Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Statistics

Master's Thesis

Statistical analysis of selected marketing database

Ushakova Elizaveta

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Elizaveta Ushakova

Economics and Management
Economics and Management

Thesis title

Statistical analysis of selected marketing database

Objectives of thesis

The goal of this thesis is to assess website users' visits within selected large marketing database.
The author will focus on
- Big Data technologies in marketing
- Development of script for repeated analyses in marketing database
- Development of simulations in real marketing database

Methodology

The theoretical basis of the research is the positioning and presentation of Big Data technology in the scientific environment: scientific articles, publications in the media, reports of consulting companies.

The methodological basis of the research is analysis and synthesis, the method of scientific abstraction and modeling, induction, and bibliographic analysis. In the course of the work, a software package will be developed in the python 3.8 programming language.

The proposed extent of the thesis

60 – 80 pages

Keywords

Big Data, marketing, decision-making process, data-driven approach

Recommended information sources

Al-Jarrah, O., Yoo, P., Muhaidat, S., Karagiannidis, G. and Taha, K., 2015. Efficient Machine Learning for Big Data: A Review. Big Data Research, 2(3), pp.87-93. doi: 10.1016/j.bdr.2015.04.001

Ashton K., 2009. That ‘‘Internet of Things’’ thing, RFiD Journal,53.

Assunção, M., Calheiros, R., Bianchi, S., Netto, M. and Buyya, R., 2015. Big Data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing, 79-80, pp.3-15. doi: 10.1016/j.jpdc.2014.08.003

Das, T., Acharjya, D. and Patra, M., 2014. Opinion mining about a product by analyzing public tweets in Twitter. International Conference on Computer Communication and Informatics. doi: http://dx.doi.org/10.1109/ICCCI.2014.6921727

Hashem, I., Yaqoob, I., Anuar, N., Mokhtar, S., Gani, A. and Ullah Khan, S., 2014. The rise of "big data" on cloud computing: Review and open research issues, 47, pp. 98-115.

Huang, Z., 1997. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. Cooperative Research Centre for Advanced Computational Systems CSIRO Mathematical and Information Sciences, pp.1-8.

Cheng, S., Zhang, Q. and Qin, Q., 2016. Big data analytics with swarm intelligence. Industrial Management & Data Systems, 116(4), pp.646-666. Doi: 10.1108/imds-06-2015-0222

Kumar, P., Das, T.K., 2013. BIG Data Analytics: A Framework for Unstructured Data Analysis. International journal of engineering and technology, 5(1), 153-156.

Liu S. et al., 2020. Scalable Topological Data Analysis and Visualization for Evaluating Data-Driven Models in Scientific Applications, IEEE Transactions on Visualization and Computer Graphics, 26(1), pp. 291-300. doi: 10.1109/TVCG.2019.2934594.

Mishra, N., Lin, C. and Chang, H., 2015. A Cognitive Adopted Framework for IoT Big-Data Management and Knowledge Discovery Prospective. International Journal of Distributed Sensor Networks, 66, pp. 1-13. doi: 10.1155/2015/718390

Expected date of thesis defence

2021/22 SS – FEM

The Diploma Thesis Supervisor

Ing. Tomáš Hlavsa, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 9. 11. 2021

prof. Ing. Libuše Svatošová, CSc.

Head of department

Electronic approval: 23. 11. 2021

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 21. 03. 2022

Declaration

I declare that I have worked on my master's thesis titled "Statistical analysis of selected marketing database" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the master's thesis, I declare that the thesis does not break any copyrights.

In Prague on 2022 ______________________________

Acknowledgement

I would like to express my deep gratitude to my supervisor Ing. Tomáš Hlavsa for valuable comments and constant attention to my work.

Statistical analysis of selected marketing database

Abstract

The growth of business models based on the collecting and processing of data arrays - "Big Data" - has been fuelled by the global expansion of Internet access and a multiple rise in computer capacity. Companies can give consumers solutions suited to their particular requirements and feedback thanks to advances in data mining and machine learning. Marketing attribution is one of the strategies for utilizing big data technology in marketing. The term "attribution model" is defined in this study. A statistical analysis of the marketing database is performed, as well as a comparison of typical attribution models, to determine which channels have the highest conversion probability. Research showed that traditional attribution methods and the Markov chain model indicated considerable disparities in the analysis. On the basis of the evaluation, recommendations are made for the use of models in certain situations. There are also suggestions for using analytics system indications to develop your own attribution model.

Keywords: Big Data, data-driven decision making, attribution models, online marketing, customer journey, omni-channel marketing

Statistická analýza vybrané marketingové databáze

Abstrakt

Růst obchodních modelů založených na sběru a zpracování datových polí – „Big Data“ – byl poháněn globálním rozšířením přístupu k internetu a mnohonásobným nárůstem kapacity počítačů. Společnosti mohou spotřebitelům poskytovat řešení vhodná pro jejich konkrétní požadavky a zpětnou vazbu díky pokrokům v dolování dat a strojovém učení. Marketingová atribuce je jednou ze strategií pro využití technologie velkých dat v marketingu. V této studii je definován pojem "atribuční model". Provádí se statistická analýza marketingové databáze a také porovnání typických atribučních modelů, aby se zjistilo, které kanály mají nejvyšší pravděpodobnost konverze. Výzkum ukázal, že tradiční atribuční metody a model Markovova řetězce naznačovaly značné rozdíly v analýze. Na základě vyhodnocení jsou navržena doporučení pro použití modelů v určitých situacích. Existují také návrhy, jak používat indikace analytického systému k vývoji vlastního atribučního modelu.

Klíčová slova: Big Data, rozhodování na základě dat, atribuční modely, online marketing, cesta zákazníka, omni-channel marketing

Table of content

List of pictures

Chapter 1. Introduction

Everyday life of a modern person has long been associated with the use of information technology, various multifunctional electronic devices - the so-called gadgets (smartphones, tablets, fitness bracelets, game consoles, etc.), banking and cash register systems, video surveillance systems, and numerous accounting systems in enterprises. All these systems record the actions of people on a daily basis, turning this data into large streams of stored information.

This development of information technology, which over the past ten years has led, on the one hand, to an explosive increase in computing power and a manifold decrease in the cost of data storage, and on the other, to an ever-increasing flow of digital data in all aspects of human life, including economic, social and cultural, caused the emergence of a new concept - "big data".

The use of the mathematical apparatus in relation to large arrays of semi-structured data formed by the daily activity of users of network resources and services, such as search services (for example, Google, Yandex), social networks (Facebook, Twitter, YouTube) and commercial services (for example, online stores and other e-commerce resources), allows to identify hidden patterns that are implicitly present in the multidimensional aspects of information behavior of people.

However, the very possession of big data does not automatically lead to improvement of processes but gives the potential for development. What matters is not so much the data as the decisions that companies make on its basis, while ignoring big data technologies, companies at the same time began to notice lost profits.

Big data analysis finds its application in marketing for forecasting demand, identifying features of consumer behavior, segmenting them, developing marketing and communication strategies. By comparing factors such as seasonality, geography of requests for a specific product, it is possible to predict seasonal demand and develop a plan for distributing the advertising budget for a specific period in a specific region.

Despite the fact that analyses based on big data are already being implemented in practice, in the scientific literature, the main attention is paid to the technical and technological aspects of the issue, methods of analysis and their application for solving practical problems. Issues of the influence of "Big Data" on decision-making, development and implementation of innovations, optimization and improvement of the efficiency of economic activities of individual subjects and the economic system as a whole are reflected less often in studies or they touch upon particular problems.

This thesis is a systematization of information in the field of big data, an analysis of the methods and approaches used today in marketing practice (advertising, media measurements, marketing research), an analysis of the boundaries and barriers standing in the way of wider implementation of technologies related to big data.

Objectives and Methodology

1.1 Objectives

Advertising campaigns grab the attention of customers from a variety of sources. The result of visiting a website or making a purchase can be due to a series of advertising redirects. Analyzing this information, it is necessary to find out the weight of each transition. This is one of the most significant challenges for online marketing today.

Despite the huge amount of information accumulated over time, modern solutions often use a too simple approach. For example, one popular approach is to assume that the very first ad shown determines the buying decision.

Hence, **the main goal of this Master Thesis work** is to evaluate marketing channels from the selected database in terms of their likelihood of driving a conversion.

With this as a fundamental goal, the research also seeks to accomplish particular objectives such as:

1 Development of script for an analysis in marketing database;

2 Reveal the influence of the channels on conversion;

3 Assessment of changes in baseline metrics when channel is disconnected.

1.2 Methodology

The theoretical basis of the research is the positioning and presentation of Big Data technology in the scientific environment: scientific articles, publications in the media, reports of consulting companies.

An analysis will be conducted on a selected marketing database, which represents the data of the clients visits to the website from different marketing channels. The data set includes 456 217 records within a period of 1/12/2017 12:00:04 AM - 14/12/2017 11:59:59 PM. The reason for choosing this particular period can be justified by the fact that the average customer journey takes about two weeks. In other words, it takes about 14 days from the client's idea of the need to purchase a product before purchasing it.

In the course of the work, a software package will be developed in the python 3.8 programming language (Appendix A). Its purpose is to compute the total conversion probabilities for each channel based on user visits to a website. These probabilities will be sorted and the result compared with classical attribution methods.

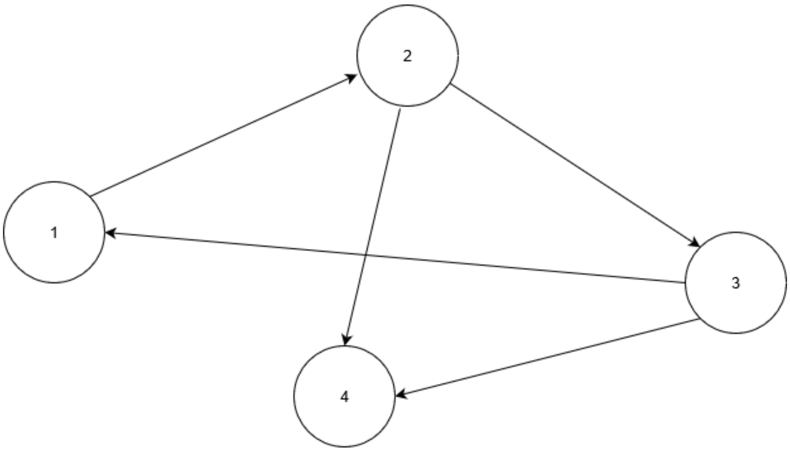
The methodological basis of the research is an attribution analysis based on Markov model. After analyzing the literature, the conclusion was made that Markov chains are the most effective and visual method for working with chains of sequences. However, there are two different ways to interpret the multichannel attribution model: based on a weighted directed graph and based on a probability matrix. The first model is excellent visually describing the model, while the second is an effective method for calculating parameters. Both models depict the same discrete stochastic process, only in different ways.

Random process

The main parts of every random process are state and transition. A state is the state of some random object in a certain period of time. Transition is the process of changing state. A random state is best described by a directed graph that illustrates a state - graph vertices in a certain period of time and transitions between states - graph edges that demonstrate the order of processes.

Picture 1 shows a directed graph that displays a model of some random process that goes between states 1, 2, 3, 4.

Picture 1. Directed graph as a model of a random process



Source: author, based on Wayne and Sedgewick, 2011

A random process is in a certain state only if its position is completely determined in terms of all the variables that characterize it. If there is a transition from one state to another, then there is a change in the variables from the original to a new one, which characterizes the new state, which determines it.

A random process is a successive transition from one state to another, previously unknown. Therefore, when modeling real systems, one often talks about the state of the system in a certain period of time and the probability of the system transition from one state to another (Pishro-Nik, 2014, pp. 540–574).

There are discrete and continuous random processes. If all states of the system belong to a countable set, then it can be said that the corresponding process is called a random process with discrete states or a discrete random process. In this case, the states of a discrete random process are mapped to a set of integer values, usually from a set of natural numbers, or other various numbers that characterize the process. Usually, the states of a discrete random process are determined by an ordinal number, while the number of possible states of the system can be finite: $A_1, A_2, A_3, \dots, A_n$ or infinite: $A_1, A_2, A_3, \dots$ A classic example would be the set of states of a dice, numbered from 1 until 6.

If the set of states cannot be numbered, then there is a random process with continuous states or just a continuous random process, which is characterized by a smooth transition from state to state and which is given as a continuous function of time: A(t). For example, the process of changing the temperature of some object can be considered as a random process with continuous states.

Since the paper considers the case of a discrete quantity, the justification will be presented later, therefore, in what follows - discrete random processes.

Random processes with discrete states

Let's denote the system states as $A_1, A_2, A_3, \dots, A_n$, while the time intervals when the state changes as $T_1, T_2, T_3, \dots, T_n$. Such a random process is called a random process with discrete time. If transitions between states are possible in any, strictly non-deterministic period of time, then such a process is a random process with continuous time.

The use of a discrete-time process model is justified in the case of transitions between states, possibly in a strictly defined time interval, or the state of the system can be generalized over certain time intervals without a large loss of significance. In this case, it is possible to map the set of system states $A = (A_n) = (A_1, A_2, A_3, \dots, A_n)$ onto the set of time intervals $T = (T_n) = (T_1, T_2, T_3, \dots, T_n)$, thus a certain function F(T) is given that describes the states of the system a in time intervals T. Such processes with discrete time are called stochastic sequences or random chains.

For clarity, such random processes with discrete states can be displayed as a directed graph, as in Picture 1, such graphs are also called transition graphs.

A transition graph is called labeled if there are transition probabilities on its edges.

There is the following classification of states an: an absorbing state is a state upon reaching which the random process ends; a non-returning state if the process of a certain number of transitions between states necessarily leaves them.

A random process is called Markov if the probability of transition between the current state and the state in the future depends only on the current one and does not depend on the previous ones. A Markov chain is a process with a discrete set of states and discrete time that satisfies the property of a Markov process.

The Markov chain implies that: $A = (A_n) = (A_1, A_2, A_3, \dots, A_n)$, where each A_n belongs to a discrete set of states, and a set of transition probabilities between states is also given:

$$P(A_{n+1} = k_{n+1} | A_n = k_n, A_{n-1} = k_{n-1}, A_{n-2} = k_{n-2}, \dots) = P(A_{n+1} = k_{n+1} | A_n = k_n)$$

Parameters of a Markov Stochastic Process

To describe a Markov random process with discrete states, the following set of parameters is used:

- The set of states $A = (A_1, A_2, A_3, \dots, A_n)$, in which a random process can be
- Transition matrix describing the probabilities of transitions between states
- Vector of initial probabilities defining the initial state of the system

For random processes with discrete time, state changes occur only at certain times $T = (T_n) = (T_1, T_2, T_3, \dots, T_n)$. Transitions between states are described by transition probabilities recorded in the transition matrix. If the transition between some states A_i and A_j is impossible, the probability corresponding to this transition is zero, respectively, that is, $P\left(A_i, A_j\right) = 0$.

Graph model

A graph is an abstract mathematical object, which is a set of graph vertices and a set of edges connecting the vertices in pairs. Events are located at the vertices of the graph, and transitions between them are on the edges (Goldin, 2014, pp. 409–413).

A graph is called directed if each of its edges has a strict direction, that is, a beginning and an end are given. Each edge can be represented as a vector that starts at one vertex and ends at another (Izhikevich, Kuramoto, 2006, pp. .448–453). An example of a directed graph is a person's route, where the set of vertices is the set of places where he has been, and the set of edges is the set of moves that he has made.

A graph is called weighted if each of its edges has some value, that is, each edge vector has its own weight. If add weights to each edge in a directed graph, then the resulting weighted directed graph can be used to create more complex models. Thus, it becomes possible not only to assess the direction of movement, but also to quantify it in some way.

A practical example is the construction of an optimal route from point A to point B, here the set of vertices is the set of intersections, and the set of edges is the set of roads. The complexity of the task is that here it is necessary to estimate not only the length of the route (the minimum length from vertex A to vertex B), but also the direction of movement (the existence of an edge connecting intermediate points in the required direction), as well as the road congestion.

Thus the idea of presenting the chains as a directed weighted graph is approached. To create a graph, there is a need to select a set of vertices and a set of edges. In case of the work scope, the set of vertices will be the set of channels, and the set of edges will be the set of transitions between them. To use Markov chains, it is needed to add two more objects to the set of vertices: 1 - conversion and 0 - last visit.

Hence, the set of vertices $A = (A_1, A_2, A_3, \dots, A_n, 1, 0)$, and the set of edges is the set of pairs from the set of vertices, the values of which are numerically equal to the probabilities of transition from one vertex to another:

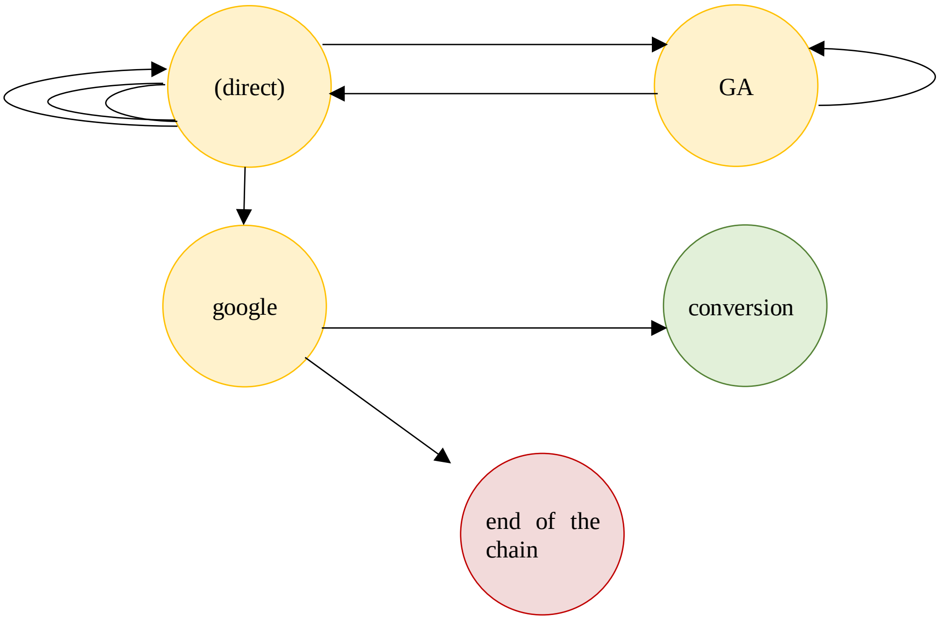
$$B = ((A_1, A_1), (A_1, A_2), \dots, (A_1, 1), (A_1, 0), (1, 0), (0, 1), (0, 0), (1, 1)) \qquad (1)$$

Source: Fewster, 2019

It should be noted that a node can return to a node, so the size of set B is $(n + 2)^{n+2}$, where n is the set of channels, and the additional two elements are the conversion - 1 and the end of the chain - 0.

Let's create a graph using the example: [(direct), (direct), GA, (direct), (direct), (direct), GA, GA, (direct), google, 1, google, 0]. This graph will have five vertices: three for channels, plus a vertex for conversion, and a vertex for end of sequence.

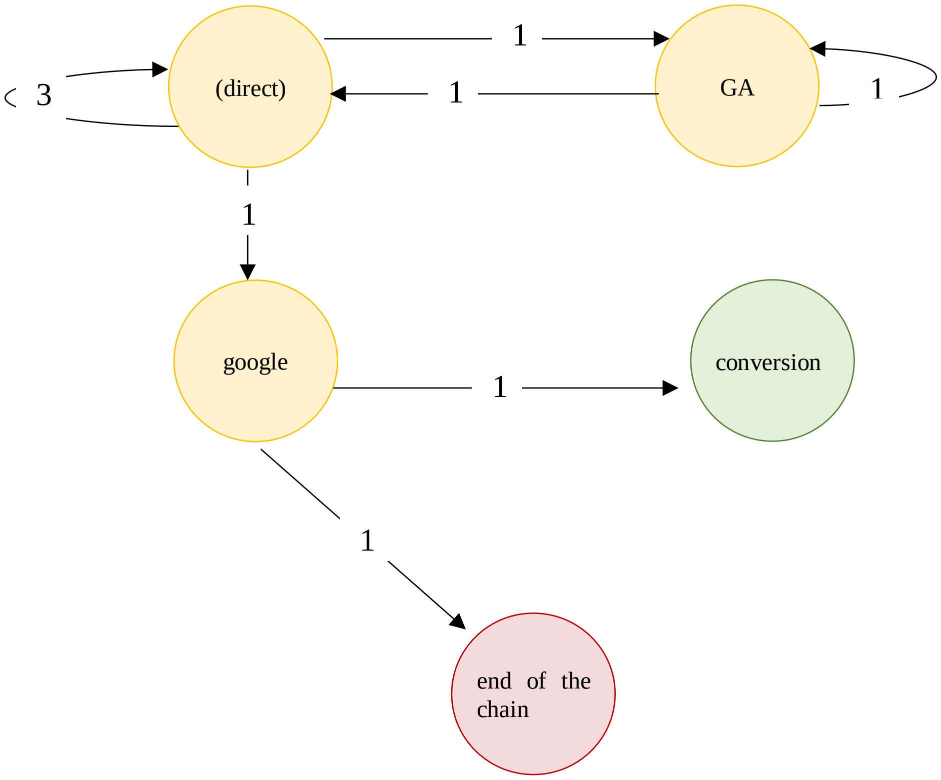
Graph 1. Directed graph based on example



Source: author, based on Peri, Sathya, Muktikanta and Singhal, Nandini, 2016

Of course, for more complex chains, when their total duration increases greatly, such a representation in the form of a graph becomes impossible. The chain can be simplified, if replace the duplicates with the weight of each chain, then the resulting graph will be directed and weighted.

Graph 2. Weighted and directed graph based on an example



Source: author, based on Peri, Sathya, Muktikanta and Singhal, Nandini, 2016

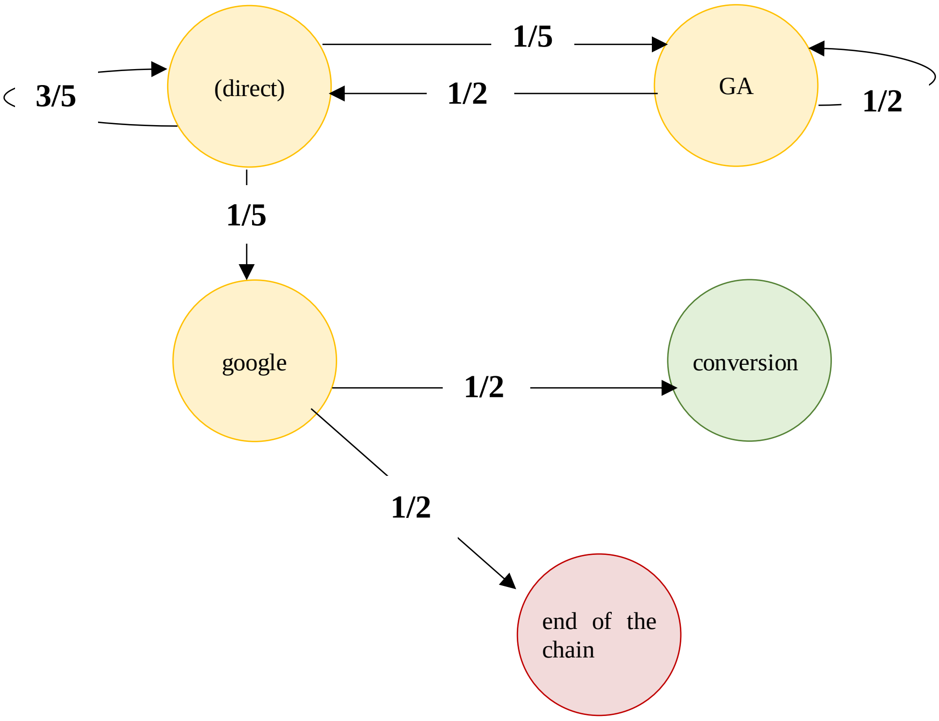
Such a graph is much more visual and with its help it is more convenient to carry out analysis. The next stage is the transformation of the number of transitions on the edge to the probability of such a transition, that is, to the probability of moving from one vertex to another.

For example, consider the top (direct). The following vertices of the graph are reachable from it: GA, google, (direct). A total of five transitions were made from this peak, and most of them (three out of five) were returnable, that is, they returned to themselves. One transition was to the top of GA and one to the top of google. If pass to the probabilities of events P((direct), (direct)), P((direct), GA), P((direct), google) as the probabilities of transition from the top (direct) to the tops (direct), GA and google respectively, then the following probabilities are obtained:

$$\mathrm{P}\big((\mathrm{direct}), (\mathrm{direct})\big) = \frac{3}{5}, \mathrm{P}\big((\mathrm{direct}), \mathrm{GA}\big) = \frac{1}{5}, \mathrm{P}\big((\mathrm{direct}), \mathrm{google}\big) = \frac{1}{5}$$

If to calculate the transition probabilities to all vertices, then the resulting graph will fully illustrate how and with what probability transitions between channels are made.

Graph 3. Directed graph with weights equal to transition probabilities based on an example



Source: author, based on Peri, Sathya, Muktikanta and Singhal, Nandini, 2016

Using a weighted and directed graph like the one in Picture 3, it is possible to get the full conversion likelihood for each channel. The following recursive formula is used for calculation:

$$P_{full}(A_i, 1) = \sum_j^{n-1} P\left(A_i, A_j\right) * P_{full}\left(A_j, 1\right) \tag{2}$$

Source: Fewster, 2019

This formula (2) is a calculation of the probabilities of transition from a given vertex to all connected to it, and from them to the original one. Then the total probability is recursively calculated from the vertices connected and the original one. This formula, as it were, calculates all the options for the paths along which transitions can be made and sums up the probabilities in this way to reach the top with a conversion. With the use of the formula, the total conversion probability can be obtained if the graph is unidirectional, that is, if there is an edge connecting the vertices A_i and A_j, but there is no edge that connects the edges in the opposite direction, then A_j and A_i. Otherwise, the formula implies a system of linear equations, with the number of unknown parameters equal to the number of graph edges in the opposite direction.

For example, let's calculate the total conversion probability $P_{full}\big((direct), 1\big)$ for the channel (direct):

$$P_{full}\big((direct), 1\big) = P\big((direct), GA\big) * P_{full}(GA, 1) +$$

$$P\big((direct), google\big) * P_{full}(google, 1) =$$

$$\frac{1}{2} * P_{full}(GA, 1) + \frac{1}{2} * P_{full}(google, 1)$$

There is no chance of returning to itself here, as this will lead us to an infinite recursion. Let's calculate the total probabilities for the $P_{full}(GA, 1)$ and $P_{full}(google, 1)$ channels.

$$P_{full}(GA, 1) = P\big(GA, (direct)\big) * P_{full}\big((direct), 1\big) = 1 * P_{full}\big((direct), 1\big)$$

Consider the total conversion probability for the top google:

$$P_{full}(google, 1) = P(google, 1) * P_{full}(1, 1) + P(google, 0) * P_{full}(0,1)$$

$$= \frac{1}{2} * 1 + \frac{1}{2} * 0$$

The probability of getting from 1 to 1 is 1 because the desired vertex has already been reached, and the probability of getting from 0 to 1 is 0 because it is a non-returning vertex. Combining all the equations into one, then get:

$$P_{full}((direct, 1) = \frac{1}{2} * 1 * P_{full}\big((direct), 1\big) + \frac{1}{2} * \frac{1}{2}$$

It is possible to express from here $P_{full}\big((direct), 1\big)$ and get the probability of 0.5.

Of course, this model is very illustrative and makes it easy to follow the pattern of user actions and get a complete map of his visits. The given graph was very simple, since there were very few vertices, the number of edges was also small, so calculating the total probability was not difficult, but the user's history of visiting the site is usually much larger and such a model becomes too much complexity. In addition, if to allow the possibility of edges of the form (A_i, A_i), that is, return the return edges to consideration, then the system of equations turns from linear to nonlinear, the solution of which is much more complicated and will require more computing power. Therefore, it is necessary to find a more efficient solution, which, although it does not have such clarity, but allows to calculate the total probability with greater performance.

Matrix model

Above a graph model was considered, which, although is very visual, but not very effective. To move to a more efficient matrix model for calculating probabilities, there is a need to return to the original set of channels.

$A = (A_1, A_2, A_3, \dots, A_n, 1, 0)$ which were used as vertices earlier. Let's make a matrix of probabilities, the elements of which will be transition probabilities: $P(A_i, A_j)$. It should immediately be discussed that the rules that were valid for the graph model remain valid for the matrix model. So, for example, the probability of any transition from the event of the end of the chain - 0 is equal to zero $P(0, A_n) = 0 \quad \forall n \in A \setminus \{0\}$, excluding itself (Fewster, 2019).

The same is true for the conversion event $P(1, A_n) = 0 \quad \forall n \in A \setminus \{1\}$. However, the probability of transition from the end of the chain event and the conversion event to themselves is equal to one $P(0,0) = P(1,1) = 1$. It should also not be forgotten that the return probabilities for the channels are equal to zero $P(A_n, A_n)\ \forall n \in A \setminus \{0, 1\}$. Thus, a square matrix of size $(n+2, n+2)$ is obtained.

Here is an example probability matrix for our example: [(direct), (direct), GA, (direct), (direct), (direct), GA, GA, (direct), google, 1, google, 0]

$$K = \begin{pmatrix} 0 & 0{,}5 & 0{,}5 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}5 & 0{,}5 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Source: author

The condition must be satisfied that for any i row of the matrix K, its sum must be equal to 1, which is quite obvious, since the probability of transition to any channel, including itself, is equal to 1:

$$\sum_{j}^{n+2} P(A_i, A_j) = 1 \tag{3}$$

Source: Przytycka, Rogozin, 2018

If in a matrix the sum of any row is equal to one, then such a matrix is called stochastic. And from this it follows that such a matrix is a transition matrix for some Markov process. The process of transition between channels can be called Markovian, since each subsequent state of the system depends only on the current one, which fully satisfies our conditions. Having determined such a transition matrix K, it is possible to calculate the transition probability $P\left(A_i, A_j\right)$, as well as the probability distribution of each channel for an arbitrary number of transitions.

Of course, the main question of interest is the calculation of the total probability of getting from an arbitrary channel to the conversion state $P_{full}(A_i, 1)$.

There is an analytical solution to this problem, provided that 1 and 0 are final states: the total transition probability for some channel A_i to 1 is equal to the element of the probability matrix to an infinite degree at position $(i, n+1)$, so

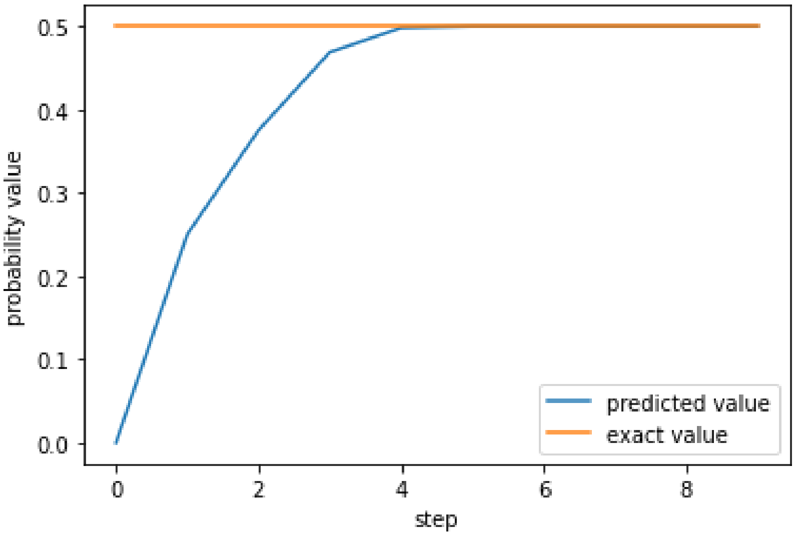
$$P_{full}(A_i, 1) = \lim_{x \to \infty} K^x(i, n+1) \tag{4}$$

Source: Hryniv, 2018

There is a rigorous proof of the possibility of calculating this expression, but it is enough to stop at a fairly close approximation. To do this, it is possible to use not an infinitely large degree, but to raise the matrix to a power of two. The advantage of raising to the power of two is that at each step one can multiply the matrix by itself.

Let me show on the example above the rate of convergence of the predicted probability to its exact value, calculated analytically for us the probability, where the step is the squaring of the matrix:

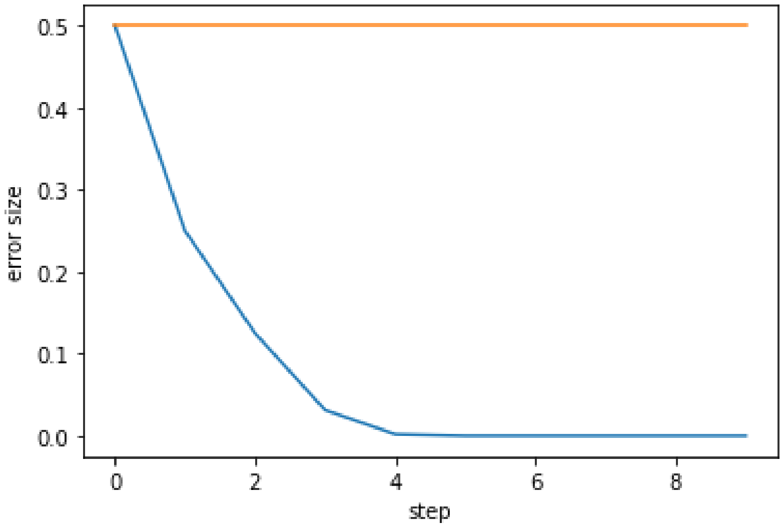
Picture 2. The value of the probability depending on the step



Source: author

As it can be seen from Picture 2, the probability value of the calculated number became equal to the real value already at step 4.

Picture 3. The size of the error depending on the step



Source: author

Picture 3 shows that the size of the error drops by about two times at each step, and after 4 it is already equal to zero, so it is possible not to continue the calculation further. Summarizing the result, the matrix model has proven to be an effective alternative to the graph model, so it can be used to calculate the total probability for large amounts of data.

As it was mentioned in the research objectives, the scope of the work will include not only the findings the total probabilities but also observe the influence of the individual channels and the changes in the baseline metrics.

Calculating the influence of channels on conversion

The magnitude of the influence $I(A_i)$ calculated on a certain channel A_i will be estimated as the volume of channels ending in conversion, where there is a target channel A_i, on the total number of channels with conversion, indicated by X:

$$I(A_i) = \frac{\left|\left\{R_j \mid A_i \in R_j, 1 \in R_j\right\}\right|}{X} \tag{5}$$

Source: Zeiler, Fergus, 2013

Naturally, for any, the quantity satisfies the following inequality:

$$0 \leq A_i \leq 1$$

When the channel is not included in any "conversion" sequence, then the influence is equal to 0 and equal to 1, provided that the channel was present in each conversion chain. Thus, it is possible to estimate the new number of conversions that occurred after the channel was deleted:

$$CV_{new} = X * \left(1 - I(A_i)\right) \tag{6}$$

Source: Zeiler, Fergus, 2013

Assessment of changes in baseline metrics when a channel is disconnected

To estimate the basic metrics, it is additionally necessary to somehow estimate the transition cost for each channel. The cost of a transition is a conditional value that shows how much it costs to go from each channel to the site. Then denote that the transition cost for each channel in the chain is $V\left(R_j\right)$.

Thus, it is possible estimate the cost of one chain as follows:

$$V(R_j) = \sum_{A_j \in R_j} V_j(A_j) \tag{7}$$

Source: Li, 2016

In this case, the total costs per channel are equal to:

$$V(A_j) = \sum_{J:A_i \in R_j} V_j(A_j) \tag{8}$$

Source: Li, 2016

In this case, the amount of expenses for advertising the site when using the channels is equal:

$$V = \sum V_j \sum_{A_i \in R_j} V_j(A_i) \tag{9}$$

Source: Abhishek, Fader, Hosanagar, 2012

If it is necessary to estimate the new cost, which was the result of the removal of a certain channel, then obviously: $V_{new} = V_{old} - V$

where $V_{new}(A_i)$ - new consumption, which was the result of the removal of the channel A_i, and V_{old} – old expenses, before deleting the channel A_i.

However, such simple judgments do not lead to the correct decision for the reason that in those Rj chains where A_i occurs, other, not remote channels could be involved in front of it. Summarizing, in order to increase the completeness of the estimate, one should take into account the previous channel, which affects the cost of chains, excluding not only the A_i channel, but also the previous one. So:

$$V_{new} = V_{old} - \sum V_j \sum_{i=0}^{A_{j-1}} V_j(A_i) \tag{10}$$

Source: Abhishek, Fader, Hosanagar, 2012

This formula means that there is not only the A_i channel removed from the chain, but also the A_{i-1} channel.

Chapter 2. Theoretical foundations of researching Big Data technologies

2.1 Development of Big Data technology in the literature

Researches related to big data are a topical area of development today. They study big data and how to extract knowledge from it. They are conducted within various disciplines and fields, such as information science, uncertainty modeling, machine learning, statistical learning, pattern recognition, data storage methods, signal processing, etc. Big data researches also have its own problems and tasks. The academic environment considers the several groups of research problems related to big data.

The first group of problems addresses the problems associated with storing and analyzing big data. They are being studied by researchers Z. Huang (1997), T. K. Das, P. M. Kumar, D. P (2013), Acharjya and M. R. Patra (2014).

The storage problem arises from the increase in the speed of new data creation in recent years. Thanks to mobile devices, the Internet of Things, the increasing availability of the Internet and other factors, the amount of information produced is growing exponentially. Due to the lack of space to store them, they are either deleted or not recorded at all. In this regard, the role of information carriers and the speed of its writing and reading increases for the availability of big data for the purpose of their analysis. Despite advances in this area, such as, for example, the proliferation of solid-state drives, the required performance of drives for processing big data has not yet been achieved (Huang, 1997, pp.1-8).

In addition to the volumes of information produced, their diversity has also increased in recent years, which significantly complicates the tasks of analyzing big data. It becomes necessary to compress or sample the data being processed, since the existing methods and algorithms do not allow them to be analyzed in a reasonable time. Automation of this process, including with the help of machine learning, is the task facing researchers (Kumar, Das, 2013, pp. 153-156).

Recent technologies such as Hadoop and MapReduce allow to collect large amounts of semi-ordered and disordered data in a reasonable amount of time. To be able to process them further, they need to be ordered. The development of data ordering algorithms is also an urgent task (Das, Acharjya, Patra, 2014).

Within the framework of the second group of research problems, the problems of extracting knowledge from big data and the computational complexity of processing big data are considered. They are being studied by researchers Al-Jarrah O. Y., Yoo P. D., Muhaidat S. and Karagiannidis G. K (2015).

Extraction and presentation of knowledge from big data is the main task of their processing. It includes several sub-tasks such as authentication, archiving, management, storage, retrieval and presentation of knowledge. The algorithms that are used to solve these problems are based, for the most part, on the theory of fuzzy sets and fuzzy logic, which are currently being actively developed.

Big data analysis can be computationally complex. The main problem in the analysis is the elimination of inconsistencies and uncertainties that are present in the datasets. Despite the fact that attempts to overcome computational complexity are implemented in most cases of processing big data sets, there is still no single method applicable to all cases. The available analysis tools have too low performance and are unable to efficiently deal with the inconsistencies, uncertainties, and computational complexity that arise when processing big data sets. There are already developments in this area, as well as new ones, mainly using machine learning. The main goal is to minimize the computational complexity(Al-Jarrah, Yoo, Muhaidat, Karagiannidis and Taha, 2015, pp. 87-93) (Yoo, Ramirez and Liuzzi, 2014, p. 50).

The third group of research problems are the problems of scalability and visualization of big data. They are being studied by such researches as Liu S., Wang D., Maljovec D., Anirudh R., Jacobs A. Recently, research in the field of big data has made it possible to achieve acceleration of their processing, against the background of an increase in the performance of processors according to Moore's law. Despite this, big data volumes are growing much faster than processor performance. In this regard, the task is to parallelize computations between different processors, including between different cores of the same processor. Methods and algorithms for parallel computing are one of the areas of research (. Liu et al., 2020, pp. 291-300).

The purpose of visualizing a big data set is to give analysts an adequate understanding of its properties, to help interpret them correctly. Visualization allows to turn large amounts of data into graphs or images that give analysts an intuitive view of their content. Modern visualization tools have poor performance, functionality and response time, which is also a research problem.

The fourth group of research problems is represented by the problems of information security associated with big data. They are studied by researchers Zhu H., Xu Z., Huang Y

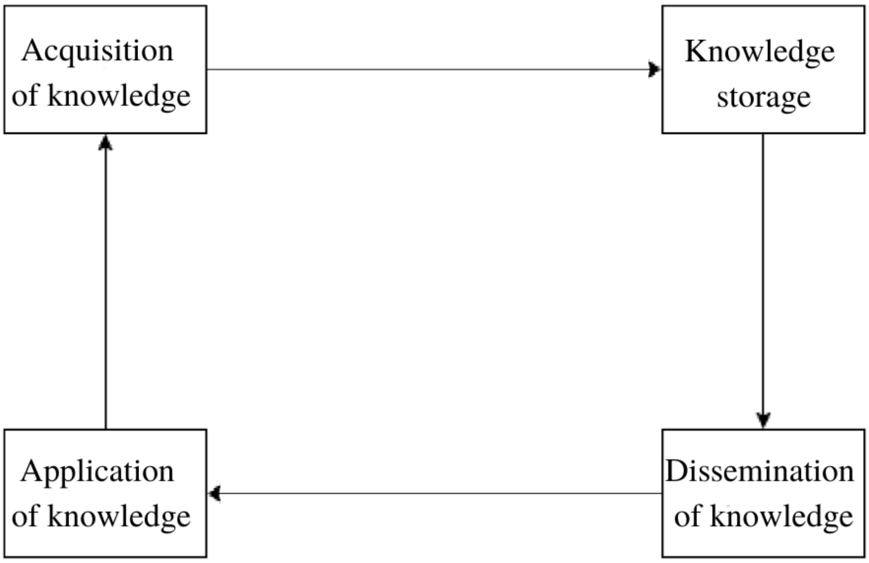
(2015). In the process of analyzing big data sets, useful knowledge is extracted from them. Different organizations have different security policies to protect confidential information, which is necessary due to the high risks associated with operations with big data (Zhu, H., Xu, Z. and Huang, Y., 2015, pp.1041-1044). Information security is becoming a problem when analyzing big data. It can be achieved using authentication, authorization and encryption techniques. Big data security measures exist in the context of computer networks, a variety of devices, real-time security tracking and checking for information leaks. Despite the development in this area, measures to ensure the security of big data still need to be improved. The main challenge is to develop a multi-layered security system that provides complete privacy protection when processing big data.

Few more relevant, but smaller-scale research areas related to big data: the Internet of Things, cloud computing, and bioinspired computing should be observed.

The Internet of Things is a concept of a computer network of physical objects equipped with built-in technologies for interacting with each other or with the external environment, which considers the organization of such networks as a phenomenon that can restructure economic and social processes, excluding from part of actions and operations the need for human participation (Ashton, 2009). In the course of the work, "things" generate a continuous stream of data. Big data technologies can be used to process it, which will allow to extract knowledge from them and improve the management of this data, which will make it possible to achieve greater efficiency of the “thing”. Researchers N. Mishra, C. Lin, and H. Chang (2015) have made great progress in this area. Picture 4 shows the big data knowledge cycle.

At the first stage, the acquisition of knowledge, knowledge is extracted using various computational methods. At the second stage, the storage of the acquired knowledge is organized in a system specially designed for this task. At the third stage, information that is significant for making a decision is extracted from the stored knowledge, and at the fourth it is used to make a decision, both by a person and by a computer (Mishra, Lin, Chang, 2015, pp. 1-13). What has been adopted may be reflected in the next iteration, determining what schemes the next extraction of knowledge will follow.

Picture 4. Cycle of using knowledge from big data



Source: author, based on Mishra, Lin, Chang, 2015

Advances in visualization have made supercomputer computing more accessible. While the user sees only the result of visualization, calculations are carried out using a computer infrastructure remote from him, the computing power of which is many times greater than the power of personal computers. This technology is called cloud computing and is the key to big data processing. The demand for it is due to the unprofitability of owning infrastructure with a large computing power on an ongoing basis, while the results of its work may be needed. Currently, there is research in this area, with an emphasis on data management and storage (Assunção, Calheiros, Bianchi, Netto and Buyya, 2015, pp. 3-15) (Hashem, Yaqoob, Anuar, Mokhtar, Gani and Ullah Khan, 2014, pp. 98-115).

Another advantage of cloud computing is the ability to store data in the cloud. If the data is stored in a different location, it becomes difficult to load and unload them for computations, which is a problem due to their large volume.

Another problem associated with the use of cloud technologies is data security. Since the data and the results of their processing must be transmitted over the network, there is a danger of their interception. Solving all of the above problems will bring big data processing to a new level.

Bioinspired computing is a technology for constructing algorithms that copy the principles of biological systems. The main advantages of these algorithms are simplicity and high speed of finding the optimal solution (Wang, Shen, 2013, pp. 1-7). Prospects for the use of bioinspired computing have been discussed by researchers C. Shi, Y. Shi, Q. Qin

(2016), and Del Ser et al (2022). According to their findings, bioinspired computing has great potential for solving big data problems.

The amount of data generated from various sources doubles every two years. Their processing opens up broad prospects for obtaining useful information but requires appropriate methods. Transforming big data into knowledge is a complex and challenging task. Data can include uncertainty, methods for overcoming which already exist (fuzzy logic, neural networks), but are far from perfect. The data are often not useful in full, which requires methods of their selection for a specific study or task. All these methods require high computing power and memory, which is also a problem. There were the main areas of research related to big data reviewed. Let’s now consider how the phenomenon of big data has been treated in the past and in the present.

2.2 Positioning of Big Data technology in the past and present

The understanding of the phenomenon of big data has changed from its inception to the present day. Let's consider the key events that characterize the shift and deepening of views on big data.

The term Big Data itself was first coined in 1997 by Michael Cox and David Ellsworth at the 8th IEEE Visualization Conference. They called the problem of big data the lack of capacity of the main memory, local and remote disks to perform virtualization (Cox, Ellsworth, 1997). And in 1998, the head of research at SGI John Mashey at the USENIX conference used the term Big Data in its modern form (SGI, 2021).

The understanding of the possibilities of big data came a little later. So, in November 2000, at the eighth world congress of the econometric community, Francis Diebold presented a report entitled "Big Data Dynamic Factor Models for Macroeconomic Measurement and Fore-casting", in which he stated that recently, science has been confronted with the phenomenon of big data and benefited from them. Big data is an increase in the quantity (and sometimes quality) of available and potentially important data, which is a consequence of high achievements in the field of recording and storing information (Diebold, 2003, pp. 115-122).

On February 6, 2001, Douglas Laney of the Meta Group (part of Gartner) issued a document describing the main problem areas associated with increased requirements for central data warehouses against the backdrop of the rapid growth of e-commerce, as well as forecasting a change in IT strategy. companies regarding approaches to building the architecture of solutions related to the storage and processing of information.

Three most important areas were identified on which to focus on solving data management issues: Volume (data volumes), Velocity (the speed of data accumulation and processing) and Variety (a variety of sources and types of data). Later, these concepts became the basis for a descriptive big data model called 3V (VVV).

Volume

The amount of data is an important factor. When working with big data, it is needed to handle massive amounts of unstructured, low-density data. Some companies can receive tens of terabytes of data, others hundreds of petabytes.

Velocity

Speed in this context is the speed at which data is received and, possibly, actions based on it. Typically, high-speed data streams go directly to RAM instead of being written to disk. Some Internet-based smart products operate in real or near real time. Accordingly, such data requires real-time assessment and action.

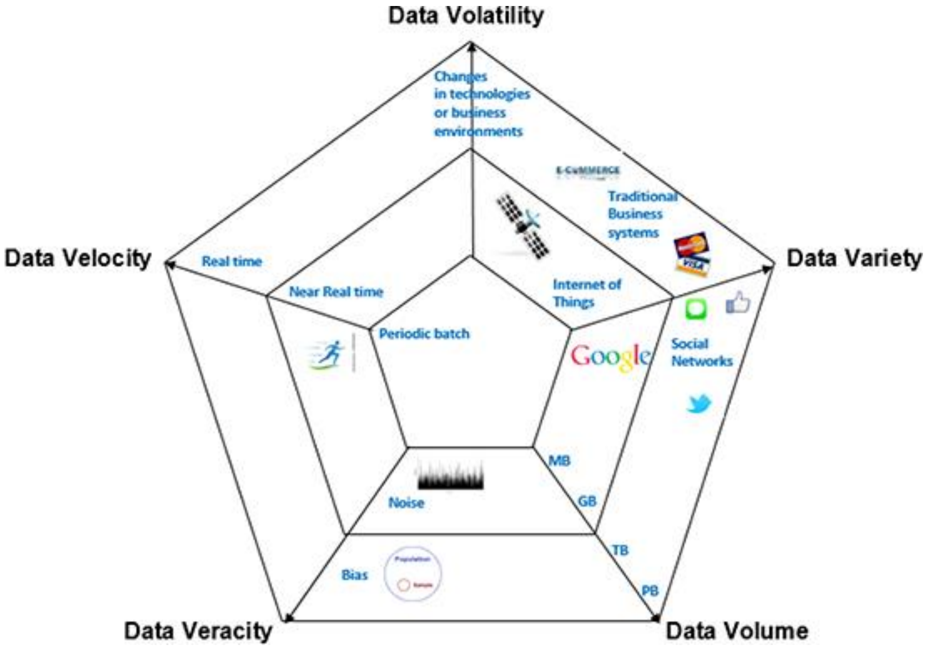
Variety

Diversity means that the data available is of different types. Traditional data types are structured and can be directly stored in a relational database. With the advent of big data, data began to flow in unstructured form. Unstructured and semi-structured data types such as text, audio, and video require additional processing to determine their meaning and support metadata.

It should be noted that these aspects were discussed without referring to the concept of big data, but these parameters described the basic principles of what is called Big Data today.

The list of characteristics (V) has grown over time, highlighting both the opportunities and the challenges that companies and organizations face when integrating big data into their existing ones manufacturing operations (Picture 5). Credibility deals with noise and bias in data and is one of the biggest challenges to delivering value and value to big data. Volatility refers to the changing technologies or production environments in which big data is created, which can lead to inaccurate analysis and results, and the fragility of big data as a data source (Banbura, Giannone, Modugno and Reichlin, 2013).

Picture 5. Characteristics of big data (5 V)



Source: Hammer, Kostroch, Quiros, 2017

The attention of the general public to big data was drawn in June 2008, when Chris Anderson published an article “The end of theory: the data deluge makes the scientific method obsolete” (Anderson, 2008). It is argued that the ever-increasing volume of data allows science to make predictions without forming a theory for this. Knowledge of the correlation between values may be sufficient to make a decision.

The widespread introduction of the term “big data” in the scientific community is associated with Clifford Lynch, the editor of the journal Nature, who prepared a special issue for September 3, 2008, with the topic “How technologies that open up opportunities for working with large volumes can affect the future of science -by data? ", which collected materials on the phenomenon of explosive growth in the volume and variety of processed data and technological prospects in the paradigm of a probable jump from quantity to quality.

In December 2008, Randal E. Bryant, Randy H. Katz and Edward D. Lazowska (2008) published the article "Big-data computing: creating revolutionary breakthroughs in commerce, science and society." It stated that like the search engines that have changed the availability of information, processing big data can change the way companies, research, medicine, and defense. Big data processing is probably the biggest advance in computing in recent years. decade. People have just begun to understand what potential big data has in all areas of life. Public investment can significantly accelerate research in this area (Bryant, Katz, Lazowska, 2008, pp. 1-8).

Views on big data in business were also taking shape. So, in May 2009, the research and consulting company Gartner published a document that predicted the growth of data stored by enterprises by 650% over the next 5 years. It also argued that if a way was found to extract useful knowledge from this data, it could lead to a revolution in business.

In 2011, the McKinsey Global Institute published a report that analyzed Big Data technologies. Big data is considered in three dimensions at once - the growth of volumes, the growth of the speed of data exchange and the increase in information diversity. According to the report, there are five main ways that leveraging big data can create value. First, big data can create significant value by making information transparent and usable at a much faster rate. Second, as organizations create and store more transactional data digitally, they can collect more accurate and detailed performance information from product inventories to sick days, and therefore demonstrate agility and improve performance. ness. Leading companies use data collection and analysis to conduct controlled experiments to make better management decisions; others use predictive data to make timely changes to their businesses. Third, big data provides tighter customer segmentation and, therefore, more accurate advertising offers. Fourth, big data analytics can dramatically improve decision making. Finally, big data can be used to improve the development of next generation products and services. For example, manufacturers use data from sensors embedded in products to create innovative after-sales service offerings, such as proactive service (preventative measures that prevent failures and breakdowns). It is also argued that big data will become a key foundation for competition, which will increase productivity and growth for individual firms. While the use of big data will have implications for different sectors of the economy, some industries are set to make big profits. IT, as well as finance and insurance are ready to significantly benefit from the use of big data (McKinsey, 2011).

In May 2012 Danah Boyd and Kate Crawford published the article "Critical Questions for Big Data". They define big data as a cultural, technological and scientific phenomenon based on the interaction of technology (increasing computing power and algorithmic accuracy for collecting, analyzing, linking, and comparing datasets) and analysis (using big data sets to identify laws for the purpose of presenting economic, social, technical and legal requirements) (Boyd, Crawford, 2012, pp. 662-679).

On October 6, 2015 it became known about the exclusion of information about big data from the Gartner "Hype Cycle 2015" report. The analysts of the company explained their decision by the fact that the concept of "big data" includes a large number of technologies that are actively used in enterprises, they partially relate to other popular areas and trends and have become an everyday working tool., And in science research related to big data dispersed within the framework of applied disciplines and areas (Stamford, Conn, 2015).

The amount of data processed globally is expected to grow exponentially. US technology company Cisco The number of devices connected to IP networks will be more than three times the global population by 2023. There will be 3.6 networked devices per capita by 2023, up from 2.4 networked devices per capita in 2018. There will be 29.3 billion networked devices by 2023, up from 18.4 billion in 2018 (Cisco, 2020). This volume is a result of the ubiquity of network and Internet transactions.

The OECD (2020) emphasizes that virtually all media and socio-economic transactions are moving to the Internet (including e-commerce and e-government); thus, petabytes of data are generated every second. Data growth is fueled by the actualization of Moore's Law, according to which even more powerful, smaller, more intelligent and less expensive devices have become available to almost any individual. In turn, this led to a reduction in the costs of collecting, processing and analyzing data. At the same time, access to data is facilitated by the spread of Internet platforms, e-commerce and the popularity of smartphones.

Stucke and Gruns point out that the speed of accessing, processing and analyzing data for some companies is approaching real time at the moment. The ability to use real-time data - a phenomenon known as "nowcasting" (Stucke, Grunes, 2016).

The concept behind this prediction is to look at a current event and use it to predict events as they happen, such as identifying an influenza outbreak due to the explosion of online flu drug searches. This approach can be used to discover potential competitors by determining the number of app store downloads and then comparing them with online usage or search preferences. The use of "nowcasting" can give an existing market participant leverage over new players.

Thus, a new distinction between Big Data and traditional data emerges: the meaning of time. The ability to process large amounts of data in real time is of inherent value, more important in some cases than the acquisition of data with a time interval, for example, when evaluating traffic information in road map applications.

The diversity of data has also increased due to the ability to collect and process, as a result, companies know not only the address of consumers (physical or IP), date of birth and gender, but also a lot of other information (family composition, eating habits, data on previous purchases, frequency and duration of visits to traditional and online stores), as well as information from other databases to form a dossier on the consumer. This will allow the retail business not only to differentiate prices, but also to target consumers with marketing and promotional materials to influence their behavior (Bańbura, Giannone, Modugno and Reichlin, 2022).

French and German antitrust authorities stress that changing consumer habits to maximize Internet use for everything from shopping to reading the news, watching movies and posting videos of themselves allows companies to record so accurately that it is possible to draw detailed and individualized conclusions about their susceptibility to commercial appeals. Flickering advertisements from the thriller "Minority Report" (2002), when a person's identity can be identified by a scan of the iris and then personalized advertising messages are urgently transmitted to him, is no longer a fantasy. This example illustrates the importance of data synthesis: big data sets merge together, information is extracted from them, and in the aggregate new information is formed, thanks to which sellers or competitors can better understand the market and work on it. Sometimes the potential for data synthesis can be developed further by combining personal data with other types of data (weather conditions, public events, inventories, or even vehicle component data collected for wear detection).

The value of Big Data is both a cause and a consequence of increasing volume, diversity and speed. While the data itself may be considered “free” (depending on the method of collection), the process of extracting information from the data generates value. The antitrust authorities of France and Germany unanimously point to the “development of new methods with which it is possible to extract valuable information from large arrays of (often unstructured) data. The OECD defines data analytics as "the technical means of generating analytical insights and empowering means to better understand, influence, or control knowledge-based information objects (e.g. natural phenomena, social systems, individuals) (OECD, 2020).

Stucke and Gruns (2016) emphasize that Big Data is closely related to what is called "Big analytics" and a phenomenon known as "deep learning technologies": computers learn to solve problems by compressing large databases by using advanced algorithms and neural networks that are increasingly reminiscent of human brain.

At this point in time, while the potential of big data may be large, the possibilities of big data for individual countries will depend on the characteristics of the country. The existence of social media, traditional businesses, administrative systems and the Internet of Things generating big data will vary. Therefore, assessing the power of big data and its potential policy application needs to take into account the availability of data and associated tools, user experience, privacy and security issues, and existing legal and technological systems.

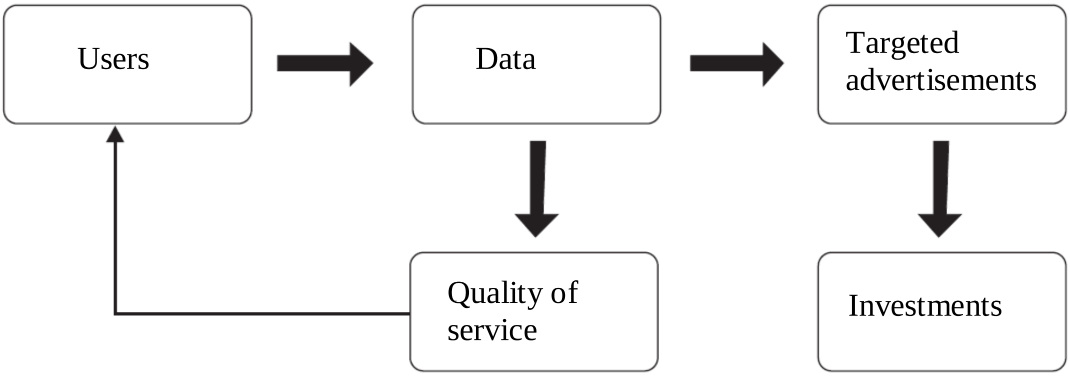
2.3 Potential competition issues associated with the use of Big Data

As the acquisition and use of Big Data becomes a key competitive parameter, companies will increasingly develop strategies to acquire and maintain an advantage in data. As Stucke and Ezrachi (2019) argue that, companies are increasingly adopting business models in which personal data is the leading input to production. Companies offer free services to individuals in order to obtain valuable personal data that will help advertisers better target ads to influence behavior. While competitive rivalry and incentives to maintain data advantage can be pro-competitive, bringing innovation to consumers and companies, some competition authorities emphasize that the network effects and economies of scale brought about by Big Data can also lead to market power and long-term competitive advantage.

Does the use of Big Data present a problem different from the use of conventional or traditional data? Local stores thrived on the knowledge of their customers. The traditional seller always builds close relationships with customers to know their preferences and offer products according to their requirements. Similarly, manufacturers use historical data to assess demand and improve products in highly competitive industries. Based on this, is it possible to argue that Big Data creates a new, previously unobserved problem for competition?

Unlike the traditional retail sector, modern business models are often characterized by data-driven network effects that can improve the quality of products or services. These data-driven network effects are the result of the two user feedback systems shown in Picture 6.

Picture 6. Feedback systems



Source: Stucke, Ezrachi, 2019

On one hand, having a wide user base, a company can collect more data to improve the quality of service (for example, creating better algorithms) and thus acquire new users - a user feedback system. On the other hand, companies can study user data to improve targeted advertising and service monetization, generating additional cash to invest in service quality and again attract more users - a feedback monetization system. Such endless systems can make it very difficult for any new market entrant to compete with "old-timers" with a wide customer base.

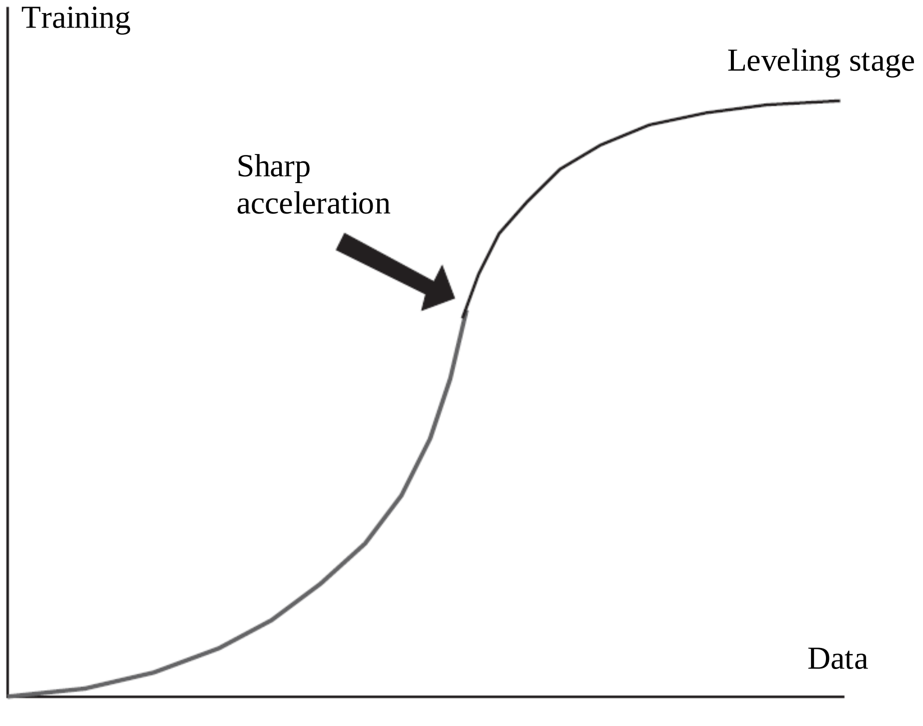
To illustrate, if a search engine only receives 1,000 searches per day, the algorithms have less data to learn from guided search results (other than more direct queries) and fewer related searches to offer to users. Low-quality search results are unlikely to attract large numbers of users from larger search engines. With fewer users, the search engine is of interest to fewer advertisers, which means fewer opportunities for users to switch to paid search results and, accordingly, low advertising revenue to expand the platform to other services.

With the acquisition of each user by a company compared to its competitors, a quality gap can arise. If users notice qualitative differences, the feedback system speeds up, attracting both new users and users of competitors' products. In markets where these network effects are driven (search engines, social networks, and community-supplied information management applications), the winner not only earns potential revenue, for example, when a user clicks on sponsored advertisements; user data also helps to improve the quality of the product, affecting the attractiveness of the product to future users and advertisers. Network effects like these can eventually narrow. However, data-driven network effects in interactive markets can amplify user acquisition and user loss.

As a result of network effects, users can become dependent on the dominant platform, even if they prefer a different platform model. For example, while Internet users may like the privacy options offered by some search engines, the major search engines provide more targeted results. Another example concerns a turn-by-turn navigation application, where a smaller application may have better performance, but the user is reluctant to use the dominant application due to the better traffic information provided by many users. The dominant platform may do nothing to be deemed anti-competitive, and yet the feedback system may reinforce dominance and prevent competitors from acquiring customers.

Another difference between modern Big Data applications and traditional business models concerns the lack of physical connection between the amount and variety of data that can be collected in the digital environment and the unlimited knowledge that can be obtained by running algorithms for extracting information from data on a range of data sets or using data synthesis.

Picture 7. Business learning curve



Source: Quan, 2022

As a result, Big Data has shifted the slope of the business learning curve (Picture 7), making the boost segment longer for a company in the Big Data market, and increasing data returns less depleted. When the big data market player finally enters the leveling stage, it has already reached such a large size that it will be difficult for any smaller player to implement competitive pressures, creating opportunities for market "bias" and winner-take-all outcomes.

Problems for competition can also arise from the fact that the cost structure of processing and using information is rather unusual, including high initial sunk costs and marginal costs tending to zero (Shapiro, Varian, 1999).

This is especially true for Big Data: information technologies for storing and processing data are very expensive. They cover large data centers, servers, data analysis software, Internet connection with advanced firewalls, highly paid workforce (computers and programmers). Once the system is fully operational, incremental data can train and improve algorithms at low cost (and thus also an element of product or service quality). This cost structure is characterized by significant economies of scale and diversification and can thus promote economic concentration if the Big Data market has a small number of participants.

Moreover, unlike "small data", where units of information provide meaningful and valuable human-understandable representations, there is little value in individual observation of Big Data. For example, data on a single click on a website is useless unless it is correlated with a billion other similar actions, which then need to be correlated with purchase decisions. To be profitable, Big Data arrays need to scale and are more often collected by the big players. Finally, other competitive challenges for competition arise from the specific structure of markets where Big Data transactions are typically supported.

Chapter 3. The concept of marketing attribution as a Big Data tool in marketing

3.1 Application of the big data technologies in marketing

The current conditions there is a segmentation of the Big Data technologies themselves and their industry specialization. From individual successful cases, the market is moving to the development of industry models for collecting, analyzing and using big data. The tools of work may be similar, but the boundaries of application, forms of organization and consequences of implementation become different. Marketing in this sense is moving from the use of big data to the transformation of marketing technologies under the influence of big data (Erevelles, Fukawa, Swayne, 2016, pp.897–904). Firstly, analytics comes to the fore, which for a long time in marketing was equal to statistical tools. Secondly, new marketing technologies are emerging at the intersection of big data and specific marketing tools (merchandising, advertising campaigns, marketing department management are changing).

There is a need to take a closer look at the problems in using big data in marketing. Firstly, there is the problem of scaling (Wright et al., 2019, pp.281–293). Big data is, as it was noted above, always a large amount of information that requires not only storage, but also constant access. Most corporate information centers were not designed for such volumes. Consequently, the company has to not only think about expanding its own corporate storage centers, but also look for ways to optimize, for example, use common storage and processing standards, transfer data to clouds.

Secondly, it is the integration of data collected earlier. For a marketer, it is important to have access to data on customers, campaigns, past marketing research. Without it, it is often impossible to build a trend, to understand the specifics of consumer behavior in the market. Storage systems for such data were never intended to be used in real time. All experts in the field of Big Data agree that Big Data technologies are meaningless if server systems cannot support real-time transactions (Kaisler et al., 2013).

Thirdly, data collection and processing systems in modern large companies resemble a real zoo. Organizations collected data for different purposes, in different ways, rarely integrated collection systems with each other. Nobody imagine that someday one will need to interact with completely unrelated systems and data stores, both inside and outside the enterprise, both for analysis and for visualization. Even when technology can provide integration and interoperability solutions, business owners are reluctant to relinquish control or require IT staff to prioritize projects based on current business interests.

Fourthly, Big Data technologies do not work without talented people (Harvard Business Review, 2020). When collecting and analyzing data, one need to ask the right questions. This is especially important in the world of big data, where there is a high probability that important data will not be received or interpreted correctly. Even well-endowed marketing departments will have a hard time buying up talented analysts from investment and financial firms. The specialists available on the market are often "sharpened" for IT projects and are not familiar with the philosophy and culture of marketing. It is important to understand that employers themselves are often not yet able to correctly formulate a request for the search for specialists in the field of Big Data and accurately assess the abilities and capabilities of existing candidates.

Fifth, there is the problem of developing a common language for discussing and working with data within the company. Businesses are used to the fact that IT specialists speak one language, marketers speak another. However, integration is vital for Big Data technologies. Consequently, companies will have to train a large number of employees in basic data skills. Only the joint work of all employees can give a cultural shift and force the entire company to use Big Data technologies correctly and effectively.

Also, for the qualitative implementation of Big Data technologies, it is necessary to constantly remember that some types of data, including financial and medical records, are subject to protection and significant regulation, which may vary depending on geography and jurisdiction (Kaisler et al., 2013). These rules may make it difficult or impossible to use some data. Companies are constantly looking for ways to overcome the problems described above and develop their own big data strategies. To overcome obstacles to the introduction of big data, they may use: interdepartmental working groups that bring together specialists from different areas who are able to work with big data; project teams or start-ups offering innovative big data tools; democratization of work with big data, i.e., the transition from complex processing systems to visualization or tools known to all such as Excel tables; new roles and statuses in companies, i.e. there are positions of directors of digital technologies or directors of marketing technologies.

The active implementation of Big Data in marketing gives companies a number of advantages (Lies, 2019, p.134) (Erevelles, Fukawa, Swayne, 2016, pp.897-904):

1. Creation of the most accurate portrait of the target consumer;

2. Predicting consumer reactions to marketing "messages" and offers one product or another;

3. Personalization of advertising messages;

4. Optimization of production and distribution strategies;

5. Creation of digital marketing and advertising campaigns;

6. Retaining a large number of customers through the least spending;

7. Getting a better idea of the company's own product, etc;

Following these advantages, the understanding of habitual marketing tools, for example, the marketing mix is enriched with new ideas:

– **Promotion**: Through data analysis, marketers can create an accurate portrait of a potential customer. Moreover, it is possible to predict the reaction of consumers to advertising.

– **Product**: Modern data processing tools can be used for product and market research. In addition, the manufacturer can view and analyze activities in the digital environment, which helps to improve the product according to the needs and desires of customers.

– **Place**: big data analysis allows to determine the most effective channels for advertising about products and the goods themselves. In particular, today it is more profitable to conduct sales online in some cases.

– **Cost**: Vendor data, financial statements, business models, etc. can be analyzed to correctly set the cost. If the target audience is quite large and “motley”, custom pricing can be used.

As it was mentioned above, the emergence of new data sources and new ways of analyzing opens up a lot of opportunities for marketers in many scopes demand forecasting, identifying consumer behavior patterns, segmenting them, developing marketing and, in particular, communication strategies. Comparing such factors as seasonality, geography of requests fora specific product, it is possible to predict seasonal demand and develop a plan for distributing the advertising budget for a specific period in a specific region (Lies, 2019, p.134).

When developing an advertising campaign, it is necessary to determine the target audience and develop the content of the advertising message. To identify the target audience, marketers can rely on both personal criteria and the advertiser's data on visitors their sites (promo site, corresponding pages of the main site). To search for a similar audience, predictive algorithms are used based on a learning sample, the so-called Look-alike Yandex launched this special type of Look-alike targeting back in 2013 (Ginny, 2021). Look-alike targets advertising to an audience that, in terms of its characteristics and behavior, is similar to the required, target audience. To understand similarity, there is a need to organize a large sample in order to more accurately characterize the behavior of the customers. Information is collected about those customers who have performed a certain action: downloaded a price list or catalog, followed a link, ordered and paid for goods, and so on.

For small volume companies sample information can be bought. For example, the Visual DNA and Weborama platforms provide various data, including user posts on social networks and their latest purchases. To increase the efficiency of Look-alike, it is useful to have at least tens of thousands of records and to have additional information providers who are really ready to share this information.

The Look-alike effect in social networks, as practice shows, will be much smaller due to the small audience coverage. To target the content of an advertising message to a specific user, personalization of email newsletters, product recommendations on the site, and dynamic retargeting are used. The essence of dynamic retargeting is that advertising is given only to those members of the community who have joined there recently. This is due to the peculiarity of the individual interests of the participants, which advertising takes into account. If a member leaves the community, then the continuation of advertising will lose its meaning.

The use of Big Data, the development of behavioral targeting led to the emergence of the RTB-auction technology - Real Time Bidding (Yang, Kang, 2017, pp.527–528). In accordance with this technology, the sale and purchase of advertising impressions takes place on the basis of an auction. The advertiser's website is tied to a specific RTB agency. The user enters the query of interest to him in the browser line. The site sends a request to view an advertisement and user data for targeting (gender, age, interests) to the RTB agency. The request is classified according to a number of parameters (data about the platform to which the entry was made, the time of entry, data about the client, etc.). Information about the client to whom the advertisement is shown is determined based on the data of the cookie and its Internet history. Advertisers participating in the auction receive a request to display ads and offer their bids. The right to display ads gets the one who offers the highest price. In this case, the winner pays not the maximum, but the second (of all those offered) the highest price. This is possible when bidders do not know each other's bids. RTB auction technology allows advertisers to control pricing and target ads only to interested consumers. This significantly distinguishes it from contextual and display advertising, when a certain number of impressions are sold.

The content of advertising, its color scheme is dynamically changing in accordance with the needs of the client, his demographic characteristics, and the level of loyalty. Regular customers can get more favorable conditions. RTB advertising is effective in the case of inexpensive brands and with a significant audience coverage. Experts do not recommend using online auction technology for prestigious brands, so as not to undermine their reputation.

Big data also determines one of the most important modern trends in marketing research. ISPs, mobile operators, marketing agencies, etc. have huge amounts of consumer data. Research companies have online panels with hundreds of thousands of respondents, and specialized agencies up to 600 thousand. Online research is most effective when testing advertising, products, price research. loyalty research. Data on consumer behavior on the Internet (search queries, pages visited, ads viewed), accumulated on the basis of online panels, is the basis for consumer segmentation and the development of both promotion and media strategies for the relevant sites on the Internet.

Thanks to the single source approach, it became possible to work with information flows from one source (from one respondent). Of considerable interest is data on the behavior and purchases of the same respondent on the Internet. To obtain such data, panel participants install specialized programs. To control the television viewing of the same participants, automated technical devices are used, the so-called man-counters or people meters. To study TV viewing data in general, the analysis of return path data of cable operators is used. Modern media measurement technology involves the integration of big data from digital tuners, Internet counters that control the behavior of Internet and TV users, as well as data from traditional sample surveys that allow determining the demographic data of these users.

The technology of single-source research is quite progressive, which makes it possible to measure the media activity of a person using a single sample (Assael, Poltrack, 1993, p.48). The same content, for example, a movie, can be seen not only on a computer screen, but also in an online cinema, the Internet, and other distribution channels - in applications on SMART TV, tablets and smartphones. Single-source research allows to track the diverse media consumption of sample participants.

An example of a successful single-source study of cross-platform TV viewing measurement is Nielsen in the USA. Not only TV viewing is measured, but also TV content viewing on various platforms and screens (DVR (digital video recorder), VOD services of pay TV operators, set-top boxes with Internet access, as well as computers, tablets and smartphones).

Big data is not a substitute for research, but it is the information basis for conducting deeper and more concise surveys. Social networks are sources of a large amount of marketing information and, above all, about users, including both their identification data and various additional data (interests, friends, shopping, etc.)

The development of big data technology has created new opportunities for integrating and enriching various information about social network users. Today, marketing pays special attention to personality recognition technologies that will allow more understand exactly what advertising and marketing materials the consumer sees and understands, as well as build the right campaign settings.

3.2 Customer journey in omnichannel marketing

Advertising campaigns capture the attention of customers through a variety of sources. The result of visiting a site or making a purchase may be due to a series of promotional redirects. Analyzing this information, it is necessary to find out the weight of each transition. This is one of the most significant tasks for online marketing today.

Despite the huge amount of information accumulated over time, modern solutions often take too simple an approach. For example, one popular approach is to assume that the very first ad shown determines the purchase decision.

According to foreign studies, the number of users who make purchases offline and online at the same time is steadily growing in the world every day. Those who buy only in online stores or only in offline sales points are becoming less and less. It is much more convenient for people to use several channels at once: offline stores, mobile applications, social networks, marketplaces.

To build effective communications with a client, a business needs to look for the most convenient and affordable ways. Omnichannel marketing helps with this because it can easily satisfy the user's need for product selection, payment advice, and product usage.

Omnichannel marketing is a type of marketing that involves the inseparable use of several communication channels to interact with the consumer.

As it was already mentioned, for most businesses, it is extremely unlikely that a user who enters a site will buy on their first visit. Usually, users come to the site several times from different sources before eventually making a purchase or submitting an application.

These sources may include (Tueanrat, Papagiannidis and Alamanos, 2021):

- organic search;
- contextual search advertising;
- advertising in social networks;
- partner networks;
- retargeting;
- mailing list;
- etc.

Omnichannel is, first of all, the consistency of the work of all channels of communication and influence on the consumer. Omnichannel allows to control every step of the client, which is very important for every business, as omnichannel customer is more likely to make purchases, show greater brand loyalty, and is also more likely to repeat purchases, because he is given more opportunities to buy in a way that is convenient for him, taking into account all the wishes and preferences (Mosquera, Olarte Pascual and Juaneda Ayensa, 2017, pp.92-114).

The more tools that affect such a consumer in a personalized way, the more he makes purchases. That is why using an omnichannel marketing strategy is an advantage for any business that wants to increase the number of loyal customers.

Due to the increase in the number of communication channels through which users receive information about the brand, it is becoming difficult for companies to maintain such a multi-channel communication with the client.

The classic sales funnel, based on two directions (traffic and outreach promo) and two KPIs (lead generation and brand awareness), as the basis of marketing, continues to be relevant and necessary for every business. But with the increase in the number of communication channels and the emergence of a variety of roles for each of them in the customer-sale chain, traditional models for evaluating each individual channel are becoming less and less effective. The role of the same channel can be different for two users.

Today, the user journey has become more confusing. It looks something like this: he googled, found a product, got to the site through advertising in the search, contacted the manager to clarify questions, and eventually took the purchase in an offline store. Tomorrow, the same or some other user went to the site from social networks, got acquainted with the range of goods, added the site to bookmarks and went about his business. And after a while, he returned from remarketing in the context and made a purchase. So, what channel led to the sale? It is clear that they all together influenced the user.

A modern client makes much more gestures and “descents” from the usual route, which, on the one hand, complicates the work of marketers, and on the other hand, shows a direct way to gain consumer confidence (Mosquera, Olarte Pascual and Juaneda Ayensa, 2017, pp.92-114).

Over time, a large number of different types of traffic sources and their sequences accumulate, which makes it difficult to analyze each of these sources in isolation from others. It is for this purpose that the rules were invented, according to which the value of the final conversion is distributed among all sources in certain proportions. And these rules are called attribution rules.

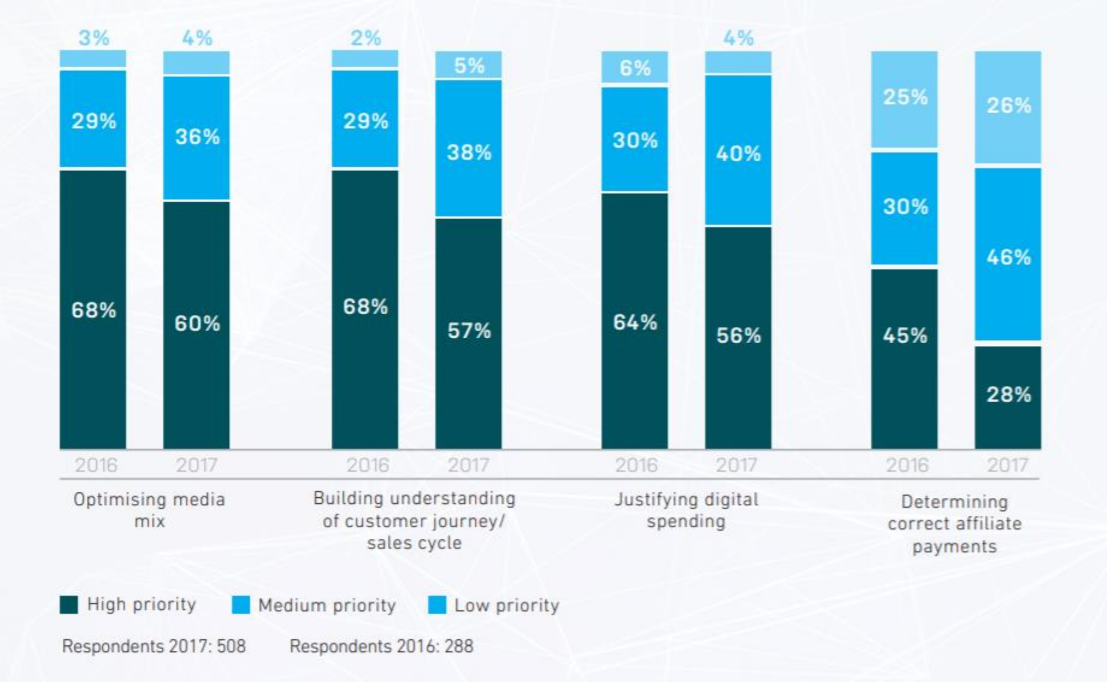
3.3 Attribution Models and it’s types

According to Gartner research, more than ⅔ of the advertising budget today falls on online channels (Pemberton, 2018). Moreover, as it was already mentioned in Chapter 1, in order to build an effective communication strategy on the Internet, it is extremely important to analyze the behavior of the target audience, taking into account the entire decision-making process. The paths along which the user passes, contacting with certain communication channels, form the so-called multi-channel sequences. At some point in the implementation of a communication strategy, a business may encounter a situation where the response to actions taken on the basis of standard analytics reports does not meet expectations. This kind of dissonance makes us think about the correctness of the distribution of priorities between the elements involved. Attribution models are used to solve such problems.

According to the Ad Roll Report (2017), 4 out of 5 companies in Europe, North America and Asia use attribution in their marketing, and 51% of organizations consider attribution the most important part of marketing.

Moreover, the report contains data on the distribution of goals for the use of marketing attribution. Picture 8 shows a number of goals, as well as changes in their priorities in one year.

Picture 8. Main goals for marketing attribution



Source: Adroll, 2017

Media content optimization is a top goal for 2017, with 60% citing this goal as "high priority" and another 36% as "medium priority".

Media content optimization outpaced building an understanding of the buy/sell cycle, which was also a top goal in 2016. Justifying digital spending ranks third in the hierarchical order, but differences in the degree of distribution in which companies prioritize these three main attribution goals remain at the same level. While budget optimization remains a key benefit, it is by no means the sole purpose of implementing attribution, as the survey data reflects.

Companies that benefit the most from attribution recognize its role in improving customer experiences. A more data-driven approach to understanding the journey to purchase can help marketers focus on delivering better messaging and content across touchpoints. Like last year, getting the right affiliate payouts is more of a secondary attribution goal, with only 28% of respondents citing it as a top priority, up from 45% last year.

Initially, the term attribution was used in Western social psychology, which refers to the logic of a person attributing some characteristics to other people (Yoo, 2013, , pp.298–300).

In the modern theory of Internet marketing, this term was introduced relatively recently and is characterized as assigning a certain significance to various communication channels in the chain that led to the conversion. In turn, an “attribution model” is nothing more than a rule or set of rules that determines how the value of a conversion is distributed among the touchpoints along the conversion path (Holmes, 2022).

In practice, attribution models are implemented as analytics tools designed to help quantify the contribution of each of the communication channels, based on specific data about site users and their conversions. Despite the unique opportunity to use standard models provided by modern automated analytics tools, the question of the appropriateness of using each of them, in different conditions, remains open and, entirely, rests with the opinion of the company's web analyst. In turn, the incorrect distribution of value between channels can affect the number of financial losses, as a result of excessive investments in an inefficient channel in the communication chain, or loss of sales or customer requests, as a result of incorrect identification of key channels or disconnection of auxiliary channels.

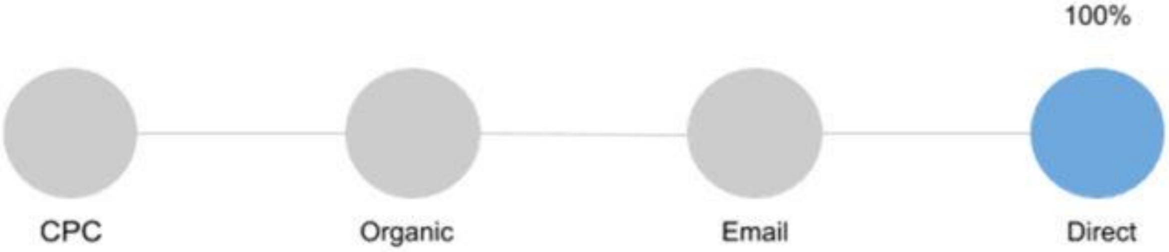
There are dozens of possible attribution models. They can be classified in different ways depending on what logic is used in the calculation. For example, if to look at what place the channel occupied in the chain before the order, then one is talking about the attribution model based on position (Time Decay, Position Based). If the calculation takes into account all the data, and not just the position of the channel in the chain, then these are algorithmic attribution models (Data-Driven, Markov Chains).

If to give all the value to only one channel that participated in the funnel, then such models are called single channel (Last Click, First Click). If the value is distributed among all channels in the chain, then these are multi-channel attribution models (Linear, Time Decay, Position-based).

Next, consider the main attribution models used in the market.

1. Last click attribution

Picture 9. Last click attribution model



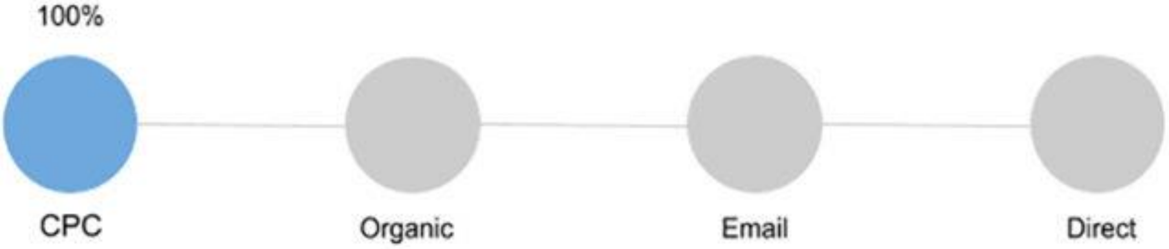
Source: Bocheva, Zhovtonizhko, Sheydayeva, 2021

Last click attribution is probably the most common and most criticized attribution model among all.

As can be seen from the figure above, this model is very simple and imprecise in nature. It assigns 100% value to the last click. If the last touch point was the result of a visitor visiting the site directly, then this model will ignore any effort put into social media, newsletters, etc.

2. First click attribution

Picture 10. First click attribution model



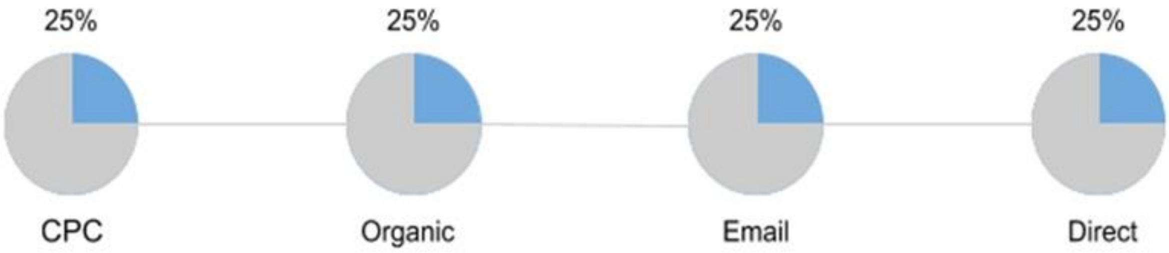
Source: Bocheva, Zhovtonizhko, Sheydayeva, 2021

This model is the opposite of the last click attribution model. She rates brand awareness efforts much higher than those that lead to specific user actions.

Of course, this model has the same problems as the previous ones. Assigning 100% value to one touch point is clearly wrong, unless the business has a very simple approach to marketing.

3. Linear attribution

Picture 11. Linear attribution model



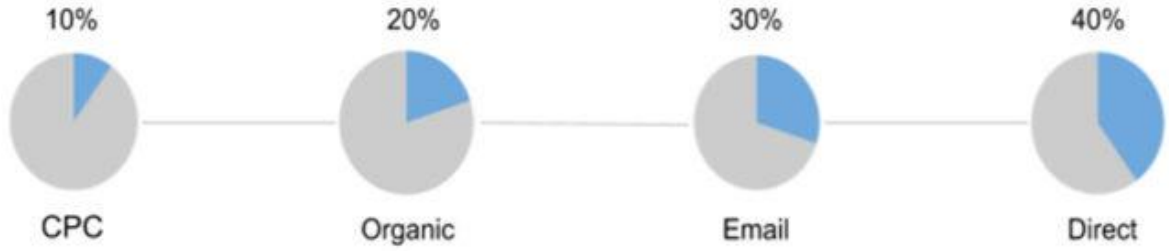
Source: Bocheva, Zhovtonizhko, Sheydayeva, 2021

The linear attribution model assigns the same value to each touch point.

There is no doubts that this is an extremely idealistic model. This is not to say that email marketing gives the same result as paid or organic traffic, as well as traffic from social networks.

4. Time decay attribution

Picture 12. Time decay attribution model

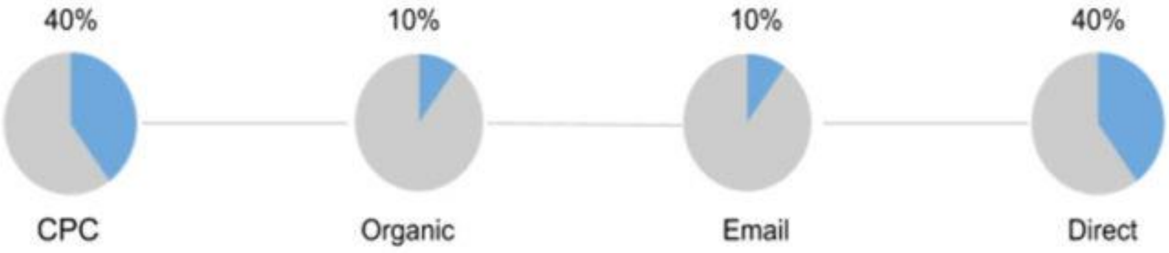


Source: Bocheva, Zhovtonizhko, Sheydayeva, 2021

The value from the transaction is distributed among the channels in increments. That is, the source that was first in the chain receives the least value, and the source that was the last and closest in time to the conversion receives the most value.

5. Position-based attribution

Picture 13. Position-based (U) attribution model



Source: Bocheva, Zhovtonizhko, Sheydayeva, 2021

This model assigns the highest value to the first and last touch points. While marketers can customize this model based on their own beliefs and an analytics data, the most common division is to assign 80% importance to the first and final touches and split the remaining 20% between the remaining touchpoints.

Having studied the standard attribution models, it is easy to assume that the situations in which they should be applied are quite specific. Thus, the last click model makes sense to apply in cases where decision-making stages are not provided. First click - if there is need to track the channels that arouse the interest of the target audience. If all links in the chain are relatively equal, a "linear" model may be used. The time decay model is applicable to analyze the effectiveness of short-term advertising campaigns or promotions, which allow to objectively evaluate the effectiveness of a particular message. In the case where the generation of interest and the final conversion are equally important, it is possible to use the model “based on position”.

Nevertheless, according to a 2017 Ad Roll study, 44% of marketers in the US and Europe use last-click attribution. At the same time, data-driven algorithmic attribution models are used by only 18% of marketers (Adroll, 2017).

72.4% of those who still use Last Click say that they do not know why they do it - it happened historically. When choosing a model, they chose the one that looks the simplest and most understandable, despite the fact that it underestimates all sessions in the chain, except for the last click (Adroll, 2017).

The development of technology has also affected marketing attribution, which has led to the emergence of analytical models. These are more advanced tools that take into account a lot of additional data in addition to the position of the channel in the chain when calculating.

Among the most interesting and productive algorithmic models are Data-Driven Attribution and Markov Chains. The first one determines the value of channels using user data and the Shelley Vector. The model does not take into account the position of the channel but makes a comprehensive assessment of how the presence of this channel influenced the conversion. It is suitable for users who need to know which campaigns are performing with the greatest efficiency, which allows them to further use the received data when planning their marketing budget.

Markov chains are an interesting tool with which one can evaluate the mutual impact of channels on sales, as well as determine which of them was the most important. It is a fairly well-known algorithm that has long been used in solving prognostic problems. Relatively recently, it began to be used in marketing. Models based on Markov chains make it possible to understand how the absence of a particular channel will affect the conversion.

Based on several researches, it can be concluded that using only attribution models it is impossible to accurately assess the impact on conversion, it is necessary to calculate several parameters. In addition, the Markov chain approach has proven to be an effective and illustrative method for estimating the likelihood of a conversion, so its use seems reasonable (Page, Brin, Motwani, Winograd, 1999) (Li et al., 2016, pp.831–848).

Chapter 4. Practical part – application of the methods

4.1. The description of the database and the technology used

The data was obtained from an attribution marketing company that chose not to identify itself and represents visits to a customer's website. The data has been anonymized for future use.

The received data, which are presented in the form of a table. The table contains columns: unique session number, source site, time of visit, unique user number, whether the visit is the first one (1 on the first visit, 0 on subsequent visits), the result of the visit (whether the call was made - 1 or not - 0). The table contains 456217 rows, 6 columns, in which 298854 are unique clients.

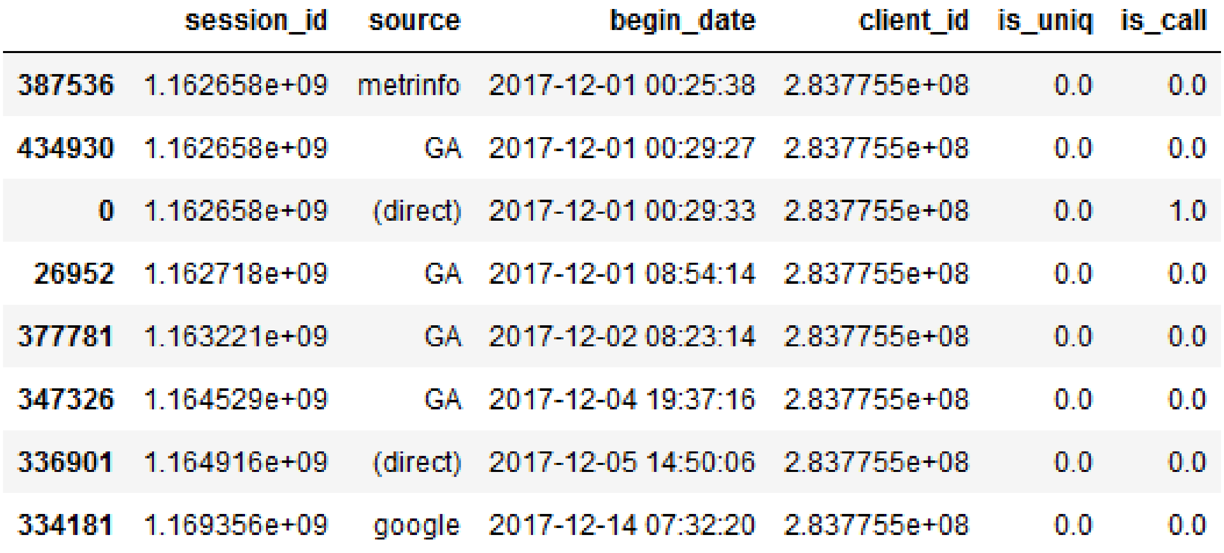
For processing, the data is grouped by a unique client ID and sorted by visit time, thus obtaining queues of visits - from the first visit to the last. If the user has performed an action, then his further visits are not tracked.

The next step is to build the transition matrix. The transition matrix is constructed as follows: the number of source-source and source-result transitions is counted, and the result is normalized. This matrix is necessary to represent the probabilities of transitions from one source to another, which is necessary for constructing Markov chains.

To work with sequences, they must be reduced to elementary sequences. An elementary sequence is such a chain, the end of which is represented by a conversion or the end of visits. Generalizing the rule for constructing elementary chains from the initial sequences of visits, it is necessary to note the main feature of the methodology for forming chains of user interaction with the site. It consists in the fact that any chain of interaction - a sequence of channels always ends with one of two events: 1 or 0. 1 at the end of the chain means a conversion, 0, in turn, means that there were no visits to the site after that. It is worth considering that event 0 is the last event in the chain and means the end of the sequence, that is, there is nothing after it, while event 1 occurs anywhere in the sequence.

Consider a typical situation with user visits to a site. The user sequentially visited the site redirected from different channels.

Picture. 14 Single user visits sorted by time

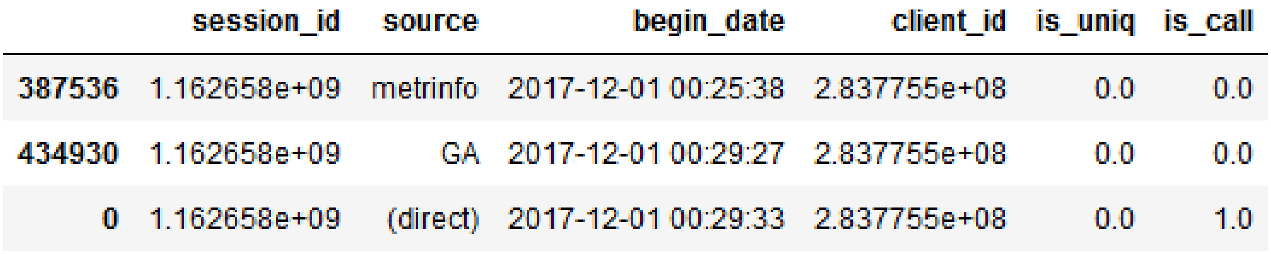


	session_id	source	begin_date	client_id	is_uniq	is_call
387536	1.162658e+09	metrinfo	2017-12-01 00:25:38	2.837755e+08	0.0	0.0
434930	1.162658e+09	GA	2017-12-01 00:29:27	2.837755e+08	0.0	0.0
0	1.162658e+09	(direct)	2017-12-01 00:29:33	2.837755e+08	0.0	1.0
26952	1.162718e+09	GA	2017-12-01 08:54:14	2.837755e+08	0.0	0.0
377781	1.163221e+09	GA	2017-12-02 08:23:14	2.837755e+08	0.0	0.0
347326	1.164529e+09	GA	2017-12-04 19:37:16	2.837755e+08	0.0	0.0
336901	1.164916e+09	(direct)	2017-12-05 14:50:06	2.837755e+08	0.0	0.0
334181	1.169356e+09	google	2017-12-14 07:32:20	2.837755e+08	0.0	0.0

Source: author's calculation based on source data

Picture 14 shows that the client visited several sources before the conversion occurred. Let's sort by the time it was visited and divide it into two elementary chains: before the conversion and after.

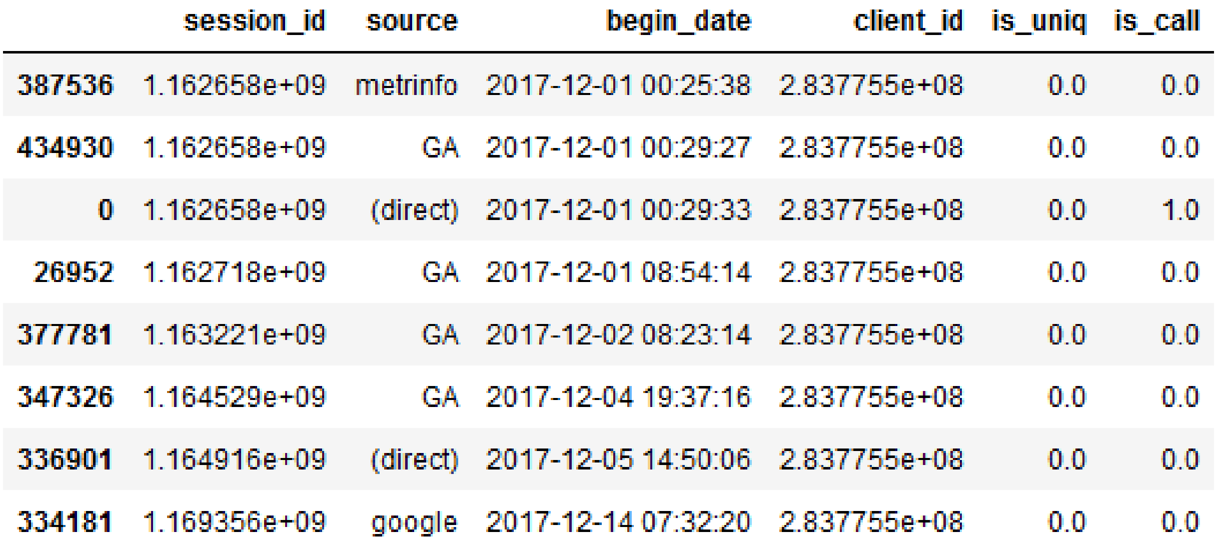
Picture. 15 First elementary string



	session_id	source	begin_date	client_id	is_uniq	is_call
387536	1.162658e+09	metrinfo	2017-12-01 00:25:38	2.837755e+08	0.0	0.0
434930	1.162658e+09	GA	2017-12-01 00:29:27	2.837755e+08	0.0	0.0
0	1.162658e+09	(direct)	2017-12-01 00:29:33	2.837755e+08	0.0	1.0

Source: author's calculation based on source data

Picture. 16 Second elementary chain



	session_id	source	begin_date	client_id	is_uniq	is_call
387536	1.162658e+09	metrinfo	2017-12-01 00:25:38	2.837755e+08	0.0	0.0
434930	1.162658e+09	GA	2017-12-01 00:29:27	2.837755e+08	0.0	0.0
0	1.162658e+09	(direct)	2017-12-01 00:29:33	2.837755e+08	0.0	1.0
26952	1.162718e+09	GA	2017-12-01 08:54:14	2.837755e+08	0.0	0.0
377781	1.163221e+09	GA	2017-12-02 08:23:14	2.837755e+08	0.0	0.0
347326	1.164529e+09	GA	2017-12-04 19:37:16	2.837755e+08	0.0	0.0
336901	1.164916e+09	(direct)	2017-12-05 14:50:06	2.837755e+08	0.0	0.0
334181	1.169356e+09	google	2017-12-14 07:32:20	2.837755e+08	0.0	0.0

Source: author’s calculation based on source data

As the result, there are two chains of visits Picture 15 and Picture 16 received, there is only interest in the sequence of channels and the result, so it is possible to get rid of the rest of the data. As a result, two elementary chains are formed:

1. metrinfo, GA, (direct), 1

2. metrinfo, GA, (direct), GA, GA, GA, (direct), google, 0

In the same way, the sequences of visits are reduced to a set of elementary chains, with which all subsequent calculations will be made. The head of the constructed chains looks like:

```
array(['GA', 'GA', 'GA', 'GA', 0], dtype=object),

array(['(direct)', '(direct)', '(direct)', '(direct)', '(direct)',

       '(direct)', '(direct)', '(direct)', '(direct)', '(direct)', 0],

      dtype=object),

array(['GA', 'GA', 'GA', 0], dtype=object),

array(['YD', 'YD', 'YD', 'YD', 'YD', 'YD', 'YD', 'YD', 'YD', 'YD', 0],
          dtype=object),

    array(['GA', 0], dtype=object),

    array(['GA', '(direct)', 0], dtype=object),

    array(['GA', 'GA', 0], dtype=object),

    array(['GA', 0], dtype=object),

    array(['YD', 'YD', 0], dtype=object),

    array(['YD', '(direct)', 0], dtype=object),

    array(['yandex', 0], dtype=object),

    array(['away.vk.com', 0], dtype=object),

    array(['GA', 0], dtype=object)
```

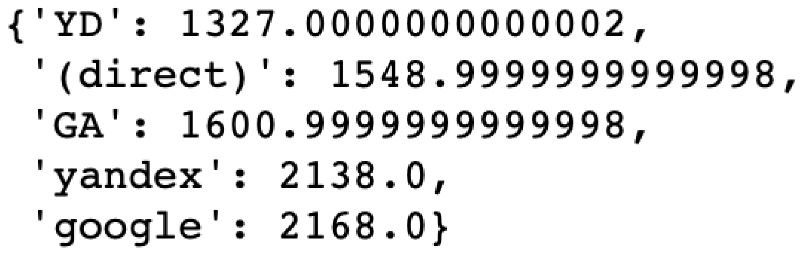
4.2. Calculation of the influence of channels on conversion

Consider a set of sequences that have been previously converted to an elementary form. Let's denote that X sequences end with 1 and - G with 0. Let's count the number of chains with and without conversion.

Obviously, the number of conversion chains is much less than the total number of chains. Indeed, X is equal to 2246 and G is equal to 298854, the ratio of the number of chains with conversion to the total number is 0.0074. Let's introduce the influence of the channel A_i on the overall conversion of the entire web resource as $I(A_i)$ for a period of time T, and R_j denote an elementary chain. The value of the influence of $I(A_i)$ in terms of a specific channel A_i will be estimated by using Formula 5.

The number of new conversions will be the reciprocal, since with an influence value of 1, the number of new conversions will be 0. Based on this fact, the significance score is inversely proportional, so there will be displayed the top 5 channels with the minimum number of new conversions:

Picture 17. Top 5 channels with the least number new conversions when deleted



```
{'YD': 1327.0000000000002,
 '(direct)': 1548.9999999999998,
 'GA': 1600.9999999999998,
 'yandex': 2138.0,
 'google': 2168.0}
```

Source: author’s calculation based on source data

The result is obvious, because the largest number of visits falls on the search results. Yandex, Google provide the lion's share of traffic on any resource. Interestingly, there was also a direct link here - these are direct transitions to the site without the participation of any third-party sites.

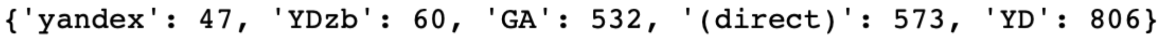
Most marketers don't like direct traffic because it doesn't show where it came from or what caused it. However, if you combine direct traffic data with multichannel GA funnel reports, you can see some insights.

For example, there is a recurring trend of growing direct traffic when running ads on Facebook. The hypothesis (supported by correlation analysis) confirmed that Facebook ads do increase direct traffic.

This may be because potential customers who see ads on Facebook are more likely to go to the site directly or through an untagged URL, or because session expiration resulted in more conversions than reported.

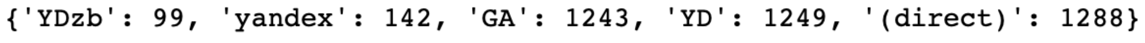
Let's calculate which channels will give the most conversions from attribution models.

Last click model:



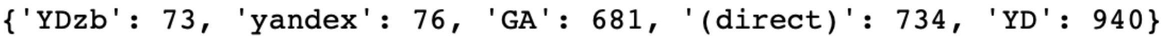
```
{'yandex': 47, 'YDzb': 60, 'GA': 532, '(direct)': 573, 'YD': 806}
```

Linear model:



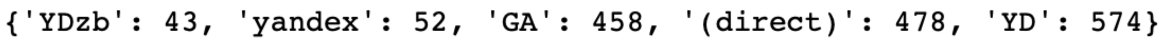
```
{'YDzb': 99, 'yandex': 142, 'GA': 1243, 'YD': 1249, '(direct)': 1288}
```

Time decay model:



```
{'YDzb': 73, 'yandex': 76, 'GA': 681, '(direct)': 734, 'YD': 940}
```

Positional model:



```
{'YDzb': 43, 'yandex': 52, 'GA': 458, '(direct)': 478, 'YD': 574}
```

Comparing the attribution models and the result, it can be noted that there is a general pattern: search results and direct took the first place in terms of the number of conversions, however, the contribution of each of them, depending on the chosen model, can vary greatly.

4.3. Estimating changes in base metrics when a channel is disabled

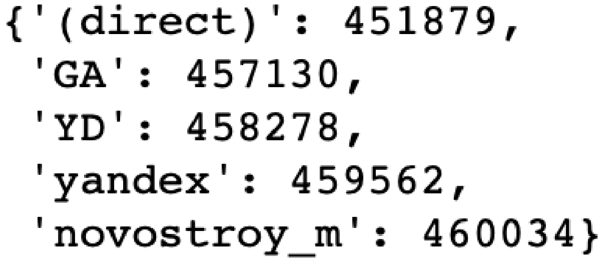
Having measured the value of new conversions, if to remove some A_i channel from all conversion chains, the question arises about the values of the most frequent metrics in advertising:

- consumption
- cost per conversion (CPA)

It is impossible to solve this problem without using additional tools, operating only with the available data, so it is needed to use some assumptions. The idea is that when a channel is removed from the R_j chain, it is interrupted. This means that if to delete the channel through which the user got to the site, then he will no longer go to it.

Of course, it is interesting to look at how the costs were distributed after the removal of a certain chain. The value is also inversely proportional. Below are presented not only the most expensive channels, but also the cheapest ones.

Picture 18. The most expensive channels

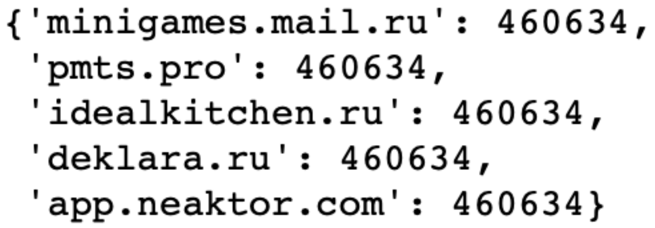


```
{'(direct)': 451879,
 'GA': 457130,
 'YD': 458278,
 'yandex': 459562,
 'novostroy_m': 460034}
```

Source: author’s calculation based on source data

Here the main trend remains the same, the channels with the highest coverage will be the most expensive, however, a new source appears, apparently a popular thematic site.

Picture 19. The cheapest channels



```
{'minigames.mail.ru': 460634,
 'pmts.pro': 460634,
 'idealkitchen.ru': 460634,
 'deklara.ru': 460634,
 'app.neaktor.com': 460634}
```

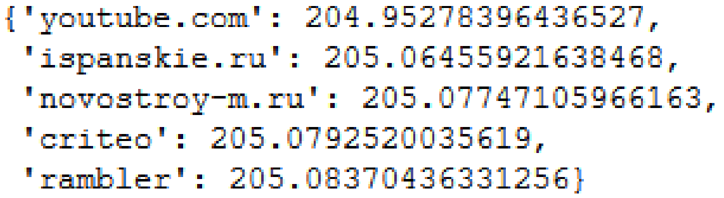
Source: author’s calculation based on source data

Increasingly, resources with low coverage will be at the bottom of the rankings.

After calculating the change in costs and identifying potential channels for removal after the removal of channel A_i , it is possible to move on to estimating the new cost of the conversion that occurs if there is no channel

Let's calculate it and identify the main trends in the cost of conversion per channel.

Picture 20. The cheapest channels in the cost per channel category

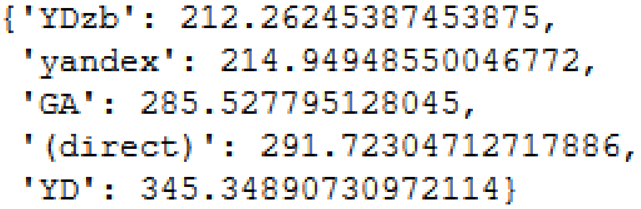


```
{'youtube.com': 204.95278396436527,
 'ispanskie.ru': 205.06455921638468,
 'novostroy-m.ru': 205.0774710596616З,
 'criteo': 205.0792520035619,
 'rambler': 205.08370436331256}
```

Source: author’s calculation based on source data

Perhaps such channels can be effective, but this requires additional calculations.

Picture 21. The most expensive channels in the cost per channel category



```
{'YDzb': 212.26245387453875,
 'yandex': 214.94948550046772,
 'GA': 285.527795128045,
 '(direct)': 291.72304712717886,
 'YD': 345.34890730972114}
```

Source: author’s calculation based on source data

The leaders are obvious here: search results provide the main flow of visitors to the site, and therefore the number of conversions per channel is the largest.

Assuming that before the channel was deleted, the conversion cost was equal to the following expression:

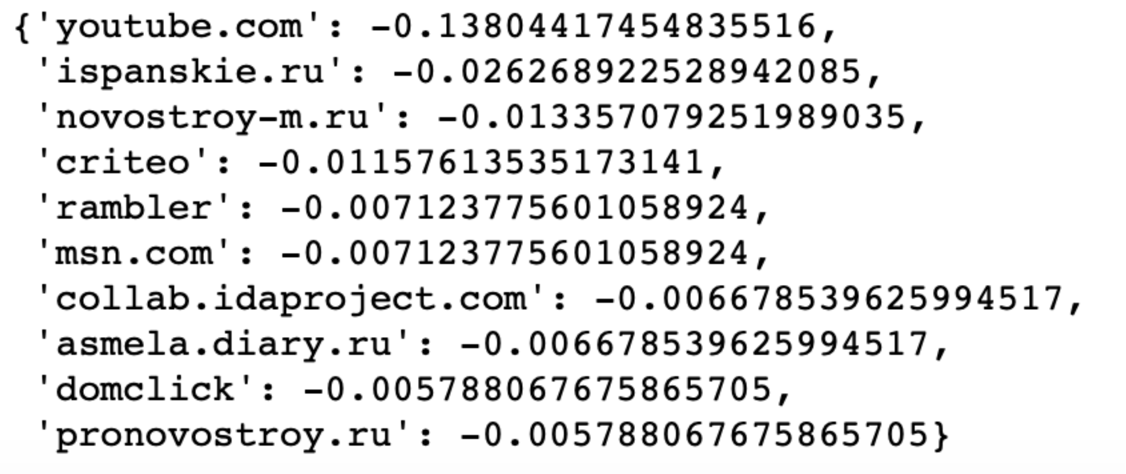
$$CPA_{new} = \frac{V_{old}}{X}$$

and using the operation of deleting a channel, it is necessary to reduce the cost of the conversion, then it is possible select channels according to the rule:

$$CPA_{new} - CPA_{old} < 0$$

Thus, if the difference between the new and old costs is less than zero, that is, the old cost is greater than the new one, then this channel can be removed and thereby reduce costs.

Picture 22. Cost reduction per channel



```
{'youtube.com': -0.13804417454835516,
 'ispanskie.ru': -0.026268922528942085,
 'novostroy-m.ru': -0.013357079251989035,
 'criteo': -0.01157613535173141,
 'rambler': -0.007123775601058924,
 'msn.com': -0.007123775601058924,
 'collab.idaproject.com': -0.006678539625994517,
 'asmela.diary.ru': -0.006678539625994517,
 'domclick': -0.005788067675865705,
 'pronovostroy.ru': -0.005788067675865705}
```

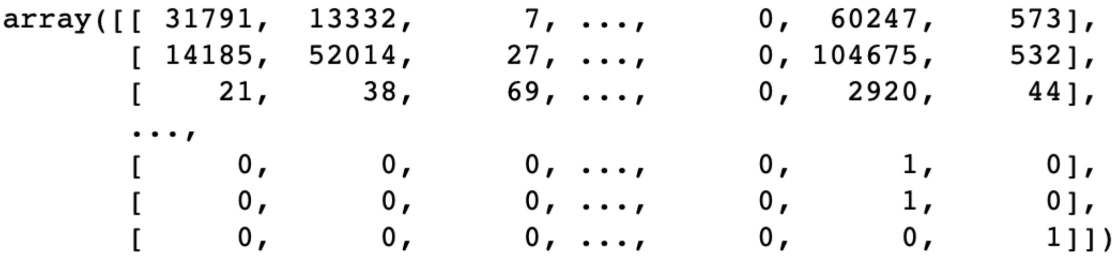
Source: author’s calculation based on source data

Picture 22 shows that deleting the youtube.com channel will make a significant contribution to reducing the cost per conversion. It should also be said that this calculation is made on the assumption that all channels have the same cost, which is fundamentally different from the real situation.

4.4. Total Probability Calculation

In the theoretical part, an efficient way to calculate the total probability in a matrix way was considered. To implement it, it is necessary to use chains, the receipt of which was considered earlier. Using chains, it is possible to count the number of transitions from one channel to another, going through each user session and adding up with the total number of transitions, it is possible to get a transition matrix V, size (n+2,n+2), where n is the number of unique channels plus two final states 0 and 1.

Picture 23. Transition matrix



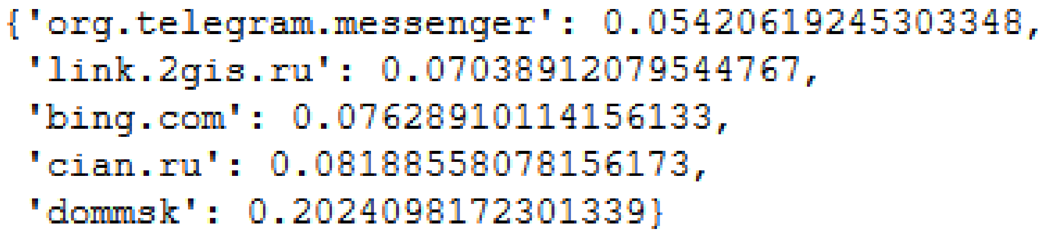
```
array([[ 31791,  13332,      7, ...,      0,  60247,    573],
       [ 14185,  52014,     27, ...,      0, 104675,    532],
       [    21,     38,     69, ...,      0,   2920,     44],
       ...,
       [     0,      0,      0, ...,      0,      1,      0],
       [     0,      0,      0, ...,      0,      1,      0],
       [     0,      0,      0, ...,      0,      0,      1]])
```

Source: author’s calculation based on source data

It is necessary to remove the diagonal elements from the resulting matrix by presenting transitions in chains of the form: $A_j, A_i, A_i, A_k \rightarrow A_j, A_i, A_k$, that is, interpreting them as one visit. The next step is to normalize the matrix row by row, this is necessary for the transition from a quantitative description of the transition to a probabilistic one. The transformation is performed in the following way: the probability of transition from the source A_i to A_j is equal to the value of the matrix $V(A_i, A_j)$ divided by the sum of the row $P(A_i, A_j) = \frac{V(A_i,A_j)}{\sum_j^{n+2} V_j}$. This formula looks quite obvious, the more transitions were made from one source to another, the more likely it is that the next time the transition will be the same. To calculate the total probability, it is necessary to raise the matrix to the power of two, let's raise the matrix to the power of 10.

Let's calculate the probabilities and sort them in descending order, thus ranking the channels.

Picture 24. Channels most likely to convert

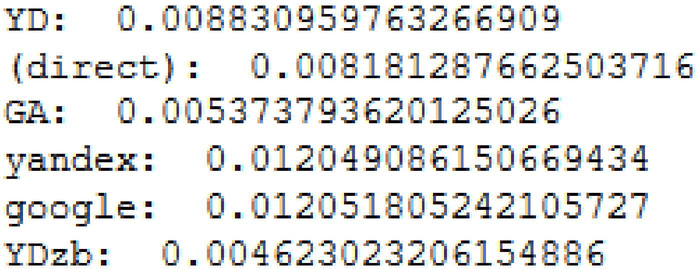


```
{'org.telegram.messenger': 0.05420619245303348,
 'link.2gis.ru': 0.07038912079544767,
 'bing.com': 0.07628910114156133,
 'cian.ru': 0.08188558078156173,
 'dommsk': 0.2024098172301339}
```

Source: author’s calculation based on source data

As it can be seen from Picture 24, the top channels by conversion probability include sites that were not in any of the lists of the most likely channels that were calculated using attribution models. Thus, it is possible to safely remove channels that not only reduce the total cost of conversion, but also get rid of some of the channels that are search results. Let's derive the total conversion probabilities of the channels that participate in the largest number of chains with conversion.

Picture 25. Full conversion rate for the most popular channels



```
YD:  0.008830959763266909
(direct):  0.008181287662503716
GA:  0.005373793620125026
yandex:  0.012049086150669434
google:  0.012051805242105727
YDzb:  0.004623023206154886
```

Source: author’s calculation based on source data

From the Picture 25, it is clear that the most popular channels do not have the highest conversion probability, which contradicts previous calculations. That is, using the formula for calculating the number of new conversions per channel and all attribution methods gave us the wrong answer.

4.5 Results

The analysis revealed significant differences between traditional attribution models and the model based on Markov chains.

To determine the conditions suitable for a particular channel importance distribution logic, a comparative analysis of the attribution models was performed.

Table 1. Comparison of attribution models

Model	Advantages	Disadvantages
Last - click	Accurately shows the visit that ended in a conversion	The contribution of other channels to interaction chains is not taken into account

First - click	Shows the channel that initiated interest in the product	The contribution of other channels to interaction chains is not taken into account
Linear	The contribution of all channels in the chain of interactions is taken into account	May underestimate more effective channels
Time decay	The contribution of all channels in the chain of interactions is taken into account; touches that lead to a conversion are considered more valuable	May underestimate more effective channels
Position based	The contribution of all channels in the chain of interactions is taken into account. The relationship of channels that initiated both interest and conversion is taken into account.	May underestimate more effective channels
Data based (Markov model)	More objective distribution of values. Every session counts.	It is a closed technology. High price. Not suitable for small companies and companies at the start, as it requires a large amount of data for analysis.

Source: author

However, it is worth remembering that even properly selected attribution models will not be able to give an objective picture of the world if there is no complete history of user interaction with the channels that participated in attracting him. There is no universal way to record the history of user interaction with advertising channels. The more methods are used, the more likely you will get an answer that is close to reality. With changes in data privacy (changes in how browsers handle cookies, IDFA and GAID restrictions), click-based attribution becomes less effective. All this requires the development of new skills, especially in relation to mobile analytics. Even with changes in the field of data privacy, it is possible and necessary to use a click as a way to record user interaction with advertising. The more fixation methods are used, the more complete will be the history of user interaction with advertising sources. Not all advertising channels lend themselves to fixation and evaluation at the level of an individual user, but this is not necessary. To solve the problem of determining the value of the contribution of a particular channel, it will often be enough to work with segments and make attribution based on them.

Conclusion

So, the concept of "big data" is complex. It includes large data arrays, ways of organizing and storing them, mathematical methods of analysis, revealing and formalizing hidden dependencies and patterns, as well as approaches to managing relevant processes at all stages - from data collection to the use of the obtained analytical materials.

Innovations related to the accumulation of big data, their effective use, are already becoming strategically important for the development of companies and increasing their competitiveness.

As it was analyzed in the scope of the work, in modern conditions, many enterprises are faced with the problem of increased market competition and the need to find new tools to promote their products and services. Today, the Internet often becomes such a tool, since it is a unique way of providing information that differs significantly from traditional media in a high level of flexibility and scale.

However, the trend of mass entry of thousands of companies into a single information space is gradually creating competitive conditions for promoting products on the Internet, in many respects comparable to off-line activities. If a few years ago the issue was the presence of the company on the network, which gave a huge competitive advantage, today, high competition forces companies to look for the most effective methods of brand communication with the consumer - the question of the quality of campaign planning arises. In this regard, the problem arises of a competent assessment of the contribution of specific communication channels to the results of the organization's integrated communication strategy on the Internet. One of the tools that can help to solve this problem is a marketing attribution.

As the conclusions of the study show, there are no universal ways to competently distribute the contribution of all channels in the interaction chain, within the framework of standard models. A smart solution may be to use own custom model that takes into account the specifics of the business.

Having clarified for the picture of the user's behavior and linking it to the most appropriate situation from those described above, you can choose the appropriate model for it as a basis. Adjustments to the model will be made in accordance with the statistical data for the site. In particular, to determine the model, indicators such as:

- Per-channel conversion rates of various stages of the funnel;

- Percentage of new users attracted by the channel;
- Bounce rates from channels;
- The number of associated conversions of each channel and its relation to the number of conversions from the channel.

So, a high rate of new users can, to a large extent, define a channel as initiating knowledge about a product or company. A low bounce rate, with a significant probability of passing through a number of stages of the funnel, may indicate the high importance of the channel as an auxiliary one. Quantifying channel conversions across the first, last, and intermediate interactions will help to more accurately allocate value between them, while the metrics described above will help show the importance of the channels more clearly.

Thus, knowing the specific data on the behavior of users on the site, as a result of interaction with one or another communication channel included in the unified communication strategy of the organization on the Internet, it seems possible to develop your own transparent attribution model that meets business objectives.

It should be noted the limitations of this study, firstly, the period of data on which the analysis was carried out was 2 weeks, since it was already indicated in the study that the average client journey fits into this period, but when considering a longer period, the results will be more accurate. Plus, the customer journey is different from the type of product, the customer journey for expensive purchases will be longer than for less expensive ones.

It is also worth remembering in general that at the moment attribution cannot work with offline conversions, which means that if a client went online, but then came to buy in a physical store, then his purchase will not be reflected in the analysis.

Further research can be directed to the research and development of methods for evaluating the effectiveness of other attribution models, such as Recurrent neural networks (RNN), as well as the development of management methods based on the attribution model, which would take into account the specifics of the influence of "big data" on decision-making processes at all management levels throughout the organization.

References

1. Abhishek, V., Fader, P. & Hosanagar, K., 2012. The Long Road to online conversion: A model of multi-channel attribution. *SSRN Electronic Journal*. doi 10.2139/ssrn.2158421

2. Adroll, 2017. The State of Marketing attribution 2017. Available at: https://www.adroll.com/assets/pdf/guides-and-reports/AdRoll-State-of-Marketing-Attribution-2017.pdf [Accessed January 25, 2022].

3. Al-Jarrah, O., Yoo, P., Muhaidat, S., Karagiannidis, G. and Taha, K., 2015. Efficient Machine Learning for Big Data: A Review. *Big Data Research*, 2(3), pp.87-93. doi: 10.1016/j.bdr.2015.04.001

4. Anderson, C., 2008. *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. [online] Wired. Available at: <https://www.wired.com/2008/06/pb-theory/> [Accessed 13 November 2021].

5. Anon, 2020. Digital transformation is about talent, not technology. *Harvard Business Review*. Available at: https://hbr.org/2020/05/digital-transformation-is-about-talent-not-technology [Accessed January 25, 2022].

6. Ashton K., 2009. That ‘‘Internet of Things’’ thing, RFiD Journal,53.

7. Assael, H. and Poltrack, D.F., 1993. Using single source data to select TV programs, *Journal of Advertising Research*, 33(1), pp. 48+.

8. Assunção, M., Calheiros, R., Bianchi, S., Netto, M. and Buyya, R., 2015. Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, 79-80, pp.3-15. doi: 10.1016/j.jpdc.2014.08.003

9. Banbura, M., Giannone D., Modugno M., and Reichlin, L., 2013. Now-Casting and the Real-Time Data Flow. *Working paper series 1564, European Central Bank, Frankfurt*. ISSN 1725-2806 (online)

10. Bańbura, M., Giannone, D., Modugno, M., Reichlin, L., 2022. *Now-Casting and the Real-Time Data Flow.*

11. Boyd, D. and Crawford, K., 2012. Critical Questions for Big Data. *Information, Communication & Society*, 15(5), pp.662-679. doi: 10.1080/1369118X.2012.678878

12. Bryant R., Katz R.H., Lazowska E.D., 2008, Big-data computing: creating revolutionary breakthroughs in commerce, science and society. Computing community,8, pp. 1-8.

13. Cox, M. and Ellsworth, D., 1997. *Application-controlled demand paging for out-of-core visualization*. doi: 10.1109/VISUAL.1997.663888

14. Das, T., Acharjya, D. and Patra, M., 2014. Opinion mining about a product by analyzing public tweets in Twitter. *International Conference on Computer Communication and Informatics.* doi: http://dx.doi.org/10.1109/ICCCI.2014.6921727

15. Diebold, F., 2003. Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting. *Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society*, pp.115-122.

16. Erevelles, S., Fukawa, N. & Swayne, L., 2016. Big Data Consumer Analytics and the transformation of marketing. *Journal of Business Research*, 69(2), pp.897–904. doi: 10.1016/j.jbusres.2015.07.001

17. Ezrachi, A., Stucke, M., 2019. *Virtual competition*. Cambridge, Massachusets: Harvard University Press. ISBN 9780674545472

18. Fewster, R., 2019. Chapter 8: Markov Chains. *Lecture notes for Stats 325*. Available at: https://www.stat.auckland.ac.nz/~fewster/325/notes/ch8.pdf [Accessed January 23, 2022].

19. Ginny Marvin, 2021. Yandex launches look-alike audience ad targeting for display campaigns. *MarTech*. Available at: https://martech.org/yandex-launches-look-alike-audience-ad-targeting-for-display-campaigns/ [Accessed January 25, 2022].

20. Goldin, G.A., 2014. Mathematical representations. *Encyclopedia of Mathematics Education*, pp.409–413. doi: 10.1007/978-94-007-4978-8_103

21. Hammer, C., Kostroch, D. and Quiros, G., 2017. Big Data: Potential, Challenges and Statistical Implications. *Staff Discussion Notes*, 17(06), p.1. doi: 10.5089/9781484310908.006

22. Hashem, I., Yaqoob, I., Anuar, N., Mokhtar, S., Gani, A. and Ullah Khan, S., 2014. *The rise of "big data" on cloud computing: Review and open research issues*, 47, pp. 98-115.

23. Holmes, K., 2022. What is attribution modelling, why does it matter and how to get started. *Ruler Analytics*. Available at: https://www.ruleranalytics.com/blog/click-attribution/attribution-modelling/ [Accessed January 25, 2022].

24. Hryniv, O., 2018. Markov chains: definition, Chapman-Kolmogorov eqn's, examples. *Durham University, Probability II (MATH 2647)*. Available at:

https://maths.durham.ac.uk/stats/courses/ProbMC2H/_files/handouts/1516MarkovChains2H.pdf [Accessed January 4, 2022].

25. Hryniv, T. & Rogozin, I., 2018. Lecture 6a: Introduction to Hidden Markov Models. Available at: https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/download/lectures/PCB_Lect06_HMM.pdf [Accessed December 12, 2021].

26. Huang, Z., 1997. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. *Cooperative Research Centre for Advanced Computational Systems CSIRO Mathematical and Information Sciences*, pp.1-8.

27. Cheng, S., Zhang, Q. and Qin, Q., 2016. Big data analytics with swarm intelligence. *Industrial Management & Data Systems*, 116(4), pp.646-666. Doi: 10.1108/imds-06-2015-0222

28. Chris Pemberton, C.P., 2018. 8 top findings in Gartner CMO Spend Survey 2018-19. *Gartner*. Available at: https://www.gartner.com/en/marketing/insights/articles/8-top-findings-in-gartner-cmo-spend-survey-2018-19 [Accessed January 25, 2022].

29. Izhikevich, E.M. & Kuramoto, Y., 2006. Weakly coupled oscillators. *Encyclopedia of Mathematical Physics*, pp.448–453. doi: 10.1016/B0-12-512666-2/00106-1

30. Kaisler, S. et al., 2013. Big data: Issues and challenges moving forward. *2013 46th Hawaii International Conference on System Sciences*. doi: 10.1109/HICSS.2013.645

31. Kumar, P., Das, T.K., 2013. BIG Data Analytics: A Framework for Unstructured Data Analysis. *International journal of engineering and technology, 5(1)*, 153-156.

32. Li, H. A. et al., 2016. Attribution strategies and return on keyword investment in paid search advertising. *Marketing Science*, 35(6), pp.831–848. doi: 10.1287/mksc.2016.0987

33. Li, H.(A. et al., 2016. Attribution strategies and return on keyword investment in paid search advertising. *Marketing Science*, 35(6), pp.831–848. doi: 10.1287/mksc.2016.0987

34. Lies, J., 2019. Marketing intelligence and big data: Digital Marketing techniques on their way to becoming social engineering techniques in marketing. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(5), p.134. doi: 10.9781/ijimai.2019.05.002

35. *Minority Report*. 2002. [DVD] Directed by S. Spielberg.

36. Mishra, N., Lin, C. and Chang, H., 2015. A Cognitive Adopted Framework for IoT Big-Data Management and Knowledge Discovery Prospective. *International Journal of Distributed Sensor Networks*, 66, pp. 1-13. doi: 10.1155/2015/718390

37. OECD, 2022. *Data-Driven Innovation*. doi: 10.1787/9789264229358-en

38. Page, L., Brin, S., Motwani, R., & Winograd, T., 1999. The PageRank Citation Ranking : Bringing Order to the Web. *WWW 1999*.

39. Peri, S., Sa, M., Singhal, Ni., 2016. Maintaining Acyclicity of Concurrent Graphs.

40. Pishro-Nik, H., 2014. Basic Concepts. In *Introduction to probability, statistics, and Random Processes*. Blue Bell, PA: Kappa Research, LLC, pp. 540–574. ISBN 978-0990637202

41. Quan, K., 2022. *Learning Curve: Definition, Theory (Graphs), and Examples*. [online] Getting People Right. Available at: <https://gettingpeopleright.com/resources/learning-curve/> [Accessed 12 March 2022].

42. S. Liu et al., 2020. Scalable Topological Data Analysis and Visualization for Evaluating Data-Driven Models in Scientific Applications, IEEE Transactions on Visualization and Computer Graphics, 26(1), pp. 291-300. doi: 10.1109/TVCG.2019.2934594.

43. SGI, 2021. [online] Static.usenix.org. Available at: <http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf> [Accessed 6 November 2021].

44. Shapiro C., Varian H., 1999. Information Rules: A Strategic Guide to the Network Economy. *Harvard Business Review Press. Boston.*

45. Stamford, Conn., 2015. *Gartner's 2015 Hype Cycle for Emerging Technologies Identifies the Computing Innovations That Organizations Should Monitor*. [online] Gartner. Available at: <https://www.gartner.com/en/newsroom/press-releases/2015-08-18-gartners-2015-hype-cycle-for-emerging-technologies-identifies-the-computing-innovations-that-organizations-should-monitor> [Accessed 13 November 2021].

46. Stucke, M. E., Grunes, A. P., 2016. Introduction: Big Data and Competition Policy. *Big Data and Competition Policy, Oxford University Press.*

47. Wang, L. and Shen, J., 2013. Bio-Inspired Cost-Effective Access to Big Data. *International Symposium for Next Generation Infrastructure, 43,* pp. 1-7. doi: http://dx.doi.org/10.14453/isngi2013.proc.42

48. Wayne, K. and Sedgewick, R., 2011. *Algorithms (4th Edition)*. 4th ed. Addison-Wesley Professional, p.430.

49. Wright, L.T. et al., 2019. Adoption of big data technology for innovation in B2B marketing. *Journal of Business-to-Business Marketing*, 26(3-4), pp.281–293. doi:10.1080/1051712X.2019.1611082

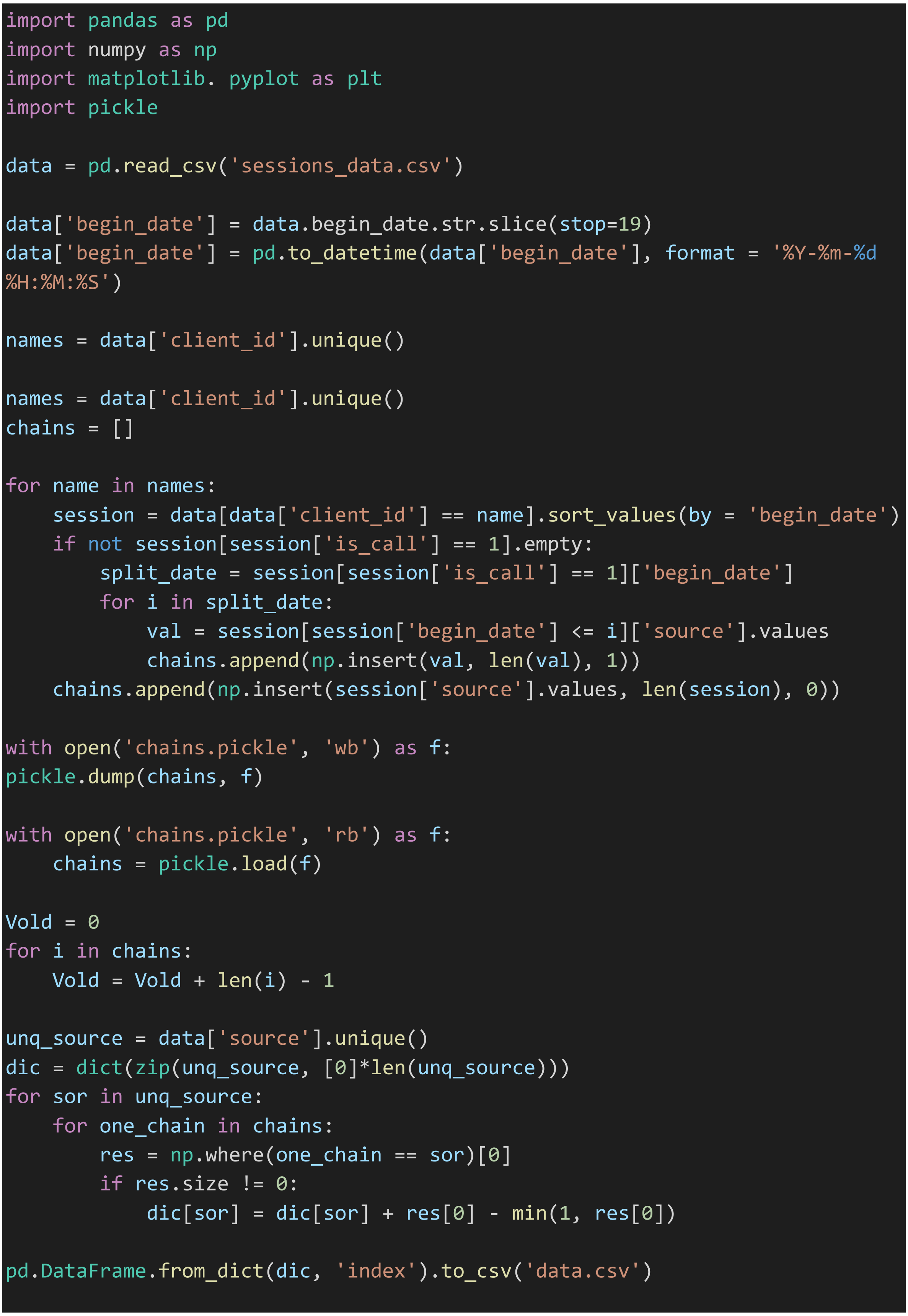
50. Yang, K.C. & Kang, Y., 2017. Big Data, consumer analytics, and real-time bidding (RTB) advertising: Emerging international policy and regulatory issues (an abstract). *Creating Marketing Magic and Innovative Future Marketing Trends*, pp.527–528. doi: 10.1007/978-3-319-45596-9_99

51. Yoo, C., Ramirez, L. and Liuzzi, J., 2014. Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine. *International Neurourology Journal*, 18(2), p.50. doi: 10.5213/inj.2014.18.2.50

52. Yoo, S.H., 2013. Cultural attribution fallacy. *The Encyclopedia of Cross-Cultural Psychology*, pp.298–300. doi: 10.1002/9781118339893.wbeccp122

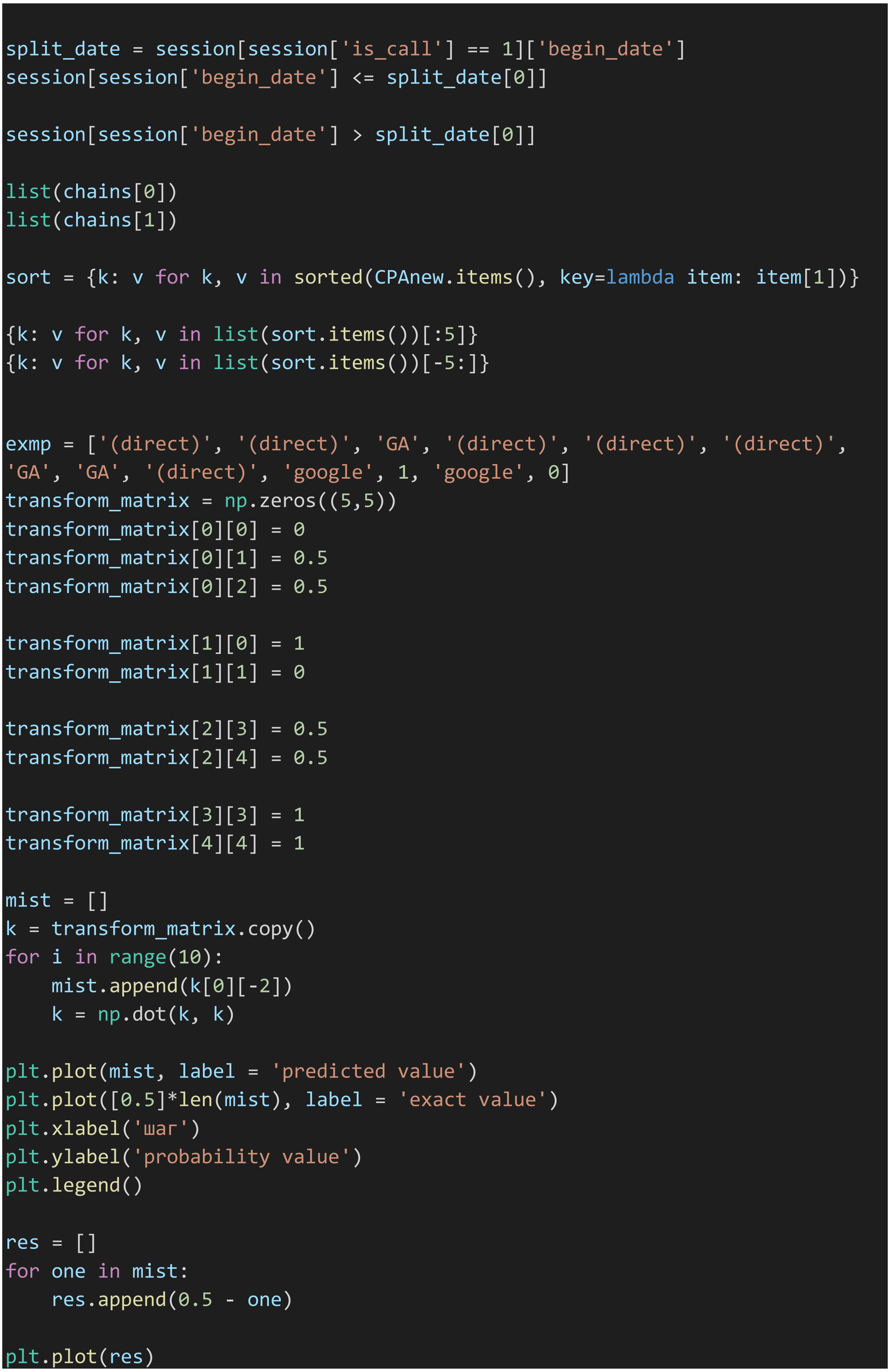
53. Zhu, H., Xu, Z. and Huang, Y., 2015. Research on the security technology of big data information. *Proceedings of the 4th International Conference on Information Technology and Management Innovation*, 55, pp.1041-1044. doi: 10.2991/icitmi-15.2015.174

Appendix A

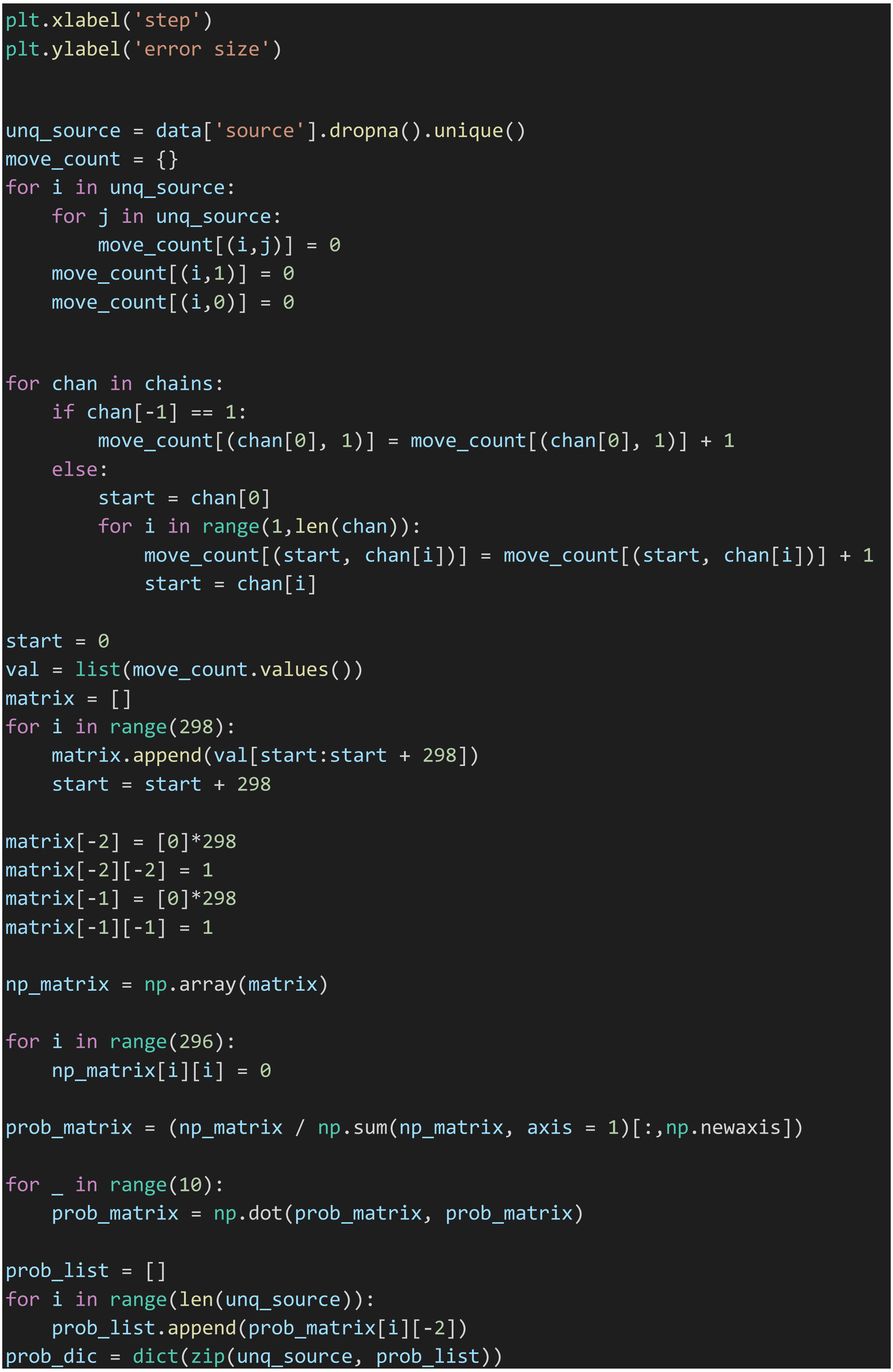


```
import pandas as pd
import numpy as np
import matplotlib. pyplot as plt
import pickle

data = pd.read_csv('sessions_data.csv')

data['begin_date'] = data.begin_date.str.slice(stop=19)
data['begin_date'] = pd.to_datetime(data['begin_date'], format = '%Y-%m-%d
%H:%M:%S')

names = data['client_id'].unique()

names = data['client_id'].unique()
chains = []

for name in names:
    session = data[data['client_id'] == name].sort_values(by = 'begin_date')
    if not session[session['is_call'] == 1].empty:
        split_date = session[session['is_call'] == 1]['begin_date']
        for i in split_date:
            val = session[session['begin_date'] <= i]['source'].values
            chains.append(np.insert(val, len(val), 1))
    chains.append(np.insert(session['source'].values, len(session), 0))

with open('chains.pickle', 'wb') as f:
pickle.dump(chains, f)

with open('chains.pickle', 'rb') as f:
    chains = pickle.load(f)

Vold = 0
for i in chains:
    Vold = Vold + len(i) - 1

unq_source = data['source'].unique()
dic = dict(zip(unq_source, [0]*len(unq_source)))
for sor in unq_source:
    for one_chain in chains:
        res = np.where(one_chain == sor)[0]
        if res.size != 0:
            dic[sor] = dic[sor] + res[0] - min(1, res[0])

pd.DataFrame.from_dict(dic, 'index').to_csv('data.csv')
```

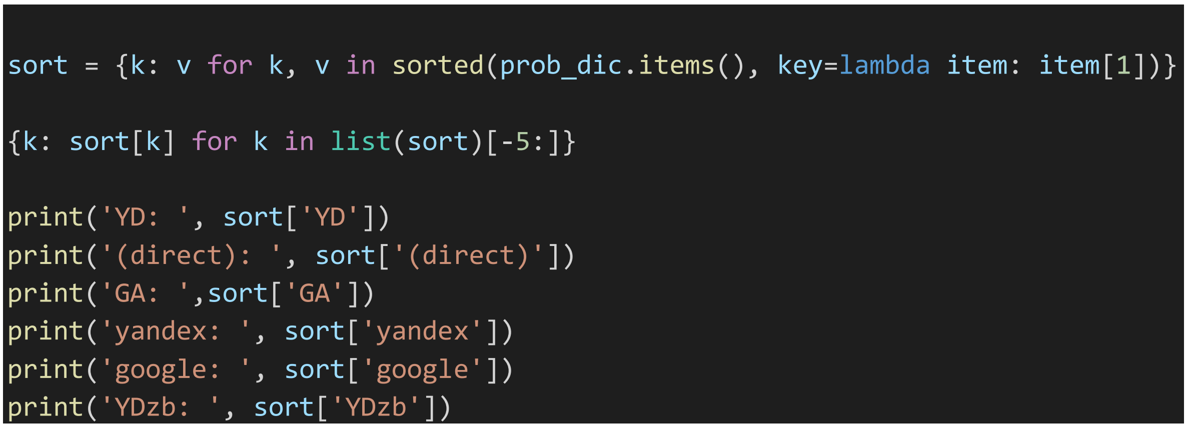
```
Vloss = pd.read_csv('data.csv')

Vlos = dict(zip(Vloss.iloc[:,0].values, Vloss.iloc[:,1].values))

unq_source = data['source'].unique()
Vnew = dict(zip(unq_source, [0]*len(unq_source)))
for sor in unq_source:
    Vnew[sor] = Vold - Vlos[sor]

with open('Vnew.pickle', 'wb') as f:
    pickle.dump(Vnew, f)

with open('Vnew.pickle', 'rb') as f:
    Vnew = pickle.load(f)

X = 0
for chan in chains:
    if chan[-1] == 1:
        X = X + 1

CVnew = dict(zip(unq_source, [0]*len(unq_source)))
for sor in unq_source:
    cur_sum = 0
    for chan in chains:
        if (chan[-1] == 1) & (sor in chan):
            cur_sum = cur_sum + 1
    CVnew[sor] = X*(1 - cur_sum/X)

with open('CVnew.pickle', 'wb') as f:
    pickle.dump(CVnew, f)

with open('CVnew.pickle', 'rb') as f:
    CVnew = pickle.load(f)

unq_source = data['source'].dropna().unique()
CPAnew = dict(zip(unq_source, [0]*len(unq_source)))
for sor in unq_source:
    CPAnew[sor] = Vnew[sor]/CVnew[sor]

pd.DataFrame(list(zip(list(CPAnew.values()), list(Vnew.values()),
list(CVnew.values()))),
             columns = ['CPA', 'Vnew', 'CVnew'], index =
unq_source).to_csv('newLabels.csv')

unq_user = data['client_id'].unique()

session = data[data['client_id'] == unq_user[0]].sort_values(by =
'begin_date')
```



```
split_date = session[session['is_call'] == 1]['begin_date']
session[session['begin_date'] <= split_date[0]]

session[session['begin_date'] > split_date[0]]

list(chains[0])
list(chains[1])

sort = {k: v for k, v in sorted(CPAnew.items(), key=lambda item: item[1])}

{k: v for k, v in list(sort.items())[:5]}
{k: v for k, v in list(sort.items())[-5:]}


exmp = ['(direct)', '(direct)', 'GA', '(direct)', '(direct)', '(direct)',
'GA', 'GA', '(direct)', 'google', 1, 'google', 0]
transform_matrix = np.zeros((5,5))
transform_matrix[0][0] = 0
transform_matrix[0][1] = 0.5
transform_matrix[0][2] = 0.5

transform_matrix[1][0] = 1
transform_matrix[1][1] = 0

transform_matrix[2][3] = 0.5
transform_matrix[2][4] = 0.5

transform_matrix[3][3] = 1
transform_matrix[4][4] = 1

mist = []
k = transform_matrix.copy()
for i in range(10):
    mist.append(k[0][-2])
    k = np.dot(k, k)

plt.plot(mist, label = 'predicted value')
plt.plot([0.5]*len(mist), label = 'exact value')
plt.xlabel('шаг')
plt.ylabel('probability value')
plt.legend()

res = []
for one in mist:
    res.append(0.5 - one)

plt.plot(res)
```



```python
plt.xlabel('step')
plt.ylabel('error size')


unq_source = data['source'].dropna().unique()
move_count = {}
for i in unq_source:
    for j in unq_source:
        move_count[(i,j)] = 0
    move_count[(i,1)] = 0
    move_count[(i,0)] = 0


for chan in chains:
    if chan[-1] == 1:
        move_count[(chan[0], 1)] = move_count[(chan[0], 1)] + 1
    else:
        start = chan[0]
        for i in range(1,len(chan)):
            move_count[(start, chan[i])] = move_count[(start, chan[i])] + 1
            start = chan[i]

start = 0
val = list(move_count.values())
matrix = []
for i in range(298):
    matrix.append(val[start:start + 298])
    start = start + 298

matrix[-2] = [0]*298
matrix[-2][-2] = 1
matrix[-1] = [0]*298
matrix[-1][-1] = 1

np_matrix = np.array(matrix)

for i in range(296):
    np_matrix[i][i] = 0

prob_matrix = (np_matrix / np.sum(np_matrix, axis = 1)[:,np.newaxis])

for _ in range(10):
    prob_matrix = np.dot(prob_matrix, prob_matrix)

prob_list = []
for i in range(len(unq_source)):
    prob_list.append(prob_matrix[i][-2])
prob_dic = dict(zip(unq_source, prob_list))
```



```
sort = {k: v for k, v in sorted(prob_dic.items(), key=lambda item: item[1])}

{k: sort[k] for k in list(sort)[-5:]}

print('YD: ', sort['YD'])
print('(direct): ', sort['(direct)'])
print('GA: ',sort['GA'])
print('yandex: ', sort['yandex'])
print('google: ', sort['google'])
print('YDzb: ', sort['YDzb'])
```