2018 42nd IEEE International Conference on Computer Software & Applications

Understanding Mindsets Across Markets, Internationally

A public-private innovation project for developing a tourist data analytic platform

Kristoffer Jon Albers
Dept. of Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby, Denmark
kjal@dtu.dk

Mikkel N. Schmidt
Dept. of Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby, Denmark
mnsc@dtu.dk

Morten Mørup
Dept. of Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby, Denmark
mmor@dtu.dk

Marisciel Litong-Palima
Department of Management, Society & Communication
Copenhagen Business School
Frederiksberg, Denmark
mpa.msc@cbs.dk

Rasmus Bonnevie
Dept. of Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby, Denmark
rabo@dtu.dk

Fumiko Kano Glückstad
Department of Management, Society & Communication
Copenhagen Business School
Frederiksberg, Denmark
mpa.msc@cbs.dk

***Abstract*— This paper presents an ongoing public-private innovation project that integrates unsupervised machine learning tools and a marketing theory, in order to analyze segment-based attitudes and behaviors of tourists. Our case study involving the major governmental tourism stakeholders emphasizes the importance of developing a user-friendly data analytic pipeline that carefully considers users' data collection procedure, easy access to the back-office computation algorithms, an interactive output data analysis workflow and its visualization. At the end of this paper, we present our vision to further develop a cloud-based tourist data collection platform.**

Keywords-component; Tourism data analysis; unsupervised machine learning; behavior prediction; data visualization; case study

I. INTRODUCTION

While the quantitative methods employed by marketing & tourism academicians have long existed based on the theory-driven positivistic perspective, a recent emergence of the Big Data trend has triggered data scientists and Artificial Intelligence (AI) researchers to enter data-driven research on tourists' behaviors and decision-making predictions. Such data-driven approach enables not only marketing & tourism academicians to integrate the quantified qualitative research method into their existing quantitative research, but also marketing & tourism practitioners to identify diverse consumer segments and their behaviors that are useful to develop segment-specific product concepts and their positioning strategies. UMAMI, Understanding Mindsets Across Markets, Internationally, is a governmentally funded project that addresses the aforementioned challenges to integrate the marketing & tourism research tradition with the contemporary machine learning technology. The project is funded for the period of 2017-2020 and consists of three public tourism authorities (Visit Denmark, Wonderful Copenhagen, Visit North Sealand), one private tourism stakeholder (Visit Carlsberg) and two universities (Copenhagen Business School and Technical University of Denmark). The ultimate scope of this project is threefold: i) identify top three new segments with potential market growth and acquire in-depth knowledge about them; ii) develop innovative tourism products and digital communication strategies targeting the three selected segments; and iii) develop a tourism data eco-system, i.e. a prototype of a segment-based data collection platform that can subsequently can accumulate and analyze tourists data across multiple markets; and integrate user-specific behavioral prediction tools that can potentially be developed by project participants after the completion of the project.

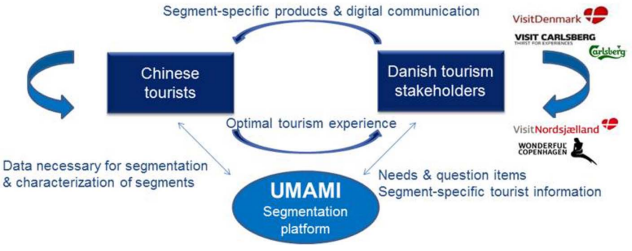

Figure 1. Concept of the UMAMI data analytic platform.

0730-3157/18/$31.00 ©2018 IEEE
DOI 10.1109/COMPSAC.2018.10221

This paper presents our initial attempt to develop an overall workflow of the tourism data analytic framework. The next section presents the overall concept and elements of our data analytic framework. Section III discusses lessons learned from our first case study. In the final section, our future perspectives and strategy to develop a segment-based tourism data collection platform will be presented.

II. A UMAMI DATA ANALYTIC FRAMEWORK

Our UMAMI data analytic framework is positioned in Eisenmann's platform-mediated networks [1] as a platform that mediates two types of stakeholders: i) tourists and ii) tourism business stakeholders including public tourism authorities and private tourism industries. As shown in Figure 1, our segment-based data analytic platform collects data from consumers in order to extract segments characterized by their personality and personal value priorities [2][3]. The tourism business stakeholders supply survey question items and/or their consumer data to be analyzed per segment. The segment-specific consumer information provided by our data analytic platform enables the tourism business stakeholders to provide segment-specific products & digital communication strategies, which will guide tourists to select optimal tourism experience suitable to their personal value priorities. The UMAMI data analytic platform consists of five elements described below.

A. *Data collection and generation of input data: Ask*

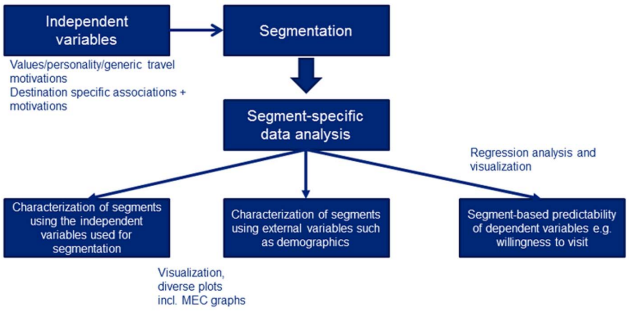

Figure 2. The UMAMI data analytic framework.

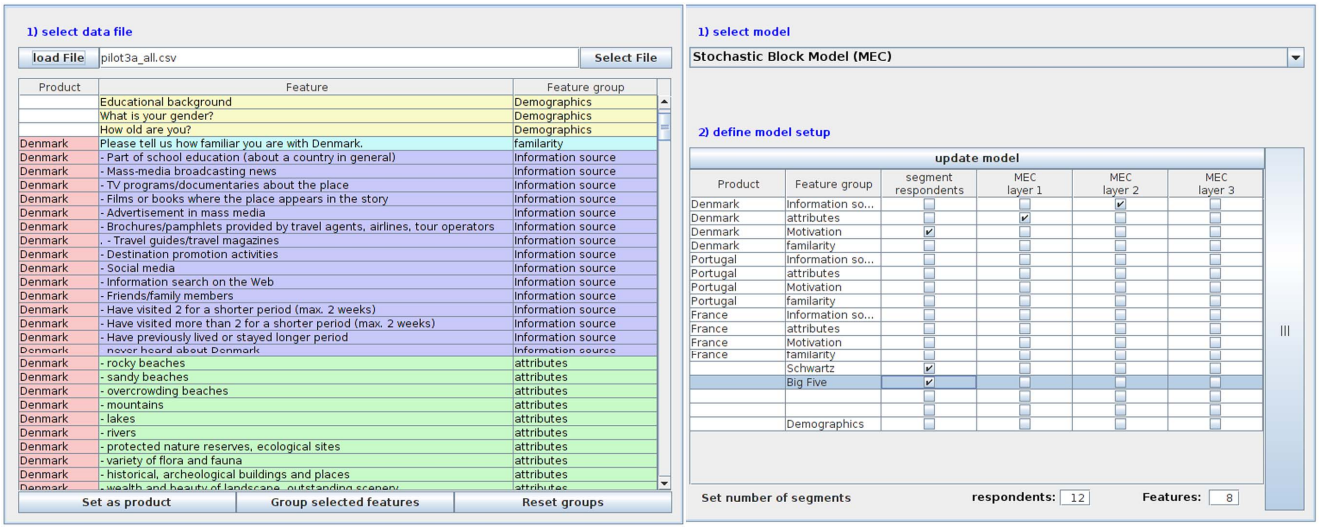

Figure 3. Screenshots, illustrating the GUI design for classifying input variables (left) and specifying independent features and model parameters for the segmentation procedure (right).

The current data analytic platform is designed for analyzing data collected from an online survey questionnaire consisting of question items classified into three types of variables (see: Figure 2). The first type of variables is independent variables to be used for segmenting respondents according to their value priorities and/or personality traits. Our current case study employs question items extracted from Schwartz theory of basic human values [3] and Big Five [2]. However, researchers have flexibility to test other question items not only value priorities and personality traits but also other scales to measure lifestyles and motivational aspect of tourists. The second type of variables is external variables used for characterizing segments such as demographics, their previous travel experience (i.e. actual past behaviors) and their knowledge structure of tourism products such as travel destinations. Finally, dependent variables are used to predict behavioral intentions such as willingness to visit a destination. This process of collecting survey responses is described as the "Ask" process displayed in Figure 4.

The responses from the online questionnaire are usually delivered in a comma-separated values (CSV) file format automatically generated from a majority of online survey administration platforms. As shown in Figure 3, the UMAMI data analytic platform provides a user friendly graphical user interface (GUI) that enables users to both classify the types of variables and specify what characteristics of a product to be analyzed per segment as well as define appropriate model parameters for the segmentation procedure, including the size distribution and number of inferred segments. Integrated with the input CSV file, these user specifications enables the segmentation of respondents based on the specified *independent* variables as well as computations for the user-defined characterization of the extracted segments

based on the specified *external* variables. Since the segmentation and the segment-specific characterization are based on the machine learning tools integrated in the data analytic framework, the process including data conversion, segmentation and characterization is categorized as "Compute" process in Figure 4. Though the status of the compute process is continuously communicated to the user by the GUI, these computations rely on external performance-optimized implementations, designed such that they can be executed on various distributed architectures, including cloud-based high-performance servers. This design enables the analysis of "bigger data" and complex segmentation procedures, without the need for the graphical interface to be run on a dedicated workstation.

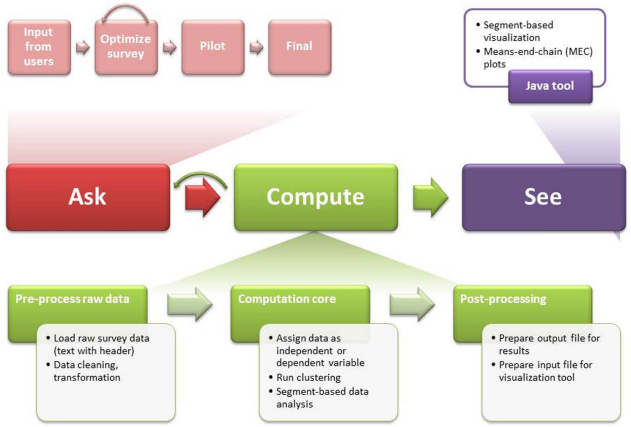

Figure 4. The UMAMI data analytic pipeline

B. *Segmentation tool: Compute*

The Compute process of the UMAMI data analytic platform integrates unsupervised machine learning tools. Our intension in the final product is to enable users to select suitable segmentation algorithms from options. In the current prototype system, the platform employs a Bayesian stochastic blockmodel that partitions respondents and independent variables (type 1 variables) simultaneously according to binary responses to the independent variables (type 1 variables based on [2] and [3]).

The model assumes that the survey data with N respondents and Q binary question items can be represented as a binary matrix $\in \{0,1\}^{NxQ}$, where $[\boldsymbol{X}]_{nq}$ is equal to 1 if respondent n answered positive to question item q. Binary matrices of this form can be interpreted as a bipartite graph, where links between the respectively N and Q nodes are given by the adjacency matrix represented by $\boldsymbol{X}$, such that $[\boldsymbol{X}]_{nq} = 1$ implies an edge between respondent n and question item q.

This interpretation allows us to employ one of the many existing Bayesian methods for graph clustering, including the stochastic blockmodel [4-6]. In the bipartite setting, the model assumes that the respondents are split into K_N groups, and the independent question items are split into K_Q groups, under the assumption that the probability of observing an edge only depends on the group membership of the respondents and question items. Mathematically we define the model as:

$$x_{nq} | z_n^{(N)}, z_q^{(Q)} \sim \mathrm{Ber}(\eta_{z_n^{(N)}, z_q^{(Q)}}) \quad (1)$$

$$z_n^{(N)} \sim \mathrm{Cat}(\boldsymbol{\omega}^{(N)}), \quad \boldsymbol{\omega}^{(N)} \sim \mathrm{Dir}(\boldsymbol{\alpha}^{(N)}) \quad (2)$$

$$z_q^{(Q)} \sim \mathrm{Cat}(\boldsymbol{\omega}^{(Q)}), \quad \boldsymbol{\omega}^{(Q)} \sim \mathrm{Dir}(\boldsymbol{\alpha}^{(Q)}) \quad (3)$$

$$\eta_{kl} \sim \mathrm{Beta}(a, b) \quad (4)$$

Where indices n and q range over all respondents and independent question items, and $x_{nq} \equiv [\boldsymbol{X}]_{nq}$ is modelled as a Bernoulli random variable with probability of η_{kl} depending only on group memberships. The group membership of respondent n is $z_n^{(N)}$, which is an integer from 1 to K_N , and similarly we have $z_q^{(Q)}$ ranging from 1 to K_Q. The model description is completed by the group-to-group connection probabilities $\eta_{kl} \in [0,1]$ and the group membership probability vectors $\boldsymbol{\omega}^{(N)}$ and $\boldsymbol{\omega}^{(Q)}$, which are normalized such that $\sum_{k=1}^{K} \omega_k = 1$. We put standard priors on both.

As with most Bayesian models, exact inference of the Bayesian posterior is intractable, but an advantage of the stochastic block model is that it belongs to the class of conjugate exponential family models, allowing us to employ efficient approximate inference algorithms. In particular, we use variational inference [7-8] over the entire set of latent variables $\boldsymbol{\theta} = \{ z^{(N)}, z^{(Q)}, \boldsymbol{\omega}^{(N)}, \boldsymbol{\omega}^{(Q)}, \boldsymbol{\eta}\}$ which minimizes the Kullback-Leibler divergence between an approximation $q(\boldsymbol{\theta})$ and the true posterior $p(\boldsymbol{\theta})$, as:

$$\mathrm{KL}\big(q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta}|\boldsymbol{X})\big) = \mathrm{E}_{q(\boldsymbol{\theta})}\left[ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{X})}\right] \quad (5)$$

In practice, this is implemented by minimizing a surrogate objective known as the ELBO, and using a simple parametric form of q where we assume that each $z_n^{(N)}$ and $z_q^{(Q)}$ is modelled a posteriori by an independent categorical distribution. We remove the remaining latent variables from the optimization problem by employing the variant known as KL-corrected variational inference [9], and solve the optimization problem using off-the-shelf optimization algorithm such as BFGS.

The output computed by the segmentation tool is converted from array to javatext as shown in the "Compute" process in Figure 4.

C. *Visualization: See – value priorities and demographics of segments*

The segmentation result of our case study is displayed in Figure 5. The yellow dots in the bottom plot indicate that a respondent "x" has selected a personal value question variable "y" such as "Being creative is important" from a presented list of multiple choice questions. For example, in the second segment (Segment 2) from the left in the bottom plot, intersections with the third and fourth variable clusters from the top are dominated by yellow color. This means that

No. 61579-00001A, Understanding Mindests Across Markets, Internationally is funded by Innovation Fund Denmark

majority of respondents belonging to Segment 2 selected most of the choices grouped as the third and fourth variable clusters (corresponding to the light red and light blue blocks in the upper plot). On the other hand, the intersection between Segment 2 and the bottom variable cluster are dominated by black dots. This means that the respondents belong to Segment 2 have not responded to a group of variables listed at the bottom (the light green variable block in the upper plot).

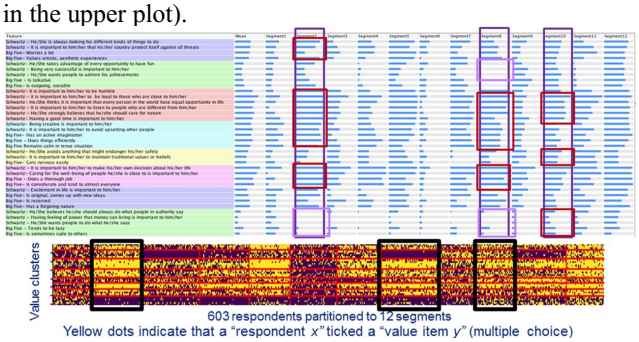

Figure 5. Screen shot of the segmentation result

This case study has analyzed 600 British respondents collected from an online survey. For example, according to the results displayed in Figure 5, Segment 2 is characterized as "open-minded", "extravert" and "growth-minded", while Segment 8 as "social-focused" and "introvert". Regarding the demographics, majority of members in Segments 2 are female frequent travelers with multi-cultural identity while members in Segment 8 are dominated by males with mono-cultural identity who are less frequent travelers.

D. Visualization: See – consumers' association with products

One of the objectives of our data analytic platform is to support tourism practitioners to develop segment-specific tourism products and digital communication strategies. For achieving this objective, the platform has integrated a classic marketing theory called Means-End Chain (MEC) theory [10][11]. While the value-based segmentation is a useful approach for developing products and their positioning strategies, Reynolds & Gutman [11] points out some weakness: "*these rather general classifications fail to provide an understanding, specifically, of how the concrete aspects of the product fit into the consumer's life* [11]". The MEC theory [10] complements this weakness by focusing on the connection between the attributes of products (the "means") and the personal values (the "ends"). The MEC is usually analyzed in the form of qualitative interview using a special interview technique called "laddering technique". As shown in the illustration left side of Figure 6, an interview typically starts with a question, "which product attributes are important for you". When interviewees selects one of the attributes displayed with a product in question, the interviewer tries to elaborate his or her question one step deeper by further asking "why is that attribute important to you?" which eventually guide interviewees to reveal their personal values connected with their choice of the product in question. These connections revealed by the laddering technique are also known as "association networks", which represent the higher-order knowledge structure of consumers represented by their product choice [11][12].

Our case study has integrated this MEC element into our online survey by modifying the qualitative interview format into a quantitative multiple choice question: "Please select all the associations you relate to (a product)". Specifically, we selected three products (travel destinations: France, Denmark and Portugal) and prepared common attributes that are generally used to measure images of a travel destination [13],[14]. Examples of attributes are, among others, "big city" "folklore", "wealth of beauty", "shopping", "bad weather", "expensive", "quiet and peaceful" describing a particular place. Following a procedure also used in the domain of concept learning and semantics [15],[16], we asked respondents to select attributes they associate with a destination, which eventually reveals respondents' knowledge structure about a destination. Subsequently, we also included another multiple choice question "which of the following items are relevant criteria in the choice of your travel destination?" to measure respondents' motivation to select a destination. Although all these elements are not directly connected, co-occurrence between product attributes and motivations to select a product establishes implicit attribute-motivation links. Our data analytic platform attempts to complete implicit MEC links by analyzing segment-specific attribute-motivation links.

To contrast and identify the particular segments of interest, the GUI allows the user to interactively explore the MEC linkage presented per inferred segment as illustrated in the right screen of Figure 6. The user specifies how many features to list in each MEC layer, which are shown and sorted according to the support of the features for the respondents in the particular segment. The user can interactively select features from the lists and will be presented with various similarity scores between the selected features and the visible features in the other MEC layers.

Since the segments have been extracted and characterized based on personal value priorities and personality traits, the segment-specific data analysis enables tourism practitioners to investigate what people who share common value priorities associate with a product and what are their motivation for selecting the product. Such insightful knowledge enables tourism practitioners to develop new products tailored to a specific segment and position the products targeting this specific segment.

E. Prediction of behavioral intensions

The last important element is to compute predictability of segment-specific behavioral intensions such as "willingness to visit a destination". Such indicator helps tourism practitioners to identify a segment with potential growth. In order to compare predictabilities of value-based segments against those of other general demographic variables such as age and educational backgrounds, the linear regression analysis has been employed. Variables with categorical data (e.g. segments, educational backgrounds) are converted into dummy variables enabling the linear regression analysis. Table 1 demonstrates that the value-based segments

extracted in this case study, used as an independent variable has an *R* score "0.275", to predict the "willingness to visit" variable: "I plan to visit Denmark as a tourist at some point in the future", while *R* scores by age and educational backgrounds are 0.144 and 0.215, respectively. The same tendencies are observed for the willingness to visit Portugal and France, too. In other words, the value-based segments indicated better predictability than other variables such as age and educational backgrounds that are usually used by tourism and marketing practitioners to target segments.

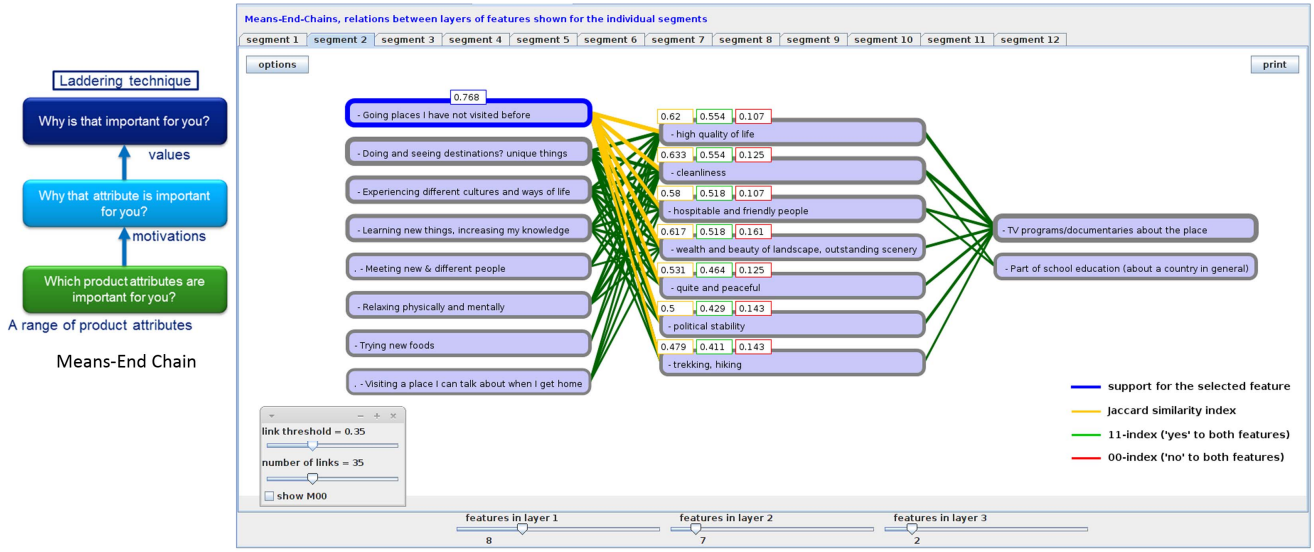

Figure 6. Screen shot of the interactive visualization of MEC links. By selecting a feature, the GUI shows the support for the feature as well as similarity metrics between the selected feature and features in other MEC layers.

TABLE I. REGRESSION ANALYSIS

Independent variables	Dependent variable	R	R Square	Adjusted R Square	Sig.
Model summary (Denmark)					
Value-based segments	I plan to visit Denmark as a tourist at some point in the future	**0.275**	**0.076**	**0.059**	**0.000**
Age	I plan to visit Denmark as a tourist at some point in the future	0.144	0.021	0.019	0.000
Education	I plan to visit Denmark as a tourist at some point in the future	0.215	0.046	0.040	0.000
Model summary (Portugal)					
Value-based segments	I plan to visit Portugal as a tourist at some point in the future	0.196	0.038	0.020	0.016
Age	I plan to visit Portugal as a tourist at some point in the future	0.034	0.001	0.001	0.411
Education	I plan to visit Portugal as a tourist at some point in the future	0.141	0.020	0.013	0.021
Model summary (France)					
Value-based segments	I plan to visit France as a tourist at some point in the future	0.216	0.047	0.029	0.003
Age	I plan to visit France as a tourist at some point in the future	0.010	0.000	0.002	0.799
Education	I plan to visit France as a tourist at some point in the future	0.207	0.043	0.036	0.000

The UMAMII data analytic framework is equipped to visualize all these results presented in this case study, i.e. characteristics of segments such as value-priorities, personal traits, demographics, implicit MEC links (association networks) and regression analysis, in a highly user-friendly manner. It enables users to interactively select segments to be analyzed and compare characteristics and predictabilities of a segment, but also knowledge structure, attitudes and behavioral intension to multiple products in a systematic manner.

III. LESSONS LEARNED FROM THE CASE STUDY

This case study attempted to demonstrate how machine learning tools can be integrated into the overall workflow of segment-specific tourism data analysis framework of which purpose is to enable tourism and marketing practitioners to identify target segments with potential growth and to develop and position a new product concept tailored to the specific segments. From this point, the current early prototype developed has been well received by industrial stakeholders involved in our project. Through this case study, several lessons have been learned. First, although the artificial intelligence technologies are expected to give huge impacts on the marketing and tourism practitioners' decision support in the future, the advanced machine learning algorithms alone cannot fulfil the requirements from the tourism and marketing stakeholders. Integration of theories of values, personality traits, tourism and marketing with the technologies has been the key for serving the objectives defined by our public-private innovation project. Second, the stakeholders have different skillsets in handling data. Therefore, user-friendly data handling and visualization is

the mandatory elements to be included in the comprehensive data analytic framework. Third, the data analytic framework should be designed to give users to flexibly select data analytic algorithms according to characteristics of input data and objectives of output data. This implies final remark that the functionality and performance of the data analytic framework depends on a context, i.e. who is collecting what types of data for what purpose. As shown in Eisenmann's platform-mediated networks, it is highly important to design a business model that suits the context in a specific business scenario. The ownership of data and analytical tools has to be carefully defined according to a context in a specific business environment. This also has brought future considerations in our ongoing project, which will be explained in the next section.

IV. Future Perspectives and Challenges

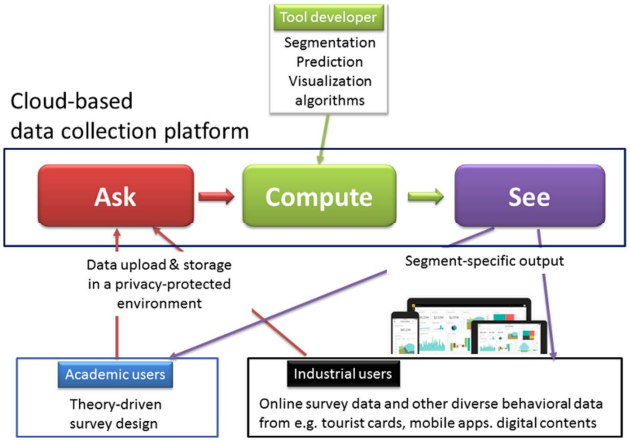

Figure 7. Future perspectives of the UMAMI data analytic platform

The current governmentally-funded public-private project is expected to develop a technological framework to be used by public and private tourism stakeholders at the end of the project. The recent emergence of a cloud-based computational environment has enabled to connect various database available from such diverse public and private business entities. For example, by integrating a short questionnaire asking about personal value priorities in an application of ticket sales at museums, amusement parks or public tourist information and web-sites, segment-based behavioral predictions of tourists may become available. Figure 8 depicts our future cloud-based tourism data collection platform integrating "Ask", "Compute", and "See" processes where computational tools are uploaded by tool developers and data can be uploaded by the respective stakeholders. The more data is accumulated in a platform and segmentation is performed using accumulated data, the more accurate will the predictability of segment-based behavioral inference be. On the other hand, the recent data protection law took effect in Europe regulate exchange of personal data across multiple business entities. Therefore, our future challenge is to investigate possible *data curation* frameworks and design a business model suitable for the context given under the current business environment.

Acknowledgment

This work has been conducted as part of UMAMI: Understanding Mindsets Across Markets, Internationally, No. 61579-00001A funded by Innovation Fund Denmark.

References

[1] T. R. Eisenmann, "Platform-Mediated Networks: Definitions and Core Concepts," Harvard Business School Module Note 807-049, September 2006.

[2] R.R. McCrae and P.T. Costa Jr., "Personality trait structure as a human universal," American psychologist, vol. 52(5), 509-516, 1997.

[3] L. Sagiv, S. Roccas, J. Clecluch and S.H. Schwartz, "Personal Values in Human Life" in Nature Human Behaviour, vol. 1, 630–639, 2017 .

[4] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First step" in Social Networks 5.2, 109-137, 1983.

[5] M. N. Schmidt and M. Mørup, "Nonparametric Bayesian Modelling of Complex Networks: An Introduction", IEEE Signal Processing Magazine, 30(3): 110-128, 2013.

[6] T. P. Peixoto, "Bayesian stochastic blockmodelling", arXivP 1705.10225 [stat.ML], 2017.

[7] C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[8] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians", in Journal of the American Statistical Association 112.518, 859-877, 2017

[9] J. Hensman, M. Rattray, and N. D. Lawrence, "Fast Variational Inference in the Conjugate Exponential Family", in Advances in Neural Information Processing Systems 25, 2888-2896, 2012

[10] J. Gutman., "A Means-End Chain Model Based on Consumer Categorization Processes" in Journal of Marketing, vol. 45, 60-72, Spring 1982.

[11] T. J. Reynolds and J. Gutman, "Laddering Theory Method, Analysis, and Interpretation" In Journal of Advertising Reearch, 11-31, February/March 1988

[12] F. ter Hofstede, A. Audenaert, J-B E.M. Steenkamp and M. Wedel, "An Investigation into the Association Pattern Technique as A Quantitative Approach to Measuring Means-End Chains," In International Journal of Research in Marketing 15, 37-50, 1998

[13] S. Baloglu and K. W. McCleary. "A Model of Destination Image Formation." Annals of Tourism Research 26(4): 868-897, 1999.

[14] A. Beerli and J.D. Martin. "Factors Influencing Destination Image." Annals of Tourism Research 31(3): 657-681, 2004.

[15] S. De Deyne, S. Verheyen, E. Ammel, W. Vanpaemel, M. J. Dry, W. Voorspoels, et al. "Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts." Behavior Research Methods, 40(4), 1030 1048, 2008.

[16] J. B. Tenenbaum, C. Kemp, T. L. Griffiths and N. D. Goodman, "How to Grow a Mind: Statistics, Structure, and Abstraction" Sience, vol. 331, 11 March 2011.