, services, and experiences that promote social impact. In particular, this project's target is to addresses how the design network and its members have changed over time by using network analysis techniques on collaboration data from OpenIDEO.

On OpenIDEO, each challenge has a problem description and stages — e.g., Inspiration, Concepting, Applause, Refinement, Evaluation, Winning Concepts and Realisation — where the community refines and selects a small subset of winning ideas, many of which get implemented or funded. During the 'Concepting' stage, participants generate and view hundreds to thousands of design ideas; in practice, the number of submissions make exhaustive review (even of the titles) impossible e.g., for a medium-sized challenge of 600 ideas, it would take a person over 25 hours to read all entries.

This project's investigation hopefully can cover multiple scales of the design network: the network itself as a whole, the community structure within that network, and how actions of individual members contribute to its overall behavior. Alternatively, another ideation website, Quirky (`www.quirky.com`) can also be explored.

The idea is that with technology that can autonomously and robustly harvest data from ideation websites, one can develop smart applications and also data analytics-based implications. For example, one could harvest all ideas and comments of all users, from ideation and design, till launch stages, and then, finding hidden patterns among successful inventions and users. In more detail, one could harvest all tags/pictures/-text from OpenIdeo or Quirky, and store all information in a structured way, and for example, tracked comprehensive information on all ideas that successfully passed through the ideation and development phases. This includes the sub-task inputs the ideas received (e.g., number of contributors, contents) from co-creators during all phase.

The challenge is to analyse the data according to interesting questions you come up with, potentially leading to new insights. You could use different approaches including visualizations, natural language processing, semantic analysis or data mining. Additionally, it is possible to build network of interactions among users (who has commented/liked/disliked the posts of whom) across times and gain some insights about the community.

12.4 Tips and suggestions

Network analysis tools are not part of a topic taught. If you want to research in this direction helpful tools are gephi and the networkx library for Python.

Project 13

Project 13: Data Science 4 E-Sports: The Case of FIFA 21 [ESPORTS]

13.1 Introduction

Project owner: Guido Bruinsma
Primary topic: Computer Vision + DM or DPV

A prominent development in the game industry that is winning ground quickly is eSports (with the Olympic status in 2024). Professional gamers focused on games such as FIFA20, CSGO, League of Legends and Rocketleague, compete in professional global competitions. Similar to professional (traditional) sports teams in games such as soccer or speed skating, eSports teams are composed of athletes, sport psychologists, cognitive scientists, nutritionists, and health specialists. Another important group of professionals forming the backbone for in-game analysis are data scientists that analyze, interpret and visualize in-game performance.

All this effort is aimed to optimize performance of the eSporters. Minor differences in the setup, training or (psychological, physiological or performance) can have major impact on in-game performance. For FIFA 21the University of Twente is teaming up with diverse parties (pro gamers and pro teams, coaches) with the aim to develop tools for eSporters to give them that edge in FIFA (`https://www.utwente.nl/en/news/2019/10/251292/esportslab-at-university-of-twente`). It does this by providing, interpreting and visualizing their in-game data.

13.2 Description of data set

The data set consists of

- 300+ and still growing mid and end game screens (printscreens) showing game performance statistics. (see Figure 13.1) for an illustration.
- Video footage of gameplay from FIFA players on youtube (e.g., Bryan Hessing `https://www.youtube.com/c/BryanHessing/videos?view=0&sort=dd&shelf_id=1`). See Figure 13.2 for an illustration.

Figure 13.1: Gameplay statistics (end screen image)

Figure 13.2: Minimap (screen cast video)

13.3 Description of challenge

The challenge has been split into two individual projects, one that bases itself on the end screens (image data) and the other on screen casts (video data):

ESPORTS-IMG Automatic extraction, analysis, and visualisation of meaningful gameplay statistics from the end screens of FIFA matches.

ESPORTS-VID Automatic extraction, analysis, and visualisation of information on gameplay (heatmap, gameplay development) that is hiding in the minimap of gameplay screencasts from Youtube or other recorded gameplay video material available from the esportslab. The minimap provides information on the whereabouts of all players during the game. Information that can be extracted can be:

- heatmap of the most played areas on the field (by the player)
- defense (in, outside 16m on own side), midfield (own side, opponent side) attack (in, outside 16m on opponent side), general
- build up tactics that lead towards specific in game moments: i.e. goals, and goals against
- shifts in player focus (soccer player that is being controlled)

13.4 Tips and suggestions

Visualisation: `https://www.futwiz.com/en/gamestats`
Neural networks and FIFA: `https://towardsdatascience.com/building-a-deep-neural-network-to-play-fifa-1`

Project 14

Project 14: Energy Disaggregation [ED]

14.1 Introduction

Project owner: Elena Mocanu
Primary topic: TS
Good to combine with DPV or DM.

Energy is a limited resource which faces additional challenges due to recent efficiency and de-carbonization goals worldwide. An important component of the ongoing process is the improvement in the building energy management systems, which account for 30-40% of the total energy demand in the developed world. Buildings are complex systems composed by a different number of devices and appliances, such as refrigerators, microwaves, cooking stoves, washing machines etc. However, there are also a number of electric sub-systems, e.g. electric heating, lighting. Even there are many influencing factors in building energy consumption, some patterns can be clearly identified and used further to improve demand side management systems and demand response programs.

Disaggregation occurs when a quantity is divided into its component parts [1]. Energy disaggregation, also referred to as a non-intrusive load monitoring (NILM) problem, is the task of using an aggregate energy signal, such as the one coming from a whole-home power monitor, to make inferences about the different individual loads of the system. Traditional approaches for the energy disaggregation problem (or NILM problem) start by investigating if the device is turned on/off, and followed by many steady-state methods and transient-state methods aiming to identify more complex appliance patterns.

Moreover, new data analytics challenges arise in the context of an increasing number of smart meters, and consequently, a big volume of data, which highlights the need of more complex methods to analyze and take benefit of the fusion information. More recent researches have explored a wide range of different machine learning methods [2-5], using both supervised and unsupervised learning, such as sparse coding, clustering or different graphical models to perform energy disaggregation.

Time series specific techniques used in the pre-processing steps (e.g. correlation between appliances [2], windowing [4], additional features or feature extraction [5]), have proven useful in improving the performance of the classification, regression or clustering methods. Still, there is an evident challenge to develop an accurate solution that could perform well for every type of appliance.

14.2 Description of data set

The Reference Energy Disaggregation Dataset (REDD) is an open dataset collected specifically for evaluating energy disaggregation methods. The REDD dataset can be downloaded from `http://redd.csail.mit.edu/` with the id/password: redd/disaggregatetheenergy.

We propose to focus on low_freq.tar.bz2 (at 1Hz) from REDD dataset. It contains detailed power usage information from six buildings over few weeks sampled at 1 second resolution, together with the specific data for all appliances of each building at 3 seconds resolution. More information about REDD dataset can be found in

J. Z. Kolter and M. J. Johnson, "REDD: A Public Data Set for Energy Disaggregation Research", In proceedings of the SustKDD workshop on Data Mining Applications in Sustainability, 2011.

14.3 Description of challenge

The challenge is to analyse the REDD dataset according to interesting questions you come up with, potentially leading to new insights. You could use different approaches including data preparation and visualization, time series analysis or data mining.

The research is aiming to facilitate the comparison of disaggregation algorithms. Also, it is possible to build models for one specific device across different buildings and gain some insights about their contribution in the total consumption (energy prediction at the device level). Finally, main applications of energy disaggregation, including providing itemized energy bills, enabling more accurate demand prediction, identifying mal-functioning appliances, and assisting occupancy monitoring, are possible.

14.4 Tips and suggestions

In addition to the REDD dataset, a good starting point for this project could be `https://github.com/minhup/Energy-Disaggregation`.

14.5 References

1 Steven Vitullo, "Disaggregating Time Series Data for Energy Consumption by Aggregate and Individual Customer", PhD thesis, 2009.

2 H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han, "Unsupervised disaggregation of low frequency power measurements," in SIAM International Conference on Data Mining, pp. 747–758.

3 J. Z. Kolter and T. Jaakkola, "Approximate inference in additive factorial hmms with application to energy disaggregation," Journal of Machine Learning Research - Workshop and Conference Proceedings, vol. 22, pp. 1472–1482, 2012.

4 A. Iwayemi and C. Zhou, "Leveraging smart meters for residential energy disaggregation," in IEEE PES General Meeting — Conference Exposition, July 2014, pp. 1–5.

5 E. Mocanu, P. H. Nguyen and M. Gibescu, "Energy disaggregation for real-time building flexibility detection", IEEE Power and Energy Society General Meeting, 2016.

Project 15

Project 15: Process discovery and analysis [PDA]

15.1 Introduction

Project owner: Faiza A. Bukhsh
Primary topic: PM

15.1.1 Deliverables

The answers for this project should be presented in a report and a presentation. The reason for doing so is that these documents are always produced when showing process owners the results of process mining analysis in real-life situations. Thus they can communicate the results for other levels of administration in the hierarchy. Therefore, to guide you in producing this report, we have defined an outline with the points your report should include. This outline is available on Canvas.

15.2 Description of data set

The dataset contains a collection of experiment results and event logs generated. The experiment comprises a job-shop scheduling problem, implemented in a discrete-event simulation model. The raw experiment results are given, from which event log files can be generated by following the steps as described in this data paper. A set of event log files, which can be constructed therefrom, is given. These event logs include the filtered part of case study as presented in the paper "An agent-based process mining architecture for emergent behavior analysis" by Rob Bemthuis, Martijn Koot, Martijn Mes, Faiza Bukhsh, Maria-Eugenia Iacob, and Nirvana Meratnia, EDOC2019.

15.2.1 Files

The raw results from the simulation model are presented in the repository [RawData/Events.txt] and the filtered event logs are stored in the repository [FilteredFiles/Experiment.xes].

RawData/Events.txt

Describes the event data generated as output of the simulation study conducted. Upon request, we can provide you with more experimental results, for that please contact the corresponding authors.

File naming convention

ExperimentXYZ.tzt, where:

- X = number of vehicles 4,5,6;
- Y = vehicle driving direction 1 = forward; 2 = backward; 3 = forward and backward;
- Z = vehicle dispatching rule 1 = random; 2 = longest waiting time; 3 = nearest vehicle.

File Format

.txt with tab-separated values.

File Content

- ID = unique identifier of event;
- Timestamp = YYYY/MM/DD HH:MM:SS.MS;
- Product = type of product Console; Helicopter; Robot followed by a unique identifier;
- Type = type of product Console; Helicopter; Robot;
- Event = activity Arrival; Drain; Drilling; Painting; Sawing; Transport; Welding;
- Status = life cycle Blocked; Complete; In progress; Start; Waiting;
- Resource = additional information about a utilized resource such as entering source (e.g., JobShop.Source; Job- Shop.Buffer), departing sink (e.g., JobShop.Drain), vehicle (e.g., AGV:1; AGV:2), and machine entrance buffer (e.g., JobShop.Welding.Input), machine process (e.g., JobShop.Welding.Machine), and machine exit buffer (e.g., Job- Shop.Welding.Output).

Additional Remarks

- All experiments are conducted with the same random seed values;
- Only one replication is conducted per experiment;
- The run length of one experiment is 24 hours;
- A warm-up period is not taken into account;
- .MUs = moveable units.

15.2.2 FilteredFiles/Experiment.xes

These are the input files for determining the quality metrics of the process models and the key performance indicators (e.g.,throughput times).

File format

Extensible Event Stream (XES).

File naming convention

XYZ.xes, similar to naming convention of [RawData/Events.txt]

15.3 Description of challenge

"In the project, you will use the skills learned while solving PM assignments. Although in general, we do not want to prescribe tools for the projects, we advise using ProM in the project. In the topic assignments, you only saw a small portion of its possibilities. Note that some of the plug-ins may be unstable so you may encounter one that doesn't work as expected. The project is about performance and conformance checking. ProM has the possibility to replay an event log on the discovered process (e.g., visual miner). The data set of the project is from a so-called job shop. In essence, the logs contain information about events carried out by machines and vehicles, of which the vehicles transport logistics goods in a manufacturing facility. . You will explore the data set and discover the process. The data contains assignments of a job over a period of time. The cases in the log contain information on the main application as well as objection procedures in various stages. Furthermore, information is available about the resource that carried out the task. One of your tasks is to find possible points for improvement in the organizational structure. For example, think about yielding a better performance while utilizing less resources. Moreover, if some of the processes will be outsourced(more costly in terms of resource, time etc) then they should be removed from the process and the applicant needs to have these activities performed by an external party before assigning a job. Management wants to know will outsourcing affects the organizational structures? Is there a best or worst experiment case? Moreover, can one experiment learn from others? The organization would like to streamline their business process and has asked you for advice (e.g., about the occurrence and start/end of events, used the resource, relations between resources, performance, etc.), process discovery and performance/conformance analysis. Write an advisory report with performance statistics, bottlenecks, etc. and present recommendations for the organization on how to improve and enhance their business process. Notice that the dataset contains the results of 27 different job shop configurations. Where 613 and 621 are two selected datasets with good performance in terms of low throughput time. Produce an advisory report and recommend one best process to the administration. "

As a bonus you can also select a third dataset of your preference and comment on it.

15.4 Tips and suggestions

Following are some tips:

(a) The report should clearly state the aim of analysis. As the target audience is managers of organizations, the introduction should indicate what these managers can find in the report in an appealing way so that the managers will be motivated to read the rest of the document.

(b) For presenting your results. The idea is that you define questions for yourself based on the project description. You should include: (i) The questions addressed; (ii) Your answer for these questions; (iii) Screenshots of mined models/results that support your answer.

Project 16

Project 16: Business Intelligence [BI]

16.1 Introduction

Project owner: Maurice van Keulen
Primary topic: DPV

This project is about providing recommendations to the management of a business. We follow the Balanced Scorecard approach to define a set of relevant business questions that is well-balanced over the various aspects of a business, hence doesn't forget or de-emphasize one or more of these aspects / perspectives.

The balanced scorecard perspectives are:

1. Financial: How do we look to our Shareholders?
2. Customer: How do our Customers See Us?
3. Internal Business Process: What should we do that is Excellent?
4. Employee and Organization Innovation and Learning: Can we continue to Improve and Add Value?

Please find more information on the Balanced Score Card and how it is to be applied, for example, from this source: S. Kaplan, "Conceptual Foundations of the Balanced Scorecard". Harvard Business School Working Paper, No. 10-074, 2010. `https://www.hbs.edu/faculty/Pages/item.aspx?num=37638`

Follow the same steps as done in the DPV topic: design a star or snowflake schema using the multidimensional modeling approach, prepare the data (extract, transform and clean) using the ETL approachand store it in a DBMS, and design and realize a visualization.

The deliverables are:

- The report explaining the design decisions of each step. Pictures of the design of the star or snowflake schema, the design of the ETL flow, and the design of the visualization / dashboard should be included in the report (screenshots are also allowed).

16.2 Description of data set

There are several data sets available the contain data about sales and other business aspects of several businesses

16.3 Description of challenge

16.3.1 Task 1: Business Questions and Multidimensional Modeling

Please choose **one** of the datasets available on Canvas. You need to come up with interesting business questions based on the Balanced Scorecard perspectives that can be answered with the particular dataset. You need one or more business questions for each perspective. *Note: Please come up with non-trivial business questions that requires a bit of thought in the datawarehouse modeling. For understanding what makes a good business problem(s), see the paper on Balanced Scorecard on Canvas.*

Based on these business questions, think of Key Performance Indicators (KPIs) as well as metrics that you think best represent a solution to the business problem. Then design a Star or Snowflake schema (OLAP model) to answer the business questions.

16.3.2 Task 2: ETL

Design a star schema for the defined business questions, create a database (data warehouse) for it, and prepare and store the data in this database.

16.3.3 Task 3: Visualization

Design and realize dashboard with visualizations that can be used to answer the defined business questions.

16.4 Tips and suggestions

16.4.1 Hint 1

Hint for parts 1 and 2 if using Sakila dataset:
`http://www.percona.com/live/mysql-conference-2012/sessions/starring-sakila-building-data-warehouses-and-bi-solutions-using-mysql-and-pentaho`

16.4.2 Hint 2

It may happen in your project that there is not enough data in the available data set. For example, you want to show a trend over several years, but there is only one years worth of data. In that case you are allowed to artificially add more data to your data set by generating it randomly. See
`http://kedar.nitty-witty.com/blog/generate-random-test-data-for-mysql-using-routines`

If you experience issues running the `populate_fk` function provided on the page referred above, please use the file `generate_random_data.sql` attached. You can use it by running the sql statement and then calling for example: `call populate_fk('sakila','film',10,'Y');`

Project 17

Project 17: Classification of incident-related image using machine learning [CIRI]

17.1 Introduction

Project owner: Estefania Talavera
Primary topic: CV or DM

In computer vision, the automatic recognition and classification of images enable humans to derive meaning information from images, videos or other visual inputs. Based on this information, people take action or give recommendations. Machine learning and computer vision have become closely related. This is given the human goal of building artificial intelligence systems that are based on learning models combined with visual input. These AI systems aim to, eventually, understand the world in a way that humans do.

The automatic recognition of natural disasters and other events that require human intervention is relevant for relief organizations. The effectiveness of the human response relies on the fast acquisition and processing of information. Nowadays, this analysis requires manual processing which is inefficient. Recently, efforts from the computer science community have been directed towards the analysis of satellite imagery, remote sensing data, among others to overcome the need for manual data processing.

Social media posts shared by people on-ground, when a natural disaster happened or are happening, provide a new source of information; time-stamps, images, textual descriptions, audio descriptions, etc. Recently, new large-scale data sets for incident recognition in the wild has been made available. The main objective of these data sets is to sever as a basis for the training of algorithms for the automatic filtering of relevant images. Detecting these images can help organize a response by the relief organizations to the event.

17.2 Data set description

The Incidents1M dataset is provided [1]. This dataset is a large-scale dataset of images describing natural disasters, damage, and incidents in the wild. It is a multi-label dataset which contains 977,088 images, with 43 incident and 49 place categories. For simplicity, in this project you are asked to work with a minimum

of 1000 samples per category. If you decide to work with more images, please go ahead, but keep in mind for the analysis of the performance of the models that unbalanced datasets are harder to train.

Link to the github repository with the more information about the dataset:
`https://github.com/ethanweber/IncidentsDataset`

[1] Weber, Ethan, Dim P. Papadopoulos, Agata Lapedriza, Ferda Ofli, Muhammad Imran, and Antonio Torralba. "Incidents1M: a large-scale dataset of images with natural disasters, damage, and incidents." arXiv preprint `https://arxiv.org/pdf/2201.04236.pdf` (2022).

17.3 Description of challenge

You will be provided with a subset of the Incidents1M dataset. The challenge in this project is to develop a classification framework for incident recognition in images. Look for suitable models to test your machine learning framework and compare their performance.

17.4 Tips and suggestions

Tips and suggestions for the report;

- The data set needs to be described and presented. Described: Add a table quantifying the dataset. Presented: Add sample images that describe the categories in dataset.
- You are suggested to include a figure of your learning framework that will help you better preset it to the reader, in combination with the text description.
- To assess generalization capabilities of your model, you are encouraged to run k-fold cross validation.
- Implement a set of metrics to assess performance, e.g. accuracy, weighted accuracy, f-score, precision, and recall.
- Presenting a confusion matrix will help you present and describe the misclassifications of your model.