Conferences
leibniz-psychology.org

Big Data in Psychology

June 7-9, 2018
Trier, Germany

Abstract Collection

leibniz-psychology.org

Session 1: **Predicting Behavior**

Thursday, June 7
2:30 pm - 4 pm

2:30 pm - 3 pm	*Victor Nozdrachev, Larisa V. Mararitsa, Pavel Novikov* "Individual differences in digital behavior: predicting the next action on the basis of the set of previous ones."
3 pm - 3:30 pm	*Brian Schwartz, Wolfgang Lutz* "Personalized predictions and clinical support tools based on big data: Development and implementation."
3:30 pm - 4 pm	*Ramona Schoedel, Sarah T. Völkel, Jiew-Quay Au, Markus Bühner, Clemens Stachl* "Digital footprints of sensation seeking: a traditional concept in the big data era."

2:30 pm - 3 pm

Victor Nozdrachev, Larisa V. Mararitsa, Pavel Novikov

"Individual differences in digital behavior: predicting the next action on the basis of the set of previous ones."

Humans are increasingly migrating to the digital environment, producing large amounts of digital footprints of behaviors. Analyzing big datasets of such footprints presents unique methodological challenges and could greatly increase our understanding of individuals. Now we have an opportunity to fix series, sequences of man's digital actions like launching applications on his smartphone for example. There are some good solutions for predicting next launched application (Baeza-Yates et al., 2015). But while studying the sequences of digital actions we can raise a question about individual particularities in such predictions of further actions on the basis of previous ones.

The goal of the present research is to establish patterns that define the required horizon of analysis of previous actions to predict further ones. To test the hypothesis about the difference of a man's individual history to be used for making a forecast with the help of analyzing launches of applications on his smartphone.

We supposed that the repertoire and intensity of a smartphone usage is determined not only by some set of previous applications, but also by applications launched a day or a week ago.

Our research was conducted with the help of "Livli" mobile application that gives a user an opportunity to analyze his digital activity and time share between different applications. The sample was 842 people with personal digital history from 7 to 139 days (with median

35) who gave permission to gather and use their digital data. The information about launching applications was analyzed with CART method.

Data processing showed that taking in account about ten directly launched applications paying attention to the order of launching and about five applications launched during a previous day is enough. Adding applications launched a previous week didn't gave a significant contribution to the reliability of a forecast.

The ability to predict next application on a basis of previous one was reached only for people who use less than 4 applications. The possibility of correct prediction of launched application found to be rather stable for each user but changed between users. We marked out a group of users with more than 50% correct predictions and another group of "unpredictable" users with less than 10% correct predictions made with the help of our method.

The research showed that people differ by the strength of a habitual way and logics of choosing next application to launch, and also have different vulnerability to accidental external influences. The question of extrapolation of the results to other people or activities as well as making a psychometrical profile of a person on the basis of digital patterns seems interesting. It uncovers promising prospects for enhancing phone-user interaction through discovering specific habit structures. Such data can be useful and find its implications in developing machine theory of mind and building human-computer interaction.

3 pm - 3:30 pm

Brian Schwartz, Wolfgang Lutz

"Personalized predictions and clinical support tools based on big data: Development and implementation."

Background. There is clear evidence that psychotherapy is effective in treating mental disorders. However, about 5–10% of clients deteriorate during treatment. To prevent patients from deteriorating, an early detection of patients at risk for a negative course is crucial to adapt the treatment strategy. Since predictions of treatment failure are more accurate when based on statistical algorithms than on therapists' judgements, we implemented an internet-based clinical assessment, decision and information tool. The decision support tool is based on research on tracking and predicting individual change using early response, therapist differences, and continuous and discontinuous patterns of change within treatments as well as between treatments. Personalized predictions as well as individualized treatment recommendations are offered based on the database of the outpatient center at the University of Trier.

Objectives. The aim of the DFG-funded study is to evaluate the improvement in treatment outcomes by continuous psychometric assessments, feedback, and a decision and clinical support tool.

Research Questions. The following questions guided our work: Is treatment progress improved by a complex prediction tool providing feedback for the therapist?

Method. The development and validation sample of the tool consisted of 1000-odd patients treated for at least 10 sessions with cognitive-behavioral therapy in the outpatient center at the University of Trier. To select variables for dropout and personalized treatment predictions, a bootstrap ranking LASSO procedure was applied. Treatment predictions are based on the 30 most similar patients of the index patient using nearest neighbor methods. Furthermore, an adaptive prediction model for treatment course was trained, updating the prediction at each session to identify patients who are not on track. These deteriorating patients trigger the clinical support tools, providing individualized treatment recommendations based on the psychometric assessments of each patient. Patients starting treatment are randomized to a feedback or a non-feedback control group to evaluate the effects of the tool.

Results. Data structures in our outpatient center, the development of the statistical models, and the implementation of this complex assessment, feedback, and decision tool are introduced. Moreover, preliminary results on the effects of the use of this tool are presented.

Conclusions. The introduced feedback and decision tool handles the session-by-session assessments and selects treatment relevant variables out of a big set of potential predictors. The tool seems to be able to improve treatment selection and adaptation and could be implemented in routine care. An outlook is given concerning further statistical approaches to improve the predictive accuracy of the tool as well as attempts to implement a dynamic, self-updating dataset for the model building process.

3:30 pm - 4 pm

Ramona Schoedel, Sarah T. Völkel, Jiew-Quay Au, Markus Bühner, Clemens Stachl

"Digital footprints of sensation seeking: a traditional concept in the big data era."

(a) Background

Why do some people go skydiving, while others read detective stories to feel aroused? Individual differences in the need for external stimulation have been described as a stable personality trait called sensation seeking. Initially proposed by Zuckerman, it refers to "seeking of varied, novel, complex, and intense sensations and experiences, and the willingness to take physical, social, legal, and financial risks for the sake of such experience" (Zuckerman, 1994, p.27). Hence, the construct of sensation seeking represents a biopsychological personality perspective and is explained by genetic, biological, psychophysiological, but also social factors (Roberti, 2004; Zuckerman, Kuhlman, Joireman, Teta, & Kraft, 1993; Zuckerman, 1994). The traditional concept of sensation seeking has stimulated a vast amount of research efforts in which we have identified three key aspects that are the basis for reasoning for our study.

First, research about sensation seeking has mainly focused on high risk activities such as extreme sports (Guszkowska & Bołdak, 2010) or criminal behaviors (Zuckerman, 2007). But according to Arnett (1994) or Roberti (2004), sensation seeking is not limited to experiencing risk per se. Rather, a certain amount of risk is accepted to obtain a level of arousal which is considered to be individually ideal.

Second and related to this, the majority of studies has dealt with an unsocialized form of sensation seeking. A term that refers to actions like alcohol and substance usage, excessive gambling, risky sexual activities, or reckless driving (see Roberti (2004) for extensive review). However, Zuckerman (1994) also postulated a non-impulsive and socialized type of sensation seeking which is for example represented by outgoing, sociable, and extraverted behaviors (Glicksohn & Abulafia, 1998). Only little work can be found in this area, even though there are some promising and interesting findings. As an illustration, chess experience and musical taste have been associated with sensation seeking levels (Joireman, Fick, & Anderson, 2002; Litle & Zuckerman, 1986).

Third, traditionally the collection of actual behavior has been very difficult and costly to achieve. Therefore, previous behavioral measures of sensation seeking have almost exclusively consisted of retrospective self-reports. Self-reported behavioral correlates like reckless driving or everyday-activities like smartphone-usage (e.g. Dahlen, Martin, Ragan, & Kuhlman, 2005; Leung, 2008) could have therefore been subject to memory biases and other self-report typical problems such as social desirability (Ziegler & Bühner, 2009).

However, the new research branch of smartphone-sensing (Harari et al., 2016) offers promising opportunities for both personality research and psychological science in general. Consumer electronics are equipped with a large number of sensors and data logging capabilities, providing various information about its user's natural everyday-activities and habits (Harari et al., 2016). There has already been some research showing that data collected in this way can be very informative about the 3 user's individual traits (e.g. Andone et al., 2016; de Montjoye, Quoidbach, Robic, & Pentland, 2013; Stachl et al., 2017a).

(b) Objectives and research question

To summarize our key points, we think that for observing objective behavioral manifestations of sensation seeking in everyday-contexts appropriate investigation methods have been missing so far. Thus, the aim of our study is to investigate the traditional concept of sensation seeking by the aids of new methods available in the increasingly digitalized world. Smartphones as collectors of digital footprints, provide data both high in frequency and dimensionality. By combining smartphone-based big data with traditional self-report measures we aim to gain new insights in the behavioral manifestations of sensation seeking. This study investigates, whether individual sensation seeking scores can be reliably predicted from behavioral markers collected via mobile sensing on conventional smartphones. The extraction of these behavioral markers is literature based and will be described in the method section.

(c) Method/Approach

This study will be pre-registered prior to analysis of the data and the pre-registration form will be available via an open science framework link in the full-length article.

Research framework and data collection procedure

This study is part of the larger, ongoing “PhoneStudy” research project, at Ludwig-Maximilians-Universität München (LMU), Germany (see Stachl et al., 2017b). This project is an interdisciplinary project between the departments of psychology, computational statistics, and media informatics at LMU. It focuses on the investigation of associations between individual differences (measured via self-reports) and a vast variety of behaviors, logged via smartphones. The logs include information about participants' mobility, app-usage, music consumption, communication behavior, and many more. Currently, the dataset grows by 1110 events per participant per day on average.

Data collection for the current study takes place between October 2017 and January 2018. The onboarding-process for study-participation (informed consent, randomized allocation of pseudonyms for app-login) takes place online and is open until the 12th of December 2017. Once completed, the participants install a specifically developed research app on their smartphones, to participate in the study. In the consequent 30-days data collection period, rich behaviorally-focused log-data and answers from a series of psychometric self-report questionnaires are collected. The self-report questionnaires have to be completed by the participants at a for them convenient time during the 30 days of data logging.

Participants

Participants are recruited by student researches during a seminar, via flyers, social networks and street recruitment. Participation is possible for German native speakers with a minimum age of 18. Participation is rewarded by giving written feedback about their individual Big Five-based personality profile and their average smartphone-usage within the 30 days of data collection. Additionally, participants enter a lottery to win online-shopping gift cards. The recruitment of participants is limited by the timeframe of the seminar (last person in 12.12.2017). Until then, as many participants as possible will be recruited. Nevertheless, a minimum final sample size of 250 participants is anticipated. At the time of abstract submission (15.11.2017), 185 participants have already been recruited.

Self-report measures

In addition to log data, a series of self-report questionnaires is administered via the PhoneStudy research app. Sensation seeking is assessed via the Impulsive Sensation Seeking (ImpSS) subscale of the Zuckerman-Kuhlman Personality Questionnaire (ZKPQ-III-R; Zuckerman, 2002). In addition, subjects are asked to indicate how much they agree to be a sensation seeker based on Zuckerman's definition (Zuckerman, 1994).

Demographics are also collected in the same procedure. Big-Five personality is assessed via the German version of the Big-Five-Structure-Inventory (BFSI; Arendasy, 2011) and the newer German version of the Big-Five-Inventory 2 (BFI-2; Danner et al., 2016). We will check the discriminant validity of our sensation seeking measure by considering personality traits based on the BFSI and the BFI-2.

Data Pre-Processing

Originally, the collected log-data exists as timestamped event-data. Each row is a registered event (e.g. call), each row is an event characteristic (e.g. outgoing, timestamp, duration, contact-hash, …). In order to use the data for prediction modeling, the raw data needs to be processed in order to extract variables. This is done via specifically created aggregation functions from an R-package which is currently under development by the Computational Statistics Institute of LMU (citation will be added later).

We first identify behavioral manifestations of sensation seeking from previous literature and transfer them to measures of how they could be represented in smartphone usage. As an illustration, it is a common finding that sensation seeking is associated with gambling (Roberti, 2004). As a consequence, we will derive the usage frequency and duration of installed gaming apps as predictor variables. An exhaustive overview of our pre-specified features will be provided in the pre-registration report. After feature extraction, all further pre-processing will be incorporated in the resampling-process of the prediction modeling – in order to avoid overfitting.

Data Analyses

Besides descriptive statistics and simple pairwise correlations, the data will be analyzed within a prediction modeling framework. Framed as a regression problem, we will predict concrete sensation seeking scores from behavioral features, extracted from smartphone-logs.

Cross-validated model fit will be evaluated based on how accurate new (unseen) samples can be predicted. As common in the predictive modeling culture (Breiman, & Breiman, 2001; Wolpert & Macready, 1997), we will compare different prediction algorithms against a common guessing baseline in a benchmark experiment. The generalized predictive ability of a regularized linear model, a random forest, a support vector machine and a xgboost model will be evaluated. These methods are sorted by decreasing interpretability and increasing flexibility and represent different trade-off levels between interpretability and expected prediction accuracy. And third, we do not want to be content with a black box model. In order to better understand which variables are important for prediction success, variable importance measures and partial dependence plots will be presented (Goldstein, Kapelner, Bleich, & Pitkin, 2014).

(d) Results/Findings

As the data collection procedure is still in progress, no results are available at this time. Results will be based on data analyses procedures specified in the section above and the pre-registered report. In order to counteract publication bias, the full-paper will be submitted/published (in case of acceptance) regardless of the results.

(e) Conclusions and Implications

Our findings highlight the potential of big data in psychological research by transferring findings mainly based on questionnaire data to everyday-manifestations of behavior. Additionally, our study can help to expand already existing literature about the association between digital footprints and user traits (e.g. Kosinski, Stillwell, & Graepel, 2013; Stachl et al., 2017a; Youyou, Kosinski & Stillwell, 2015). A more specific practical implication of predicting sensation seeking based on smartphone-logging data could for example be to recognize traits, associated with risk-taking, for the development of smartphone-based intervention programs.

Session 2: **European Association for Data Science (EuADS)**

Thursday, June 7
4:30 pm - 6 pm

4:30 pm - 5:15 pm	*Ernesto William De Luca* “Integrating Textbook Resources and Tools.”
5:15 pm - 6 pm	*Myra Spiliopoulou* “Analysing how patients with a chronical disease discuss about treatments in a social self-help platform.”

4:30 pm - 5:15 pm
Ernesto William De Luca
“Integrating Textbook Resources and Tools.”

Background. The Georg Eckert Institute (GEI), a key institution in the field of textbook research, hosts the largest international textbook collection worldwide. The GEI has a track record spanning over 40 years of conducting research projects, working with international partners in bodies such as international textbook commissions, and publishing its research. Over the last 20 years, the Institute has digitalised a large proportion of its library holdings and initiated field-specific data collections and open-access publications. As a member of the German Leibniz Association and as a leading research institution in the field, the GEI regards the development and maintenance of research infrastructure for international textbook and educational media research as a core part of its mission.

Objectives. At the moment we are integrating all our data and meta data formats into one repository and cooperating with international partners to incorporate global data on textbook research from other institutions. Our aim is to create an interoperable, standardized, multilingual access to all available data on textbook research across formats. The data include, but are not limited to, open-access texts, web pages, databases, library data, curricula, TEI-tagged textbooks, textbook reviews, research project descriptions, and data on institutions and researchers in the field.

Research question(s) and/or hypothesis/es. The GEI is currently pursuing a three-way approach, effectively building three ‘roads’ which will eventually merge into one wider road. The project “WorldViews” is integrating data available via GEI applications and websites, while “GLOTREC” (Global Textbook Resource Centre) is creating an international, multilingual reference and information architecture for textbooks. Finally, the project “Children and their World” develops novel, reusable instruments for the access and analysis of large digital corpora, namely textbook contents.

Concluding these projects, the three ‘roads’ will merge to become a system which will be integrating all available data on textbooks and textbook research worldwide.

Method/Approach. The increasing internationalisation of databases calls for multilingual and even multiple-alphabet accessibility. This challenging task is part of the GLOTREC project, which has recently reached the milestone of completion of the International TextbookCat, serving as a multilingual hub to data on textbooks from three different library catalogues, held decentrally (GEI OPAC, Edisco and MANES). We are currently working on the integration of more data, more languages, more alphabets and more data types.

While our principal focus is data integration and representation at present, we are committed throughout the process to two fundamental principles:

1. Usability. Simply collecting and providing the data would not meet the needs of our target group.
a) We wish to represent the data in a way that is helpful to users with multiple needs. Our aim in this context is to present the data in an ‘integrated but separate’ state and show users what kind of data they have found and what is the source of it.

b) We will enable multiple points of entry. Further, all existing applications, such as databases and document repositories, will be maintained. The Repository of Textbook Research will serve as a hub to these applications as well as to their data.

c) We are attempting to create meta data schemata that are simultaneously as coherent as possible and as specific as necessary for the various included datasets. The Challenge is to present datasets as diverse as textbook data, research articles and digitalised textbooks.

2. Openness. We believe we have reached at least the two-star level of Tim Berners-Lee’s (2006) 5‑star deployment scheme for Open Data.
a) We are working on making at least the GEI-owned data in the Repository available in non-proprietary formats.

b) Furthermore, we are using URI wherever possible and linking the data with our own and other datasets.

c) As well as creating and maintaining our repository, we are mirroring our data to repositories of a wider thematic scope. The texts contained within our own academic repository are delivered to other repositories, for example Leibniz Open, via OAI-PMH.

The presentation will cover the Georg Eckert Institute’s path to the integration of all available data on textbooks and textbook research within a single repository. The data are in various formats and stem from international and multilingual sources. We have already created a middleware for GEI-generated data, which deals with data standardisation and homogenisation, and are hosting a multilingual catalogue, named International TextbookCat, within which we integrate three international textbook catalogues from Italy, Spain and Germany. Furthermore, in the project “Children and their World”, we are creating Natural Language Processing methods such as Topic Modeling for the analysis of large digitalised historical textbook corpora.

Results/Findings. By presenting this work at the conference, we hope to spread our idea of integrated and linked open data in textbook research to other fields and international research communities and to provide a role model for the gradual implementation of long-term projects. The presentation covers novel technical approaches for multilinguality and user centred design in digital humanities as well as the representation of diverse data sets of multiple formats, scope, and (qualitative and quantitative) depth in one repository.

Conclusions and implications (expected). Tools of the Digital Humanities have shown to be successful in supporting research on books.

For instance, our institute provides several tools for doing research on textbooks and curricula. Unfortunately, not all institutes have the possibilities and the qualified staff to set up their own Digital Humanities architecture. Hence, we enhanced our successfully applied tools to be ready to cover additional data from all around the world.

Our objective, then, is to create a multilingual, multi-alphabet, multi-dataset five-star-openness repository on textbook-related data whose content is available via multinational and multi-thematic applications. The proposed presentation will detail a role model for data integration and representation in the 21st century, covering the milestones already achieved and those still ahead of us. Another focus of the presentation will be the user perspective; in this way, we hope to demonstrate the uniqueness and novelty of the interface, scope and scale the Repository of Textbook Research will offer.

5:15 pm - 6 pm

Myra Spiliopoulou

“Analysing how patients with a chronical disease discuss about treatments in a social self-help platform.”

Background. Tinnitus is a currently incurable condition, in which patients hear sounds without an external stimulus. Baguley and McFerran (2013) report a tinnitus prevalence between 10 and 15%. Probst et al (2017) have analysed tinnitus patients from an outpatient tinnitus clinic, users of a mobile application for tinnitus self-help and participants of the self-help social forum TinnitusTalk, and identified profile differences among the users of mHealth and social platform and the patients registering via the outpatient clinic. Social platforms allow for the exchange of information and insights among patients, and can thus contribute to patient empowerment and self-help. However, a newcomer to an active social platform may have difficulties in deciding which discussion threads to follow and in getting an overview of how specific items (eg treatments) are perceived among other patients.

Objectives. We investigate the potential of data science for the analysis of social data in a self-help platform, where patients discuss about treatments, elaborating on personal experiences and contributing facts they have acquired from various sources. We study to what extend data science can contribute to organizing discussions on treatments into threads and to characterizing the attitude of the patients towards the treatments over time.

Research questions. How can we identify the treatments that are subject of discussion in a social platform for patients? How can we assess the sentiment and subjectivity of the texts that refer to treatments while minimizing human annotation effort? How can we quantify sentiment and subjectivity into a score and capture the evolution of this score over time? How can we compile a time-oriented overview of this information?

Materials and Methods. We analysed the platform TinnitusTalk.com, acquiring data from 9 subfora, as of July 2017. We report on our method TinnitusTalkMonitor, which encompasses a crawler over the platform's fora, a workflow for the semi-automated identification of treatments on the basis of an initial list of treatment names, a component that processes sentences and identifies those that refer to treatments, a multi-target classification algorithm that characterizes sentences on sentiment and subjectivity on the basis of a small handcrafted sample of annotated sentences, a mechanism that propagates the multit-target classifier's prediction on the unlabeled data, a scoring mechanism for the characterization of a treatment's dominant sentiment and subjectivity at each timepoint, and a front-end visualization tool that encompasses a summary view over all treatments and a detailed view for each treatment.

Results and Outlook. We have shown that our approach can extract treatments and associate them with sentiment and subjectivity, achieving good performance over the validation subsample. The front-end of the tool is operational and provides an overview of the frequently discussed threads and details on each treatment. We are currently extending our approach for a platform that does not concentrate solely on patients with a given disease.

CITATIONS

Baguley D, McFerran D (2013) Tinnitus. Lancet 382:1600–1607

Probst T et al (2017) Outpatient tinnitus clinic, self-help web platform, or mobile application to recruit tinnitus study samples? Front Aging Neurosci 9:113

This work has been published:

Dandage S. et al. (2018) Patient Empowerment Through Summarization of Discussion Threads on Treatments in a Patient Self-help Forum. In: Maglaveras N., Chouvarda I., de

Session 3: **Methodological Issues (Part 1)**

Friday, June 8
10:30 am - 12 pm

10:30 am - 11 am	*Pia Tio* “Estimating cross-source relationships from wide big data using component- and networks-analysis.”
11 am - 11:30 am	*Yucheng Zhang* “Integrating Split/Analyze/Meta-Analyze (SAM) Approach and Multilevel Framework to Advance Big Data Research in Psychology: Guidelines and an Empirical Illustration via Human Resource Management Investment-Firm Performance Relationship.”
11:30 am - 12 pm	*Anna-Sophie Ulfert, Mona Rynek, Henrike Peiffer, Elisa Clauss, Prof. Dr. Thomas Ellwart* “Theoretical and methodological issues in collecting and handling big data in organizational research.”

10:30 am - 11 am

Pia Tio

“Estimating cross-source relationships from wide big data using component- and networks-analysis.”

Background. Network analysis has successfully been applied to many different sources of psychological data, including personality, cognitive performance, and clinical symptoms. While investigating these different areas in their single domains is useful, a better understanding of their structure requires an integrated analysis with several sources of information. Performing such integrated analyses offers the opportunity to explicitly investigate relationships between two or more of such sources (cross-source relationships). However, this often requires wide big data sets containing information about individuals from multiple sources. Although such data are becoming more and more commonplace, estimating a network using big data is not without its challenges. The dimension of the dataset, often containing more variables than observations, hinders accurate estimation of relations, even when some form of regularisation (e.g. lasso penalty) is used. Reducing the number of variables would be a straightforward way to remove (or at least reduce) this problem, except that we do not yet know which variables are involved in cross-source relationships. An additional challenge is that wide big data contains data from different sources that inherently may have different characteristics. For example, indicators of cognitive performance are expected to correlate much higher with one another than with indicators of gene expression. Applying network analysis to such data without taking this difference into account again leads to inaccurate estimation of the wrong relationships.

Objectives. Developing a statistical network procedure that allows for the accurate estimation of cross-source relationships from wide big data.

Hypothesis. We propose that the combination of regularized simultaneous component analysis with the network framework will perform better in accurately estimating cross-source relationships compared to only using network analysis.

Approach. When estimating unique cross-source relationships from wide big data, a statistical procedure is needed that a) combines data from different sources where each source may have different characteristics, b) selects variables to determine which variables are involved in cross-source relationships, c) estimates these unique cross- source relationships, and d) presents results that can be interpreted in a meaningful, substantive way. While the network framework (also known as graphical models) is apt at estimating interpretable relationships, it is not suitable to handle the first two requirements.

To single out cross-source relations, we need to disentangle the common sources of variation shared between the different data blocks (with each block containing the data or variables of one source) from the sources of variation that are specific for a single or a few data blocks only. To do this, we perform a so called sparse DIStinctive and COmmon Simultaneus Component Analysis decomposition of the data (sparse DISCO SCA), a method that was developed for the integrated analysis of multi-source data with the specific aim of separating block-specific sources of variation from common sources of variation.

We therefore introduce Sparse Network And Component model (SNAC), a two-step component-network model for estimating cross-source relationships from wide big, multivariate Gaussian distributed data. First, sparse DISCO SCA is used to reveal the underlying common and source-specific sources of variation; second, network analysis is applied on the common sources of variation only.

To investigate whether including the sparse DISCO SCA analysis as a pre-processing step results in more accurate cross-source relationship estimates, we performed a simulation study. We manipulated the sparsity, the strength of the cross-source relationships relative to the strength of the course-specific relationships, and the ratio of number of individual to the number of variables. Network analysis performance was compared to SNAC analysis performance.

SNAC analysis has been implemented in R; the R-package also contains a toy example using real data.

Results. SNAC analysis outperforms regular network analysis when estimating cross-source relationships from wide big data. The difference is most pronounced when the big data contains more variables than observations.

Conclusion. SNAC reveals more accurate relations between cross-source data sources. This is the first step towards bridging the gap between data from different structures or different locations in a hierarchy like genetic, biological and behavioural data.

11 am - 11:30 am

Yucheng Zhang

"Integrating Split/Analyze/Meta-Analyze (SAM) Approach and Multilevel Framework to Advance Big Data Research in Psychology: Guidelines and an Empirical Illustration via Human Resource Management Investment-Firm Performance Relationship."

Though big data research has undergone dramatically development in the past decades, it was mainly applied in the disciplines such as computer science and business. However, insufficient psychology research applied big data to examine psychology research issues.

One of the major challenges is that most psychology researchers may not have sufficient knowledge about big data analytical techniques rooted in computer science. This paper integrates Split/Analyze/Meta-Analyze (SAM) approach and multilevel framework to illustrate how to use SAM approach to address multilevel research question with big data. In specific, we first introduced SAM approach, and illustrate how to implement SAM approach to integrate two big datasets at firm-level and country-level respectively. We then discuss the theoretical and practical contributions, we also proposed the future research directions for psychological scholars.

11:30 am - 12 pm

Anna-Sophie Ulfert, Mona Rynek, Henrike Peiffer, Elisa Clauss, Prof. Dr. Thomas Ellwart

“Theoretical and methodological issues in collecting and handling big data in organizational research.”

During the past decade, “big data” has become an increasingly popular term in research and practice alike. An increasing number of organizations is now interested in using big data for employee management, improvement of organizational commitment, and organizational decision making (Bassi, 2011). In the field of human resources (HR) the use of big data analytics is commonly referred to as HR analytics or people analytics (Sullivan, 2009). In psychological research, theoretical and methodological approaches for implementing big data analysis in organizational settings, have been widely discussed and improved over recent years (e.g. Hastie et al., 2009; Kuhn & Johnson, 2013; Harlow & Oswald, 2016). However, while many organizations claim the intention of using big data analysis in their HR practices, the actual use of such methods remains limited to few large organizations, despite well researched analytics approaches. It has been argued, that this discrepancy may be due to HR departments often lacking basic knowledge on how to implement big data projects (e.g. for data driven decision making; King, 2013). With a lack of theoretical and methodological knowledge, organizations are initially confronted with problems of asking the *right questions* and using the *right methodological approaches* for utilizing big data analysis.

In order to systematically investigate theoretical and methodological difficulties practitioners are faced with when handling big data, we conducted three case studies in different organizations (two different industrial companies, and one consulting firm). The aim of these three case studies was to gain an understanding of (1) how organizational HR data is conceptually structured, (2) the type of questions organizations address in their analysis, and (3) how stakeholders handle the complexity of big data. Results indicate that difficulties with implementing big data projects in organizations often result from insufficient theoretical and methodological knowledge of how an organization’s data is structured and how it can be analyzed. Based on the three case studies, we discuss the following topics:

1) Validity of data: Errors occurring when analyzing big data in the field of HR

2) Empirical evidence for questions organizations address in big data analysis (e.g. predicted turnover)

3) Communication between stakeholders of different professions (e.g. researchers and organizations)

In furtherance of using organizational HR data in research, we discuss best practice examples as well as a process model of how to implement people analytics projects in organizations.

Session 4: **Applications: Consumer Psychology (Part 1)**

Friday, June 8
2:30 pm - 4 pm

2:30 pm - 3 pm	*Anat Rafaeli, Monika Westphal, Galit Yom-Tov, Daniel Altman, Shelly Ashtar, Michael Natapov, Neta Barkay* “Big Data and Customer Emotion Dynamics: Automated Analyses in Chat Services.”
3 pm - 3:30 pm	*Marco Biella, Cristina Zogmaister, Stefania Ceolato, Elena Parozzi* “Twitter Twitter on the wall, which university's the fairest of them all? Exploring brands' social perception on social media using Big Data.”
3:30 pm - 4 pm	*Christian Mumenthaler, Ulf J. J. Hahnel, Tobias Brosch* “Taking advantage of Twitter data to investigate sentiments towards environmental issues during the 2016 U.S. presidential election.”

2:30 pm - 3 pm

Anat Rafaeli, Monika Westphal, Galit Yom-Tov, Daniel Altman, Shelly Ashtar, Michael Natapov, Neta Barkay

“Big Data and Customer Emotion Dynamics: Automated Analyses in Chat Services.”

Background and objectives. Emotions are dynamic, changing as perceptions and experience change (cf., van Kleef, 2014; Weiss & Cropanzano, 1996). Emotions are important in customer service, reflecting customer reactions to the service. Yet emotion research mostly relies on observations and self-report (Donaldson & Grant-Vallone, 2002), thus assessing emotion at a *specific time point*, and in relatively small samples. We suggest instead automated sentiment analysis as an unobtrusive tool for detecting customer emotions at *various time points* in large-scale data. Since available sentiment tools were developed to analyze *overall* emotion in structured texts (e.g., movie reviews), we make an adaptation to the *dynamic* domain of online chat service. We then use the adapted tool to test hypotheses on customer emotion dynamics in 390,483 interactions (8,526,814 individual text messages) of a Telecommunication company, conducted October-December 2016. Our paper illustrates the merit of automated sentiment analysis tool for research on psychological elements of customer service. Using such a tool, we show that customer emotion dynamics provide valid predictions of service quality.

Our theoretical contribution is in depicting emotions as dynamic notions that *evolve over time* (Filipowicz et al., 2011). We presume that customers request service because of a service failure, and so begin a service interaction with negative emotions. Requesting service requires people to spend time and effort on something they feel should not have happened. We predict that as the interaction ends customer emotions will be more positive, assuming service agents resolve the problem, and create service recovery:

Hypothesis 1: Customer emotions during interactions are dynamic, evolving from initial negative (service failure logic), into more positive emotions (service recovery logic).

We further suggest that the improvement in customer emotion during an interaction connects to customer assessment of service quality. If initial negative emotion does not change, we presume the customer issue was not resolved, and the customer is less likely to rate the interaction as satisfying and effective, than if the improvement in emotion during the interaction is substantial:

Hypothesis 2: *Differences in customer emotions dynamics in successful vs. unsuccessful interactions are valid indicators of service quality: The magnitude of change in emotion from negative (in the start) to positive (at interaction end) reflects service quality.*

Method – Part 1: Adjusting a Sentiment Analysis Tool. The accuracy of available sentiment analysis tools in detecting emotion in chat-interactions is limited, because available engines were designed for structured texts (e.g. movie reviews), while service interactions are spontaneous and unstructured. We therefore began by adjusting a sentiment analysis tool to the context of customer service, and validating it with a dataset of chats from multiple service domains. We made three core adjustments: (1) Words that have special meanings in the service domain context were either excluded (e.g. support, confirm, approve) or added to the lexicon (e.g. cancel, legal, waiting). (2) Words that have special meaning in a specific brand business context were either excluded (e.g. advanced, premium, unlimited) or added (e.g. missed, limited, complex). (3) Words such as *well, right, ok,* misspellings of emotionally charged words, slang and obscenities in previous lexicons were adjusted for customer language.

Several rounds of human coding confirmed that our tool outperforms previously available automatic detection models (Stanford RNTN, LIWC, and SentiStrength in the precision of detecting negative emotion (p<0.001). In detecting positive emotion, it is better than most of the models and comparable to SentiStrength.

Method – Part 2: Testing Hypotheses. We tested *Hypothesis 1* in the full dataset of 390,438 interactions (including at least ten messages) by standardizing all interactions to ten sections, and averaging the sentiment of all customer messages in each section. We then calculated the average positive and negative sentiment in each of the ten sections. *Hypothesis 2* we tested on a subset of 286,671 interactions, for which customers completed a post-interaction survey (73% response rate); responses included two service quality questions: Service Recovery (SR: "Was your service need resolved in this interaction?" (Yes/No)), and Customer Satisfaction (CSAT: "Please rate your satisfaction with the service you received" (1="Very unsatisfied" to 5="Very Satisfied")).

Findings. Initial customer emotions were significantly more negative (Mean=-0.131) than those at the end (Mean=0.497; p<0.001), supporting *Hypothesis 1*. Our tool identified that terms in the early sections indicated customer frustration and service failure (e.g., *error, problem, issue, wrong, lost, unable, invalid, cancel, mistake, incorrect),* while terms at the end suggest customer positivity and service recovery (*thank(s), good, help, great, works, correct, appreciate, nice, happy, best*).

Hypothesis 2 was supported by a logistic and an ordinal regression, showing that customer sentiment in each section predicts service quality, (SR: $\chi 2(10)$=28481.386, p<0.001; CSAT: $\chi 2(10)$=44725.318, p<.001). The effect of sentiment scores of later (vs. earlier) sections was significantly larger (SR: β=0.83 and 0.97 of sections 9-10, vs. β =0.27 and 0.07 of sections 1-2; CSAT: β=0.90 and 1.10 of sections 9-10, vs. β=-1.71 and -1.27 of sections 1-2), supporting the prediction that emotion dynamics during an interaction reflect service quality.

Conclusions and implications. We introduce a new approach to studying customer emotion dynamics and to assessing service quality in chat services. By offering a new model for automatic assessment of customer emotions, particularly relevant for analyzing large-scale data, we provide a promising replacement of self-report and observations. Providing objective, unobtrusive assessments of customer service that build directly on customers' actual expressions, brings important implications: Identifying customers whose emotions do not improve when the interaction ends can help managers intervene before a service situation escalates. A system of alerts, for example, when customer sentiment stays negative, can trigger notification that something is wrong.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-reported bias in organizational behavior research. *Journal Business and Psychology*, *17*(2), 245–261.

Filipowicz, A., Barsade, S., & Melwani, S. (2011). Understanding emotional transitions. *Journal of Personality and Social Psychology*, *101*(3), 541–556.

van Kleef, G. A. (2014). Understanding the positive and negative effects of emotional expressions in organizations: EASI does it. *Human Relations*, *76*(6), 1145–1164.

Weiss, H. M., & Cropanzano, R. (1996). Affective Events Theory. *Research in Organizational Behavior*, *18*(1), 1–74.

3 pm - 3:30 pm

Marco Biella, Cristina Zogmaister, Stefania Ceolato, Elena Parozzi

"Twitter Twitter on the wall, which university's the fairest of them all? Exploring brands' social perception on social media using Big Data."

Background. Every University is becoming more and more like a brand. In fact, universities compete against each other to attract scholars, public or private funds, and students. For this reason, it is of growing importance to understand how universities are perceived by various stakeholders, put differently, their brand image. So far, brand image has typically been investigated through interviews, questionnaires and focus groups. Today, thanks to their increasing diffusion, the internet and social media are becoming useful information sources to understand how universities are perceived.

Objectives. The aim of this work is to test a new method to understand universities' social perception and to investigate important facets of such perception using users' digital footprint in social networks. Such method provides quantitative metrics that allow for a rigorous investigation without losing brand image's content and the shared representation used by users to explore the social environment.

Method/Approach. One basic step for every work that aims to understand a given phenomenon is to observe it and to place it in context. To correctly frame every phenomenon a timely assessment phase is necessary. Such need brought us to focus on the comparison between our target (the universities) and other accounts of the social environment.

Our goal is to measure some dimensions, in order to map the positioning of our target along them. Such dimensions are defined a priori, based on the nature of the different universities. Data have been obtained using a validated methodology (Culotta & Cutler, 2016; Permalink: http://dx.doi.org/10.1287/mksc.2015.0968). We used a modified version of the technique and we applied it to the universities under investigation.

A difference with Culotta and Cutler's work is the nature of our targets. Indeed, Culotta and Cutler investigated how various consumer brands (cars, apparels, foods and beverages, personal care), while universities are non-profit organizations. Universities are perceived along different attribute dimensions (e.g. research, employability, teaching, ecc…) in comparison with profit organizations. Therefore we are extending Culotta & Cutler's work to the investigation of social perceptions of a very different type of targets.

Following Culotta and Cutler, we employed online behaviors to track social perception dimensions. We focalized on a large group of prototypical accounts, used as benchmarks of the attribute dimension, and compared them to targets accounts. The online behavior that we considered was the act of following a prototypical (i.e., attribute related) or target account (i.e., Universities). We relayed on available online data in contrast with data generated by means of direct question such as those collected using classical brand image methods.

The main advantage of this kind of data is its high ecological validity. In fact, they are free behaviors and not answers to direct questions. The spontaneous nature of online behaviors makes them very valuable and interesting. Moreover, online behaviors allow the researcher to understand how a given target is represented in the subjects' mind because the behavior is performed without external influence such as in the case of responding a questionnaire.

The proposed method is based on three stage. In the first stage, attribute dimensions are defined according to the target under investigation. In the second stage, target and prototypical accounts are selected. Target accounts are the ones of interest while prototypical accounts are accounts that show characteristics is desirable to have in order to have a high score on the dimension. The selection of prototypical accounts will determine a dimension's benchmark, therefore, it allows for the construction of nuanced or ad-hoc dimensions. In our work, the selection of prototypical accounts was negotiated among the authors. In the third stage, the similarity among target and prototypical accounts of each dimension is computed. During this phase, a score for each target along each dimension is produced. Such score is a purely quantitative variable drawn upon the overlap between the target and the prototypes' followers. The degree of overlap is based on the computation of a Jaccard Similarity Coefficient that summarizes the ratio of the intersection over the union of the followers' sets. The Jaccard Similarity coefficients between the target and each prototype are then weighted to produce a unique social perception score along the given dimension.

The method is based on two assumptions. First, to follow prototypical accounts of a given dimension implies that the follower is interested in such dimension. It means that accounts followed by the same user share some feature of the dimension. Second, if an account is followed by many users interested in a specific dimension the account followed is perceived as related to the dimension.

Results/Findings. The intersection between followers of a target account and sets of followers of prototypical accounts is meant to capture and quantify the interest in the dimension of the follower of the target. Such intersections can be aggregated in order to obtain a score for each target along each dimension, therefore, the score can represent how targets are perceived along the dimensions of interest.

Conclusion and Implications. Once a profile for each target on different dimension had been obtained, it is possible to map the environment where targets are perceived. Moreover, our method allows for an estimation of the relative importance of each follower. Such estimation provides useful insights for understanding which users drive a target's social perception and which are to acquire in order to improve the target's perception on a specific dimension.

3:30 pm - 4 pm

Christian Mumenthaler, Ulf J. J. Hahnel, Tobias Brosch

"Taking advantage of Twitter data to investigate sentiments towards environmental issues during the 2016 U.S. presidential election."

Over the last few years, social media platforms have had an increasing role in social movements, political campaigns and everyday life around the world. Millions of people broadcast their thoughts and opinions allowing social scientists to study individual voluntary reactions and sentiments on a wide range of topics. Although accessing to this data provides incredible possibilities for social science research, scientists have to be aware of their limitations and how to circumvent them.

Among other social media platforms, Twitter is widely used to investigate psychological and social processes, mostly because this platform provides a relatively easy access to the data with several Application Programming Interfaces (APIs). These APIs allow to extract a variety of information ranging from specific user information to general trends reflecting the discussion in the twitter sphere. Although useful, researchers should only draw measured conclusions based on these data. In fact, given the large scale of data available on Twitter servers, some APIs do not return all the available information but just a sample of it, without providing the parameters taken into account to perform such sample. However, properly using the combination of some APIs allow us to overcome this limitation by first, using non-textual data to investigate the structure of the users' social network allowing the identification of a specific sample, and then focus on the content of the message that provides affective and semantic information on a particular issue over time.

Several studies have investigated the use of Twitter data in a variety of political contexts, such as predicting election outcomes based on social media content (see Gayo-Avello, 2013) or observing difference in language usage between liberals and conservatives (Sylwester et al., 2015). In this paper, we take advantage of using both type of information extracted from Twitter, non-textual information to identify a sample of interest, and the content of each message posted to investigate the evolution of affective reactions towards a special topic over time. Particularly, we focused on the interest and sentiments towards environmental issues of Democrat and Republican supporters during the 2016 U.S. presidential election. We predicted that the interest in climate change will depend of specific events related to environmental issues occurring during the 2016 U.S. presidential election. We also predicted a polarization of sentiments towards environmental issues of Democrat and Republican supporters after the 2016 U.S. presidential election.

A social network analysis of the official Twitter accounts of the Republican and Democratic US Congressional Parties, and both official party candidates revealed a sample of approximatively 130K Twitter accounts susceptible of representing persons with strong Democrats (60K users) or Republicans (70K users) values. For this sample, we collected all tweets posted over the period of one year and identified tweets related to climate change by selecting only those containing the terms "climate change" or "global warming". In order to control for a possible spread of emotionality affecting different topics after the elections, we designated migration as our control topic. The final database contained approximatively 125K tweets referring to climate change and 420K referring to migration collected from a sample of 130K users sent from July 1st, 2016 to July 1st, 2017.

Results revealed that the numbers of climate change tweets posted by Democrats users were influenced by specific timely events related to environmental issues that occurred during the 2016 US presidential election. For instance, Democrats tweeted more about climate change than about migration during the first presidential debate, during the appointment of the new director of the EPA, and when the United States withdraw of the Paris Agreement. However, the interest of Republicans for climate change was only stronger than for migration at the time of the withdraw of the Paris Agreement. We interpret these results as reflecting a strong willingness from Republicans to refer to climate change with more positive words and by doing so, decreasing concerns about this matter, while Democrats accentuate the danger and fear linked to that issue. Interestingly, the pattern inverse was observed when the discussion was about migration, decreasing of 7% for Republicans and increasing of 12% for Democrats.

Twitter provides the largest accessible information about human behavior to date. Although accessing and interpreting these data could be challenging, it provides a tremendous opportunity to conduct social science research. In this study, we suggested a reliable and accurate way of gathering twitter data increasing the ability to examine social data on a variety of topics, over short or long periods of time.

Session 5: **Applications: Consumer Psychology (Part 2)**

Friday, June 8
4:30 pm - 6 pm

4:30 pm - 5:15 pm	*Michela Vezzoli* “Should they stay or should they go? Predict whether consumers are going to leave their energy supplier.”
5:15 pm - 6 pm	*Nikolas Zöller, Ingo Wolf, Tobias Schröder* “Political Ideology and Affective Attitudes Towards Mobility Innovation: Comparing Survey Responses and Twitter Data.”

4:30 pm - 5:15 pm

Michela Vezzoli

“Should they stay or should they go? Predict whether consumers are going to leave their energy supplier.”

Background. The Italian energy market is going through a liberalisation process. After many years of monopoly by the Enel supplier, it will be entirely liberalised from 2019. This full liberalisation will be the result of a process that has begun in 1999. In the initial phases of the liberalisation, the new companies entering the Italian market were mainly focused on attracting new customers. Now, the attention has shifted toward retaining existing ones. This change requires the building of statistical models to predict which clients are intentioned to churn (i.e., to leave the company), the understanding of the reasons behind this intention, and the development of strategies aimed at customer retention.

In this era of Big Data, these goals are made possible by the fact that companies are flooded with a massive amount of very different type of data (mainly, social demographic information and past interactions between client and company) that could allow a better understanding of the complex psychological dynamics of churn behaviour.

Objectives. Since the literature on churn prediction in the energy market is still absent, the primary aim of this study is to develop an initial churn prediction model in the electricity market. To attain the goal we have used data mining and machine learning methodologies and techniques. Secondly, we aim to detect what information about consumers are more predictive in this particular context and therefore, to shed some light on the reasons that lie behind churn behaviour.

Research question(s) and/or hypothesis/es. This study aims to explore and understand energy consumer behaviour through a data-driven approach. One of the essential characteristics of big data is the variety of information we can hold about consumers. The modelling of this amount of information could allow us to recognise the hidden value of consumer's characteristics that have never been considered before, possibly because of the theory-driven approach that typically characterises psychological research. On the other hand, modelling may confirm the value of some other features that were detected by theory-driven models of consumer behaviour.

Method/Approach. To build a predictive model we have used a data mining approach, which is an analytic process to find unknown relationships between the information about consumers and their future behaviour. In the first phases of the process, we have dealt with the issue of data quality: data needed to be cleaned and prepared for being modelled. After this stage, we ended up with a dataset composed of 81836 consumers, each of which owns one electricity domestic contract. The set of predictors consists of demographic (e.g. age), account (e.g. length of the contract), behavioural (e.g. the number of contacts the consumer had with the company) and socio-economic information (e.g. whether the customer lives in a wealthy or struggling area). In the second phase, the data were modelled with two machine learning algorithms: decision tree (CART and C5.0) and logistic regression.

Each model was first built on a training set (n = 57269; 70% of the dataset); then the best-trained model was tested on a set of unseen data (n = 24544; 30% of the dataset). To assess the model performance, we used the Area Under the ROC Curve metrics.

Finally, as the distribution of instances in the criterion variable is highly uneven (8% churners and 92 % non-churners), we had to handle the class imbalance issue. Class imbalance may cause churn models to break down because of the lack of information. To solve the uneven distribution within the criterion variable, we used the SMOTE resampling, which oversamples the minority class through the K-Nearest Neighbour graph. The SMOTE training dataset is composed of 53328 consumers. On this resampled dataset, we trained and tested the same algorithms we used for unbalanced data.

Results. We built six different models, each of which is different because of the algorithm used and whether or not the minority class was resampled. Logistic regression on unbalanced dataset reached the best performance (AUC on the training = 0.67 and AUC on the test = 0.68). In addition to providing the best performance, logistic regression achieved more stable results between training and testing than other models did.

For each feature, we calculated the odds ratio for assessing the direction and the strength of the predictive relationship between each predictor and the criterion. The features that increase the likelihood of churn are the regional area, the type acquisition channel, the type of the contract and whether the consumer has already churned on previous contracts. Furthermore, the features that decrease the likelihood of future churn are the length of the contract, the subscription to a loyalty program and having received cross-sell offers.

Discussion. Although the level of performance of the model was modest, we have achieved our primary objective. This model enabled us to reach a first understanding of consumer churn behaviour. We have found that subscription to a loyalty program is one of the most important predictors of our model. Being loyal, that is to be committed to the company over time, regardless of changes in competitors pricing or changes in the external environment, has crucial psychological importance. We do not know whether loyalty affects the likelihood of churn directly or if it has an impact because loyal consumer shares some characteristics which decrease the likelihood of churn. Therefore, we should figure out if loyalty is a protective factor against churn, what is its psychological role and whether the various components of loyalty may have a different impact on churn behaviour. The results we have reached open to a wide range of research opportunities. More specifically, we could try to expand the model towards consumer segments (e.g. domestic gas consumer or B2B), use different algorithms for building models, or understand more deeply the causal relationship between a specific predictor (e.g. loyalty) and churn behaviour through various experiments.

5:30 pm - 6 pm

Nikolas Zöller, Ingo Wolf, Tobias Schröder

“Political Ideology and Affective Attitudes Towards Mobility Innovation: Comparing Survey Responses and Twitter Data.”

Background

The effect of emotionally charged partisan stereotypes on information processing about political issues has long been acknowledged and studied [6, 9]. In the present work, we use traditional survey data and Twitter data to analyze how party preferences and party membership constrain attitudes towards different modes of transport; currently an important, yet also ideologically contentious issue in light of the global transformation towards a more sustainable, urbanized society.

Objectives and Research Questions

To elucidate the relationship between political partisanship, emotions, and affective attitudes towards mobility, we use a mixed-methods approach, combining classical social-science approaches with novel sentiment-analysis techniques of a corpus of topical tweets. Comparing the two methodologies we analyze what the limitations and opportunities of either approach are and whether the results of the online survey are reproducible via Twitter data, a data source that is easily accessible with little resources. Particular research questions are: How do party preferences constrain affective attitudes towards mobility? How can we automatically extract sentiments from mobility related tweets? Do sentiments towards mobility change over time and what does the time series for sentiments extracted from twitter data look like for particular transport modes? How do affective attitudes derived from survey data compare to sentiments derived from twitter data?

Method/Approach

We conducted a representative online survey (N=6047) whose participants rated affective meanings of different transport mode options on bipolar adjective scales in the dimensions evaluation (E, positive - negative), potency (P, weak - powerful) and arousal (A, calm - exciting), a well established three-dimensional model of affective meaning [3]. Participants rated the more traditional transport options e.g. bicycle, car, and public transport, as well as innovative mobility concepts e.g. electric cars, autonomous driving and carsharing. We then analyzed the ratings as a function of participants' party preferences, also assessed in the survey as a separate item.

As a second data source we used German Twitter data recorded from January 2016 to January 2017 which amounted to more than 300 million tweets and retweets. The data was filtered by a keyword list (i.e. "electric car", "autonomous driving", "car sharing" etc.) and a German language classifier to identify topical tweets. To identify party affiliated Twitter users we used the Twitter handles of politicians in the German parliament and created disjoint sets for each party from the politicians' followers' networks. For the five parties in the German parliament these range in numbers from 27927 Twitter ids for the christian social union (CSU) to 176436 for the green party (Bündnis 90 die Grünen). To gain insight into the emotional charge of the collected topical tweets we developed a novel three-dimensional sentiment analysis algorithm based on a mix of supervised deep learning and dictionary methods, where we treat each EPA-dimension separately.

After preprocessing and tokenization the tweets are mapped to word embeddings trained on the whole corpus of German tweets with the word2vec [4] algorithm. The word embeddings are then channeled to a combination of convolutional and gated recurrent neural networks for feature extraction and the concatenated feature vectors are finally fed to a fully connected soft-max layer for classification, an algorithmic architecture inspired by [8]. For supervised training of the neural network we used existing EPA – dictionaries that in total amounted to 2753 rated words and also conducted an online survey to obtain additional data. This data consisted of 500 words selected by frequency from mobility related tweets that were each rated by 30 participants as well as 1500 mobility related tweets that were labeled by 10 people each. During training we accounted for unbalanced class data by upsampling, used dropout regulation to prevent overfitting and used an automatic dictionary expansion [1]. On a nine-point scale we thereby obtain accuracies on our test data sample of 0.681, 0.667 and 0.686 for E, P and A ratings with an average absolute mean error per class label of 0.077, 0.073 and 0.070.

Results

For reasons of conciseness we restrict ourselves here to the case of flexible carsharing. Figure 1 exemplary shows the EPA-profiles for this transport mode differentiated for supporters of the five political parties that are represented in the German parliament. For both methodological approaches we see clear differences across party typologies, illustrating the correlation between political ideology and affective attitudes towards transport modes. The EPA-profiles obtained from the survey data agree with general partisan stereotypes, e.g. the supporters of the Green party (Die Grunen), which has a progressive and eco-friendly image, rate the innovative transport mode of flexible carsharing as more positive, more powerful and more exciting than the supporters of the other parties. On the other end of the spectrum are the supporters of the Christian Social Union (CSU), the most conservative party in the German parliament, who in contrast rate flexible carsharing as negative, weak and less exciting.

The EPA profiles obtained from the time-averaged Twitter data show the same relative tendencies between parties in all three affective dimensions evaluation, potency and arousal except for the supporters of The Left Party (Die Linke). Here the distinct nature of the communication medium Twitter and the limits of the automatic sentiment analysis approach of Twitter data become apparent for several reasons. To begin with, the demographics of Twitter are not representative of the demographics of the general population with strong biases in gender and age in the distribution of users [5]. However, the main reason for the deviation is the event and news driven communication behavior of Twitter users. Looking at the tweets of supporters of The Left Party that have been rated particularly negative, we find that a huge amount of these are retweets concerned with a particular event. One such example is the news that BMW, the company behind the carsharing service DriveNow passed on user data of a DriveNow user to the public prosecution office. Since data protection is a topic the political left feels strongly about this news was shared extensively among the Left Party's Twitter network and consequently affected the EPA-statistics disproportionately. This event driven nature is also reflected in the time series of derived sentiments for flexible carsharing for the party die Linke depicted in Figure 2. The strongest deflections in the sentiment curves can be assigned to single events.

Another significant difference between the results of the two data sources are the consistently higher values on the arousal scale in the EPA profiles generated with Twitter data. Two distinct sources of this effect can be identified. One being that Twitter users use the service instantaneously, while the situation in which a survey is taken is much more controlled. Secondly, the higher arousal values simply result from the increased probability for users to post messages about topics they are excited about.

Conclusions and Implications

We found strong empirical evidence for the effect of party preference on affective attitudes towards transport modes both in online survey data and in Twitter data. By using a three-dimensional instead of the commonly applied one-dimensional (E) sentiment analysis approach we highlighted some of the possibilities, limitations and characteristics of Twitter as a data source. In particular, with restrictions we were able to reproduce relative tendencies in the attitudes of party affiliated groups obtained from survey data, where the deviations can be explained to a huge extent through the event driven communication behavior. Another interesting finding are the observed consistent higher arousal states in the Twitter data. Furthermore, a promising feature of twitter as a data source is the possibility to obtain time series of attitude changes in real time, where the fluidity of affective attitudes becomes observable.

A possible continuation of this work would be to quantitatively analyze the biases that affect the Twitter sentiment analysis and in consequence develop techniques to counterbalance their influence. A starting point could be the systematic application of different filtering strategies to the tweet corpus and the decomposition of the survey data into different demographic sets for a subsequent comparative analysis.

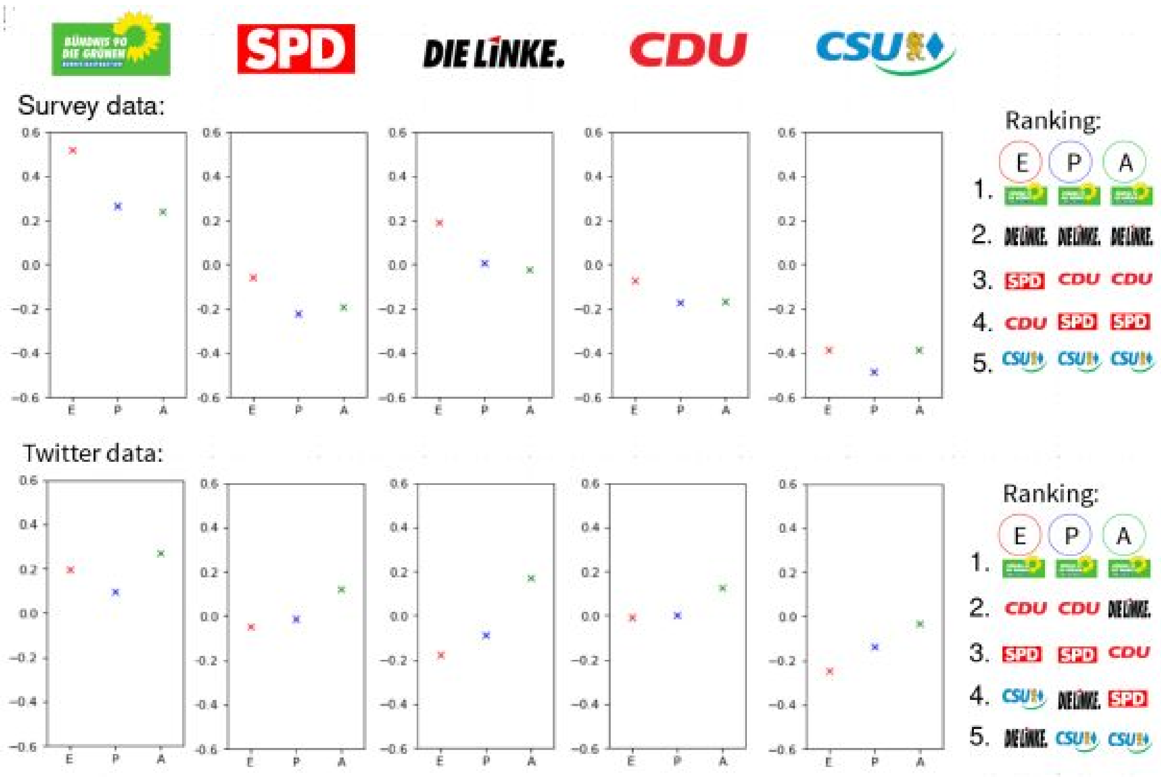

Figure 1: EPA profiles for the transport mode flexible carsharing obtained from survey data and via automatic sentiment analysis of Twitter data. The results are differentiated by party preferences for the different parties in the German parliament.

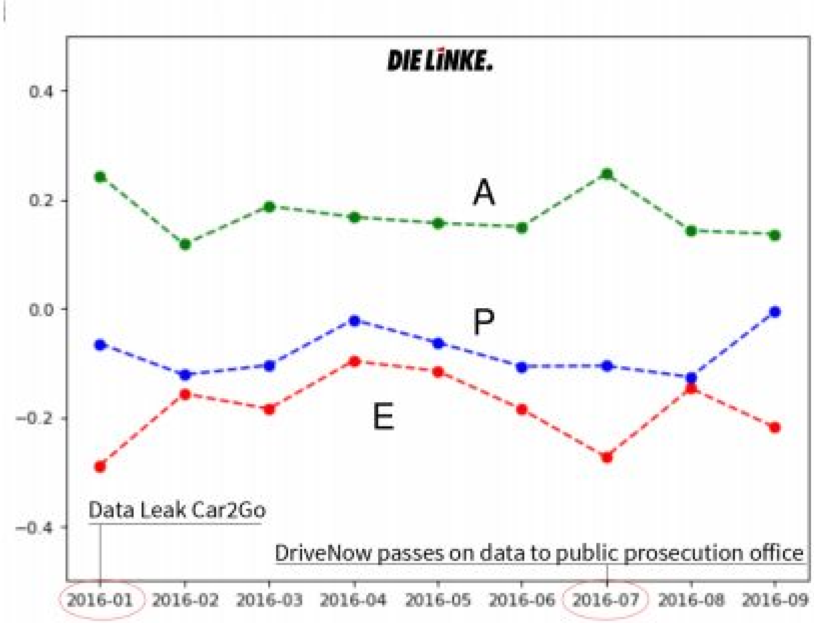

Figure 2: Sentiment values towards flexible car sharing extracted from the tweets of twitter users that follow politician from the party ”Die Linke” and no other party.

References

[1] Areej Alhothali and Jesse Hoey. Semi-Supervised Affective Meaning Lexicon Expansion Using Semantic and Distributed Word Representations. 2017.

[2] Jens Ambrasat et al. “Consensus and stratification in the affective meaning of human sociality”. In: Proceedings of the National Academy of Sciences 111.22 (2014), pp. 8001–8006.

[3] David R Heise. Surveying Cultures: Discovering Shared Conceptions and Sentiments. Hoboken: Wiley, 2010.

[4] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: arXiv preprint arXiv:1301.3781 (2013).

[5] A. Mislove et al. “Understanding the Demographics of Twitter Users”. In: ICWSM. 2011.

[6] Wendy M. Rahn. “The Role of Partisan Stereotypes in Information Processing about Political Candidates”. In: American Journal of Political Science 37.2 (1993), pp. 472–496.

[7] David S. Schmidtke et al. “ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words”. In: Behavior Research Methods 46.4 (2014), pp. 1108–1118.

[8] Dario Stojanovski et al. “Finki at SemEval-2016 Task 4: Deep Learning Architecture for Twitter Sentiment Analysis”. In: (2016), pp. 154–159.

[9] Drew Westen. The political brain: The role of emotion in shaping the fate of the nation. Public Affairs, 2007.

Session 6: **Methodological Issues (Part 2)**

Saturday, June 9
10:30 am - 12 pm

10:30 am - 11 am	*André Bittermann* “Extracting latent topics in large text corpora by taking metadata into account: Do empirical studies address psychological research topics differently than nonempirical studies do?”
11 am - 11:30 am	*Shuai Yuan, Katrijn Van Deun, Kim De Roover* “Cluster-wise Sparse Simultaneous Component Analysis: A Novel Method for Clustering Analysis on Multi-source Data.”
11:30 am - 12 pm	*Florian Pargent, Johannes Albert-von der Gönna* “Predictive Modeling with Psychological Panel Data.”

10:30 am - 11 am

André Bittermann

“Extracting latent topics in large text corpora by taking metadata into account: Do empirical studies address psychological research topics differently than nonempirical studies do?”

Background. Facing large text corpora, the extraction of information using traditional text analysis techniques such as thematic analysis or content analysis is challenging. Specifically, the identification of latent topics within hundreds of thousands of research database records is an endeavor which strongly benefits from automatization. Topic modeling, an unsupervised computational content analysis technique, reduces the text to various dimensions which are referred to as topics. Topic models based on latent Dirichlet allocation (LDA) have increasingly been used in psychological research and scientometrics. The more recently introduced structural topic model (STM) is an advancement of LDA since it allows for the inclusion of covariates and the examination of their effects on both, topic prevalence and content. In the context of scientometric investigations, metadata such as the study type (i.e., empirical or nonempirical study) can be used as such a covariate to gain further insight into psychological publication output patterns.

Objectives. Applying STM, differences between empirical and nonempirical psychological publications were examined regarding topic prevalence (i.e., how frequently the topic is addressed by publications) and content (i.e., the terms that constitute the topic).

Method. Psychological publication output in the German-speaking countries containing

German- and English-language publications from 1980 to 2016 documented in the PSYNDEX database was analyzed. STM was applied to a corpus of N = 237,625 publications. As input, the publications' controlled terms were used (a standardized vocabulary of keywords in psychology). Analyses were conducted using the "stm" package for the R programming language.

Results. Of all publications included, 42.9% were classified as empirical studies and 57.1% as nonempirical, respectively. Based on the publications' standardized keywords, 300 topics were identified. For 103 topics (34.3%), a significant difference in topic probability between empirical and nonempirical documents could be found. Topics that were most likely addressed by nonemprical studies referred to specific aspects of certain therapeutic approaches, psychological theories, psychoanalytical theory, father-son relations, and leadership. Most likely addressed by empirical studies were topics referring to specific aspects of academic achievement, electroencephalography, visual memory, and psychological testing. Drawing upon the topic referring to leadership as an example, it is shown that terms such as "employee attitudes," "supervisor-employee interaction," or "leadership style" were more likely addressed by empirical studies, whereas terms such as "human resource management," "management methods," or "organizational development" showed a higher probability of occurrence in nonempirical publications. Regarding this topic's prevalence, nonempirical publications had a more strongly increasing trend from 1980 to 2016 compared to empirical studies.

Discussion. In conclusion, the findings demonstrate the usefulness of structural topic modeling for investigating psychological research topics by taking the publications' metadata into account. It was shown that empirical and nonempirical publications differ regarding topic prevalence and content. This may indicate current blind spots of empirical evidence as well as topics that could be of interest for research synthesis approaches. Methodological limitations, possible applications, and future perspectives are discussed.

11 am - 11:30 am

Shuai Yuan, Katrijn Van Deun, Kim De Roover

"Cluster-wise Sparse Simultaneous Component Analysis: A Novel Method for Clustering Analysis on Multi-source Data."

Background. Psychological studies more and more often yield multi-source data, which consists of novel blocks of data (e.g. genetic data) and traditional blocks of data (e.g. survey data) collected from the same sample. Multi-source data could offer researchers valuable insights into the complex social mechanisms where several influences act together. To also take into account the possibly significant individual differences in such complex social mechanisms, clustering methods are needed for the analysis of multi-source data.

Objectives and research questions. Fully revealing the composite mechanisms underlying multi-source data is challenging, since the appropriate clustering methods should simultaneously detect both the subgroups and the cluster-specific associations or mechanisms. Additionally, in empirical practice, the methods should also be able to handle high-dimensional datasets which might include huge amounts of irrelevant information. The existing methods could not adequately address the clustering problems because they cannot advise cluster-specific mechanisms and (or) because they cannot deal with high-dimensional data. To facilitate clustering analysis on multi-source data, the objective of the current paper is to develop and evaluate a novel clustering method, which groups the observations that possess the same mechanisms while extracts cluster-specific linked variables that collectively suggest these mechanisms. To facilitate its wide applications, the novel methods should also demonstrate convincing performance in handling high-dimensional data. Furthermore, an efficient algorithm is developed for the estimation of the novel clustering method and a large-scale simulation study is conducted to examine the robustness and boundary conditions of the applications of the method.

Approach. In the current project, we have introduced and developed the Cluster-wise Sparse Simultaneous Component Analysis (CSSCA) method. CSSCA originates from the simultaneous component analysis (SCA) framework , a popular family of methods for multi-source data analysis. Though the original SCA is interesting for multi-source data integration, the interpretation of SCA becomes daunting when the data is high-dimensional. This drawback motivated the proposal of sparse simultaneous component analysis (SSCA), which results in components that can be interpreted on the basis of a (small) subset of the variables. Furthermore, to distinguish the common and distinctive mechanisms underlying multi-source data, the sparse DISCO-SCA is further proposed to take this challenge. Finally, CSSCA method is developed as a cluster-wise extension of sparse DISCO-SCA, which applies the same principal as cluster-wise regression and cluster-wise SCA. We further develop a multi-start alternating algorithm for CSSCA analysis, with pre-specified parameters. We conduct a simulation study to examine the performance of CSSCA analysis in different parameter settings. More specifically, the performance of the CSSCA method is compared when (1) the *noise level* is low versus high, (2) the *average congruence level* is low versus high, and (3) the *proportion of between-cluster variance* is low versus medium versus high. We further compare the performance of CSSCA with iCluster, a popular clustering method for multi-source data which is based on mean levels exclusively.

Results. From the results of the simulation study, we could conclude that the CSSCA method, together with its algorithm, is not sensitive to local minimum. Further, CSSCA consistently shows strong performance, quantified by excellent recovery of cluster memberships (indicated by *ARI*) and cluster-specific loading matrices (indicated by *GOCL*), in all conditions except for the very specific conditions where the substantially larger amount of total variance is explained by noise, other than the covariance structure. It is understandable though, since the exceeding amount of noise masks the *true* covariance structure and makes up a seriously biased one, which will be consequently used to detect unknown subgroups and further result in erroneous partitions. In the performance comparison between CSSCA and iCluster, we find that CSSCA clearly outperforms iCluster method in various conditions, especially when the proportion of between-cluster variance is small or medium. When the proportion of between-cluster variance is large, and the noise level is small, both methods exert very strong performance. . Nevertheless, under very specific conditions where substantially larger amount of total variance is explained by noise than the covariance structure, as discussed before, the performance of CSSCA would not be satisfactory and worse than iCluster's performance.

Conclusions and implications. In this paper, we develop and present a novel clustering method, CSSCA method, for clustering on multi-source data. The subsequent simulation study demonstrate the convincing performance of the method in all conditions but when a large amount of total variance is explained by noise. The superior performance of CSSCA over the existing iCluster in most simulation conditions also confirms the additional value of CSSCA method in analyzing multi-source data. To ensure and broaden the applicability of the CSSCA method, future researches are expected to develop and incorporate an efficient variable selection procedure and to ease the constraints of equal number of components over all possible clusters.

11:30 am - 12 pm

Florian Pargent, Johannes Albert-von der Gönna

"Predictive Modeling with Psychological Panel Data."

Background:

Longitudinal panels include several thousand participants and variables. Traditionally, psychologists use linear models to analyze small subsets of this data.

Objectives and Research Questions:

Methods from predictive modeling deal with large quantities of data and should also be considered. We illustrate these techniques on exemplary variables from the German GESIS Panel, while describing the choice of data preprocessing, imputation of missing values, model classes, resampling techniques, hyperparameter tuning, and performance measures.

Methods:

In analyses with about 2000 subjects and variables each, we predict panelists' gender, sick days, an evaluation of president Trump, income, life satisfaction, and sleep satisfaction. Elastic net and random forest models which are heavily used in machine learning were compared to dummy predictions in benchmark experiments.

Results and Discussion:

While good performance was achieved, the linear elastic net performed similar to the nonlinear random forest. Elastic nets were refitted to extract the ten most important predictors. Their interpretation validates our modeling approach and illustrates how both predictive and exploratory research questions can be pursued. Further modeling options are discussed.

Post-Conference Workshop

Mike Cheung

“Testing model driven hypotheses with Big Data: Applications with R.”

Saturday, June 9

2:30 pm - 6 pm