# Probabilistic modeling

**Keywords or keyphrases**:

"probabilistic models come", "today probabilistic models", "model different probabilistic", "basic probabilistic models", "probabilistic model different", "probabilistic modeling learn", "probabilistic models exists", "talking probabilistic modeling", "probabilistic model model", "types probabilistic models", "probabilistic models", "probabilistic models wanted", "create probabilistic model"

**Topics**:

"random probability distribution", "random distribution probability", "distribution random probability", "random distribution could", "probability random probabilistic", "random distribution probabilistic"

**Summary**:

We will talk about generative models which help to synthesize data with the same characteristics as original one. Based on observations, we can describe the real objects by some mathematical abstraction. This abstraction may rarely meet the real object and may differ from it with one or another degree of accuracy. robability measure, natural, axiomatic and unnatural. Natural is based in a classical way of probability as a frequency. Axiomatic allows introducing a measure based on mathematical apparatus. Perverted in all other cases when you really want to quantify uncertainty but are not clear how to do it exactly. In order to describe such a system, we need to have an endless number of characteristics. Each of them is a multi-dimensional function. The easiest way is to work with a random event. And it has only one characteristic of probability. But unfortunately, very often, we do not have enough knowledge on distribution form. In the case of ideal situation, we can find model distribution for our task. In other cases, we need to apply converted distributions, which means to transform some models so that we can work with them with our data. First kind of these distributions is truncated distribution. Another mechanism is suing of distributions. Generative approach: Take some data, build the model and create as many synthetic data as you want. If we don't have a model but have data, then we need to build a model on this data, or work with sampling to create a synthetic data. Simulation of continuous variable could be done in different ways. ion density with a rectangle and drop random points, which are characterized by random numbers gamma 1 and gamma 2. The more complex the shape of the distribution function, the lower the efficiency. There are specialized algorithms for converting tails based on the box-muller theorem. The models discussed above are for time series, means discrete time processes. If we want to describe a random function with continuous time, we can use models of the deterministic function of a random argument. For example, for Gaussian or Normal, and for exotic logs, peculiar and autoregressive models are needed.

**Extracted text**:

Good afternoon. We present the next lecture from the methodology of translational research and we'll talk about probabilistic modeling. You will learn today where do probabilistic models come from? What kind of basic probabilistic models exists and what could be created based on these models? How to work with these models easy and effectively? We will talk about generative models which help to synthesize data with the same characteristics as original one. While talking about the probabilistic modeling, we should not forget that it is just one of the chapters of mathematical modeling. Based on observations, we can describe the real objects by some mathematical abstraction. Unfortunately, this abstraction may rarely meet the real object and may differ from it with one or another degree of accuracy. You may think, what is the problem? If we have data, we can refine the model and get something more relevant. However, if this data differs from each other by size, direction or even a type, another question is raised. How can we describe these differences and consider them while building our model? In this situation, people would say about the problem of uncertainty. A uncertainty is a factor that does not allow definite determinations of the object's characteristics due to limited data and knowledge of reliability of information about it. More often, people face with epistemological uncertainty. For example, we know an exact value but we could not calculate or measure it. In this case, we talk about an error between calculated and real value, that this is mostly the topic of metrology. Another type of uncertainty in particular, aleratory uncertainty, describes object and systems for which the word the exact value does not exist at all. In that case, we see a task of description of this variability of this data. For example, by patterns, intervals, spaces, fractions, etc. Probabilistic uncertainty may be related to different phenomena. In micro scales, for example, we can talk about Heisenberg-Conser-Tenty principle, when the object could not be referred to the fixed point in the space of time. It exists there with some degree of probability. Lager scale is described by random effects related to a thermal motion of molecules and diffusion processes. Turbulence can also be a source of randomness or wave processes in continuous medium. If we talk about living beings, then biological mutation can also be a source of the randomness or different behavior aspects in relation to the social norms and regulations. We may also find sources of randomness in digital systems, for example in informational entropy. While speaking about the sources of uncertainty, you should remember that very often, emphasized and sporadic things can mimicry to uncertainty. Improcized is connected to the situation when we could not introduce any probability. For example, many. What is it? Something which is more than 5, 8 or several from air 3 to 8? By this way, we can describe improcized things. Sparadic is another thing. It refers to even quite rare, we even could not estimate the sources of their appearance. They could be completely determined, however, we could not say when and why they will appear again. We try to apply to them probabilistic description anyway, even though there is no probability here. How to digitize uncertainty? Or in another words, how to create probabilistic model? If we have some abstract object, then we need to determine probabilistic space which could be described by three values, i, b and p. i determines space of elementary events which is observed with a very simple experiment with the coin. Talehead Edge. Next component is b, Boral Sigma algebra. Quite complex construction, which is very important in understanding of how from elementary events more complex could be created. The third component is a probability measure. If we determine these three components, it is in the bag. We have described our model. What do we need to understand next? Is how to create such descriptions? There are three different ways of the imposition of probability measure, natural, axiomatic and unnatural. Natural is based in a classical way of probability as a frequency. When we have observable repetitive phenomenon which leads to different results under identical conditions, the law of large numbers. Axiomatic allows introducing a measure based on mathematical apparatus which is an experiment is not possible. Choice of the separators is determined by nature of the phenomena or by the research. Perverted in all other cases when you really want to quantify uncertainty but is not clear how to do it exactly. Measure p is assigned by analogy with other probabilistic objects. How to describe a probabilistic model? Let's start from the simplest model, model of the random event. Imagine that you ask someone, what is the probability of meeting a dinosaur out in the street? One say, probably meet or probably not. There are two ends with probability of 50 to 50%. Another say, it is impossible and the event A will have probability 0%. Complimentary event, if we do not meet a dinosaur out in the street, probability is equal to 1. Even handling rules, addition or multiplication are determined by sigma algebra. These things are well known to us from classical course of probabilistic theory, mathematics and our real life. As we have said before, natural definition of the random event could be given through the frequency of outcomes. That means that when we have an occurrences of random event, heads, tails, edge, we can describe this frequency as a ratio of successfully realizations to N and we get exact assessment of probability. Good news, always with the increase of number experience, this assessment close to the real probability. However, we have a bad news as well. The convergence rate is quite limited and inversely proportional to the root of N. Model of random event is the simplest one. More complex is a model of random variable. When we talk not only about the probability of the event, but about the probability P to fall within the concrete integral AB, which could be continuous or discrete. For these variables, the main characteristic is the distribution. For discrete variables, we talk about the discrete distribution. For continuous variable, we talk about the cumulative distribution function and survival function, which means 1-distribution function. Why does it important? Sometimes, especially when we consider extreme events, it could be more interesting to work not with X, which is less than some value, but with the inverse of X, which is more than some value. Both distribution function and survival function are monotonic functions and the density of distribution is a positive value. Natural definition of the random variable could be given through the frequency of events, which fall within concrete integral. We talk about the value of the distribution function as a certain probability and evaluate it in the same way through the sample. But in this situation, instead of the probability of an event, we can see that the probability of falling within the interval. Now, we'll talk about the model of the multivariate random variable. We see two-dimensional model, which means that probability falls not into a specific interval, but within concrete space. The shape of the space could also be different. In this case, basic and comprehensive characteristic is joined distribution function. For continuous variables, we can use joined probability density function. However, we should remember that any two-dimensional distributions can create two one-dimensional distribution, which are called marginal. It is hard to say something new while talking about the natural definition of distribution density. This is the same idea as for one-dimensional distribution. However, we talk here about not falling into the interval, but about falling into space, for example, into the rectangle. And here, we deal with not histogram, but 3D-bar diagram. Unfortunately, in that case, we should mention the curse of dimensionality. Imagine that we have 100 values and histogram with 10 columns, 10 values for each column. If we have 100 values and 3D diagram with 10 by 10 columns, which means 100 columns and 1 value for a column. Probability for a single value could not be calculated. So, in case we want estimate multi-dimensional objects, our sample must be very big. Let's consider now a model of random function process or time series. Here, we talk about some characteristic, which behaves as one-dimensional random value during each moment of time t. Which models could be described depending on t? Random sequence. When instead of time, we have just a number of steps. Another one is a stochastic process, when t is a subset of continuous timeline. Time series, which is count down on the timeline. Please, be careful as a random sequence could not be confused with time series. With time series, interval could be different, while with a random sequence, we have only a number and no interval at all. If we talk about comprehensive characteristics of such kind of process, then we have here a family of distributions. One-dimensional, two-dimensional, three-dimensional, etc. And this is also a very bad news. In order to describe such a system, we need to have an endless number of characteristics. Each of them is a multi-dimensional function. What can we do in such situations? We can try to simplify a model. Solomon said, the thing that has been, it is that which has shall be. And that which is done is that which shall be done. And there is no new thing under the sun. In fact, he was the first one who introduced possibility to determine process probability characteristic based on their trajectories. For example, we observe all possible varieties of random processes development. We collect the processes and build on their basis a section at time T. And then we use the same approach as for their random variables. But as we cannot observe all possible events developing in parallel, we may replace the consideration of all sections with the consideration of one trajectory of the process. We assume that the process can consequently consider the paths of all sections that could be at time T. If it is possible, we can say about ergodicity. Next step refers to a stationarity, maintaining the conditions of a statistical experiment. Which means that for any moment of time distribution will remain the same. But to have a terrible news, there are not so many stationary processes in the world. And B, stationarity is often ensured only on a limited interval T. Perhaps I've completely upset you with the bad news about how complexity of the model characteristics grows with the complexity of the model. However, if you could not work with a probabilistic model, then you just did not create it correctly. And how to do it will explain to you right now. What do we mean saying work easily? The object is described by a small number of characteristics. All characteristics are independent of each other when other possible. In that case, we will be able to change any of them and do not think what to do with others. If the characteristics are dependent, then they can be expressed through each other, ideally cause a fact. Descriptions of complex objects could be attributed to a set of simpler ones. And the dimension of the object should be minimal. The easiest way is to work with a random event. And it has only one characteristic of probability. Random variable is characterized by a distribution function, which is a multi-dimensional object. If you want to work with it, you need to have simplified characteristics that helps to estimate form of the distribution function. It could be at least three categories. Measures of central tendency, measures of variation, measures of distribution form. There are also several types of aggregated characteristic that based on central moments or based on quantile. Many different distributions have already been described, normal, uniform, exponential, log normal, and others, as well as the characteristics such as moments, quantiles, and others. But unfortunately, very often, we do not have enough knowledge on distribution form that we want to use. In the case of ideal situation, we can find model distribution for our task. We are lucky one of them. In other cases, we need to apply converted distributions, which means to transform some models so that we can work with them with our data. And at the same time, we do not lose their form. First kind of these distributions is truncated distribution. When we cut apart of the interval and normalize what we have. Another mechanism is suing of distributions. Sometimes, we need to use mixture distributions as some of different distributions. It may be thought that the most simple case to describe the probability is to do it with random events. Because in that case, if the events are dependents, we can introduce the definition of the conditional probability. Probability of event A, appearing given that another event B has occurred. As consicuence, we can formulate the multiplication theorem of probability or probability of joint occurrence A and B. If A and B are independent, then the probability of the joint occurrence is equal to the multiplication of their probabilities. From this, we can formulate law of total probability for any event A with different scenario. For random variables as well as for conditional probability, we can introduce conditional density. It will allow to as to formulate factorization of joint probability distribution. It is a very good use as instead of two-dimensional object, we get a multiplication of two one-dimensional objects. In this graph, you can see conditional and marginal distributions. Now, we talk about covariance and correlation. Covariance, second mixed moment of the distribution, which reflects the measure of linear connection between two variables. Covariation coefficient, normalized version of covariance. It is divided by multiplication of two dispersion of two processes. In this picture, you can see forms of two-dimensional diagrams of dispersion or graphical description of two-dimensional samples for different meanings of correlation coefficient. For correlation close to zero, the cloud looks like an ellipseoid, with increasing correlation coefficient the cloud tends to angle bisectors. Now, let's talk about regression. Initially, regression was a conditional expectation. Essentially, conditional expectation of a set of conditional probability distributions. However, if you evaluate it by sampling, regression is stable function, blue squares in the picture. And it's not comfortable to work with that table function. The next step is to identify models that could approximate this table function without losing a probabilistic meaning. The most simple approximation is the linear regression. Absalom here I can explain how far the data points are from the regression line. And it is not a random error, but a measure of the variability, which cannot be explained by the dependence of two variables. And coefficient A is directly proportional to the coefficient of correlation. However, as we can see from the picture, conditional expectation could also be a non-linear. Consequently, we can introduce a non-linear regression where we may use phi from Pc. The model could be developed as Heteras said this ethnicity in regression, where epsilon can have different characteristics and dispersion for the different meaning of Pc. Last model is combined distribution. When we talk about two variables, regression becomes more visible. However, if we have many variables, 1,000, 100,000, how can we create regression dependence for such number of characteristics? One of possible solutions, dimensionary reduction. There are several possible ways to do it. One of them, method of the principal components. When we talk about models of random processes, then for their simplification, we look for possibility of transformation that allows to investigate them with already known probability instruments. In particular, it could be a reduction to stationary stochastic process. In that case, we can look for simple models, additive and multiplicative model, where we need to find a trend which change across time. Other solutions, which consider different ways of non-stashianarity, also exist. For example, when they say that our process could be described as serious. The composition is certain basic non-random functions, such as signs, co-signs, where the co-efts themselves are stationary random processes. This instrument is suitable for different ways of non-stashianarity. However, you need to guess a class of non-stationary process. So, we looked at different variants of probabilistic models, and wanted to do next. Next, we need them to synthesize the data. If we do not have an opportunity to use analytical operators, we may recreate the probabilistic phenomenon and with synthetic data, find anything we need. Let's look at the generative approach. Take some data, build the model and create as many synthetic data as you want. What kind of generation mechanism exists? First, based on prior knowledge of the probabilistic model. If the model describes by normal distribution, then you can take a generator of normal distribution. And that's it. If we don't have a model but have data, then we need to build a model bait on this data, or work with sampling to create a synthetic data. It is a basis of Monte Carlo method, synthesis of probabilistic processes, and phenomena using random numbers. If we talk about random numbers, we need to understand that they are quite rare. For example, when we dimple with source of the entropy caused by physical parameters. However, we can use pseudo random numbers generators. Sometimes we can apply blended mechanism, which have an algorithm of pseudo random numbers, and then we can add source of the entropy. For example, process a temperature, which is changing over time. Anyway, nowadays there are many random number generators, but it's mean property, it is recurrence interval. Let's look how to use pseudo random numbers to generate the most simple object, a random event. We have some event with probability P of A, and interval from 0 to 1. We throw a random point and look where it falls. If it falls within the interval from 0 to P of A, the event is realized. When it falls within P of A to 1, the event is not realized. The mechanism is basic for generation of all other types of probabilistic models. Imagine that we simulate discrete random variables. The same thing here. However, our interval from 0 to 1 is divided not into two parts, but into a given number of intervals, which are determined by the type of distribution law. Thus, we divide our interval from 0 to 1 into subintervals, corresponding to the probability of height of each distribution column, and look, where pseudo random number falls. And accordingly, we believe that our random variable took exactly this value n. Simulation of continuous variable could be done in different ways. The simplest is a generalization of the previous algorithm for a situation where we can assume that our distribution is represented as a piecewise constant approximation. A histogram. So we select the number of the column where the variable falls as for the discrete distribution. Then we calculate its position inside this column based on its uniform distribution. However, this approach is not very good in terms of both productivity and quality, because piecewise smoothing is by no means the only way to take into account the features of the distribution itself. However, this approach is not very successful in terms of both productivity and quality, because piecewise smoothing does not always allow taking into account the peculiarities of the distribution itself. We can also often apply inverse function method. In short, if we want to get a random variable with a given distribution law, let's take this distribution function inverse and apply to a random number uniformly distributed from 0 to 1. However, not any distribution function could be easily reversed. For example, exponential function could be while log normal not. Therefore, for such laws, there are other approaches. One of them. A geometric method based on distribution density. We outline distribution density with a rectangle and drop random points, which are characterized by random numbers gamma 1 and gamma 2. In this point falls under the distribution line, this is suitable. If it falls over this distribution line, it does not. When we collect all the gamma 1 values that fall under the distribution, we will have a sample of the required law. But the more complex the shape of the distribution function, the lower the efficiency. For certain times of distributions, there are specialized algorithms that use previously known probabilistic properties. For example, to model normal or Gaussian random variables, we use the central limit theorem, which says that if we put together random numbers, then, accordingly, we get a distribution law that it's close to normal. There is only one problem. Generating sensitivity to distribution tails. It is clear that the greater the deviation from the center, the worse the algorithms work. The more members should be given in this. There are specialized algorithms for converting tails based on the box-muller theorem, where the two Gaussian, normal random variables are simulated at once. If we talk about multi-dimensional random variables, then we can use the same idea as for generation of one-dimensional variables. First of all, the inverse function method is generalized by the conditional distribution method. When the user idea of the possibility of multi-dimensional distributional factorization as the multiplication of one-dimensional and conditional. After that, we can successfully simulate the value of x using the inverse function method. We simulate the value of the one-dimensional random variable f of x. For x, we determine the conditional distribution f of y over x and simulate the value of the one-dimensional random variable y. For the obtained x, y, we determine the conditional distribution f of x, y, simulate the value of the one-dimensional random variable z, etc. Similarly, we can use the geometric method, but we will deal with not the graph of the distribution density, but with the graph of the surface and consider the case of values above or below this surface. Bad news, the greater the screening, the higher the dimension of space. Let's talk about process dynamic. Here we'll also can see some models. For random events, for example, one of the most interesting and simplest one is Markov chain. In fact, it characterizes a set of states or the possibility of transition of each of these states to another within one time step. In order to describe these possibilities, we entrant use transition probability matrix. The generative algorithm, in this case, means that we take one row or one column of our table and based on it, we generate into which state our system will transfer within next step. Then, the trajectory of the states of the Markov chain is built. For random variables, we need another models. In particular, they could be based on different types of regression relation as dependencies their own background. Which says how, based on the background, to approximate or extrapolate what await us in the future. Outer-aggressive models. In fact, this is a regression from its background. Where we have a deterministic path that describes a certain smooth regularity in the form of a sum, and epsilon of T as a noise. However, in some situation, it is considered reasonable to take into account that the noise can be correlated or taken into account that the noise can affect all subsequent steps. And then, you may introduce our autoregressive moving average model. We are on the right side, you find not just random noise, but the sum of random noise for previous time intervals with certain coefficients. Those, the taking to account in its background. The next step is dynamic regression. When an addition to noise will also have a certain add-to process in the right path, which also influence on the evolution of the trajectory of our simulated process. Such models are simple and clear, but there are problems that do not allow them to be applied everywhere. This is due to the fact that since they are linear, they work well for additive distribution laws. For example, for Gaussian or Normal, and for exotic logs, peculiar and autoregressive models are needed. The models discussed above are for time series, means discrete time processes. If we want to describe a random function with continuous time, we can use models of the deterministic function of a random argument. For example, in the form of a right stationary process, the composition in a harmonic basis, where random phases belong to uniform distributions. Such models can be developed using the apparatus of canonical or non-canonical decompositions, including application-oriented methods similar to the principal components which are also looked above. Why do we need all kinds of models for generating data with given probabilistic properties? In fact, for one thing, to create different composite probabilistic objects that describe the influence of the real world. Moreover, this can be not just basic models, but random graphs, automata, agents with probabilistic behavior. Let's look at the actual problems of modeling the distribution of COVID-19. Two probabilistic objects are needed. First, the contact network, who the infection spreads. The second object is the Markov chain, which describes the infections of the person himself. How long does the incubation period last for him? What variability does he get six, or will be asymptomatic? This is all regulated by the probabilistic model. And further, combining these two models and describing each of nodes of this graph using the Markov chain, we can go to the results described on the right. Briefly, what did we get from the lecture? Do you want to do with probability? That's good. But, are you sure that you have it here? Find the source of randomness. Any probabilistic descriptions could be reduced to random events. Are you sure that it wouldn't make your life more difficult? Choose the right probabilistic model. Many different probabilistic methods are described in the books and implemented by software. Are you sure that they match your task? Check the conditions for applicability of probabilistic models. Your task is difficult and conditions of applicability are quite questionable, who prevents you from using generated approach. Create a Monte Carlo simulation and knock yourself out. Choose a dataset from the link from the slide or other sites. Create one-dimensional sample size, 65 plus X, where X die of your breath. Explain why you can apply probability modeling here. Choose generated methods for data reproducing by the same law as the dataset. Generate data sample by the same size as the original one. You may do it in any software. So, original X and synthetic Y samples. Put dots X, Y coordinates or which have the same numbers in the samples to the graph. Looks like Y equal to X send the report. Looks like something else? Well, try to find the mistake.

# System oriented analysis and informational modeling

**Keywords or keyphrases**:

"translational research", "concepts systems", "systems oriented methodology", "translational research methodology", "research methodology lecture", "methodology lecture devoted", "course translational research"

**Topics**:

"system analysis components", "system analysis description", "system analysis slide", "system slide description"

**Summary**:

Lecture is devoted to basic concepts of system-oriented analysis and information modeling. The triad relates object, system and model. The goal is a non-existent but desired state of the system. A problem is a set of questions to be considered with a given set of initial information. Systems-oriented analysis is a methodology for studying complex problems in theory and practice. The word system and many related ideas were known in the ancient world. A system is called complex if it is a non-linear multi-scale system. This system's elements sub-systems allocated on different levels or scales. The morphological description represents a system as a table, A, B, R, V, Q. A is a set of system components and their properties. B shows the affiliations of the inhabitants to the external environment. R represents the relationship between inhabitants that predator prey connections. System oriented analysis is unreasonable without considering the system's resources. movement is not possible without the participation of human resources including the bus driver. organizational resources are required for the ground public transport network to function. The building industry has a process that they call information modeling. It uses a digital representation of both physical and functional aspects of the object and the analysis.

**Extracted text**:

Hello, we continue our course in Translational Research Methodology with a lecture devoted to basic concepts of system-oriented analysis and information modeling. We will start with basic concepts of the system-oriented analysis and recall some historical aspects. After that, we will consider system structure and then look at different methods of system-oriented analysis and we will finish with information modeling. Let's start with the first part, basic concepts of system-oriented analysis and a bit of history. This slide contains two pictures. The right picture presents a car. The left picture shows the sets of components of this car. I hope that after the lecture, you will easily understand the key differences between these two pictures. Let's move on to the basic terms and concepts. The triad presented on this slide relates object, system and model. The object is a real entity to be learned. The system here is a formalized object, a structured entity with internal and external relations. And the model, that's a representation of an object or a process with rules, notations and components in a card with some given assumptions. What is a system? Well simply, a system is a set of components plus the relations between them and the external environment. Every component of a system may be considered as a separate system with its own components and relations. These components are called subsystems. Slide 8 presents an example of a system, science. Science is a system for acquiring, verifying, fixing, storing and updating knowledge. We may consider branches of science, including physics, biology, geology as some of its subsystems. And note that any knowledge exists in the form of a system only. The examples of the systems are computer science and synergetics. Consider synergetics. Synergetics studies general ideas, techniques and patterns of organization of different objects and processes. The invariants or unchanging entity of them. The next important concepts of system-oriented analysis are the goal and the objective. The goal is a non-existent but desired state of the system. The goal is the answer to the question, what is the system for? And what is an objective? What difference is there between goal and objective? goals are about the big picture and objectives are all about detailed tactics. So goal is long-term outcome, while objective looks at what gets done in smaller steps. To distinguish objectives from goals, we need to understand the context and the scale under consideration. For example, when we prove a lemma or preposition in order to solve a problem, that's an objective. If we try to prove the more general theorem, that's a goal. Please, look at some other examples and answer the question. Which cases would we consider the goal and cases would we consider the objective? The next concept is the problem. A problem is a question or a set of questions to be considered with a given set of initial information. And the problem statement defines a goal or objective. Possible resources, limitations, constraints, strategies to achieve the goal or intermediate states of that object must pass through. This slide shows several classes of problem. To define the problem in any class requires factual data. For example, to solve problems in linear programming, we need both the objective function and also the factual constraints. A system operates to achieve a goal. If the goal is constant, then the system is set to be functioning. If the goal changes, however, we say that the system is developing. So we have two new concepts here, the functioning of the system and the development of the system. Emergence and synergistic effects are to crucial concepts of system-oriented analysis. The properties of the system and the properties of its components are not the same thing. The overall system emerging from its components is called emergence. Combining individual components into a holistic system, we get new effects that are not found in any of the components. These new effects are called synergistic effects. They come from the synergy between the components. A system has an internal environment which contains its components and relations. From there, a system interacts with external environment, including with other systems. The technological boundary is a space where the external interactions emerge. For example, an Auntafreej or a cell membrane. And so, what is system-oriented analysis? It is a methodology for studying complex problems in theory and practice. System-oriented analysis is the family of concepts, methods, procedures and technologies for the study, description and implementation of objects or processes. We've looked at some of the basic concepts of systems-oriented analysis. Let's touch on a bit of history. The word system and many related ideas were known in the ancient world. Aristotle, for example, said, the whole is greater than the sum of its parts. Contributions to systems-oriented methodology was made by philosophers and other thinkers throughout history. Bacon, Hegel, Kant and numerous others offered new ways to structure and advance knowledge. Many of those ideas became incorporated into the practice of modern science. The foundations of contemporary systems-oriented analysis are well established and are used in various fields of science. New science approaches, including interdisciplinary approaches, can be a source of new methods of system-oriented analysis. Let's move to Part 2, system exploration, structure and relations of a system. The system description is a formalization of a system's elements, subsystems, interrelations, functioning and a set of all its states. If the system's descriptions can be reduced to a linear structure, the system is called an easily formalizable system. If it is impossible to express a system description in known terms, notations and concepts, we call it a poorly formalized system. This slide presents the general scheme of system description. This scheme shows that there are two types of relations, internal relations between the system's components and external relations between the system and its environment. Therefore, there are two types of system descriptions, their internal description and their external description. Their internal description represents internal resource flows between a system's elements from the inputs to the outputs. The external description describes this entity as a black box with inputs and outputs and its interactions with the external environment, including other systems. The system's description contains information about its structure. The structure of a system is an organization of a set of its components and relations between them. This slide shows two basic topologies of system structure, a linear topology and hierarchical topology. The idea hierarchical and matrix topologies are particular kinds of network topology more generally. This slide shows some examples of a system structure identification. In particular, the underground railways communication network structure is a network structure. Please see the picture in the right. Let's talk about system complexity. A system is called complex if it is a non-linear multi-scale system. This system's elements sub-systems allocated on different levels or scales. One level of a complex system may contain so many components that it is not feasible to consider all of them together. Relations are possible not only between distant elements, but these elements may be located on different levels scales. A human offers an example of a complex system. We may see at least six scales from their atomic scale to their social level. That example is a matrix of colorite pixels. If there are fewer of them, then we see them well. If there are tens of thousands or more in the same area, then we notice the content of the image, not the pixels. We see the main building of it more university, instead of 1.4 million pixels in this slide. How can we describe a system? This slide shows some forms of system description which are often used in system-oriented analysis. A morphological description describes a set of system components and a set of relations between them. And in the biological description describes informational relationships inside the system and informational connections between the system and the external environment. The morphological description represents a system as a table, A, B, R, V, Q. A is a set of system components and their properties. B is a set of relations between the system and the external environment. R is a set of relations inside the system. V is a system structure. Q is a description of system according to the given notations. Let's consider morphological description of the microecco system as an example. Imagine that this microecco system contains the following inhabitants as its components. Man, tiger, hawk, pike, sheep, gazelle, wheat, boar, clover, wool, snake, a corn, crucial car. X and Y are subsets of this species, predators and prey respectively. B shows the affiliations of the inhabitants to the external environment. R represents the relationship between inhabitants that predator prey connections. The system structure V can be represented as a table. It is it on the stand that one in the cell means that the species in the troll is prey for the species in that column. For example, a vowel can be prey for a hawk. Slide number 14 shows the structure of microecco system V represented as an oriented graph. The log-covered terrain equations are a way to represent the system description to you. It is a mathematical description of the system populations. So, we come to the third part, methods and means of systems oriented analysis. This slide presents two basic approaches to system oriented analysis, formal and conceptual. Formal approaches are rarely used in pure form. Methods from exact signs are used to varying extents depending on the goal of the system oriented analysis. The list of methods is not very complete. As I said, the range of methods is continually expanding, especially from the influence of in new crosscutting technologies. System oriented analysis is unreasonable without considering the system's resources. Real energy, information, human, organization, space, time are seven basic resources types. Space and time resources are occurring in their pure form during resource analysis. The other types, though usually appear in combinations with each other. What is the resource analysis? A resource analysis is the identification of resource flows and resource transformations between the system's competence and the system and its external environment during its state's transitions. Resource analysis allows us to detect pros and cons in the resource flows and transformations and to prepare related managerial decisions. Materials found in the natural world that have practical use and value for humans. Material resources include wood, glass, which comes from sand, metals and plastics which are made from natural chemicals. Energy is the ability to work. There are many different types of energy such as kinetic energy, potential energy, light, sound and nuclear energy. One form of energy may be converted into another without violating a law of thermodynamics. Their information resources reflect the order, structure and self-organization of matter. Humans are the most critical and unique economic and social resource of society. We cannot imagine the functioning of any institution, for example, a fried forruda on internet provider without humans. The organization resources. Groups or societies defines their organizational structures including the institutions of human society and its superstructures. Organization of the system is associated with cause and effect relationships in the system. Special space relating to the position, area and size of things. For example, a student's workplace requires two square meters in the classroom. Time is the measured measurable period during which an action, process or condition exists or continues. Suppose that a student is required to get to one of the buildings of a university. The next slide, slides 51 and 52, lists the aspects of the resource analysis. For example, the student's movement is not possible without the participation of human resources including the bus driver. In addition, organizational resources are required for the ground public transport network to function. After enumerating the basic resources, we need to identify and analyze resource transformations during state transitions. One convenient way to do this is to use a table as shown on the right hand side of the slide. It shows the resource transformations which correspond to the system component, the student and the state transition, the bus trip. Slide number 54 shows a possible improvement. A student could move far or all of the learning process to hold. The goal of system-oriented analysis is to improve the object process. And the method of system-oriented analysis includes actions aimed both at identifying the system and also at formulating its description. And it sprows and counts, justifying and developing grounds for managerial solutions. This slide lists the steps of optner's method from identification of symptoms to assessment of the consequence of solution implementation. And so, the last part, information modeling. We speak a great deal about data and information. And it is important to remember that what we are really working with is information that can be represented for storage, for communication, for processing. And it is important to remember that to mean something, data and information have to be interpreted. The building industry has a process that they call information modeling. It uses a digital representation of both physical and functional aspects of the object and the analysis. And it covers phases of the object's life cycle. The steps of information modeling are presented in this slide. These steps follow from the basic principles of modeling. Therefore, we start from the object statement, carry out model development using a choosing method and then finish model documentation. The next slide, slide number 61 and slide number 63, lists of methods of information modeling in computer science. I want to waste your time reading out a precise definition. You can find it in the cited sources. As you will see, each method reflects processes in a particular branch of computer science, for example, in web technologies or knowledge basis. The last slide presents an outcome of information modeling. The ER entity relationship model of trained schedules. This model represents the physical and functional aspects. The physical aspects are the trains, drivers and stops. The functional aspects are the routes and the schedules. We can note that one railway stop may be on several routes and that anyone route contains several stops. So, the relationship between stops and routes is many to many. Other relations of this model are one to many. Thank you for your attention.

# Factographical informational analysis

**Keywords or keyphrases**:

"course methodology translational", "translational research lecture", "lecture framework", "lecture consider", "lecture methods working", "second lecture methods", "research lecture calls", "research lecture", "methodology translational research", "lecture methods"

**Topics**:

"research also may", "may research example", "may research also", "research may find", "research may also", "research may networks"

**Summary**:

Alexandra Klimova: Literature review is one of the most important parts of scientific work. When you enter to unfamiliar research area, you may need to find answers to very simple questions. Two most popular interdisciplinary bibliographic online databases are Skopus and Weppel Science. Web of Science Co-Collection allows defining search field. You can make your search by topic, author, publication name, funding agency, organization, hands, etc. Can group and rank records in a result set by extracting data values from a variety of fields. First part of the lecture is about methods of working with scientific literature. While preparing the review, the first thing to keep in mind is a type of you expected result. Literature review is also an essential part of any scientific paper. It is better to consider the source of paper for example first paper is the list is published at archive. When you decide the selected paper is worth mentioning your review, you start to collect the bibliographic link. The basic way to write down all the details is not so convenient because you may want to collect all your links in a single format to facilitate further work with them. To identify the most important authors for each topic, you may use means of network analysis. In this study, we analyze the medical keyword networks from international conference on computational science. The plot illustrates the dynamics of normalized betweenness, centralness, in of several selected keywords. Such method may be used to find the topics which have influenced a given community of researchers throughout the time.

**Extracted text**:

Good afternoon. We would like to present the first lecture in the framework of the course methodology of translational research. The lecture calls methods of pathographic information analysis. My name is Alexandra Klimova and we prepare this lecture with my colleague Claudia Bacchina, associate professor of the Digital Transformation Department. Doing a literature review is one of the most important parts of scientific work. If you come to a new scientific area, you may want to view the overall research performance, find out about impact of the leading scientist institutions, scientific sources involved in this research. Then you get a concrete research question and you need to dig deeper. How to do it in both ways? We will try to describe it in this lecture. When you enter to unfamiliar research area, you may need to find answers to very simple questions. What country publishes the most papers in this research area? Which authors wrote about this topic in the last five years? What is the modern research trends here? Well, we can find information to answer those questions. Two most popular interdisciplinary bibliographic online databases are Skopus and Weppel Science. Which contain abstracts and citation databases. Google's color is another massive database of scholarly literature that allows users to assess information and keep up with new research as it comes out. One more interesting source of information is Research Gate. This is a social network for scientists. Here you may ask an author to share with you a non-open access paper. However, finding information is not a simple task, considering the fact that almost 2 million papers are published every year. Let's look at the example. Complex networks. Complex networks is a relatively young and active area of multidisciplinary scientific research inspired by the study of real-world networks such as computing networks, biological networks, technological networks, brain networks, and social networks. Topic Complex Network gives around 296,000 results in Google's color, 49,700 in Science Direct and 19,900 in Weppel Science Coal Collection. Key word Complex Network also gives more than 6,500 results. It is clear that researchers today have started to analyze larger and larger amount on scientific literature, thus they need to use more advanced analytical instrument and tools. One of such tools is SyWOW, which is used for research performance assessment and strategic planning based on scope of data. It offers access to research performance of more than 17,300 institutions and their associated researchers from 231 nations worldwide. It allows to visualize your research performance, benchmark relative to peers, develop strategic partnerships and identify and analyze new emerging research trends. What is also important, you may find your own research area. Research area can represent strategic priorities, research trends, or any other interesting topic based on different components used to create it, search term, entities, competencies etc. In order to define research area, you need to choose this option in SyWOW Tombowr. Then define search terms, apply filters, and finally save your research area. A opportunity to define your research area allows following its performance. You have an opportunity to see the most important key words in a research area with the word cloud visualization. You may also look at the overall research performance, including scholarly output, use count, citation count, and other metrics. To get more detailed analysis and access impact of different stakeholders included to their research, look at another categories such as institutions, countries and regions, authors, and scope sources. Slightly presents the results for research zone, which includes two scientific areas, machine learning and complex networks. There are five top countries and regions in this research area, top 10 authors and six top scope sources. And we go further to another analytical instrument, which is provided by Clarionway Tenelitics. Web of Science Co-Collection is a collection of over 20,000 peer-reviewed high-quality scholarly journals published worldwide in over 250 social science and humanities disciplines. Conference proceedings and book data are also available. Web of Science Co-Collection allows defining search field. You can make your search by topic, author, publication name, funding agency, organization, hands, etc. You can also combine different keywords and phrases or add another search field by using search operators such as end, or nor, etc. You may customize a time span by choosing publication period, function, analyze the results or analysis of results. Can group and rank records in a result set by extracting data values from a variety of fields. These two helps to find the most prevalent authors in a particular field of study or generate a list of institutions rendered by record found based on the search query. Each item contains a lot of information, title, abstract, author keywords, autumn names, funding information. You may also find a link to full text library holdings of Google's color, information about citation is also available. All cited references are indexed and searchable by a cited reference search. As I said before, analyze results function allows you to group and rank records in a result set by extracting data values from a variety of fields. It permits and analysis of the records by various data points and visualizations like Tremac or BarGraph. The field option for the analysis of web of science categories, publication years, document types, organization enhanced, funding agencies, authors, source titles, book series titles, meeting, etc. etc. If you compare results obtained for keywords, machine learning and complex networks, with what we have got by using SAVAL, you will see almost identical results. A powerful approach to analyze a large variety of bibliometric networks is visualization. What's fever is a software tool for constructing and visualizing bibliometric networks. These networks may, for instance, include journals, researchers or individual publications, and they can be constructed based on citations, co-satations or co-authorship relations. It also offers text-mine functionality that can be used to construct and visualize co-curance networks of important terms extracted from the body of scientific literature. The tool was developed in the Center for Science and Technology Studies of the University of LADEN. It works with data from different sources such as web of size, scope, cross-reference, etc. and it quite easy to use. In order to start, analyze your data, you need to choose create a map based on bibliographic data, find source of information and choose type of analysis and counting method. Let's look at the example of visitor digital transformation. Digital transformation has already became a quiet buzzword, triggering different disciplines in research and influence in practice, which leads to independent research streams. Those streams include management and business, computer science, education and economics. It is relatively young and active area and started from 2013. The number of publications is increasing significantly. The tool has defined several zones where we can definitely find different research streams. Random blue one describes digital economy and organizations. Green, technology and yellow education. However, we can see several terms such as literature, theory, good. That could be deleted as they didn't help much our classroom process. Keywords removing might be done during keywords verification option. Overlaid visualization can be used to show developments over time. A color bar is shown in the bottom right corner of the visualization. Colors indicate years when a specific term was more frequently used in a framework of digital transformation terms. So yellow color shows more recent keywords associated with digital transformation. Therefore example, machine learning, artificial intelligence, smart city, disruptive technology. You can also create a keyword called Curien's Map. Keyword can be extracted from the title and abstract of the publication. All they can be taken from the author's applied keyword list over publication. Well, second part of the lecture is about methods of working with scientific literature. Compared to the first part, this is not about finding actual trends in a quite general stand-up study, but about creating rather complete representation and understanding of a specific branch. While preparing the review, the first thing to keep in mind is a type of you expected result. This may be a review article when one tries to systematize and summarize the current state of the art research. Together with some criticism and knows for further findings. Literature review is also an essential part of any scientific paper. For example, for the conference paper in computer science, this is a necessary chapter usually after introduction or before conclusion. Finally, this may be a part of scientific report when you prepare a review for justification of choice of your methods and models. According to the type of the review, it may have different length and a number of sources from 8 to 10 to 100 and more. In this part of the lecture, we will consider different strategies and practical advises about how to prepare and select literature sources, how to systematize and compare them as well as suggestions for the tools which may be used to facilitate this process. Three most popular types of review are state of the art review, comparative review and critical review. First type is aimed at describing the most recent and popular methods of your specific narrow field of study. For example, algorithms for particular problem statement and typically consists of several sentences in each method describing the essence of the result obtained. A comparative review is similar but typically it is used to show that you, investigation has some significant scientific novelty. This type of review requires preparing some system of features which are used as aspects of methods comparison. Usually a method proposed by the author is also added to such comparison to show that it is, it has some properties which are not presented in the previous study. Finally, critical review is a deep study of the field when the goal is not only in unerate all the methods and to compare them but also make some conclusions and suggestions about the impact of these methods and the way of further development of the field. Let's consider the use case of performing literature search suppose that we want to find some algorithms to create low dimensional representations or embeddings of a specific type of graph having vectors of attributes associated with the notes. Such graphs are called attributed graphs. So our first guess about selecting the right keyword for the literature search would be attributed graph embeddings. Let's try to find something in Google's code. You have different filters in the left part of the user interface. Usually for the algorithms, basic period of consideration is equal to five years. It is better to consider the source of paper for example first paper is the list is published at archive. This may mean that it is now during the review process. Most of the papers also have PDF attached so you may have access to the full text. Finally, good idea is to consider papers with a lot of citations. A good practice else so to look at the list of papers which size some state of the art paper. If the paper about verse in our example. From this list for our particular case, we may find more recent works of the same authors methods for another problem statements or another literature review. This is a valuable thing because probably the authors did most of the work for us. Also, good practice is to keep notes on the conferences and journals which promote this field of study and this may be a good choice for your own publications. Then you look at the full text of the related papers and so read the part with the review. You may consist in authors systematic comparative review of the previous methods which allows you to create and to maintain your own picture of the state of the art method. Paper is about algorithms usually contain not only quantitative based on text features but also qualitative comparison of algorithms. This qualitative comparison is performed based on the experimentation with different algorithms on the same dataset. Some fields have also benchmark datasets which are usually used to compare the algorithms. Sometimes especially for the new problem statements, benchmark datasets are not available. However, to compare your solution with previous solutions, even for your own dataset is usually good idea as it makes your contribution more solid. Finally, look at the reference list of your featured research. Again, it allows you to form a view of places when this try to study mainly published and about the most prominent people in the field. There is a way to check if the conference or a journal works considering for the publication. For conferences, you may use conference relics website. The highest trend is A-asterix, top rated worldwide conferences. The conferences usually have acceptance rate not more than 20% and B and C conferences are not so significant as A or A-asterix. Sometimes, a conference has no rank. It may be because of its low quality or because its novel and has no deserved interpretation yet. Regarding the journals, you may use the same manga website to check quartet of the journal. For example, Q1 is the highest impact factor and the dynamics of the characteristics. When you decide the selected paper is worth mentioning your review, you start to collect the bibliographic link. The basic way to write down all the details is not so convenient because you may want to collect all your links in a single format to facilitate further work with them. Google's color supports important bibliographic links in a different format. You may create files containing several links to the same format and attach them, for example, to tech document. In this case, your bibliography will be automatically available to the main text of your review. Another option which will also work for Microsoft of the solutions is called Mendeling. It is highly recommended by our research team as it allows you to keep all of the papers which you find interesting to create and maintain the library with source files and full bibliographical links. It also supports folders different ways of importing and exporting papers information from files to manual input, working with different bibliographical styles and cloud storage. Plug-in for Microsoft work is available. These are three ways of importing information to Mendeling. First of all, is manual. When you type all information in different fields, second one is searching by an existing DOE, digital object identify. Third one is import from files, for example from Viptex. What plugin is a part of bibliography panel? You may choose desired bibliographical style. Source is added either by search panel or by selecting the paper in the main main delay application. Formatted bibliography is inserted by the button of the panel. Let's consider now most sophisticated ways of investigating literature sources. One of the examples is created so called co-authorship networks. At this slide, a network for a Russian journal oil industry is depicted. A node is an author, and the link denotes that these two authors are co-authors in at least single paper. This network also is attributed. Attributes are shown by different colors. For each author, we prepare a set of keywords from his or her papers. We find the most popular keyword for each author. This keyword determines predominant topicality of works over given author. So the attribute means the predominant topicality. In this graph, one can see the clusters of authors with the same topicality, medical clusters. So, unit of methods you may visually identify medical clusters. This slide represents a subgraph of the previous graph filtered by a single topic, machine learning. To identify the most important authors for each topic, you may use means of network analysis, for example, calculated different centrality measures. Next example are key work networks when you create a graph, when a node is a keyword, and link connects two keywords from the same paper. In this study, we analyze the medical keyword networks from international conference on computational science. The plot illustrates the dynamics of normalized betweenness, centralness, in of several selected keywords. Such method may be used to find the topics which have influenced a given community of researchers throughout the time. You may also consider dynamical co-authorship network. In this example, also from ICCS conference, you may observe the widening of a gand core of the network with time. That means that we may observe the formation of the community over computational scientists. You may create such kind of networks by manual programming, for example, using Python Networks library, all with special tools, for example, site net explorer. Another example is VOS viewer, allowing to create colon maps and cluster bibliographic networks. As an example of systematic comparison, let's consider this table with the list of methods. You may see that amount of bases for comparison types of inputs for the method, data set for experimentation and another domain specific features, for example, number of clusters may be used. Another format of systematic comparison is a table when solutions of methods are in rows and feature sign columns and cells up plus and minus sides. This compact format is often used in scientific reports. For a comprehensive review, one may also use graph of methods similar to shown in this picture. Here, know the methods and edges denote the fact that these methods are compared by someone. Such kind of graph allow to distinguish state of the art methods. To conclude, we have two different general ways of working with bibliographic information. First one is from a view about some general field and second one is to justify your assumptions and to deepen your knowledge about some narrow field. Different tools may also be used to facilitate collecting and processing bibliographic information. Now you task. So, what you need to do is choose topic based on Russian Science Foundation Research field. Here is only. Then you need to register in schools of size while using your corporate email. Analyze research topic. You have to provide screenshots from size while or web of signs. You need to indicate 10 most cited authors in the research area, five top sources of this research area, 10 most frequent keywords, and top 100 countries regions in this research area. Install was here, make analysis of keywords, and provide a screenshot. Choose one direction of the research field, find five papers using different ways of search for information, use Mendelay to create a reference list. And here is the example how to use this Russian Science Foundation Research field. So, if you go from the link, you will see the list of the research fields. And based on your first letter, you choose the first level on the research field. Second level will be based on your name. Here is an example from the category mathematics computer and system sciences. And third level is based on your choice. You will find the list of these categories and subcategories in our Google classroom folder. Example, Peter Smith. So, based on his first letter from his family name, as it will be mathematics computer and system thinking. His name is Peter P. So, we can choose network technologies. And the third level will be knowledge management technologies. So, I wish you good luck with your future study in this course and also good luck with the task to be performed. Thank you for your attention. Goodbye.