Chapter 2

Strategic motivation for data augmentation

When visiting a marketing conference, it is very likely that at some point the allegory of the "Tante-Emma-Laden"[1] is mentioned. People in the middle of the 20th century shopped in such a small general store. "Tante Emma" stands for customer focused individual information. Through her personal contact to all customers, she exactly knew her customers' needs, wants, and preferences. However, "Tante Emma's" capabilities were not primarily due to her outstanding targeting abilities. Rather, mobility, variety, and competition were so limited that people did not have a choice.

Today, the competitive environment is much tougher. People are more mobile and better informed, so that loyalty has become a valuable asset. Personalized, individualized, and relevant communication has become the superior marketing goal, because marketers and managers have recognized the that a customer directed communication increases loyalty and boosts sales. Such communication is only possible, if the companies have knowledge on their customers, much like "Tante Emma". But today, the face-to-face

[1] In English, "Tante-Emma-Laden" is translated as "mom-and-pop store" or corner shop. The literal translation of "aunt Emma" personifies the seller as a friendly and knowledgeable counterpart.

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_2

information acquisition process is largely replaced by digital means. Every transaction process is recorded, every scan of a loyalty card produces data, and every click on a website can be tracked. One of the main challenges today is the collection and integration of data in order to form a holistic picture and to supply customers with as relevant information as possible.

In this chapter, we analyze the strengths and weaknesses of companies regarding individual communication. We name the opportunities and threats arising from the economic, technological, legal, sociological and psychological environment. Data augmentation can be a solution to the information lacks of companies. With a SWOT analysis, we eventually motivate data augmentation as a relevant strategy for marketing and management today.

2.1 Internal conditions within the company

The internal conditions of data augmentation in database marketing are divided into the three sub parts marketing goal, practice, and measures. In the following, we describe the trend of customers having become the focus of marketing, because treating customers individually leads to a maximum profit for the company. Accordingly, knowledge about the customers needs to be acquired. This knowledge is used in marketing practice for targeting customer segments and addressing them personally. The success of marketing campaigns is measured by conversions, which quantify the success of the marketing efforts. The internal conditions are an important factor in the SWOT analysis that we conduct in chapter 2.4. From them, strengths and weaknesses of the company are derived, facilitating the benefit of data augmentation for the marketing strategy.

2.1.1 Customer focus as a major marketing goal

Only relevant marketing communication attracts the attention of customers. A customer is a person who has already had a contact with a company, e.g.

bought a product or service, has taken interest in doing so, or subscribed for promotional communication. Relevance is achieved by attractive offers for products customers are interested in – presented to the customers at a suitable point in time, when they are receptive to these offers. However, with the low costs of advertising and direct marketing today, especially for online communication, customers are often contacted with all kinds of irrelevant promotions. Customers ignore these advertisements (Cuthbertson & Messenger, 2013). If the information conveyed by companies is not relevant, the overall attention to a company's marketing activities decreases. Relevance is not only important for the company itself, but also in its competitive environment. The so-called consumer addressability, the ability to customize communication based on individual customer information, has been shown to be a competitive advantage (Chen & Iyer, 2002).

Relevance can only be achieved, if customers are treated individually tailored to their interests and needs. The perspective of marketing and communication has changed from product to customer centric (Kelly, 2007; Garg, Rahman, & Kumar, 2010). Customers are regarded from a 360° perspective: buying behavior, demographic and socioeconomic profiles, lifestyles, attitudes, and media exposure (Kamakura & Wedel, 1997). With regards to the customers, markets and target groups can be defined and identified, offers can be created and communicated, and buys and follow-up buys can be initiated (Meffert, 2000, p. 12).

The customer focus aims at fostering each individual customer relationship in a way that it is maximally profitable for the company. Loyalty is one of the key aspects, so that customer value and equity are benchmarks for the operative and strategic directions of marketing (Helm & Günter, 2006; Huldi, 2002). An important marketing goal is to retain and to develop customers. Whereas the focus of retaining customers is on preventing them from defecting to competing market participants, the focus of developing customers is to increase their value for the company. Both implicate a constant effort, especially in so-called "non-contractual" relationships, where

every new sale has to be earned (McCrary, 2009). An increase in revenue is for example achieved by exploiting an individual's willingness to pay. Exemplary actions for these strategies are cross-selling, up-selling, and increasing the efficiency of customer contacts (Ang & Buttle, 2009).

These goals can only be achieved, if companies know as much as possible on their customers (Behme & Mucksch, 2001). Data is the basis for treating customers individually in order to increase their performance and value. It is gathered, analyzed, interpreted, and used for customer differentiation by a database marketing team. An extensive database is necessary, providing all information needed to offer the best product at the best time to the best price. The optimal allocation of resources is one of the main goals of database marketing. The demand for more information, more individualization, and more personalization is in the interest of both the customers and the companies. Customers expect offers tailored to their expectance. The companies expect efficient media usage and high conversion rates.

Data augmentation is one database marketing tool to acquire the data necessary for customer differentiation. It can provide information not otherwise available regarding product preferences, general purchasing power, preferred communication channels, life-cycle related information indicating the right time to offer a product, and much more. It enables a better description of the customers, complementing the 360° view necessary to provide relevant marketing communication.

2.1.2 Targeting in marketing practice

In order to implement individual customer strategies, a direct marketing approach is needed. Scovotti and Spiller (2006, p. 199) define direct marketing as follows: "Direct marketing is a data driven interactive process of directly communicating with targeted customers or prospects using any medium to obtain a measurable response or transaction via one or multiple channels". In direct marketing, every communication event is dedicated to a target

group, a selected group of prospective customers for which the cost-benefit ratio of marketing is highest. Targeted advertising is the essence of below the line marketing. In contrast to above the line media, wastage can be avoided (Bruhn, 2009, p. 191; Greve, Hopf, & Bauer, 2011).

When customers are selected for a marketing campaign, it is referred to as targeting. Targeting is the ability to differentiate customers based on data in order to provide them individual and personalized offers. The word is derived from the verb "to target", which means to reach exactly the group at which an offer is aimed. It has been a research topic since the late 1990's. Dong, Manchanda, and Chintagunta (2009) give a good overview on the first studies concerning targeting. Targeted marketing reduces the cost of production, distribution, and promotion (Bull & Passewitz, 1994). Hopf (2011) and Kelly (2008) state that targeting is valuable to consumers, because it gives them the sentiment that products are immediately available, findable, accessible, understandably prepared, and personally selected. They gain the impression that products and benefits are real and original, that someone cares for them, and that their data is safe at all times.

The vision of targeting is to develop it to a one-to-one marketing process, in which every customer receives relevant products, services, and information at the optimal point in time (Link & Hildebrand, 1993, p. 29). However, such a perfect targeting is possible with a disproportional effort (Freter, 1997, p. 46). The next best solution is to find meaningful customer segments, which can be reached through consequent data usage (Liehr, 2001). In order to profitably treat these segments, they have to be identifiable, quantifiable, addressable, and of adequate size (Rapp, 2002a, p. 67). They should be stable, responsive in a way that response is similar among segment participants, and actionable in a way that the marketing strategy is consistent with the company objectives (Hattum & Hoijtink, 2010).

Individualization is needed on every touch point. While (e-)mail is the most important channel for direct marketing (Bult & Wansbeek, 1995), many of the marketing tools today have the possibility to personalize and

target (Iyer, Soberman, & Villas-Boas, 2005); e.g. newsletter, websites, and mobile or social platforms. As channels become more fragmented, with the new ones not necessarily replacing, but adding to the old ones, an overall customer view is needed (Rhee, 2010). Much of the communication to the customers is done electronically and is controlled through so-called contact optimization technologies. These are based on certain business rules and contact policies, and are mainly fed by customer data (J. Berry, 2009).

Targeting has the goal to select or segment customers. *Selecting customers* refers to choosing the best target group for a given offer or communication, often given certain constraints like target group size, budget, or required conversion rate. *Segmenting customers* refers to allocating customers to several target segments in order to distinguish them, usually in terms of offers, prices, or creative appeal. Both aim at reaching a maximal conversion rate. The conversion rate is calculated by the number of conversions, divided by the number of recipients of a marketing campaign. In the first case, customers are selected according to *one specific* value of each selection variable to reduce the target group size. In the latter case, the optimal distribution of customers *among all* variable values is of interest.

Data is the basis for targeting, thus all targeting related areas of research focus on getting better knowledge from data; e.g. customer segmentation, data mining, and data fusion. The customer base is a crucial core asset, because it has a central function in the success of the company (Wirtz, 2009, p. 54). Core assets in general are distinguished by having an intrinsic value, being rare, not imitable, and not substitutable (Barney, 1991). The efficient use of customer data is a core competence of a company. The combination of the core asset customer database and the core competence database marketing can lead to a significant competitive advantage.

To purposefully segment customers is not easy. These segmentations are only possible if sufficient data is available (Behme & Mucksch, 2001). The information is sufficient, if it is relevant to the purchase behavior, if it has informative value in terms of media and channels to be used and accessibil-

ity, and if it supports the identification and measurement of customers. It is profitable and not volatile (Freter, 1997, p. 90ff). Marketing data often faces the problem of missing data in one way or another (Kamakura & Wedel, 2000). The classical market segmentation criteria used for segmentations, such as demographics and other identification data, are usually not directly relevant to the purchase behavior (Brogini, 1998, p. 113ff). A brand and its product are not preferred over competitive products because of demographic criteria like age and gender (Petras, 2007), but because they have a certain function and benefit for the customer. Descriptive data collected with the customers is often incomplete, inconsistent, and mostly aged (Kelly, 2007). Consequently, intelligent typologies aligned with the customer purchase behavior are coming into focus (Homburg & Sieben, 2005; Leitzmann, 2002). Such information is available from transaction data (Kelly, 2007).

But transaction data also ignores important information. So-called soft facts are worth knowing in order to optimally reach customers with campaigns. Customer databases have limited information on the following data categories (Breur, 2011; Dialog Marketing Monitor, 2012; Liehr, 2001).

- *General characteristics and preferences of the customers:* needs and motives, characteristics of the customers, information on media usage, attitudes towards daily routine, work life, leisure time, and family
- *Product and purchase related preferences:* product purchase motivation in the competition environment, attitudinal and evaluative data, such as quality perception and brand advocacy
- *Post-purchase behavior and opinions:* satisfaction with products and services bought, likelihood to recommend

In order to predict future customer preferences, today's data needs to be analyzed (Putten, 2010, p. 16) and enriched. While some targeting goals can already be achieved by mining the existing data, some of the necessary information is only available elsewhere. DWHs usually contain

hard facts only. With these, it is difficult to conform to the requirements of successful customer segmentation. However, it is possible to acquire them externally and augment them to the customer database (Hippner, Rentzmann, & Wilde, 2002). Missing information can be retrieved from various sources, e.g. market research or other internal and external sources.

2.1.3 The marketing measure: conversion

The main KPI for targeted campaigns is the conversion. A conversion is the reaction to a marketing communication, e.g. a sale, a response, the participation in a raffle, or any other specified desired customer activity. The desired action is specified in advance and has to be measurable. Conversions can only be calculated for direct marketing campaigns with a clearly defined and identifiable target group (Rossi et al., 1996). Reactions need to be able to be traced back to the customer.

Conversions are the ultimate goal of marketing. They pay off in terms of sales, qualifying sales leads, and building customer relationships (Roberts & Berger, 1999, p. 9f) and are monitored closely. The increasing challenges in the economic framework, as described in chapter 2.2.1, require advertising efficiency (Laase, 2011). One goal of marketing is to improve targeting performance, or targetability (Chen, Narasimhan, & Zhang, 2001). The ability to target customers can have a higher and more durable impact on marketing performance than other marketing activities (Chen et al., 2001).

For planning purposes, marketers try to predict the conversion probability for individuals, for specific segments, or for the campaign as a whole. Those customers with the highest conversion probability are selected for direct marketing campaigns, taking into consideration the costs. Ideally, a return on investment can be calculated from a model for individual customers (Ratner, 2001a). In every target group selection, there are more and less prospective customers. The overall conversion probability is a mixture of the relative concentration of target customers (Smith, Boyle, & Cannon,

2010) in a selection, depending on the individual conversion probability of the prospective customers versus the individual conversion probability of the ones mistakenly selected. The challenge of predicting conversions is to identify the most profitable customers, rather than those being most likely to respond to promotional offers (McCrary, 2009).

The conversion probability is calculated taking into account as many factors as possible. It can be divided into a baseline probability of purchasing and time, contact, and purchase history related probability factors (Moe & Fader, 2004). The baseline probability of purchasing is relatively stable and can be attributed to the characteristics of customers. The others factors are volatile and change depending on the context. The marketing instruments used to stimulate it are different from those stimulating the "ad hoc" conversion probability. Data augmentation results are generally able to improve the knowledge on the baseline probability of purchasing, rather than on purchase situation related factors.

2.2 External conditions around the company

In the following, the external conditions for data augmentation in marketing are explicated. The economic environment poses several challenges that companies need to adjust to. The technological development offers opportunities that companies can turn to their account, with risks arising from missing out the trends. The legal environment provides the framework for all company activities. Respecting the borders is mandatory. The sociological and psychological environment comprises expectations, needs, apprehensions, and anxieties of the consumers. We have outlined the external conditions already in our first data augmentation study (Krämer, 2010). It is retraced and adjusted here where applicable. The external conditions are relevant to the SWOT analysis that we conduct in chapter 2.4. Therefrom, chances and risks are derived, leading to explicit potentials and limitations for the application of data augmentation in marketing.

2.2.1 Economic environment

The economic environment poses many challenges that companies need to cope with. Companies of all branches are faced with increasing complexity and dynamics (Huldi, 2002) and the number and efficiency of competitors aggravate the competition environment (Link, 2000). Frequently mentioned challenges are cost pressure, diminishing marginal utility, increasing speed of innovation, shortened product life cycles, and raising product homogeneity (Tiedtke, 2000). These developments are not new, but they keep governing many management decisions (Hippner, Leber, & Wilde, 2002). Due to the increasing opening and liberalization of markets and the resulting international expansion strategies, more market segments exist. Through globalization, new key markets are made accessible, and successful business models are transferred abroad (Meffert & Bruhn, 2009, p. 457). Only if realistic potentials for the own company are recognized and effectively implemented, is it possible to successfully operate and increase the company value on the long run (Huldi, 2002).

As a result, all company divisions are striving for success and have to deliver measurable results (Blattberg et al., 2008, p. 458f). Especially in marketing, where the central tasks are mainly determined by other divisions, the transparency on all steps and expenditures is important (marketinghub, 2009). For cost cutting reasons, advertising budgets are examined carefully. Online media, which feature both low costs and high transparency in terms of advertising impact, are chosen increasingly often. The percentage of advertising expenses dedicated to online advertisements is raising constantly (Dialog Marketing Monitor, 2009). The newer trends of mobile marketing and social media marketing are equally expedient and are increasingly used for marketing purposes (Dialog Marketing Monitor, 2012).

Several strategies have been developed to encounter the economic challenges. The success probability of individual activities and their measurement is raised by the coordinated and targeted steering of database mar-

keting (Schweiger & Wilde, 1993), so that the overall allocation of resources is improved. With the markets being fragmented, the product homogeneity and the willingness to switch between sellers, companies must differentiate their products and personal communication (Schweiger & Wilde, 1993; Blattberg et al., 2008, p. 7). The increasing need for individualization in conjunction with price competition and complexity, but also an unmanageable number of customers and the consequent inability of companies to address all these customers in person, is encountered with the standardization and automation of products, services, and communication actions. The so-called mass customization concepts comprise modulated features and communication packages, which can be individually designed, but are then provided automatically (Piller, 2006). Mass customization is supposed to live up to the needs and wants of the customers, while at the same time being cost efficient (Meffert & Bruhn, 2009, p. 459ff).

2.2.2 Technological environment

Through the technological development in the past years, data collection and usage is enabled and impelled. New information and communication technologies have replaced the personal communication between companies and customers (Meffert & Bruhn, 2009, p. 460). Most of today's transactions, call center inquiries, and other contacts at various touch points are recorded electronically (Breur, 2011; Kelly, 2007). These electronic footprints allow for detailed customer analysis and a holistic customer picture. The wealth of customer data has been recognized and referred to as the gold of the 21st century (Hebestreit, 2009; Singh, 2013). Many of these developments pose the possibility of individual content and addressability (Bensberg, 2002, p. 164f). Based on data, behavior can be analyzed, learned from, and triggered. Nowadays, most companies have integrated CRM systems and a high-capacity DWH (Hippner, Rentzmann, & Wilde, 2002).

Because of the increasing capacities and the progressive digitalization, the data volume stored at companies is constantly growing (Baker, Harris, & O'Brien, 1989; Behme & Mucksch, 2001, p. 9). New data sources in terms of new channels, new technologies, and new customer touch points are continuously emerging (Breur, 2011; Hipperson, 2010). At the same time, real-time CRM is possible today enabling greater intelligence through better performance of analytics, growing data volumes, and higher speed of deployment (Acker, Gröne, Blockus, & Bange, 2011). Storage costs are rapidly decreasing (Breur, 2011; Dull, Stephens, & Wolfe, 2001; Kelly, 2007). Network connections for technically linking sources are inexpensive (Bleiholder & Naumann, 2008). Likewise, the cost of data collection is decreasing.

The usage of online media has rapidly increased in the last 20 years. As reported by the *ARD/ZDF Online Study*, 76% of the German population used online media in 2012 (Eimeren & Frees, 2012), as opposed to 44% in 2002 (Eimeren, Gerhard, & Frees, 2002) and 7% in 1997 (Eimeren, Oehmichen, & Schröter, 1997). As a saturation of consumption is almost reached, the growth started to level in 2010 (Eimeren & Frees, 2012) and has been below one percentage point from 2011 to 2012 (tns Infratest, 2012). Additionally, mobile devices have taken the form of little computers, containing just as much information as personal computers with its internet access and many more useful applications. Social media is regarded as another media channel with its own rules and conditions. All of these media experience a rapid dispersion and are quickly adapted especially by young people (Dialog Marketing Monitor, 2012). The establishment of new devices leads to a multiplication of usage situations and new market models.

2.2.3 Legal environment

There are three major legal questions related to data augmentation in database marketing: Is data augmentation legally sound? Which data categories are permitted to be augmented? Who may be contacted with a

targeted campaign after having augmented the data? The following explanation is based on German law and would have to be reviewed for applications in other countries.

Privacy is very important for all data augmentation activities (Breur, 2011). The Bundesdatenschutzgesetz (BDSG), the German federal data protection act, governs all questions related to personal data, i.e. data related to natural persons (BDSG, § 3) directly or indirectly enabling the recognition of individuals (Arndt & Koch, 2002). It guards every individual from being affected in his or her personal rights (BDSG, § 1). The *preventive ban subject to permit* applies, meaning that collecting, processing, and using personal data is acceptable only, if a law explicitly permits or mandates it, or if the concerned person has agreed (Arndt & Koch, 2002). The BDSG was adapted by the federal government and is effective since September 2009 (Schaar, 2011). The amendments comprise, amongst others, rules and regulations regarding address trading, market research and opinion polls, and third party address processing (Eickmeier & Hansmersmann, 2011). Companies have to respect the BDSG as soon as they collect, process, or use data with data processing equipment (BDSG, § 1).

Two data categories do not adhere to the BDSG. Anonymous data enables the recognition of individuals only with a disproportionate high effort in terms of time, cost, and manpower (BDSG, § 3). Anonymous data can be used for developing the optimal marketing mix without concerns regarding data protection (Freter, 1997, p. 446). Likewise, aggregated data is not considered personal data, because no conclusion can be drawn from them on individuals. However, if an individual person is collated to a particular group on which certain details are known, this is considered a personal reference (Arndt & Koch, 2002).

It is explicitly appreciated by the second amendment to the BDSG to fuse data with the objective of avoiding unnecessary advertising in targeted marketing (BDSG, § 28). Companies are allowed to store a reasonable amount of additional data on their customers in order to use it for the

selection of targeted campaigns (Plath & Frey, 2009). The notion of *store* relates to modifying existing data in a way that rightfully collected data is added (Däubler, Klebe, Wedde, & Weichert, 2010, p. 461). The notion of *rightfully collected* means that the data was collected with the person concerned (BDSG, § 4), at a place where the data is publicly available, or where the responsible authority would be allowed to publish them (BDSG, § 28). The notion of *a reasonable amount* has to be regarded for each case individually. The collection of data directly from a person requires a permission (BDSG, § 4), unless a law explicitly permits or mandates it. There are constraints to the collection of additional data, which require a considerable care of so-called sensitive data. Sensitive data comprises race and ethnical family background, political opinions, religious beliefs, union memberships, health, and sexuality (BDSG, § 3).

The legitimacy of data collection is not equal to the legitimacy of personal contact by advertising. To this effect, the BDSG was even tightened by the new amendments. While postal mailings are still admissible without a permission of the recipient (Lambertz, 2009), commercial emails or SMS require an active consent, referred to as opt-in (Pauli, 2009). It means that a tick box on a website, for example, cannot be prefilled, but has to be actively ticked by the customer. This consent may not be interlinked with the general signing of a contract with the company (BDSG, § 28).

The basis for all questions related to competition law is the Gesetz gegen den unlauteren Wettbewerb (UWG), the German act against unfair practices. It protects the competitors, consumers, and other market participants. It prohibits, amongst other things, unfair and misleading business activities, as well as unacceptable disturbance (UWG, § 1/3/5/7). While the BDSG and the UWG pursue different purposes, they come to an agreement in terms of advertisement (Plath & Frey, 2009). Exceptions to the BDSG are stated in the UWG, such as the personal contact by email in connection with the purchase of a product or service, the advertisement of similar goods and services, or if the customer is fully aware of how to object

to the personal contact (UWG, § 7). For these exceptions, no permission is necessary before contacting a customer personally by email. Above all, the UWG prohibits marketing actions suitable to exploit the inexperience of children and teenagers in terms of business (UWG, § 4).

Another category of data, online usage data, is regulated in the Telemediengesetz (TMG), the German telemedia act. Online websites are permitted to save online usage data for advertising and market research purposes. These profiles, however, can only be saved in pseudonomized form (TMG, § 15). Once a unique identifier is missing, the data does not fall under the regulations of the BDSG.

2.2.4 Sociological and psychological environment

Since the 1980's, an increasing information overload has been observed, meaning that only about two percent of information a customer is confronted with is perceived (Kroeber-Riel, 1988). Especially advertisement is ignored: It is only perceived by customers, if it is relevant and if there is a benefit related to it (Mahrdt, 2009, p. 12f; McKay, 2009). The perception probability rises with increasing accordance between the communicated benefit and the personal interests of the customers (Schweiger & Wilde, 1993). Direct marketing is supposed to separate spam from real information (Roberts & Berger, 1999, p. 15). Spam is not directed to any particular target group. Rather, it aims at sending out as many emails as possible via spamming botnets (Kanich et al., 2008).

Before deciding for a product, customers are by all means interested in information. When deciding for a product, they expect an offer to be superior to others in terms of characteristics, quality, and price (Link, 2000). But customers do not only rely on information provided by companies and media. They select which information they want to receive through which channel (Leitzmann, 2002). The internet makes products and prices transparent (Feinberg et al., 2002) and is becoming increasingly popular as an

information source, thereby reducing information asymmetries (Bensberg, 2002). A self-reliant information search is possible. Recommendations and reviews from other customers, so-called consumer-to-consumer (C2C) communication, are becoming stronger influences for buying decisions. Nurturing consumer communities, facilitating conversations, and listening to what people say will have a much stronger focus in the future (Hipperson, 2010).

As especially online information is getting more individualized, and personalized, data security is becoming a more delicate topic. Customers want to know which personal information is saved and used. Data security topics are regularly discussed with great passion in public. A German survey in 2009 showed that most people do neither trust the government nor private companies when it comes to data security (Insitut für Demoskopie Allensbach, 2009). There is a limit to how much data can be collected from a single person (Baker et al., 1989), as customers would like a company to hold as little data on them as possible (Ozimek, 2010).

Nevertheless, to publish personal information on the web is both contemporary and common: it is perceived as welcome, if not desirable, to present oneself on the web (Wagner, 2010). It is noticeable that customers are motivated to articulate their preferences, opinions, and interests, if they are directly related to individual products or personal benefits (Hippner, Leber, & Wilde, 2002; Tiedtke, 2000). In future, advertisers expect an increasing digital media competence, which comprises the self-selection of advertising contents and the involvement of customers in the product development process (Dialog Marketing Monitor, 2009).

2.3 Sources for data augmentation

There are various established and potential data sources available for data augmentation in database marketing. The availability of the sources is the most important chance in our SWOT framework in chapter 2.4. Data is collected at various touch points and through various channels. Sources for

data augmentation can be data publically available, data available within the company, or data dedicatedly accumulated for data augmentation purposes. The notion of external sources does not refer to *where* data was collected. External sources might well be available in-house, e.g. from a website, from an in-house survey, or from a social media application. However, they are considered external, because the data cannot directly be merged to customer profiles. An algorithm or pattern is necessary in order to augment the customer database with the external data.

While some of the sources are established data augmentation sources, others are not commonly used yet, e.g. those not being identical to the customer database or a representative sample thereof. Uncertainty arises as to whether these potential sources can be used for data augmentation. The focus of this paper is to analyze the different characteristics of data augmentation sources and their implications on the validity and significance of the results. That way, also potential sources not commonly used today can be evaluated regarding their suitability for data augmentation.

The most important traditional and newer data sources are depicted in the following, highlighting their specific potential for data augmentation. Features and downsides of the source types are explained. We compare all sources regarding access and availability, usefulness of available variables, data preparation effort necessary to augment the data, costs, and timeliness. To know the characteristic features helps in recognizing them in the methodological framework described in chapter 4.

2.3.1 Public and official sources

Public and official sources are defined by fully covering or representing a national population. These sources fully include or represent the customers (of that country), but also more people from a bigger population. They are easily accessible and methodologically sound, i.e. the quality of the data is high and data collection has taken place in a statistically proper way. It

is usually carried out by the federal statistical office or be renown market research institutes. Examples of these providers in Germany are *forsa*, *Institut für Demoskopie Allensbach*, or *TNS*. Publically available sources not fulfilling the criteria of fully including or representing the customers are not considered here.

National census data Official statistics or national census data are at hand for every registered person in a country. Common information categories interesting from such register data are income, education, housing, employment, consume propensity, and living condition information. Census data is only collected once in a few years, which makes it an impractical sources for data augmentation. Additionally, census data is often not available in enough detail regarding link variables to be used for data augmentation. A census source is usually impractical and unnecessary as a sample (Powell, 1997, p. 67). However, it is a thinkable source and is mentioned here in order to contrast it to other sources.

Market media studies It is possible to obtain market research data from market research institutes or publishing companies, such as *Axel Springer AG*, *Bauer Media Group*, and *FOCUS Magazin Verlag GmbH*. These market media studies comprise a broad range of information categories, and thus pose many information opportunities. The comprised information can be divided into socio-demographic, psychographic, consumer preferential, interest, and behavioral information. Variables are predetermined, but big market research studies include so many variables that they usually satisfy the information needs. External market research has been used and described for marketing purposes. For example, Putten et al. (2002a, p. 3) used a survey on a whole branch and Krämer (2010) used a German population representative market media study.

As public sources are not designed to fit the individual purposes of data augmentation in marketing, they need to be critically examined in terms of their suitability regarding concepts and definitions of the link variables. There is data harmonization effort related to using public sources for data augmentation purposes. Furthermore, some comprehensive so-called single source studies often use specific data fusion methods in order to reduce response burden (Gilula et al., 2006; Kamakura & Wedel, 1997). To use these fused sources for another data augmentation requires critical reflection.

Public market research information is generally easy to obtain, well-conditioned, and the market research data market is very transparent. The quality is generally high, but nevertheless should be assessed depending on the issuing institute, because market research vendors are not neutral sources, but profit-oriented companies. The qualitative judgment is dependent on the individual information goals of the companies (Hippner & Wilde, 2001). General qualitative assessment criteria are utility (e.g. as a basis for decision), completeness (in terms of information sought), timeliness, and verity. Additionally, accuracy and reliability are relevant if the data source is a sample (Berekoven, Eckert, & Ellenrieder, 2009, p. 24ff).

2.3.2 Company-owned sources

Company-owned sources are only interesting for data augmentation, if they are not already incorporated in the infrastructure of the customer DWH and equipped with a unique identifier. There are two reasons why information can be unavailable on customers. Either the information is not available for some customers, but for others. Then it is technically possible to obtain the information, but the customers without data have not had any transactions from which that data can be derived. Or there are data sources that are not connected with the customer DWH, so that the information is not available for any of the customers.

Existing customer DWH In general, the existing customer DWH is the recipient unit. However, data augmentation can be a tool, if information is available for a certain subgroup of customers only. This kind of data augmentation, also referred to as scoring, is used in order to get information on all customers, although only a fraction of the customers has a value for a specific variable. The results of these augmentations are used in order to acquire new customers for specific products or for cross-selling purposes. This data is not only available in real-time and free of cost, but also derived from the analytical systems optimized for database marketing purposes, so that data preparation efforts are minimal. The usefulness of variables augmented by scoring is limited, because variables are not entirely new to the customer DWH, but only new for a subgroup of the customers.

Operational data Some of the operational data is not (yet) available to analytical systems in a way that they can be used in database marketing. In many companies, data is collected at various touch points. The operational systems of companies were formerly not laid out to serve the database marketing purpose. There might also be reasons for not making data available on a personal level, for example confidentiality. Common examples range from call center reportings and in-store information to web analytics data. Whenever these sources have sufficient link variables, they can be augmented to the customer database instead. The data can be made available in real-time, thus preventing problems of timeliness.

Online tracking data A special kind of operational data are online sources. The company website can be tracked and much information can be derived from surfing behavior. Tracking customer transactions was formerly only possible if they had a loyalty card or if the kind of transaction required the transaction to be saved. Today, all transactions in the internet or even unfinished transactions and nonbinding product searches are traceable. This data can be reused to offer products and target customers.

Furthermore, many variables not directly linked to purchase behavior are easily and accurately available online (Moe & Fader, 2004). Even if it is possible to directly link the surfing behavior to customer profiles, for example if customers are logged in to a website, this can be not allowed due to legal constraints. More details regarding legal requirements can be found in chapter 2.2.3. Consequently, surfing data is collected on an anonymous basis and can be used by data augmentation. The data preparation effort can be high for tracking data, because information needs to be converted from an unstructured form to relational data structures, in order to be used for the data augmentation methods presented here.

Company-owned sources have a high overlap with the customer database, but might not capture all customers. To that effect, the donor unit is a subset of the recipient unit. Depending on whether the source is the existing customer DWH or an operational or online tracking source, the availability of suitable link variables differs. Whether the available information is useful needs to be decided from case to case. If no confidentiality constraints exist, data augmentation is only one, sometimes short-term, strategy to making this data available. Another more durable strategy would be to link the data directly to the customer database with a unique identifier.

2.3.3 Accumulated sources

All sources specifically created and designed for data augmentation purposes are referred to as accumulated sources. Because accumulated sources are designed by a department looking for specific insights, basically any information can be comprised. They are a way of obtaining information from a customer sample not otherwise collectable in a regular business relationship, e.g. information on education, occupation, and households (Hattum & Hoijtink, 2008a). Other interesting variables are referred to as marketing mix related reaction parameters, because they are able to segment customers with similar reactions to marketing mix instruments (Freter, 1997,

p. 92). Accumulated sources can include descriptive variables like attitudes, perceived image of the company, or customer satisfaction (Liehr, 2001).

Representative customer surveys A survey is usually carried out by a third party service provider and consists of a representative subset of the customers that has to answer a questionnaire asking for the target variables. Because of data privacy regulations, the answers cannot be matched on an exact basis with the customer profiles. The only way to receive target variables for individual customers is to augment it. Market research information has the advantage of anonymity. Surveys reduce desirability bias, compared to information stated in front of a company, so that the collected information is possibly more valuable and more truthful.

Volunteer surveys An inexpensive alternative of conducting a representative survey is to conduct a volunteer survey. Customers self-select who wants to be interviewed. Volunteer surveys have a cost advantage, because no complicated arrangements have to be made in order to achieve representation. However, Pineau and Slotwiner (2003) showed that results from volunteer online surveys cannot easily be used to draw inferences on the overall population. They used different typical marketing categories to indicate differences between the internet community recruited by volunteer surveys and the overall population, thus implying that the conditional independence assumption is not valid for these sources in general. The error in terms of representation in the context of volunteer survey is referred to as self-selection bias (Hudson et al., 2004).

Social media data Social media brings new opportunities for data categories, e.g. all kinds of personal and sometimes sensitive information are collectable. Especially sentiments about products and services or purchase intentions are present in social media. Extracting and interpreting this information, e.g. by text mining, can have tremendous effects on sales

(Breur, 2011). Also, the degree of social connectedness has an impact on purchase probabilities (Naseri & Elliott, 2011). Casteleyn, Mottart, and Rutten (2009, p. 439) stated that the "heartbeat of today's society" becomes obvious in networks like *Facebook*. Often, social networks have more accurate and much more detailed data on the (social media) population than other empiric research institutes. The information, however, is only accessible to selected researchers (Heinrich, 2011). Companies need to cope with other solutions, like data available from public social media profiles or social media applications, with which the users are asked for their permission to access distinct data categories. That way, transparency about used information is given and companies are legally enabled to collect private social media data. Groups like these are not of representative nature (Casteleyn et al., 2009).

Surveys and social media data differ in terms of their characteristics. For surveys, variables can be freely defined, so that information potentials are unlimited. One of the advantages of surveys is that formats can be chosen in accordance with the customer database. The survey usually includes a subset of the customer database. Costs can be high, if the survey is supposed to be representative. A drawback of market research data is that it is conducted at a single point in time. It should be processed and augmented right after being analyzed, because most data augmentation models do not account for time differences and problems related to this.

To use social media data can lead to high data preparation efforts. The information available needs to be interpreted. The liking of a brand page, for example, can mean a lot of things, only one of them being the fact that someone intends to purchase this brand. Because there is no single best way of obtaining social media data, access can be summarized as being limited or at least complicated. The cost of social media data varies accordingly. Timeliness is not a problem, as data extraction can take place just in time for data augmentation.

A number of American and German studies have been conducted to show correlations between social media activities and various behavioral characteristics. If target variables are correlated with the fact that someone uses social media channels, these sources can lead to biased augmentation results. Kutter (2013) showed that this risk is low. However, it is worth examining the conditional dependencies between sources and target variables, as described in chapter 4.1.4.

People update their social profiles in order to maintain a good image of them online – much more than in a customer database. It should always be kept in mind that social media profiles are designs of what people want to express about themselves. They do not necessarily reflect what people are like, but more what they want to be like (Casteleyn et al., 2009). Joining groups or liking pages are acts of self-portrayal. It can generally be assumed that the stated information is accurate (Abel, 2011). Restrictions apply in terms of verity and social desirability. We do not go into detail on which restrictions apply to the usage of social media data. The topic is discussed in the respective media psychological literature.

2.3.4 Comparison of sources

Whether a source is generally suitable for data augmentation depends on several factors. It is a precondition that a source needs to be accessible and available, needs to include useful target variables and link variables able to predict these target variables. This is explained in more detail in chapter 4.2. However, sources can differ in how easily they are accessible, in how many target variables are included, and in how much data preparation effort is necessary in order to harmonize the link variables of source and recipient unit. Sources are the more suitable, the less they cost and the more recent data is. The more obstacles there are, the higher the resistance to attempt a complicated data augmentation project. The influence of the overlap between the source and the customer database on the data augmentation

results is subject of this paper. It is explained in more detail in chapter 4 and examined in chapter 6 and chapter 7. The sources listed are compared in table 2.1 regarding these factors.

	Public sources		Company-owned sources			Accumulated sources		
Criterion	Census data	Market media	Existing DWH	Operat. data	Tracking data	Repr. survey	Volunt. survey	Social media
Access and availability	+/–	+	+	+/–	+/–	+	+	+/–
Usefulness target variables	–	+	+/–	+/–	+/–	++	++	+/–
Readiness link variables	–	+/–	++	–	–	++	++	+/–
Costs*	+	+/–	+	+/–	–	–	+/–	+
Availability data and updates	–	–	+/–	+	+	–	–	+
Overlap with customer database	full	full (repr.)	none	partial	partial	full (repr.)	sub-group	partial

* Low costs result in a high rating

++ perfect
+ good
+/– specific
– limited

Table 2.1: Rating of available data augmentation sources

Access and availability can be limited or change over time. Survey data, e.g. from market media studies or accumulated surveys, as well as the existing customer DWH are more easily accessible than national census data or those sources that need to be technically connected with the customer DWH, such as operational, tracking, or social media data. Official statistics, for example, would be a great source of very precise and substantial information, but are not available to private enterprises. Sources less available involve other parties or systems. External market research studies, especially the German market media studies, are useful, but are subject to changes that cannot be influenced. Additionally, legal constraints apply to the use of many sources. Social media data are easily accessible, but need to be handled with care when it comes to data privacy issues. However,

none of the sources really has restricted access or unavailability – otherwise they would not be potential data augmentation sources.

The usefulness of available target variables is case-specific. In general, the more target variables there are in a source, the more profitable the augmentation results. Accumulated surveys, both representative and volunteer surveys, are perfect in terms of useful target variables, because the target variables of the survey can be defined by the company. Market media studies are not self-designed, but contain so many variables that they are considered good in terms of the usefulness. All other sources only contain specific variables, while national census data is rather limited in terms of target variable variety.

Definitions and formats of the link variables can be designed for accumulated surveys, as well as for existing customer DWH data used for scoring purposes. The preparation effort can be a precluding criterion. While official statistics and survey data are pre-screened and structured, web or social media data need to be reduced and transformed before augmenting them. National census, operational, and tracking data needs to be critically examined in terms of availability and readiness of link variables. Only if suitable link variables are given, at least after data preparation, a source can be used for data augmentation.

Costs for the acquisition of data augmentation sources can vary, depending on the expenses related to collecting or buying data, connecting and preparing it. Company-owned sources are generally less costly than public sources in terms of acquisition, but more costly in terms of connection and preparation. This can be a make-or-buy decision, because the costs need to be compared to the cost of collecting data in-house, if possible. Unfavorable formats like in tracking data can lead to additional expenditures. Costs are also related to the process of data augmentation and needed technology. Accumulated data is more costly, if a survey is conducted, than if data only has to be "tapped", like social media data.

In theory, sources can only be augmented without error, if the data of the source and the customer database are observed at the same point in time (D'Orazio et al., 2006, p. 4). Customers' ideas, needs, and behaviors change continuously and augmented data ages (Ozimek, 2010). Thus, the length of time between data collection date and data augmentation influences the quality of the augmentation results. The utility of augmented data decreases over time (Even, Shankaranarayanan, & Berger, 2010; Ozimek, 2010). The error related to the usage of outdated information is referred to as timeliness error. Data permanently available and updated, like operational from company-owned sources or user generated content from social media sources, is generally more beneficial for data augmentation purposes than survey data, which is only collected once or at most yearly. For returning marketing problems and tasks, it is more valuable to acquire sources that can be augmented at any point in time with recent information. As the information in the customer database changes just as quickly when variable values are added or adapted during business processes, a so-called on-demand computation can be established, for which data sources are maintained and updated separately and only brought together at the time when augmented information is needed (Jiang et al., 2007).

2.4 Implications for data augmentation

From the external and internal conditions and the available information sources, a strategy is derived for data augmentation in database marketing. A SWOT is conducted in order to illustrate which strengths of the company can be used in order to benefit from opportunities and encounter risks, as well as which weaknesses can be turned into strengths by exploiting external chances. It shows which risks should be avoided. Data augmentation is the logical deduction from these considerations and can be implemented as solution for many problems. An overview of the factors relevant in the SWOT is given in table 2.2.

SWOT	Helpful	Harmful
Internal origin	*Strengths* Customer DWH for collecting all relevant information for customer communication. Knowledgeable database marketing team Customers can be targeted directly and individually	*Weaknesses* Customers cannot be differentiated on a one-to-one basis Not enough information available to purposefully segment customers Conversion probability is not known before implementing a marketing campaign
External origin	*Opportunities* Several source types available including information worth knowing Technological progress enables affordable data storage and efficient data usage Customers expect benefits and tailor-made offers, are interested in information facilitating their purchase decision	*Threads* Economic situation is afflicted with dynamics, complexity, competition intensity, and market fragmentation Personal data is particular protected by law and must be handled with care Customers are price-sensitive, tired of receiving irrelevant communication, and anxious about what happens to their data

Table 2.2: SWOT analysis implying strategic suggestions for data augmentation

The chances offered by additional sources can help in overcoming the lack of segmentable customer information and turning it into strength. Customers expect differentiated marketing tailored to their needs. Only if communication is targeted to highly selected groups, it reaches a maximum of acceptance (Hattum & Hoijtink, 2008a; Laase, 2011). However, there exists a discrepancy between desired and received information (Liehr, 2001). In order to offer a surplus to self-generated information and in consideration of the information overload already in place, advertising contents have to be relevant. With the increasing need to know more about the customers in order to address them individually, the necessity of combining different sources is raising (Kamakura & Wedel, 1997). This is made possible by data augmentation. It is anticipated that the successful outcome of a data augmentation leads to a better overall customer experience (Breur, 2011).

Relevance is not only of interest to the customers, for which personal benefits, time and cost savings, and quality are important. It is also in a company's' interest. Every customer contact has an economic potential in terms of return on investment (e.g. revenue increase, customer profitability, competitive advantage). The customer specific treatment and communication is thus directly creating value for the company (Rapp, 2002b).

By using the technological advancements in database marketing, companies can increase their position in the competition environment. Thanks to storage capacities and advanced analytical systems, data can be processed in shorter time. Those companies able to react to the customer needs better and faster than the competition can gain a significant competitive advantage (Fogarty, 2008). Today, technology and know-how enabling data augmentation are available. The database marketing know-how itself becomes a core asset, because it builds on a conglomerate of interdigitating activities not easy to imitate (Schweiger & Wilde, 1993).

The trends in media usage and technological advancements can be used for incorporating augmentation results in the regular communication process. The increase of online communication channels supports the cost saving and efficiency goals. The internet has low variable costs, while globalization and technical progress drive the demand of online applications. The more intensive customers use the internet, the bigger the amount of detailed information and opinions usable for improving communication and services. With the help of database marketing, modulated tailored communication instruments can be developed, automatically assigning suitable offers to customers on the basis of their information and contact history, distributed independent from channels. Data is the basis for individualization, automation, and mass customization.

Data augmentation is especially useful, because it respects the legal requirements. Data used for data augmentation purposes can be anonymous data, so that no direct inference to individuals is possible. Certain data sources containing unique identifiers must not be used, even if it was possi-

ble. On the other hand, to use data in order to deliver relevant advertising is explicitly appreciated by law, appraising the general approach of data augmentation. When moving within the legal borders, companies are able to create a lasting competitive advantage.

The ability to target customers can overcome the thread of customers being anxious about their data. Although all activities of companies, including data augmentation, are legally sound, there is a general anxiety of customers regarding data protection. Companies should therefore take these concerns seriously, address them, and establish trust, whenever marketing activities are planned. Trust is a combination of controllability, transparency, and security of action, which are facilitated by quality and stability (Winand & Pohl, 2000). Database marketing helps to create trust by implementing a clean permission handling, which is updated in real time, and by providing information for individual customer contacts. Once the customers recognize their preferences in the offers, they are apt to articulate further interests. To realize the one-to-one vision by addressing customers with perfectly individualized and personalized offers could disconcert customers, rather than satisfying them. Chen and Iyer (2002) showed that it might not be desirable for every company in a competitive environment to perfect individualization. To aim at micro segments that are augmented with external data therefore seems to be a reasonable goal.

All in all, data augmentation is a smart strategy in the current marketing environment. It adds information to the customer database not otherwise available, while respecting existing law and customer concerns. Technological advancements and internal developments like powerful CRM infrastructures and knowledgeable database marketing teams favor this approach. The additional data can be used to reach the marketing goals of individualization and relevance, leading to a competitive advantage.