Usability of web scraping of open-source discussions for identifying key beliefs

Galit Gordoni[a], Holger Steinmetz[b] and Peter Schmidt[c]

[a]Academic college of Tel Aviv-Yaffo/Israel
[b]Leibniz Institute for Psychology Information, Trier/Germany
[c]University of Giessen/Germany

Pre-Conference Symposium: Big Data in Psychology 2019

outline

1. Introduction

2. Theoretical background

- Big Data adoption: The role of beliefs
- The Theory of Planned Behavior (TPB): Application and methodological challenges
- Web scraping: An alternative data collection method

3. Case study: Big Data adoption behavior in Israel

- Web scraping vs. self-report open-ended questionnaires
- Initial results

4. Summary and outlook

1. Introduction

- Research project on Big Data adoption in organizations.

- The project applies the Theory of Planned Behavior (TPB):

The role of beliefs in the decision process of Big Data adoption

→ potential merits of integrating web scraping
in the initial stages of a TPB – driven research design

2. Theoretical background

Big Data adoption – The role of beliefs

- The Big Data vision
- Adoption of Big Data technologies in organizations
- The unexplained gap between shared beliefs and actual adoption behavior
- Lack of research on the role of beliefs in the decision-making process

(Ekbia et al., 2015; Raguseo, 2018; Shin, 2016)

TPB: Application to Big Data adoption

This study applies the most fruitful and well replicated psychological behavior theory—the Theory of Planned Behavior (TPB; Ajzen, 1991).

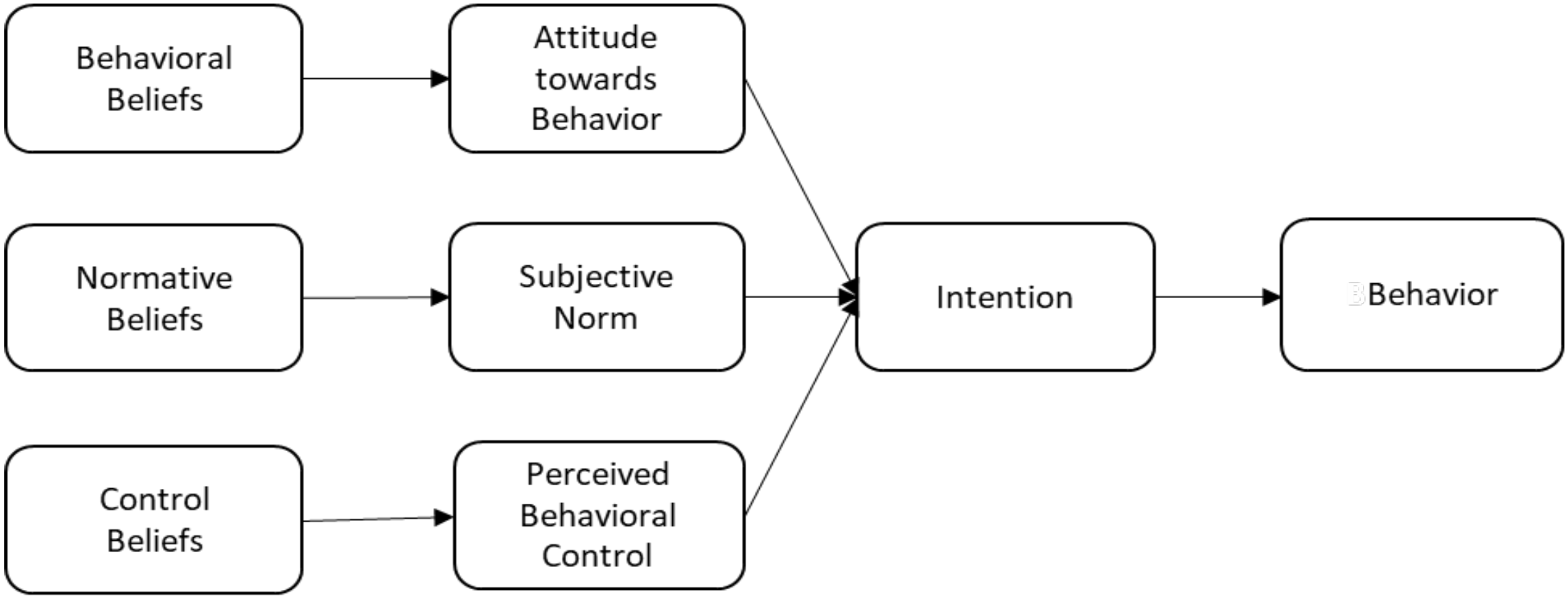
The main strength of the TPB is that it is well suited to integrate:

the cognitive foundation of motivational and decision processes (i.e., the beliefs), with:

- attitudes,
- perceptions of social legitimization,
- efficacy
- feasibility of the behavior in question

(Bamberg & Möser, 2007; de Leeuw et al., 2015; Klöckner, 2013).

Theoretical framework: Theory of Planned Behavior (TPB)



The Theory of Planned Behavior (TPB) – Core model (Ajzen, 1991; Fishbein & Ajzen, 2010)

TPB: Application to Big Data adoption

Study aims:

- To identify key beliefs of decision makers about potential benefits, costs, social expectations, barriers, and facilitators of the adoption behavior.
- To use the identified beliefs as explanatory variables in the TPB based process model.
- To add knowledge relevant for practical endeavours to change Big Data adoption rates.

Added value:

Application of a general theory for explanation instead of ad-hoc explanation with selected variables

TPB: Application to Big Data adoption and methodological challenges

Recommended first step in a TPB application study (Fishbein & Ajzen, 2010):

Elicitation study for identifying key beliefs

Requirements

- To identify readily accessible beliefs ⟹ using open-ended questionnaire
- To represent the diversity of the population ⟹ Representative sample

Common procedure

Qualitative study – content analysis of responses to open-ended questions

Methodological challenges

small *n*, interviewer effect, topic complexity and topic sensitivity effects

Web scraping: An alternative data collection method

Scraping data sources for assessing psychological constructs

- Free and open-source data
- Big dataset in minimal time and budget constraints
- User generated data and metadata
- Behavioral data: clicks, likes, comments, shares, etc.
- Self-motivated content creation
- Unobtrusive (no interviewer effect)
- Development of standardized procedures (Landers et al., 2016)
- Traditional analytic techniques used in psychology can be applied

(Farnadi et al, 2016; Gucciardi, 2017; Moessner et al. 2018)

3. Case study: Big Data adoption behavior in Israel

Web scraping was conducted by a well-known Israeli research firm specializing in web scraping of users' discussions in open source websites (Buzzilla)*

- The firm does not scan private profiles
- All the information is scanned according to the Israeli law and to Facebook regulations
- The firm is a GDPR compliant: When scanning Facebook it does not take or save personal information of the users involved in the discussion

* The firm is the exclusive provider of social listening monitoring and research to all governmental offices and agencies in Israel. Data collected by the firm is used for example in scholar's policy guidelines studies conducted in research organizations in Israel (see, for example: Rafaeli et al., 2018).

Case study: Research methodology

1. Creation of sampling frame:

- Mapping the web for all the relevant data sources.

  Inclusion definition: Professional discussions on the topic of Big Data

Challenges: Identifying the target population units and size of target population in an ongoing changing web environment

Procedure: Creating Boolean Queries with relevant key words.

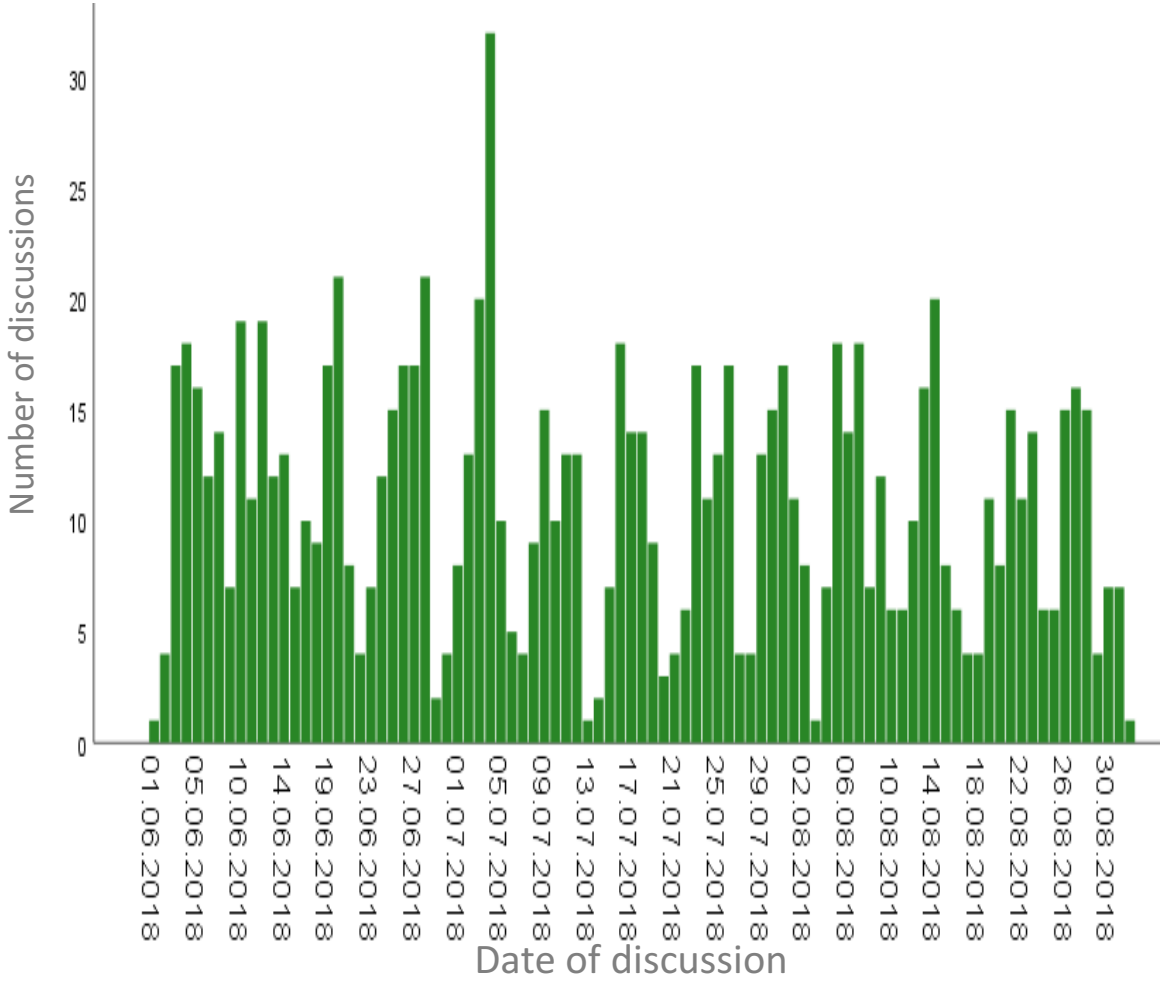
Examples:

"Data mining"
"Machine learning"
"artificial intelligence"

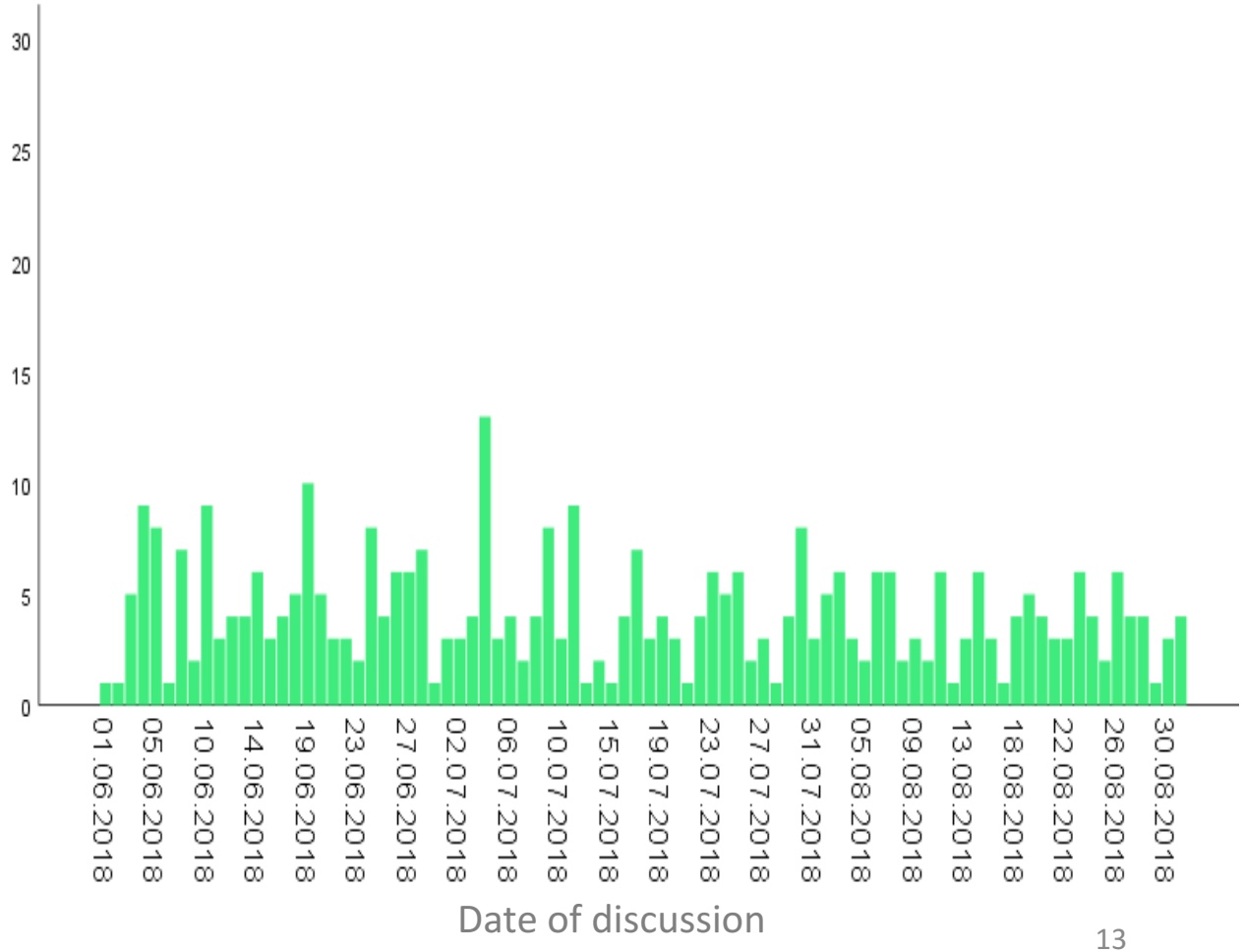
Availability of data: All open-source Israeli web sites

Case study: Big Data adoption behavior in Israel
Daily discussions volume – all population
(June-August 2018)

General discussions (*N*=987, M_{daily}=11)



Authentic discussions (*N*=359, M_{daily}=4)



Case study: **Research methodology**

2. Sampling procedure – defining arenas' distribution

- Articles
- Social networks
- Forums
- Blogs
- Twitter

Aim:

Selection of a quota sample, according to arena size, for identifying key beliefs concerning Big Data adoption in Israel

Challenges:

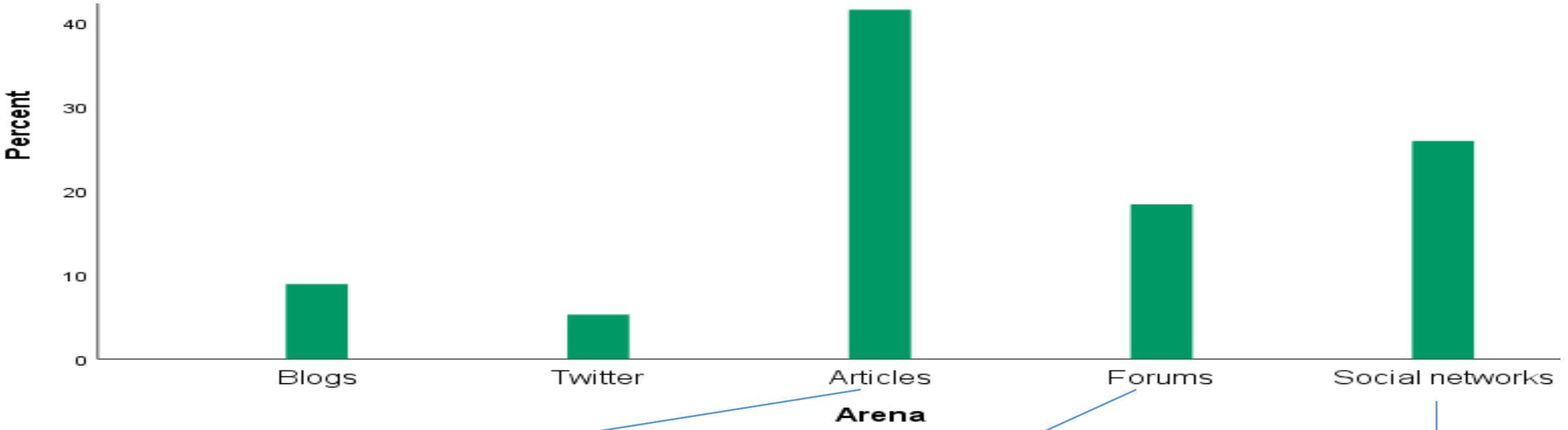
Ongoing changes in arenas' characteristics

Procedure:

- Producing sampling quotas according to volume of discussions, across arenas, in a defined time period
- In each arena: Sampling discussions based on relevancy rating of each result

Case study: Big Data adoption behavior in Israel

Arenas' distribution of Big Data authentic discussions (n=353)



149 articles in 14 leading news websites

Section of publication:

- Business
- Technology
- High-tech
- International market

Forums: 66 discussions

Main forums:

Rotter (n=19)

Fxp (n=18)

Facebook pages (93 discussions)

Main pages include:

- Job search groups
- The Innovation Authority of Israel

Number of comments (M=37.91, SD=114.47, range (min(2)-max (1325))

Case study: Big Data adoption behavior in Israel
Identifying stakeholders of Big Data – sample (n=148)-examples

Participants characteristics: self – definition (n=129)	**Main topics**	**Added value for theory-driven belief research**
Programmers (30%)	• Feasibility of developments in the fields of Big Data and AI	Potential people that will advance or support the behavior (normative beliefs)
High-tech employees (16%)	• Future development of the Israeli High-tech industry and its international status • The departure of skilled workers	
Employees of small and medium companies that integrate Big Data (19%)	• Concerns of technological unemployment • Barriers in application: costs and existing infrastructure	Potential people that will not support the behavior (normative beliefs)

Case study: **Research methodology**

Content analysis of discussions (n=148)

Aim: Determining the most frequently expressed beliefs

Methodology:

counting the number of times a given comment had been produced

1. Identifying semantic units
   - advantages/ disadvantages associated with big data adoption
   - significant others that could support or oppose big data adoption
   - factors that could impede or facilitate the adoption of big data
2. Classification of the semantic units into generic categories
3. Reclassification of the semantic units
4. Calculation of number of semantic units in each category

(de Leeuw et al., 2015; Patch, 2005)

Web scraping - Initial results
Salient behavioral beliefs (advantages and disadvantages)

Advantages: **Category (n of semantic units=130)**	**Comments Quote - examples**
Gain insights and make better decisions (n=61, 47%)	"……In the system I work with, in the psychometric evaluation center there is a reduction in the number of factors from 50 to 10, the factors prove to have meaning…." (Twitter- A comment in a discussion on the use of machine learning for decision making)
Disadvantages (n of semantic unit=83)	
Fears and concerns – mainly concerning unethical use and potential unemployment (n=41, 49%)	"Bad, so bad. Enables mainly upgrade of those with power and in social control on behalf of the more "simple" people." (A comment to an article in a news site on adoption of machine learning in organizations)

Case study: Big Data adoption behavior in Israel Self-report open-ended questionnaires

Participants:

Israeli MBA students and undergraduate Information Systems students in professional and managerial positions

- 25 respondents
- Gender: 56% women (n=23)
- Age: Mean (29.3), SD (4.7), min-max (24-45)
- Education: 32% MBA students, 68% IS undergraduate students
- Job position in the organization:17 different job positions (n=22)

Setting: Self- administrated open-ended questionnaire completed in the context of a 20-minute classroom session during march 2019.

Analysis: Content analysis was carried out and sub-categories were developed to capture the emerging themes according to the TPB model (based on de Leeuw et al., 2015 and Patch, 2005)

Initial results - Salient behavioral beliefs (advantages and disadvantages)

Category (n of semantic unit)	Sample Quote	Respondent characteristics
Advantages		
Advancing adaptability to clients needs (n=7)	• Fitting advertisements to those that can be interested in them	• Data analyst in web advertising company
Disadvantages		
Ethical issues (n=10)	• Collection of private and personal data on the employees • Prohibited use of sensitive information • Huge amount of data, in part very personal, transferred in the company and exposed to many employees	• Economist in a telephone company • Application manager in the High-tech industry • Learning developer in an insurance company

Question warding:
What do you see as the advantages/disadvanteges, for you, of advancing the usage of big data in the organization you work in, during the upcoming year?

4. Summary and outlook

4. Summary and outlook

Web scraping: Complementary or alternative data collection for testing psychological research questions?

Additional steps:

- Comparison of results of both methods of data collection, and integration of combined results in the following steps of the TPB research design
- Identifying external data sources for validating the results

  Examples: Representative surveys of the target population
  - Survey of decision-makers in France companies (Raguseo, 2018)
  - Survey of decision-makers in German companies (Commerzbank AG, 2018)
  - Survey on AI/cognitive computing in the U.S. (IBM, 2018)

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211.

Bamberg, S., & Möser, G. (2007). Twenty years after Hines, Hungerford, and Tomera: A new meta-analysis of psycho-social determinants of pro-environmental behaviour. Journal of environmental psychology, 27(1), 14-25.

De Leeuw, A., Valois, P., Ajzen, I., & Schmidt, P. (2015). Using the theory of planned behavior to identify key beliefs underlying pro-environmental behavior in high-school students: : Implications for educational interventions. Journal *of Environmental Psychology, 42*, 128-138.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., ... & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, *66*(8), 1523-1545.

Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., ... & De Cock, M. (2016). Computational personality recognition in social media. *User Modeling and User-Adapted Interaction*, *26*(2-3), 109-142.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. Psychology Press.

Gucciardi, D. F. (2017). Mental toughness: progress and prospects. *Current Opinion in Psychology*, *16*, 17-23.

Klöckner, C. A. (2013). A comprehensive model of the psychology of environmental behavior—A meta-analysis. *Global environmental change*, *23*(5), 1028-1038.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological methods*, *21*(4), 475-492.

Moeser, G., Moryson, H., & Schwenk, G. (2013). Determinants of online social business network usage behavior—applying the technology acceptance model and its extensions. *Psychology*, *4*(04), 433-437.

Patch, C. S., Tapsell, L. C., & Williams, P. G. (2005). Overweight consumers' salient beliefs on omega-3-enriched functional foods in Australia's Illawarra region. *Journal of nutrition education and behavior*, *37*(2), 83-89.

Rafaeli, S., Leck, E., Albo, Y., Oppenheim, Y., & Getz, D. (2018). An Innovative Approach for Measuring the Digital Divide in Israel: Digital Trace Data as Means for Formulating Policy Guidelines. *Samuel Neaman Institute for National Policy Research*.

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, *38*(1), 187-195.

Shin, D. H. (2016). Demystifying big data: Anatomy of big data developmental process. *Telecommunications Policy*, 40(9), 837-854.