Being proactive – analytics for predicting customer actions	**Проактивность - аналитика для прогнозирования действий клиента**
D D Nauck, D Ruta, M Spott and B Azvine	
Customers should be at the heart of most businesses, and in particular service providers such as BT. In order to serve our customers better, we regularly introduce reliable processes and procedures to improve interaction with our customers, which is known as customer relationship management. Typically, organisations collect and keep large volumes of customer data as part of their processes. Analysis of this data by business users often leads to discovery of valuable patterns and trends that otherwise would go unnoticed and that can lead to prioritisation of decisions on future investments. Current tools available to business users are limited to visualisation and reporting of data. What is needed is modelling customer behaviours to be able to build future scenarios. More advanced tools and techniques have been available for a number of years but have not been developed for the business community due to the level of expertise required to use them. In this paper we present a number of tools and techniques developed for business users to perform advanced analysis on customer data. The tools can be used to perform sensitivity analysis, what-if analysis and impact analysis, all of which are aimed at prediction and simulation of future customer actions. The paper also covers application of the tools to real customer data and reports on some of the results obtained.	Клиенты должны быть в центре большинства компаний, и в частности поставщиков услуг, таких как BT. Чтобы лучше обслуживать наших клиентов, мы регулярно внедряем надежные процессы и процедуры для улучшения взаимодействия с нашими клиентами, что называется управлением взаимоотношениями с клиентами. Как правило, организации собирают и хранят большие объемы данных о клиентах в рамках своих процессов. Анализ этих данных бизнес-пользователями часто приводит к обнаружению ценных моделей и тенденций, которые в противном случае остались бы незамеченными и которые могут привести к расстановке приоритетов при принятии решений о будущих инвестициях. Текущие инструменты, доступные бизнес-пользователям, ограничены визуализацией и составлением отчетов о данных. Что необходимо, так это моделирование поведения клиентов, чтобы иметь возможность строить будущие сценарии. Более продвинутые инструменты и методы были доступны в течение ряда лет, но не были разработаны для бизнес-сообщества из-за уровня знаний, необходимых для их использования. В этой статье мы представляем ряд инструментов и методов, разработанных для бизнес-пользователей для выполнения расширенного анализа данных клиентов. Инструменты могут использоваться для выполнения анализа чувствительности, анализа «что, если» и анализа воздействия, все из которых направлены на прогнозирование и моделирование будущих действий клиента. В документе также рассматривается применение инструментов для реальных данных клиентов и отчеты о некоторых полученных результатах.
1. Introduction	1. Введение
Customers are at the heart of most businesses and this is especially true for service providers like BT. In order to serve their customers better, businesses have to introduce reliable processes and procedures for the interaction with their customers, which is known as customer relationship management (CRM). Typically, organisations collect and keep customer data as part of their processes. Therefore, data forms the core information source for CRM. In addition to process data, other forms of data are used, such as results from market research, demographic information, or surveys that provide customer	Клиенты находятся в центре большинства компаний, и это особенно верно для поставщиков услуг, таких как BT. Чтобы лучше обслуживать своих клиентов, предприятия должны внедрять надежные процессы и процедуры для взаимодействия со своими клиентами, которые известны как управление взаимоотношениями с клиентами (CRM). Как правило, организации собирают и хранят данные клиентов как часть своих процессов. Таким образом, данные образуют основной источник информации для CRM. В дополнение к

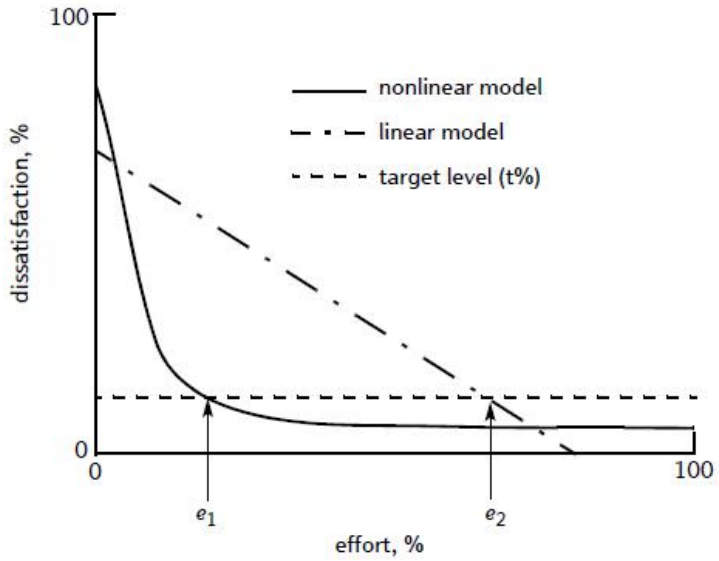
feedback.	данным процесса используются другие формы данных, такие как результаты исследований рынка, демографическая информация или опросы, которые обеспечивают обратную связь с клиентами.
Process data reveals what customers do and survey data reveals what customers say they do. There is frequently quite a difference. For example, when analysing churn, service providers often find that a large percentage of customers who leave have previously said that they are very satisfied with the service.	Данные процесса показывают, что делают клиенты, а данные опроса показывают, что клиенты говорят, что они делают. Часто есть большая разница. Например, анализируя отток, поставщики услуг часто обнаруживают, что большой процент клиентов, которые уходят, ранее говорили, что они очень довольны обслуживанием.
Customer analytics can be divided into three main areas.	Клиентскую аналитику можно разделить на три основные области.
• Customer segmentation	• сегментация клиентов
A limited number of customers are surveyed and this data is combined with account information and demographic data. By using methods like cluster analysis a number of segments are created and interpreted. The interpretation of the analysis results leads to a number of customer segments which are labelled with intuitive descriptions such as, for example, ‘traditionalists’ or ‘technophiles’. Then each account is mapped to one of the segments based on available data. This mapping obviously has to rely on less information than the segmentation exercise, because the vast majority of customers are not surveyed. Customer segmentation therefore is always imperfect.	Опрошено ограниченное количество клиентов, и эти данные объединяются с данными учетной записи и демографическими данными. Используя такие методы, как кластерный анализ, создается и интерпретируется ряд сегментов. Интерпретация результатов анализа приводит к ряду клиентских сегментов, которые обозначены интуитивно понятными описаниями, такими как, например, «традиционалисты» или «технофилы». Затем каждая учетная запись сопоставляется с одним из сегментов на основе доступных данных. Это отображение, очевидно, должно опираться на меньшее количество информации, чем на сегментацию, поскольку подавляющее большинство клиентов не опрошено. Поэтому сегментация клиентов всегда несовершенна.
Predicting customer actions	Прогнозирование действий клиента
Based on historic information about actions by individual customers, we try to predict likely customer actions in the future. This kind of analysis is mainly based on process data that records events and interactions with customers. This kind of predictive analytics can only be done if a sufficiently long customer history is available and there is relevant interaction with the customer. This can be difficult for businesses, like, for example, satellite TV providers where the only interaction is typically payment of a monthly bill, assuming there is no automatic surveying of viewing behaviour. Telecommunications providers, in contrast, have better insights because they can see how customers are using the service and their operation requires a lot of effort in providing and maintaining services, and that leads to regular customer interaction on a large scale.	Основываясь на исторической информации о действиях отдельных клиентов, мы пытаемся предсказать вероятные действия клиентов в будущем. Этот вид анализа в основном основан на данных процесса, которые записывают события и взаимодействия с клиентами. Этот вид прогнозной аналитики может быть выполнен только в том случае, если доступна достаточно длинная история клиента и существует соответствующее взаимодействие с клиентом. Это может быть затруднительно для предприятий, таких как, например, поставщики спутникового телевидения, в которых единственным взаимодействием обычно является оплата ежемесячного счета, при условии, что нет автоматического отслеживания поведения при просмотре. Поставщики телекоммуникационных услуг, напротив, лучше понимают, потому что они могут видеть, как клиенты используют услугу, а их работа требует больших усилий по предоставлению и обслуживанию услуг, что приводит к

	регулярному взаимодействию с клиентами в больших масштабах.
• Understanding customer views	• Понимание взглядов клиентов
Most large businesses run some form of customer surveys. Businesses want to understand how their brand image is perceived, whether customers are satisfied with certain products or the company in general, or whether customers are happy to recommend the products and services they use. Businesses also want to understand what potentially drives satisfaction or loyalty and how such drivers can be influenced. Survey data has to be treated with caution because responses can be influenced by factors not covered by the survey, e.g. personal situation of the interviewee, interview situation and interaction with the interviewer, competitor activities. It is also important to keep in mind that surveys reflect what customers say and not necessarily what they actually do or are about to do.	Большинство крупных компаний проводят опросы клиентов. Компании хотят понять, как воспринимается имидж их бренда, удовлетворены ли клиенты определенными продуктами или компанией в целом, или же клиенты с удовольствием рекомендуют продукты и услуги, которые они используют. Компании также хотят понять, что потенциально стимулирует удовлетворение или лояльность и как на них можно повлиять. К данным обследования следует относиться с осторожностью, поскольку на ответы могут влиять факторы, не охваченные опросом, например, личная ситуация интервьюируемого, ситуация интервью и взаимодействие с интервьюером, деятельность конкурента. Также важно помнить, что опросы отражают то, что говорят клиенты, а не обязательно то, что они на самом деле делают или собираются сделать.
In this paper we will be concerned with the latter two aspects of customer analytics — predicting customer actions and understanding customer views.	В этой статье мы рассмотрим последние два аспекта анализа клиентов - прогнозирование действий клиентов и понимание взглядов клиентов.
2. Data analysis issues	2. Проблемы анализа данных
Customer analytics is essentially concerned with analysing data and requires standard techniques from areas like statistics [1], data mining [2], machine learning [3] and intelligent data analysis [4, 5]. In our work in customer analytics we particularly encountered issues around data quality and the selection of appropriate analysis methods.	Клиентская аналитика в основном занимается анализом данных и требует стандартных методов из таких областей, как статистика [1], извлечение данных [2], машинное обучение [3] и интеллектуальный анализ данных [4, 5]. В нашей работе по анализу клиентов мы особенно сталкивались с проблемами качества данных и выбора подходящих методов анализа.
2.1 Data quality	2.1 Качество данных
Most large established businesses run a huge number of legacy systems collecting data in different formats. This data is frequently not collected with analysis in mind and therefore important attributes can be missing. Data fusion across different legacy systems can be extremely difficult and often requires a lot of manual intervention and data cleansing. If customer survey data is involved we often see missing values because customers may refuse to respond to questions or questions are not asked in every survey. From our experience, it is not rare to have up to 75% missing values in such a data set. Poor data quality requires a big effort in data cleansing and pre-processing. It is important, for example, not simply to discard data records with missing values, but to check whether the fact that values are missing carries some hidden meaning that could be relevant to the analysis. Systematically missing data can introduce spurious	Большинство крупных устоявшихся предприятий используют огромное количество устаревших систем сбора данных в разных форматах. Эти данные часто не собираются с учетом анализа, и поэтому могут отсутствовать важные атрибуты. Слияние данных между различными унаследованными системами может быть чрезвычайно сложным и часто требует большого ручного вмешательства и очистки данных. Если используются данные опроса клиентов, мы часто видим пропущенные значения, потому что клиенты могут отказаться отвечать на вопросы или вопросы не задаются в каждом опросе. Исходя из нашего опыта, нередко бывает, что до 75% пропущенных значений в таком наборе данных. Низкое качество данных требует больших усилий по очистке и

relationships into the data. For example, discovery algorithms might detect a relationship between attributes that are missing in the same records and thus lead to the identification of wrong influence factors.	предварительной обработке данных. Например, важно не просто отбросить записи данных с пропущенными значениями, но и проверить, несет ли тот факт, что значения отсутствуют, какое-то скрытое значение, которое может иметь отношение к анализу. Систематически отсутствующие данные могут вводить в данные ложные связи. Например, алгоритмы обнаружения могут обнаруживать связь между атрибутами, отсутствующими в одних и тех же записях, и, таким образом, приводить к выявлению неправильных факторов влияния.
Another data quality problem can be introduced by aggregation functions. If summarised or averaged data is used where individual records would actually be required, relationships in the data can be lost or spurious relationships can be introduced. Basing an analysis on, for example, weekly summaries or averages is often done for one of two reasons — to reduce the amount of data or because some parts of the data are only available as summaries.	Еще одна проблема качества данных может быть связана с функциями агрегирования. Если обобщенные или усредненные данные используются там, где фактически потребуются отдельные записи, отношения в данных могут быть потеряны или могут быть введены ложные отношения. Анализ на основе, например, еженедельных сводок или средних значений часто выполняется по одной из двух причин - для уменьшения объема данных или потому, что некоторые части данных доступны только в виде сводок.
Table 1 gives a simple example of how averaging can hide an existing relationship. Assume we have $y = x$ 2 and that we measure x and y in two consecutive weeks and compute weekly averages. While from the individual (x, y) records we can clearly see the functional dependency of y on x, this relationship is not visible in the weekly averages of x and y. It is as easy to construct examples, where x and y are unrelated but the averages show a spurious relationship. This can also be done for other aggregates like the median.	В таблице 1 приведен простой пример того, как усреднение может скрыть существующие отношения. Предположим, у нас есть y = x 2 и что мы измеряем x и y в течение двух последовательных недель и вычисляем средние значения за неделю. В то время как из отдельных записей (x, y) мы можем ясно видеть функциональную зависимость y от x, эта зависимость не видна в еженедельных средних значениях x и y. Привести примеры так же легко, где x и y не связаны, но средние показывают ложные отношения. Это также может быть сделано для других агрегатов, таких как медиана.
Table 1 An example where averages hide a relationship present in the data (see text).	Таблица 1 Пример, в котором средние значения скрывают отношение, присутствующее в данных (см. Текст).

	first week		second week	
Number	*x*	*y*	*x*	*y*
1	1.00	1.00	1.00	1.00
2	2.00	4.00	4.00	16.00
3	3.00	9.00	1.00	1.00
average	2.00	4.67	2.00	6.00

2.2 Choosing appropriate analysis methods	2.2 Выбор подходящих методов анализа
Due to a lack of expertise and tools, data analysis is often done in a too simple or even naive way. Linear models are the most frequently used analysis methods, because they are easy to understand. A linear model assumes that one or more dependent variables are determined by a linear combination of mutually independent variables. Additionally, in many methods from linear statistics there is an implicit assumption about normally distributed values. Both assumptions — mutual independency and normal distribution — are frequently not valid for real-world problems. Linear models cannot take compensatory or reinforcing effects into account. Especially in customer analytics, we observe different types of dependencies between all variables and these dependencies can be nonlinear in nature. Assuming a linear relationship without checking for nonlinear dependencies can obviously lead to wrong conclusions.	Из-за нехватки опыта и инструментов анализ данных часто делается слишком простым или даже наивным способом. Линейные модели являются наиболее часто используемыми методами анализа, потому что их легко понять. Линейная модель предполагает, что одна или несколько зависимых переменных определяются линейной комбинацией взаимно независимых переменных. Кроме того, во многих методах линейной статистики существует неявное предположение о нормально распределенных значениях. Оба предположения - взаимная независимость и нормальное распределение - часто не подходят для реальных проблем. Линейные модели не могут учитывать компенсирующие или усиливающие эффекты. Особенно в аналитике клиентов мы наблюдаем различные типы зависимостей между всеми переменными, и эти зависимости могут быть нелинейными по своей природе. Предположение о линейных отношениях без проверки нелинейных зависимостей может привести к ошибочным выводам.
For example, consider the simplified scenario in Fig 1. We assume that by increasing the effort we can drive down dissatisfaction. Let's further assume the real model is the nonlinear saturation curve. If we instead use a linear function to represent the dependency between dissatisfaction and effort, we might end up with the depicted linear function. This would give us the wrong assumption that we could reduce dissatisfaction down to zero by putting in enough effort instead of the more realistic saturation that would settle on a certain level of effort. It also could — in this example — lead us to assume that in order to reduce dissatisfaction to a target level of *t*% we would have to input the effort *e*2 instead of the smaller effort *e*1 required by the nonlinear model.	Например, рассмотрим упрощенный сценарий на рис. 1. Мы предполагаем, что, увеличивая усилия, мы можем уменьшить неудовлетворенность. Далее давайте предположим, что реальной моделью является нелинейная кривая насыщения. Если вместо этого мы используем линейную функцию для представления зависимости между неудовлетворенностью и усилием, мы можем получить изображенную линейную функцию. Это дало бы нам неверное предположение о том, что мы можем уменьшить неудовлетворенность до нуля, приложив достаточные усилия вместо более реалистичного насыщения, которое установилось бы на

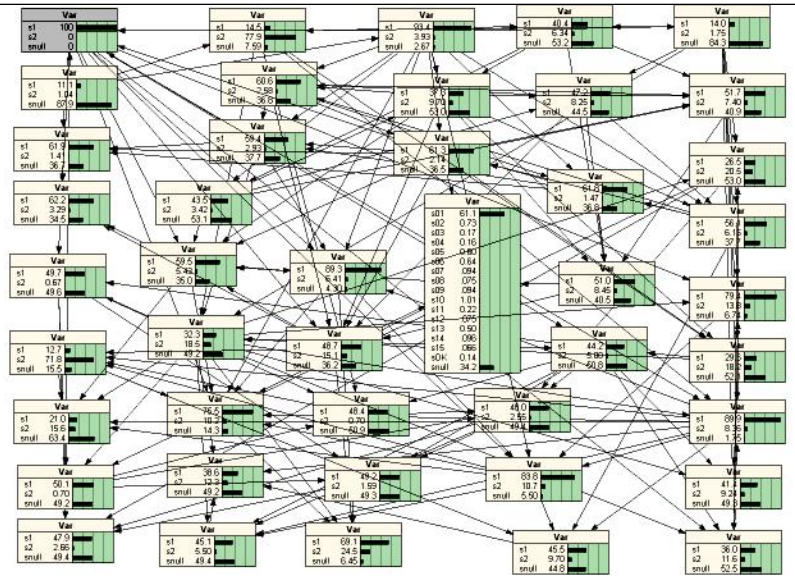
	определенном уровне усилий. Это также могло бы - в этом примере - привести нас к предположению, что для уменьшения неудовлетворенности до целевого уровня t% нам придется вводить усилие e2 вместо меньшего усилия e1, требуемого нелинейной моделью.
 Fig 1 A linear and a nonlinear model.	Рис. 1 Линейная и нелинейная модель.
In a typical customer analytics scenario we usually see some of the following methods applied depending on the skills of the analyst. These methods can all reveal useful information, but they also have limits that the analyst must consider.	В типичном сценарии клиентской аналитики мы обычно видим некоторые из следующих методов, применяемых в зависимости от навыков аналитика. Эти методы могут раскрыть полезную информацию, но они также имеют ограничения, которые должен учитывать аналитик.
A typical analysis would involve looking at the frequency distributions of all variables. From that we can learn in which areas there may be a problem, for example, in customer satisfaction. Given the right software, it is possible to drill down and look at the replies of certain customer groups based on demographical information or the replies to particular questions. From that we may be able to learn, for example, that users who complain about the layout of our Web portal are also more likely to complain about not finding the hyper-links than customers who are happy with the layout. The problem with that approach is that the analyst will only discover what he or she is looking for and that multidimensional dependencies will be overlooked, because they cannot be visually represented.	Типичный анализ будет включать рассмотрение частотных распределений всех переменных. Из этого мы можем узнать, в каких областях может возникнуть проблема, например, в удовлетворенности клиентов. При наличии подходящего программного обеспечения можно детализировать и просматривать ответы определенных групп клиентов на основе демографической информации или ответов на конкретные вопросы. Из этого мы можем узнать, например, что пользователи, которые жалуются на макет нашего веб-портала, также с большей вероятностью будут жаловаться на отсутствие гиперссылок, чем клиенты, которые довольны макетом. Проблема с этим подходом состоит в том, что

	аналитик обнаружит только то, что он или она ищет, и что многомерные зависимости будут игнорироваться, потому что они не могут быть представлены визуально.
Another typical analysis is to look at correlations between the different variables. For example, if we look at customer satisfaction surveys, we may find that many questions in a survey are highly correlated, meaning the higher the satisfaction or dissatisfaction in one area, the higher the satisfaction or dissatisfaction in another area is likely to be. A correlation analysis, however, assumes a linear relationship, and its result is therefore only relevant if the assumed linear dependence actually exists. This type of analysis can overlook nonlinear relationships and cannot detect multidimensional dependencies.	Другой типичный анализ заключается в рассмотрении корреляции между различными переменными. Например, если мы посмотрим на опросы удовлетворенности клиентов, мы можем обнаружить, что многие вопросы в опросе сильно коррелированы, то есть чем выше удовлетворение или неудовлетворенность в одной области, тем выше вероятность удовлетворения или недовольства в другой области. Корреляционный анализ, однако, предполагает линейную зависимость, и поэтому его результат имеет значение только в том случае, если предполагаемая линейная зависимость действительно существует. Этот тип анализа может пропустить нелинейные отношения и не может обнаружить многомерные зависимости.
In order to understand the quantitative influence of the result for one question on the overall satisfaction, we can use a functional model. In statistical analysis we typically see linear regression being used for this purpose. However, linear regression assumes that the individual variables are independent of each other and that a linear dependency between the independent variables and the target variable actually exists. For each value of an independent variable, the distribution of the dependent variable must be normal. These constraints are very strong and they are usually never fulfilled.	Чтобы понять количественное влияние результата для одного вопроса на общую удовлетворенность, мы можем использовать функциональную модель. В статистическом анализе мы обычно видим, что для этой цели используется линейная регрессия. Однако линейная регрессия предполагает, что отдельные переменные не зависят друг от друга и что линейная зависимость между независимыми переменными и целевой переменной действительно существует. Для каждого значения независимой переменной распределение зависимой переменной должно быть нормальным. Эти ограничения очень сильны, и обычно они никогда не выполняются.
Especially the independence assumption is typically not realistic at all. One way to alleviate this problem is to run a principal component analysis and using the uncorrelated principal components as inputs to the regression analysis. However, it should be obvious that a linear regression model is unsuitable for modelling nonlinear relationships and it completely ignores mutual relationships between the inputs and thus cannot take compensatory or reinforcing effects into account.	Особенно предположение о независимости, как правило, вообще не реально. Одним из способов решения этой проблемы является проведение анализа главных компонентов и использование некоррелированных главных компонентов в качестве входных данных для регрессионного анализа. Однако должно быть очевидным, что модель линейной регрессии не подходит для моделирования нелинейных отношений и полностью игнорирует взаимные отношения между входными данными и, следовательно, не может принимать во внимание компенсирующие или усиливающие эффекты.
In order to be more flexible and less constrained with the choice of a functional model, we can look at nonparametric nonlinear methods. They are not based on implicit distribution or independence assumptions and can model high-	Чтобы быть более гибким и менее ограниченным выбором функциональной модели, мы можем взглянуть на непараметрические нелинейные методы. Они не основаны на

dimensional nonlinear relationships. Methods like decision trees, neural networks, fuzzy systems, neuro-fuzzy systems, support vector machines, etc, are known from areas like intelligent data analysis and machine learning. Some of these methods, such as decision trees and fuzzy systems, are rule-based and can be used to obtain information about the nature of the modelled relationships. Other systems, e.g. neural networks or support vector machines, are only suitable for making predictions because the way they represent relationships is not easily interpretable.	предположениях о неявном распределении или независимости и могут моделировать многомерные нелинейные отношения. Такие методы, как деревья решений, нейронные сети, нечеткие системы, нейро-нечеткие системы, машины опорных векторов и т. Д., Известны в таких областях, как интеллектуальный анализ данных и машинное обучение. Некоторые из этих методов, такие как деревья решений и нечеткие системы, основаны на правилах и могут использоваться для получения информации о природе моделируемых отношений. Другие системы, например нейронные сети или машины опорных векторов подходят только для прогнозирования, потому что способ, которым они представляют отношения, нелегко интерпретировать.
These kinds of nonlinear model are very powerful, but they are unidirectional. That means, they can only be used to compute the impact on one or more previously selected target variables when some independent variables or drivers change.	Эти виды нелинейных моделей очень мощные, но они однонаправленные. Это означает, что они могут использоваться только для вычисления воздействия на одну или несколько ранее выбранных целевых переменных, когда изменяются некоторые независимые переменные или драйверы.
However, we also want to understand the effects on all the other drivers. To compute the impact that any variable has on any other variable we can use a multidimensional probabilistic model. Such a model is not restricted by linear dependencies or global independence assumptions. A suitable probabilistic model is a Bayesian network that can represent arbitrary probabilistic relationships between any number of variables.	Однако мы также хотим понять влияние всех остальных драйверов. Чтобы вычислить влияние, которое любая переменная оказывает на любую другую переменную, мы можем использовать многомерную вероятностную модель. Такая модель не ограничена линейными зависимостями или предположениями о глобальной независимости. Подходящей вероятностной моделью является байесовская сеть, которая может представлять произвольные вероятностные отношения между любым числом переменных.
In the next section we look at a software tool that is based on Bayesian networks and that has been used by us to analyse customer data.	В следующем разделе мы рассмотрим программный инструмент, основанный на байесовских сетях и использованный нами для анализа данных клиентов.
3. iCSat	
iCSat (intelligent customer satisfaction analysis tool) is a Java-based client/server platform for analysing customer data. Its initial focus was the survey data about customer satisfaction, but in fact iCSat is a generic tool and has subsequently been applied in other domains as well.	iCSat (интеллектуальный инструмент анализа удовлетворенности клиентов) - это клиент-серверная платформа на основе Java для анализа данных клиентов. Первоначально основное внимание уделялось данным опроса об удовлетворенности клиентов, но на самом деле iCSat является универсальным инструментом, который впоследствии был применен и в других областях.
3.1 Bayesian networks	3.1 Байесовские сети
iCSat uses Bayesian networks [6, 7] to model dependencies between all variables available in a data set. A Bayesian network is represented as a graph	iCSat использует байесовские сети [6, 7] для моделирования зависимостей между всеми переменными, доступными в наборе

where each node represents a variable and connections between the nodes represent direct conditional dependencies. Each node displays a probability distribution over the possible values of the variable represented by the node given the current state of the whole network. A Bayesian network is a convenient way of representing a high-dimensional probability space by exploiting conditional independence between variables. Nodes that have no direct connections are conditionally independent. We only need to represent conditional probabilities between connected nodes.	данных. Байесовская сеть представлена в виде графа, где каждый узел представляет переменную, а соединения между узлами представляют прямые условные зависимости. Каждый узел отображает распределение вероятностей по возможным значениям переменной, представленной узлом, учитывая текущее состояние всей сети. Байесовская сеть - это удобный способ представления многомерного вероятностного пространства путем использования условной независимости между переменными. Узлы, которые не имеют прямых соединений, являются условно независимыми. Нам нужно только представить условные вероятности между связанными узлами.
A Bayesian network exploits the fact that a joint probability distribution $p(x1,\ldots, xn)$ can be rewritten by applying the chain rule of probability as:	Байесовская сеть использует тот факт, что совместное распределение вероятностей p (x1,…, xn) можно переписать, применив цепочечное правило вероятности следующим образом:
$p(xi1,\ldots, xin)\ p(xi1\ xi2, \ldots, xin) \cdot p(xi2\ xi3,\ldots, xin) \cdot \ldots p(xin)$	$p(xi1,\ldots, xin)\ p(xi1\ xi2, \ldots, xin) \cdot p(xi2\ xi3,\ldots, xin) \cdot \ldots p(xin)$
Where $\langle i1, i2, \ldots, in\rangle$ is an arbitrary permutation of $\langle 1, 2,\ldots, n\rangle$. In addition, we will often find the distribution of a variable xjk can be described conditional on a set of parents Πik that is substantially smaller than $(xi(k+1),\ldots,xin)$ and that renders xik independent from $(xi(k+1),\ldots,xin)$, that is:	Где 〈 i1, i2,…, in 〉 - произвольная перестановка 〈 1, 2,…, n 〉 . Кроме того, мы часто находим, что распределение переменной xjk можно описать условно для набора родителей Пik, который существенно меньше (xi (k + 1), ..., xin) и который делает xik независимым от (xi (k + 1), ..., xin), то есть:
$p(xik\ xi(k+1),\ldots, xin) = p\ (xik\ \Pi ik)$.	$p(xik\ xi(k+1),\ldots, xin) = p\ (xik\ \Pi ik)$.
For example, if we have a data set of 10 variables where each variable can assume two possible values, then we have 210 = 1024 possible combinations of values. In a probabilistic model we would have to maintain one value for each combination resulting in 1024 probabilities. If we manage, for example, to represent the 10 variables in a Bayesian network such that no variable has more than two parents and one variable has no parents at all (root node), then we would have to maintain only 2 + (9 × 23) = 74 probabilities. By computing within the Bayesian network we can still calculate the probabilities for all 1024 possible value combinations, but we are not required to compute all of them up-front and store them.	Например, если у нас есть набор данных из 10 переменных, где каждая переменная может принимать два возможных значения, то мы имеем 210 = 1024 возможных комбинаций значений. В вероятностной модели мы должны были бы поддерживать одно значение для каждой комбинации, что привело бы к 1024 вероятностям. Если нам удастся, например, представить 10 переменных в байесовской сети таким образом, чтобы ни у одной переменной не было более двух родителей, а у одной переменной вообще не было родителей (корневой узел), то нам пришлось бы поддерживать только 2 + (9 × 23) = 74 вероятности. Вычисляя в байесовской сети, мы все еще можем вычислить вероятности для всех 1024 возможных комбинаций значений, но мы не обязаны вычислять все их заранее и сохранять их.
In order to obtain a Bayesian network we must first define its structure, i.e. determine which nodes are connected to each other, and then we must provide the conditional probabilities that describe the dependencies between the connected nodes. For small problems it would be possible that an expert does	Чтобы получить байесовскую сеть, мы должны сначала определить ее структуру, то есть определить, какие узлы связаны друг с другом, а затем мы должны предоставить условные вероятности, которые описывают зависимости между связанными узлами. Для небольших

this manually. However, for larger problems and for non-experts, it is usually impossible to specify a Bayesian network from scratch. We have therefore implemented a powerful structure-learning algorithm into iCSat that can learn the connections within a Bayesian network automatically from data [8]. iCSat then uses a commercial Bayesian library (Netica) to represent the network and to learn the probabilities from data [9].	проблем было бы возможно, что эксперт делает это вручную. Однако для более крупных проблем и для неспециалистов обычно невозможно определить байесовскую сеть с нуля. Поэтому мы внедрили мощный алгоритм изучения структуры в iCSat, который может автоматически изучать соединения в байесовской сети по данным [8]. Затем iCSat использует коммерческую байесовскую библиотеку (Netica) для представления сети и изучения вероятностей на основе данных [9].
Bayesian networks are used by inputting observations or assumptions into some of the nodes and then studying the changes in all other nodes. As an example, assume we are looking at a set of customers who ordered a certain service in the last three months and who have subsequently been surveyed. Assume further that we are interested in the relationship between speed of provision and overall satisfaction with the service. Say, the node for provision time displays the options ‘same day’, ‘next day’ and ‘2 or more days’. By setting the value of that node to ‘same day’ we can immediately see the satisfaction distribution for all customers who experience a same day provision.	Байесовские сети используются для ввода наблюдений или предположений в некоторые из узлов, а затем для изучения изменений во всех других узлах. В качестве примера предположим, что мы рассматриваем группу клиентов, которые заказывали определенную услугу в течение последних трех месяцев и которые впоследствии были опрошены. Предположим далее, что нас интересует взаимосвязь между скоростью предоставления услуг и общей удовлетворенностью обслуживанием. Скажем, узел для предоставления времени отображает опции «тот же день», «следующий день» и «2 или более дней». Установив значение этого узла равным «в тот же день», мы сразу увидим распределение удовлетворенности для всех клиентов, которые получают обслуживание в тот же день.
Because a Bayesian network can reason in all directions we can also run a scenario where we set the satisfaction level, say, to ‘extremely satisfied’ and then look at the distribution of provision times. Since all the nodes in the network change their distributions when we enter some value in any other node, we can also see the impact of changes in satisfaction or provision times on all other variables at the same time.	Поскольку байесовская сеть может рассуждать во всех направлениях, мы также можем запустить сценарий, в котором мы устанавливаем уровень удовлетворенности, скажем, «чрезвычайно доволен», а затем просматриваем распределение времени предоставления. Поскольку все узлы в сети меняют свои распределения, когда мы вводим какое-либо значение в любой другой узел, мы также можем наблюдать влияние изменений в сроках удовлетворения или предоставления на все другие переменные одновременно.
Large Bayesian networks can be very complex and difficult to work with in a what-if analysis (see Fig 2). User interfaces of standard software tools typically focus on the network structure and expect users to navigate through it. Users are also expected to create a network manually, which is basically impossible for domains where they have little or no knowledge about mutual dependencies between variables and are actually looking for a model helping to detect relevant relationships in the first place.	Большие байесовские сети могут быть очень сложными и трудными для работы в анализе «что если» (см. Рис. 2). Пользовательские интерфейсы стандартных программных средств обычно ориентированы на структуру сети и ожидают, что пользователи будут перемещаться по ней. От пользователей также ожидается, что они будут создавать сеть вручную, что в принципе невозможно для доменов, где они практически не знают о взаимозависимостях между переменными и в действительности ищут модель, помогающую

	обнаружить соответствующие отношения.
 Fig 2 A Bayesian network.	Рис 2 Байесовская сеть.
Business users require an intuitive and highly automated interface to Bayesian networks to benefit from the advantages that these models provide. In the following section we describe the iCSat platform that was implemented to support the analysis of customer satisfaction data.	Бизнес-пользователям требуется интуитивно понятный и высоко автоматизированный интерфейс с байесовскими сетями, чтобы воспользоваться преимуществами, которые предоставляют эти модели. В следующем разделе мы опишем платформу iCSat, которая была реализована для поддержки анализа данных об удовлетворенности клиентов.
3.2 Analysing customer data with iCSat	3.2 Анализ данных о клиентах с помощью iCSat
iCSat has been implemented to allow business users to easily analyse customer data, build models, run what-if scenarios, identify drivers, and set targets for them.	iCSat был внедрен для того, чтобы бизнес-пользователи могли легко анализировать данные клиентов, строить модели, запускать сценарии «что если», определять драйверы и устанавливать для них цели.
In order to build a new Bayesian network model, the user first loads a data description and a data set. The data description describes the meaning of each variable and its values and represents typically a survey, where a variable is a question and a value is a possible response. The data set is a table or spreadsheet where variables are organised in columns and each row or record represents data of one customer. The values in the data must represent mutually exclusive categories. Numerical data would have to be discretised first.	Чтобы построить новую модель байесовской сети, пользователь сначала загружает описание данных и набор данных. Описание данных описывает значение каждой переменной и ее значений и, как правило, представляет собой опрос, где переменная является вопросом, а значение - возможным ответом. Набор данных представляет собой таблицу или электронную таблицу, где переменные организованы в столбцы, и каждая строка или запись представляет данные одного клиента. Значения в данных должны представлять взаимоисключающие категории. Числовые данные должны быть сначала дискретизированы.

iCSat stores data sets and descriptions in an associated database. Once data is available, the user can start the modelling process by merely selecting the variables to be included in the new model, providing a name for the new model and starting the automatic model generation process.	iCSat хранит наборы данных и описания в связанной базе данных. Как только данные станут доступны, пользователь может начать процесс моделирования, просто выбрав переменные, которые будут включены в новую модель, указав имя для новой модели и запустив процесс автоматической генерации модели.
iCSat first learns the structure of the Bayesian network. This is done by using the CB algorithm suggested by Singh and Valtorta [8]. This algorithm starts with a fully connected undirected graph and uses χ 2 tests to determine conditional independence of connected nodes. If two nodes are conditionally independent the connection between them is removed. Then several rules and heuristics are used to direct the edges such that a directed acyclic graph (DAG) is created. The DAG is used to create a topological order of the nodes. This order is used as an input to the K2 algorithm by Cooper and Herskovitz [10] that computes for each node the list of parent nodes and thus creates the structure of a Bayesian network. After that we use the Netica library [9] to compute the conditional probability tables of the network. Netica is also used to represent the Bayesian network and run all the computations within the network.	Сначала iCSat изучает структуру байесовской сети. Это делается с помощью алгоритма CB, предложенного Сингхом и Валтортой [8]. Этот алгоритм начинается с полностью связанного неориентированного графа и использует χ2-тесты для определения условной независимости связанных узлов. Если два узла условно независимы, связь между ними удаляется. Затем несколько правил и эвристик используются для направления ребер так, что создается ориентированный ациклический граф (DAG). DAG используется для создания топологического порядка узлов. Этот порядок используется в качестве входных данных для алгоритма K2 Купера и Херсковица [10], который вычисляет для каждого узла список родительских узлов и таким образом создает структуру байесовской сети. После этого мы используем библиотеку Netica [9] для вычисления таблиц условной вероятности сети. Netica также используется для представления байесовской сети и выполнения всех вычислений в сети.
The learning algorithm iterates by starting with tests for zero order conditional independence, and then continues with 1st order tests and so on until an upper limit is reached, no more connections can be deleted, or a newly generated network is not better than the previous one. The complexity of the algorithm depends on the order of the independence tests and increases exponentially with the number of nodes. Therefore the algorithm is stopped after a maximum of 3 iterations, i.e. after 2nd order tests have been completed (an *n* th order test has to look at all subsets of cardinality *n* of the set of nodes connected to the two nodes whose independence is to be tested).	Алгоритм обучения повторяется, начиная с тестов на условную независимость нулевого порядка, а затем продолжается с тестами 1-го порядка и т. Д., Пока не будет достигнут верхний предел, больше не может быть удалено ни одного соединения или новая сеть не лучше предыдущей. , Сложность алгоритма зависит от порядка проверки независимости и экспоненциально возрастает с увеличением количества узлов. Поэтому алгоритм останавливается после максимум 3 итераций, т. Е. После завершения тестов 2-го порядка (тест n-го порядка должен проверять все подмножества мощности n набора узлов, соединенных с двумя узлами, независимость которых должна быть испытано).
After that automatic learning process has been completed, a model can be loaded for analysis. iCSat provides a very intuitive GUI (Fig 3) where the user sees two columns of bar charts. In the left column the charts can be manipulated to represent input data. In the right column the predictions of the model are displayed. The charts in the prediction column contain two sets of bars to compare the predictions to the original distribution found in the data. Both	После того, как этот автоматический процесс обучения завершен, модель может быть загружена для анализа. iCSat предоставляет очень интуитивно понятный графический интерфейс (рис. 3), где пользователь видит два столбца гистограммы. В левом столбце можно манипулировать графиками для представления входных данных. В правом столбце отображаются прогнозы модели.

columns can be fully configured to contain only the variables in which the user is interested. The following operations can be carried out on a model.	Диаграммы в столбце прогноза содержат два набора столбцов для сравнения прогнозов с исходным распределением, найденным в данных. Оба столбца могут быть полностью сконфигурированы так, чтобы содержать только те переменные, которые интересуют пользователя. Следующие операции могут быть выполнены на модели.
Sensitivity analysis	Анализ чувствительности
This automatically identifies drivers and their degree of impact on a selected variable (Fig 4). This information allows users to concentrate experiments on variables that actually impact on particular target variables.	Это автоматически определяет драйверы и степень их влияния на выбранную переменную (рис. 4). Эта информация позволяет пользователям концентрировать эксперименты на переменных, которые фактически влияют на конкретные целевые переменные.
What-if analysis	Что-если анализ
This understands how changes in the distributions of some variables affect all other variables. There are two modes — impact analysis and target setting. In impact analysis users can set some variables to particular values to, for example, simulate particular customer groups and see predictions for value distributions in other variables. In targetsetting mode, users can specify a new frequency distribution for some variables and see the impact on the distributions of other variables.	Это понимает, как изменения в распределении некоторых переменных влияют на все другие переменные. Есть два режима - анализ воздействия и постановка цели. В анализе воздействия пользователи могут устанавливать для некоторых переменных определенные значения, например, для имитации определенных групп клиентов и просмотра прогнозов для распределения значений в других переменных. В режиме задания целей пользователи могут указать новое распределение частоты для некоторых переменных и увидеть влияние на распределения других переменных.

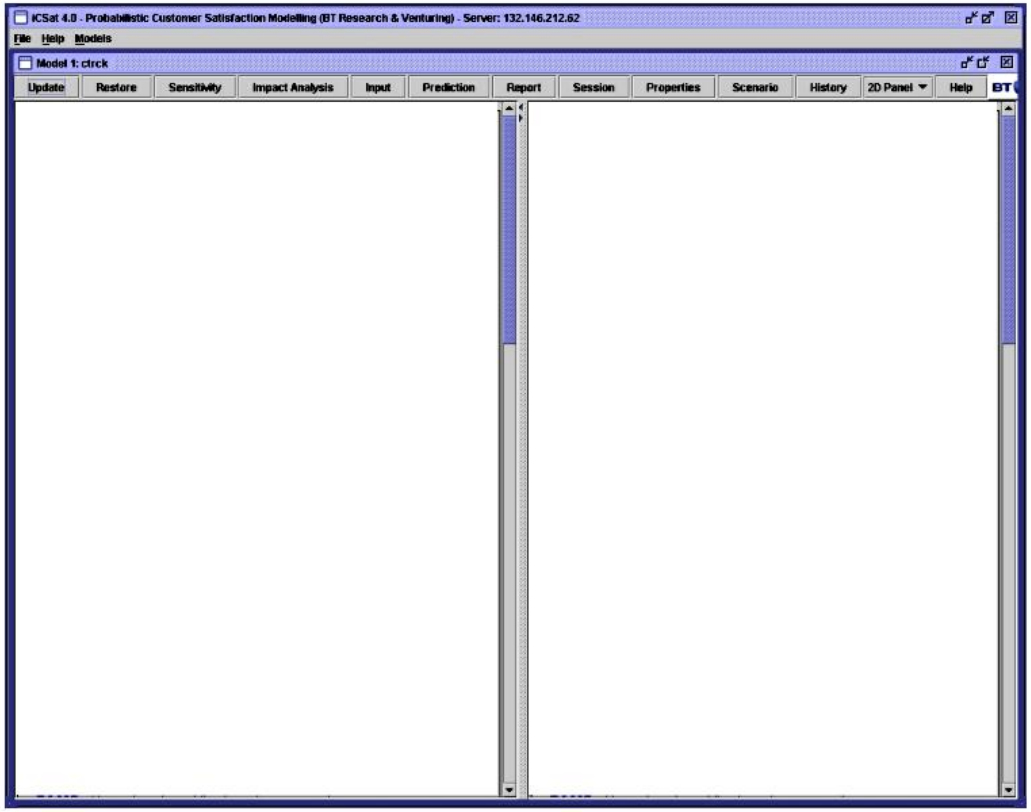


Fig 3 The graphical user interface of iCSat (variable names have been removed for confidentiality reasons).

Рис. 3 Графический интерфейс пользователя iCSat (имена переменных были удалены из соображений конфиденциальности).

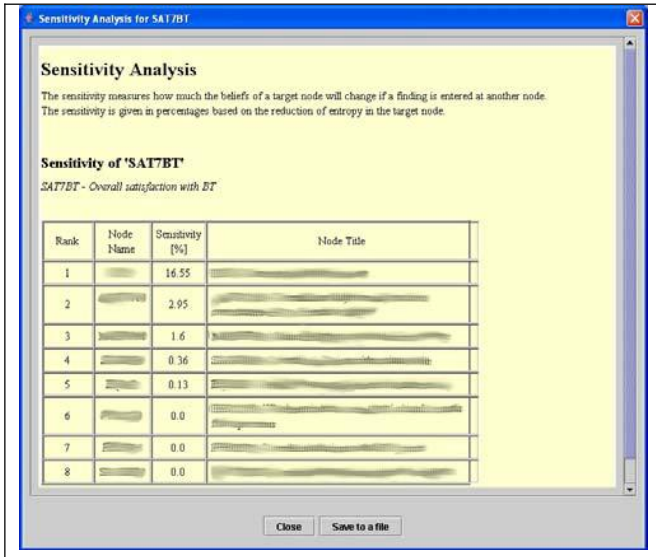


Fig 4 Result of a sensitivity analysis (variable names are obscured for confidentiality reasons).

Рис. 4 Результат анализа чувствительности (имена переменных скрыты по соображениям конфиденциальности).

Report This creates a table of the current settings for import into office documents. • History This allows saving what-if scenarios, reloading them and comparing them side-by-side. • Data aggregation This reduces the complexity of a data set by combining groups of values to new values. • 3-D View This operation represents the otherwise separate input and output charts in one 3-D chart (Fig 5). All analyses can also be carried out in the 3-D view.

Отчет При этом создается таблица текущих настроек для импорта в офисные документы. • История. Это позволяет сохранять сценарии «что если», перезагружать их и сравнивать их рядом. • Агрегирование данных. Это снижает сложность набора данных путем объединения групп значений в новые значения. • Трехмерный вид Эта операция представляет отдельные входные и выходные диаграммы в одном трехмерном графике (рис. 5). Все анализы также могут быть выполнены в трехмерном представлении.

4. Application of iCSat

4. Применение iCSat

iCSat has been used by us in different data analysis scenarios and it is also used by several business users across BT. Below we give four examples of where the tool has been applied. Because of confidentiality reasons we cannot provide any details of the analysis results.

iCSat использовался нами в различных сценариях анализа данных, а также несколькими бизнес-пользователями в BT. Ниже мы приводим четыре примера применения инструмента. По причинам конфиденциальности мы не можем предоставить какие-либо подробности результатов анализа.

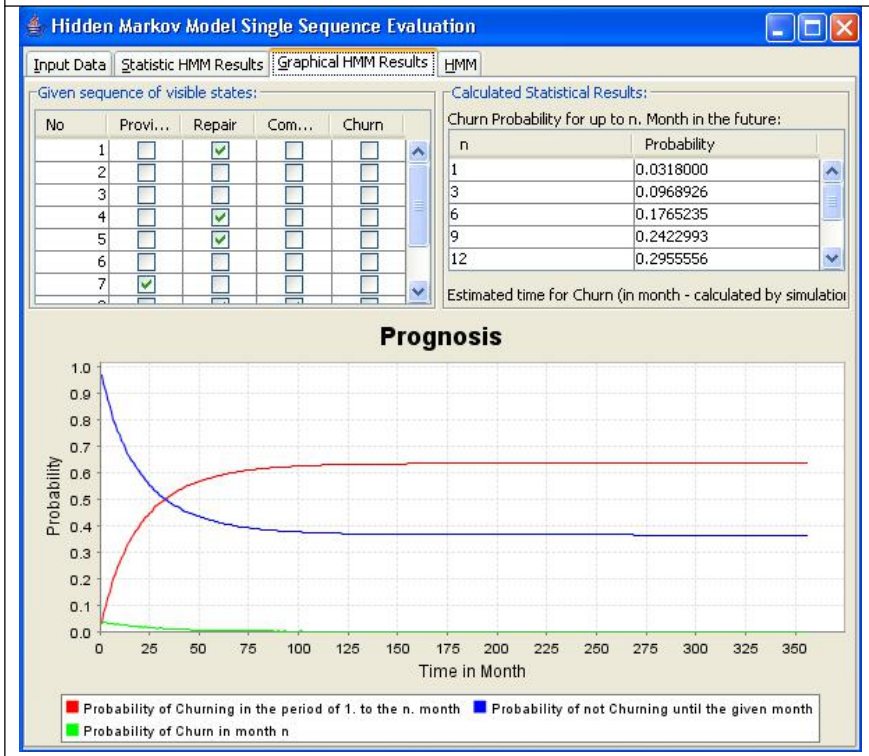
Fig 5 The 3D interface of iCSat (variable names are hidden for confidentiality reasons).

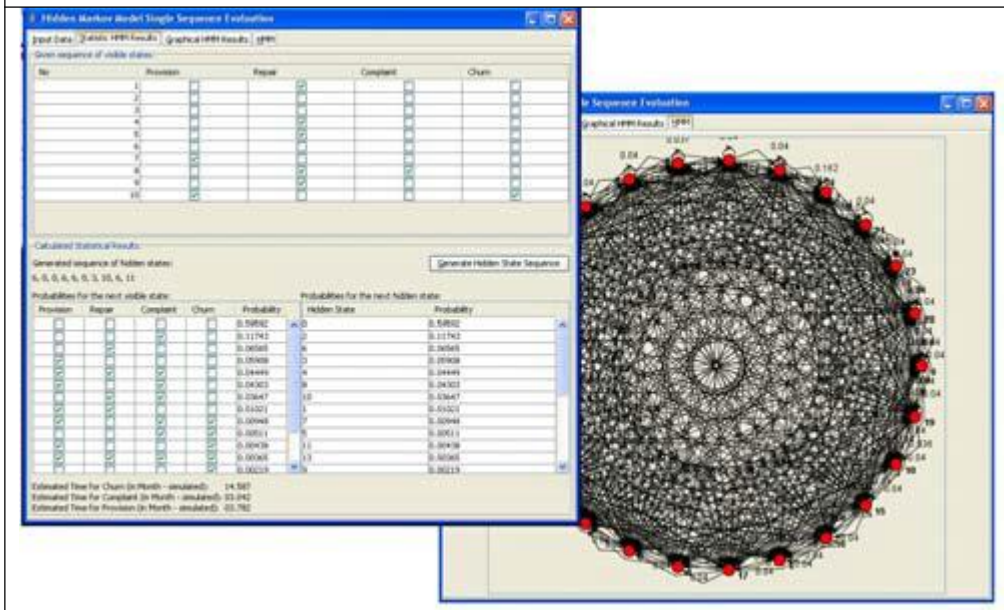
Рис. 5 3D-интерфейс iCSat (имена переменных скрыты по соображениям конфиденциальности).

• Customer satisfaction analysis iCSat has been initially conceived for this purpose. By using data from customer surveys we have identified the main drivers of satisfaction and analysed their impact.	• Анализ удовлетворенности клиентов iCSat изначально был задуман для этой цели. Используя данные опросов клиентов, мы определили основные факторы удовлетворения и проанализировали их влияние.
• Identifying customers in jeopardy	• Определение клиентов в опасности
We used iCSat to identify which circumstances in a repair process can lead to complaints by customers. Finding relevant process parameters allowed us to provide a simplified model for operational systems that allows customer services to obtain an early warning if a process for a customer is about to go wrong. A trial revealed that by intervening at the right time complaints could be substantially reduced.	Мы использовали iCSat, чтобы определить, какие обстоятельства в процессе ремонта могут привести к жалобам клиентов. Поиск соответствующих параметров процесса позволил нам предоставить упрощенную модель для операционных систем, которая позволяет сервисным службам получать раннее предупреждение, если процесс для клиента собирается пойти не так. Исследование показало, что благодаря своевременному вмешательству количество жалоб может быть существенно уменьшено.
• Target setting	• Установка цели
Combining process data with customer survey data allowed business users to understand the impact of certain process parameters on customer satisfaction. By setting a satisfaction target, target distributions for process parameters could be computed and used for setting internal performance goals.	Объединение данных процесса с данными опроса клиентов позволило бизнес-пользователям понять влияние определенных параметров процесса на удовлетворенность клиентов. Установив целевой показатель удовлетворенности, можно рассчитать целевые распределения для параметров процесса и использовать их для

	определения внутренних целей производительности.
• Field force performance	• Производительность полевой силы
By using data from field force operations we looked at the impact of certain regional factors on job performance.	Используя данные полевых операций, мы рассмотрели влияние определенных региональных факторов на производительность труда.
iCSat is a generic data analysis tool that can be applied to any domain if the provided data is categorical. We are in the process of adding new functionality to the tool and in the near future it will be integrated into iCAN — a platform for intelligent customer analytics.	iCSat - это универсальный инструмент анализа данных, который можно применять к любому домену, если предоставленные данные являются категориальными. Мы находимся в процессе добавления новой функциональности в инструмент, и в ближайшем будущем он будет интегрирован в iCAN - платформу для интеллектуальной аналитики клиентов.
5. Intelligent customer analytics	5. Интеллектуальная клиентская аналитика
iCAN is a Java-based client/server software package for the purpose of intelligent monitoring, analysis and prediction of customer behaviour, and its changes, during a complete customer life cycle with a service provider. iCAN is connected to a data warehouse with customer and process data and can work with both live and off-line data to provide analytics, build and apply various predictive models, and display the results. iCAN is designed in an open form which allows it to incorporate new analytical models at any time. All models covered by the tool provide full visualisation of their internal structure and offer user-friendly interfaces to carry out predictive experiments on customer behaviour.	iCAN - это программный пакет клиент-сервер на основе Java, предназначенный для интеллектуального мониторинга, анализа и прогнозирования поведения клиента и его изменений в течение всего жизненного цикла клиента с поставщиком услуг. iCAN подключен к хранилищу данных с данными клиентов и процессов и может работать как с оперативными, так и с автономными данными, чтобы предоставлять аналитику, создавать и применять различные прогнозные модели и отображать результаты. iCAN разработан в открытой форме, что позволяет в любое время включать новые аналитические модели. Все модели, охватываемые этим инструментом, обеспечивают полную визуализацию своей внутренней структуры и предлагают удобные интерфейсы для проведения прогнозных экспериментов по поведению клиентов.
The current research prototype includes two analytical methods — hidden Markov models (HMMs) [11] and decision trees. Decision trees [3] are capable of making static predictions of events with no related timing information, while HMMs can make timestamped event predictions. Both models can use either live data stored in remote databases or off-line data from a data warehouse. The models are supported by some standard statistical data analytics and visualisation capabilities. Models can be stored in a database and can be applied to different data sets at any time.	Текущий исследовательский прототип включает в себя два аналитических метода - скрытые марковские модели (HMM) [11] и деревья решений. Деревья решений [3] способны делать статические предсказания событий без связанной информации о времени, в то время как HMM могут делать предсказания событий с метками времени. Обе модели могут использовать либо живые данные, хранящиеся в удаленных базах данных, либо автономные данные из хранилища данных. Модели поддерживаются некоторыми стандартными статистическими возможностями анализа данных и визуализации. Модели могут быть сохранены в базе данных и могут быть применены к различным наборам данных в любое время.
Markov models represent a family of stochastic methods focused on the analysis of temporal sequences of discrete states. In traditional first-order HMMs, the states are hidden from observation, but each of them emits a number of	Марковские модели представляют собой семейство стохастических методов, ориентированных на анализ временных последовательностей дискретных состояний. В традиционных HMM

observable variables which could take either discrete or continuous values. As in all Markov models the current state of an HMM depends only on its previous state which means no prior history of sequence evolution has any effect on the current state. An HMM is fully described by its parameters, which are the transition probabilities between hidden states and the probabilities or probability densities of the emission of observables. Given a coarse model structure there are three central issues in hidden Markov models:	первого порядка состояния скрыты от наблюдения, но каждая из них испускает ряд наблюдаемых переменных, которые могут принимать либо дискретные, либо непрерывные значения. Как и во всех моделях Маркова, текущее состояние HMM зависит только от его предыдущего состояния, что означает, что никакая предшествующая история развития последовательности не имеет никакого влияния на текущее состояние. HMM полностью описывается своими параметрами, которые являются вероятностями перехода между скрытыми состояниями и вероятностями или плотностями вероятностей излучения наблюдаемых. Учитывая грубую структуру модели, в скрытых марковских моделях есть три основных вопроса:
• the learning problem — concerns calculation of model parameters based on the training observations of visible symbols,	• проблема обучения - касается расчета параметров модели на основе обучающих наблюдений видимых символов,
• the evaluation problem — assumes that the HMM model is fully defined and concerns the evaluation of the probability that a particular sequence of visible states was generated by the model,	• проблема оценки - предполагает, что модель HMM полностью определена и касается оценки вероятности того, что конкретная последовательность видимых состояний была сгенерирована моделью,
• the decoding problem — assumes the HMM and a set of visible observations while asking for the most likely sequence of hidden states that led to these observations.	• проблема декодирования - предполагает наличие HMM и набора видимых наблюдений при запросе наиболее вероятной последовательности скрытых состояний, которые привели к этим наблюдениям.
Once trained, an HMM can be used for a variety of applications, wherever sequential data and its future evolution is in question. HMMs have started to be appreciated in business analytics. Recent cross-sale models include HMM as one of the most successful approaches for recommending products to those customers who are most likely to buy them given their recent purchases [12]. In the iCAN software platform we use HMMs to solve the much more general problem of customer life-cycle modelling where the distinctions between hidden states are less sharp and the observables are available mostly in a continuous form that is difficult to process. It assumes that a customer develops a variable behaviour path which starts from subscribing to a service offered by a business and ends when he decides to cancel the service.	После обучения HMM можно использовать для различных приложений, где бы ни возникали последовательные данные и их дальнейшее развитие. HMM начали ценить в бизнес-аналитике. Последние модели кросс-продаж включают HMM как один из наиболее успешных подходов для рекомендации продуктов тем покупателям, которые с наибольшей вероятностью приобретут их, учитывая их недавние покупки [12]. В программной платформе iCAN мы используем HMM для решения гораздо более общей проблемы моделирования жизненного цикла клиента, когда различия между скрытыми состояниями менее четкие и наблюдаемые доступны в основном в непрерывной форме, которую трудно обрабатывать. Предполагается, что клиент разрабатывает переменный путь поведения, который начинается с подписки на услугу, предлагаемую бизнесом, и заканчивается, когда он решает отменить услугу.
The HMM has been implemented in the standard discrete form with a pair of Viterbi and Baum-Welch algorithms [11] used to find the unknown parameters	HMM был реализован в стандартной дискретной форме с помощью пары алгоритмов Витерби и Баума-Уэлча [11], используемых для

of the hidden states of the network. The model assumes that customer events or experiences represent observable customer variables generated from unknown (hidden) behavioural states that the customers find themselves in. The model assumption is that the customers, who behave similarly, i.e. are in the same hidden behavioural state, that is they have the same distribution of the likely events/experiences, have the same distribution of emission probabilities in the visible states.	поиска неизвестных параметров скрытых состояний сети. Модель предполагает, что события или события клиента представляют собой наблюдаемые переменные клиента, сгенерированные из неизвестных (скрытых) поведенческих состояний, в которых находятся клиенты. Модель предполагает, что клиенты, которые ведут себя одинаково, то есть находятся в одном скрытом поведенческом состоянии, то есть они имеют одинаковое распределение вероятных событий / событий, имеют одинаковое распределение вероятностей выбросов в видимых состояниях.
In addition to this general analysis of the population of customers, an HMM can also be applied to a single customer, for example, to analyse the churn risk. iCAN provides the latest historical data for the selected customer and the time evolution of the cumulative churn risk is presented graphically and numerically as shown in Fig 6.	В дополнение к этому общему анализу совокупности клиентов, HMM также может применяться к одному клиенту, например, для анализа риска оттока. iCAN предоставляет последние исторические данные для выбранного клиента, а временная динамика совокупного риска оттока представлена графически и численно, как показано на рисунке 6.
 Fig 6 HMM churn analysis for a single customer.	Рис. 6 Анализ оттока HMM для одного клиента.
Looking beyond just a churn prediction, the iCAN HMM can also be used for a	Если смотреть не только на прогнозирование оттока абонентов, то

comprehensive what-if scenario covering all the events encoded in the data (Fig 7). Given the trained HMM the user can make next step predictions based on a configuration of events and subject to user-specified simulated past-event sequences. In all cases, predictions are delivered in the form of probabilities of event configurations, immediately calculated using learned model parameters. Additionally, the model also returns the estimated average time to the next event in all categories provided by the data. In case of churn, this statistic constitutes the estimated remaining lifetime with the service provider.	iCAN HMM также можно использовать для комплексного сценария «что если», охватывающего все события, закодированные в данных (рис. 7). С учетом обученного HMM пользователь может делать предсказания следующего шага на основе конфигурации событий и в соответствии с заданными пользователем имитированными последовательностями прошлых событий. Во всех случаях прогнозы предоставляются в виде вероятностей конфигураций событий, которые сразу же рассчитываются с использованием параметров изученной модели. Кроме того, модель также возвращает приблизительное среднее время до следующего события во всех категориях, представленных данными. В случае оттока эта статистика составляет приблизительный оставшийся срок службы поставщика услуг.
 Fig 7 The iCAN HMM provides comprehensive what-if analysis.	Рис. 7. Модуль iCAN HMM обеспечивает всесторонний анализ «что, если».
Because iCAN provides automatic model building and can generate predictions on customer level automatically, it can easily serve as an analytical platform or add-on for modern CRM systems like Siebel.	Поскольку iCAN обеспечивает автоматическое построение моделей и может автоматически генерировать прогнозы на уровне клиента, он может легко служить аналитической платформой или дополнением для современных CRM-систем, таких как Siebel.
6. Conclusions	6. Выводы
Customer analytics is an extremely important area for large businesses. Barriers to customer analytics are, typically, bad data quality and lack of expertise in analytics. Bad data quality can arise from inconsistent legacy systems and the	Клиентская аналитика является чрезвычайно важной областью для крупного бизнеса. Препятствиями для анализа клиентов обычно являются плохое качество данных и отсутствие опыта в аналитике.

fact that data gathering is usually done without a subsequent analysis in mind. One way of addressing data quality is to move from outdated legacy systems to a central corporate data model and repository on top of which modern CRM solutions can operate. A lack in analytics expertise can be addressed by using highly automated intelligent tools that provide advanced analytics but with an intuitive interface. Business users are domain experts not data analysis experts. Therefore we need tools that support them and allow them to focus on their job.	Плохое качество данных может возникнуть из-за несовместимых устаревших систем и того факта, что сбор данных обычно выполняется без учета последующего анализа. Одним из способов решения проблемы качества данных является переход от устаревших унаследованных систем к центральной корпоративной модели данных и хранилищу, поверх которых могут работать современные решения CRM. Нехватка аналитических знаний может быть решена с помощью высокоавтоматизированных интеллектуальных инструментов, которые обеспечивают расширенную аналитику, но с интуитивно понятным интерфейсом. Бизнес-пользователи являются экспертами в области, а не экспертами по анализу данных. Поэтому нам нужны инструменты, которые поддерживают их и позволяют им сосредоточиться на своей работе.
analytic tools that are automated to an extent that business users do not need to worry about the analytical methods being used but can easily run scenarios, test assumptions and discover relevant information. Software like iCSat and iCAN is based on our research results in automated intelligent data analysis [4]. While iCSat has already been rolled out for business use, we are continuing to develop new analytical methods to capture the complete customer life cycle in iCAN.	аналитические инструменты, которые автоматизированы до такой степени, что бизнес-пользователям не нужно беспокоиться об используемых аналитических методах, но они могут легко запускать сценарии, проверять допущения и находить соответствующую информацию. Программное обеспечение, такое как iCSat и iCAN, основано на результатах наших исследований в области автоматизированного интеллектуального анализа данных [4]. Несмотря на то, что iCSat уже развернут для использования в бизнесе, мы продолжаем разрабатывать новые аналитические методы, чтобы охватить полный жизненный цикл клиента в iCAN.
References 1 Sheskin D: 'Handbook of Parametric and Nonparametric Statistical Procedures', Third Edition, Chapman and Hall/CRC Press, Boca Raton, Florida (2003). 2 Fayyad U M, Piatetsky-Shapiro G, Smyth P and Uthurusamy R: 'Advances in Knowledge Discovery and Data Mining', AAAI Press and MIT Press, Menlo Park, CA and Cambridge, MA (1996). 3 Mitchell T M: 'Machine Learning', McGraw-Hill, New York (1997). 4 Nauck D, Spott M and Azvine B: 'SPIDA — a novel data analysis tool', BT Technol J, 21, No 4, pp 104—112 (October 2003). 5 Berthold M and Hand D J (Eds): 'Intelligent Data Analysis: An Introduction', Springer-Verlag, Berlin (1999). 6 Pearl J: 'Probabilistic reasoning in intelligent systems: networks of plausible inference', Morgan Kaufmann Publishers Inc, San Francisco, CA (1988).	

7 Heckerman D and Wellman M P: ‘Bayesian Networks’, Communications of the ACM, 38, No 3 (March 1995).
8 Singh M and Valtorta M: ‘Construction of Bayesian Network Structures from Data: a Brief Survey and an Efficient Algorithm’, International Journal of Approximate Reasoning, 12, pp 111—131, Elsevier Science, New York (1995).
9 Netica User Manual V1.05, Norsys Software Corp (1997) — http://www.norsys.com/
10 Cooper G F and Herskovitz E H: ‘A Bayesian method for the induction of probabilistic networks from data’, Machine Learning, 9, pp 309—347 (1992).
11 Rabiner L R and Juang B H: ‘An introduction to hidden Markov models’, IEEE Signal Processing Magazine, 3, Issue 1, Part 1, pp 4—16 (January 1986).
12 Li S, Sun B and Wilcox R T: ‘Cross-selling Sequentially Ordered Products: An Application to Consumer Banking Services’, Journal of Marketing Research, 42, No 2, pp 233—240 (2005).