Chapter 1

REALISM AND INSTRUMENTALISM: CLASSICAL STATISTICS AND VC THEORY (1960–1980)

1.1 THE BEGINNING

In the history of science two categories of intellectual giants played an important role:

(1) The giants that created the new models of nature such as Lavoisier, Dirac, and Pasteur;

(2) The giants that created a new vision, a new passion, and a new philosophy for dealing with nature such as Copernicus, Darwin, Tsiolkovsky, and Wiener.

In other words, there are giants who created new technical paradigms, and giants who created new conceptual (philosophical) paradigms. Among these, there are unique figures who did both, such as Isaac Newton and Albert Einstein.

Creating a new technical paradigm is always difficult. However, it is much more difficult to change a philosophical paradigm. To do this sometimes requires several generations of scientists.[1] Even now one can see the continuation of the old paradigm wars in articles discussing (in a negative way) the intellectual heritage of the great visionaries Charles Darwin, Albert Einstein, Norbert Wiener, and Isaac Newton.

My story is about attempts to shift one of the oldest philosophical paradigms related to the understanding of human intelligence. Let me start with the vision Wiener described in his book *Cybernetics*. The main message of this book was that there are no

[1] Fortunately scientific generations change reasonably fast, about every ten years.

big conceptual differences between solving intellectual problems by the brain or by a computer, and that it is possible to use computers to solve many intellectual problems.

Today every middle school student will agree with that (five scientific generations have passed since Wiener's time!). However, 50 years ago even such giants as Kolmogorov hesitated to accept this point of view.

1.1.1 The Perceptron

One of the first scientific realizations of Wiener's idea was a model of how the brain learns introduced by Rosenblatt. He created a computer program called the "Perceptron" and successfully checked it on the digit recognition problem. Very soon Novikoff proved that the Perceptron algorithm (inspired by pure neurophysiology) constructs a hyperplane in some high-dimensional feature space that separates the different categories of training vectors.

It should be mentioned that models of how the brain generalizes and different pattern recognition algorithms both existed at the time of the Perceptron. These algorithms demonstrated success in solving simple generalization problems (for example Selfridge's Pandemonium, or Steinbuch's Learning Matrix).

However, after Rosenblatt's Perceptron and Novikoff's theorem, it became clear that complex biological models can execute very simple mathematical ideas. Therefore it may be possible to understand the principles of the organization of the brain using abstract mathematical arguments applied to some general mathematical constructions (this was different from analysis of specific technical models suggested by physiologists).

1.1.2 Uniform Law of Large Numbers

The Novikoff theorem showed that a model of the brain described in standard physiological terms ("neurons," "reward and punishment," "stimulus") executes a very simple mathematical idea — it constructs a hyperplane that separates two different categories of data in some mathematical space. More generally, it minimizes in a given set of functions an empirical risk functional.

If it is true that by minimizing the empirical risk one can generalize, then one can construct more efficient minimization algorithms than the one that was used by the Perceptron. Therefore in the beginning of the 1960s many such algorithms were suggested. In particular Alexey Chervonenkis and I introduced the optimal separating hyperplane that was more efficient for solving practical problems than the Perceptron algorithm (especially for problems with a small sample size). In the 1990s this idea became a driving force for SVMs (we will discuss SVMs in Chapter 2, Section 2.3). However, just separation of the training data does not guarantee success on the test data. One can easily show that good separating of the training data is a necessary condition for the generalization. But what are the sufficient conditions?

This led to the main question of learning theory:

When does separation of the training data lead to generalization?

This question was not new. The problem, "How do humans generalize?" (What is the model of induction? Why is the rule that is correct for previous observations also correct for future observations?) was discussed in classical philosophy for many centuries. Now the same question — but posed for the simplest mathematical model of generalization, the pattern recognition problem — became the subject of interest.

In the beginning of the 1960s many researchers including Chervonenkis and I became involved in such discussions. We connected this question with the existence of uniform convergence of frequencies to their probabilities over a given set of events. To find the conditions that guarantee the generalization for the pattern recognition problem, it is sufficient to find the conditions for such convergence.

Very quickly we constructed a theory for uniform convergence over sets with a finite number of events (1964) and in four years we obtained the general answer, the necessary and sufficient conditions for uniform convergence for any (not necessarily finite) set of events. This path is described in *EDBED*.

What was not known at the time EDBED was written is that the uniform convergence describes not only sufficient conditions for generalization but also the necessary conditions:

Any algorithm that uses training data to choose a decision rule from the given admissible set of rules must satisfy it.

It took us another 20 years to prove this fact. In 1989 we proved the main theorem of VC theory that states:[2]

If the necessary and sufficient conditions for uniform convergence are not valid, that is, if the VC entropy over the number of observations does not converge to zero,

$$\frac{H_P^{\Lambda}(\ell)}{\ell} \longrightarrow c \neq 0,$$

then there exists a subspace X^* of the space R^n whose probability measure is equal to c,

$$P(X^*) = c,$$

such that almost any sample of vectors $x_1^*, \ldots, x_k^*$ of arbitrary size k from the subspace X^* can be separated in all 2^k possible ways by the functions from the admissible set of indicator functions $f(x, \alpha),\ \alpha \in \Lambda$. (See also *EDBED*, Chapter 6 Section 7 for the definition of VC entropy).

This means that if uniform convergence does not take place then any algorithm that does not use additional prior information and picks up one function from the set of admissible functions cannot generalize.[3]

[2]Below for the sake of simplicity we formulate the theorem for the pattern recognition case (sets of indicator functions), but the theorem has been proven for any set of real-valued functions [121;140]. Also to simplify formulation of the theorem we used the concept of "two-sided uniform convergence" discussed in *EDBED* instead of "one-sided" introduced in [121].

[3]This, however, leaves an opportunity to use averaging algorithms that possess a priori information about the set of admissible functions. In other words VC theory does not intersect with Bayesian theory.

If, however, the conditions for uniform convergence are valid then (as shown in Chapter 6 of *EDBED*) for any fixed number of observations one can obtain a bound that defines the guaranteed risk of error for the chosen function.

Using classical statistics terminology the uniform convergence of the frequencies to their probability over a given set of events can be called the *uniform law of large numbers* over the corresponding set of events. (The convergence of frequencies to their corresponding probability for a fixed event (the Bernoulli law) is called the law of large numbers.)

Analysis of Bernoulli's law of large numbers has been the subject of intensive research since the 1930s. Also in the 1930s it was shown that for one particular set of events the uniform law of large numbers always holds. This fact is the Glivenco–Cantelli theorem. The corresponding bound on the rate of convergence forms Kolmogorov's bound. Classical statistics took advantage of these results (the Glivenco–Cantelli theorem and Kolmogorov's bound are regarded as the foundation of theoretical statistics).

However, to analyze the problem of generalization for pattern recognition, one should have an answer to the more general question:

> *What is the demarcation line that describes whether the uniform law of large numbers holds?*

The obtaining of the existence conditions for the uniform law of large numbers and the corresponding bound on the rate of convergence was the turning point in the studies of empirical inference.

This was not recognized immediately, however. It took at least two decades to understand this fact in full detail. We will talk about this in what follows.

1.2 Realism and Instrumentalism in Statistics and the Philosophy of Science

1.2.1 The Curse of Dimensionality and Classical Statistics

The results of successfully training a Perceptron (which constructed decision rules for the ten-class digit classification problem in 400-dimensional space, using 512 training examples) immediately attracted the attention of the theorists.

In classical statistics a problem analogous to the pattern recognition problem was considered by Ronald Fisher in the 1930s, the so-called problem of discriminant analysis. Fisher considered the following problem. One knows the generating model of data for each class, the density function defined up to a fixed number of parameters (usually Gaussian functions). The problem was: given the generative models (the model how the data are generated known up to values of its parameters) estimate the discriminative rule. The proposed solution was:

> *First*, using the data, estimate the parameters of the statistical laws and

Second, construct the optimal decision rule using the estimated parameters.

To estimate the densities, Fisher suggested the maximum likelihood method.

This scheme later was generalized for the case when the unknown density belonged to a nonparametric family. To estimate these generative models the methods of nonparametric statistics were used (see example in Chapter 2 Section 2.3.5). However, the main principle of finding the desired rule remained the same: first estimate the generative models of data and then use these models to find the discriminative rule.

This idea of constructing a decision rule after finding the generative models was later named the *generative model of induction*. This model is based on understanding of how the data are generated. In a wide philosophical sense an understanding of how data are generated reflects an understanding of the corresponding law of nature.

By the time the Perceptron was introduced, classical discriminant analysis based on Gaussian distribution functions had been studied in great detail. One of the important results obtained for a particular model (two Gaussian distributions with the same covariance matrix) is the introduction of a concept called the Mahalanobis distance. A bound on the classification accuracy of the constructed linear discriminant rule depends on a value of the Mahalanobis distance.

However, to construct this model using classical methods requires the estimation of about $0.5n^2$ parameters where n is the dimensionality of the space. Roughly speaking, to estimate one parameter of the model requires C examples. Therefore to solve the ten-digit recognition problem using the classical technique one needs $\approx 10(400)^2C$ examples. The Perceptron used only 512.

This shocked theorists. It looked as if the classical statistical approach failed to overcome the curse of dimensionality in a situation where a heuristic method that minimized the empirical loss easily overcame this curse.

Later the methods based on the idea of minimizing different type of empirical losses were called the *predictive (discriminative) models of induction*, in contrast to the classical *generative models*. In a wide philosophical sense predictive models do not necessarily connect prediction of an event with understanding of the law that governs the event; they are just looking for a function that explains the data best.[4]

The VC theory was constructed to justify the empirical risk minimization induction principle: according to VC theory the generalization bounds for the methods that minimize the empirical loss do not depend directly on the dimension of the space. Instead they depend on the so-called capacity factors of the admissible set of functions — the VC entropy, the Growth function, or the VC dimension — that can be much smaller than the dimensionality. (In *EDBED* they are called *Entropy* and *Capacity*; the names VC entropy and VC dimension as well as VC theory appeared later due to R. Dudley.)

[4]It is interesting to note that Fisher suggested along with the classical generative models (which he was able to justify), the heuristic solution (that belongs to a discriminative model) now called Fisher's linear discriminant function. This function minimizes some empirical loss functional, whose construction is similar to the Mahalanobis distance. For a long time this heuristic of Fisher was not considered an important result (it was ignored in most classical statistics textbooks). Only recently (after computers appeared and statistical learning theory became a subject not only of theoretical but also of practical justification) did Fisher's suggestion become a subject of interest.

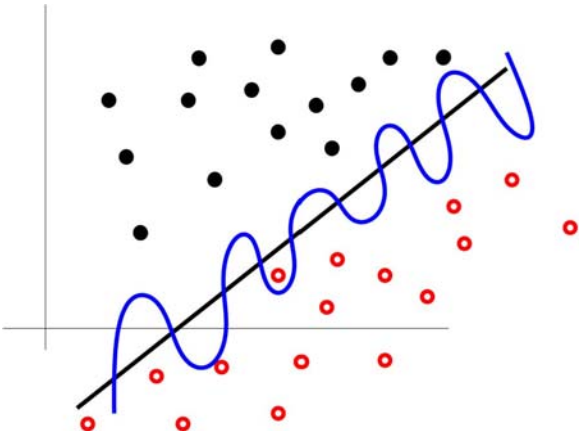

Figure 1.1: Two very different rules can make a similar classification.

Why do the generative and discriminative approaches lead to different results? There are two answers to this very important question which can be described from two different points of view: technical and philosophical (conceptual).

1.2.2 The Black Box Model

One can describe the pattern recognition problem as follows. There exists a black box BB that when given an input vector x_i returns an output y_i which can take only two values $y_i \in \{-1, +1\}$. The problem is: given the pairs $(y_i, x_i), i = 1, \ldots, \ell$ (the training data) find a function that approximates the rule that the black box uses.

Two different concepts of what is meant by a *good approximation* are possible:

(1) A good approximation of the BB rule is a function that is close (in a metric of functional space) to the function that the BB uses. (In the classical setting often we assume that the BB uses the Bayesian rule.)

(2) A good approximation of the BB rule is a function that provides approximately the same error rate as the one that the BB uses (provides the rule that predicts the outcomes of the BB well).

In other words, in the first case one uses a concept of closeness in the sense of being close to the *true function* used by the BB (closeness in a metric space of functions), while in the second case one uses a concept of closeness in the sense of being close to the accuracy of prediction (closeness in *functionals*). These definitions are very different.

In Figure 1.1 there are two different categories of data separated by two different rules. Suppose that the straight line is the function used by the black box. Then from the point of view of function estimation, the polynomial curve shown in Figure 1.1 is

very different from the line and therefore cannot be a good estimate of the *true BB* rule. From the other point of view, the polynomial rule separates the data well (and as we will show later can belong to a set with small VC dimension) and therefore can be a good *instrument* for prediction.

The lesson the Perceptron teaches us is that sometimes it is useful to give up the ambitious goal of estimating the rule the *BB* uses (the generative model of induction). Why?

Before discussing this question let me make the following remark. The problem of pattern recognition can be regarded as a generalization problem: using a set of data (observations) find a function[5] (theory). The same goals (but in more complicated situations) arise in the classical model of science: using observation of nature find the law. One can consider the pattern recognition problem as the simplest model of generalization where observations are just a set of i.i.d. vectors and the admissible laws are just a set of indicator functions. Therefore it is very useful to apply the ideas described in the general philosophy of induction to its simplest model and vice versa, to understand the ideas that appear in our particular model in the general terms of the classical philosophy. Later we will see that these interpretations are nontrivial.

1.2.3 Realism and Instrumentalism in the Philosophy of Science

The philosophy of science has two different points of view on the goals and the results of scientific activities.

(1) There is a group of philosophers who believe that the results of scientific discovery are the real laws that exist in nature. These philosophers are called the *realists*.

(2) There is another group of philosophers who believe the laws that are discovered by scientists are just an instrument to make a good prediction. The discovered laws can be very different from the ones that exist in Nature. These philosophers are called the *instrumentalists*.

The two types of approximations defined by classical discriminant analysis (using the generative model of data) and by statistical learning theory (using the function that explains the data best) reflect the positions of realists and instrumentalists in our simple model of the philosophy of generalization, the pattern recognition model. Later we will see that the position of philosophical instrumentalism played a crucial role in the success that pattern recognition technology has achieved.

However, to explain why this is so we must first discuss the theory of ill-posed problems, which in many respects describes the relationship between realism and instrumentalism in very clearly defined situations.

[5]The pattern recognition problem can be considered as the simplest generalization problem, since one has to find the function in a set of admissible *indicator* functions (that can take only two values, say 1 and -1.

1.3 Regularization and Structural Risk Minimization

1.3.1 Regularization of Ill-Posed Problems

In the beginning of the 1900s, Hadamard discovered a new mathematical phenomenon. He discovered that there are continuous operators A that map, in a one-to-one manner, elements of a space f to elements of a space F, but the inverse operator A^{-1} from the space F to the space f can be discontinuous. This means that there are operator equations

$$Af = F \tag{1.1}$$

whose solution in the set of functions $f \in \Phi$ exists, and is unique, but is unstable. (See Chapter 1 of EDED). That is, a small deviation $F + \Delta F$ of the (known) right-hand side of the equation can lead to a big deviation in the solution. Hadamard thought that this was just a mathematical phenomenon that could never appear in real-life problems. However, it was soon discovered that many important practical problems are described by such equations.

In particular, the problem of solving some types of linear operator equations (for example, Fredholm's integral equation of the second order) are ill-posed (see Chapter 1, Section 5 of *EDBED*). It was shown that many geophysical problems require solving (ill-posed) integral equations whose right-hand side is obtained from measurements (and therefore is not very accurate).

For us it is important that ill-posed problems can occur when one tries to estimate *unknown reasons from observed consequences*.

In 1943 an important step in understanding the structure of ill-posed problems was made. Tikhonov proved the so-called inverse operator lemma:

> *Let A be a continuous one-to-one operator from E_1 to E_2. Then the inverse operator A^{-1} defined on the images F of a compact set $f \in \Phi^*$ is stable.*

This means that if one possesses very strong prior knowledge about the solution (it belongs to a known compact set of functions), then it is possible to solve the equation. It took another 20 years before this lemma was transformed into specific approaches for solving ill-posed problems.

In 1962 Ivanov [21] suggested the following idea of solving operator equation (1.1). Consider the functional $\Omega(f) \geq 0$ that possesses the following two properties

(1) For any $c \geq 0$ the set of functions satisfying the constraint

$$\Omega(f) \leq c \tag{1.2}$$

is convex and compact.

(2) The solution f_0 of Equation (1.1) belongs to some compact set

$$\Omega(f_0) \leq c_0 \tag{1.3}$$

(where the constant $c_0 > 0$ may be unknown).

Under these conditions Ivanov proved that there exists a strategy for choosing $c = c(\varepsilon)$ depending on the accuracy of the right-hand side $||\Delta F||_{E_2} \leq \varepsilon$ such that the sequence of minima of the functional

$$R = ||Af - F||_{E_2} \tag{1.4}$$

subject to the constraints

$$\Omega(f) \leq c(\varepsilon) \tag{1.5}$$

converges to the solution of the ill-posed problem (1.1) as ε approaches zero.

In 1963 Tikhonov [55] proved the equivalent theorem that states: under conditions (1.2) and (1.3) defined on the functional $\Omega(f)$, there exists a function $\gamma_\varepsilon = \gamma(\varepsilon)$ such that the sequence of minima of the functionals

$$R_\gamma(f) = ||Af - F_\varepsilon||^2_{E_2} + \gamma_\varepsilon \Omega(f) \tag{1.6}$$

converges to the solution of the operator equation (1.1) as ε approaches zero.[6]

Both these results can be regarded as "*comforting ones*" since for any ε (even very small) one can guarantee nothing (the theorems guarantee only convergence of the sequence of solutions).

Therefore, one should try to avoid solving ill-posed problems by replacing them (if possible) with well-posed problems.

Keeping in mind the structure of ill-posed problems our problem of finding the *BB* solution can be split into two stages:

(1) Among a given set of admissible functions find a subset of functions that provides an expected loss that is close to the minimal one.

(2) Among functions that provide a small expected loss find one that is close to the *BB* function.

The first stage does not lead to an ill-posed problem, but the second stage might (if the corresponding operator is unstable).

The realist view requires solving both stages of the problem, while the instrumentalist view requires solving only the first stage and choosing for prediction any function that belongs to the set of functions obtained.

Technically, ill-posed problems appear in classical discriminant analysis as soon as one connects the construction of a discriminant function with the density estimation problem.

By definition, the density (if it exists) is a solution of the following equation

$$\int_a^x p(x')dx' = F(x), \tag{1.7}$$

[6]There is one more equivalent idea of how to solve ill-posed problems proposed in 1962 by Phillips [166]: minimize the functional $\Omega(f)$ satisfying the conditions defined above subject to the constraints

$$||Af - F||^2 \leq \varepsilon.$$

where $F(x)$ is a cumulative distribution function.

Therefore to estimate the density from the data

$$x_1, \ldots, x_\ell$$

means to solve Fredholm's equation (1.7) when the cumulative distribution function $F(x)$ is unknown but the data are given. One can construct an approximation to the unknown cumulative distribution function and use it as the right hand side of the equation. For example, one can construct the empirical distribution function

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \theta(x - x_i), \tag{1.8}$$

where

$$\theta(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ 0 & \text{if } u < 0 \end{cases}.$$

It is known from Kolmogorov's bound for the Glivenco–Cantelli theorem that the empirical distribution function converges exponentially fast (not only asymptotically but for any set of fixed observations) to the desired cumulative distribution function. Using the empirical distribution function constructed from the data, one can try to solve this equation.

Note that this setting of the density estimation problem cannot be avoided since it reflects the definition of the density. Therefore in both parametric or nonparametric statistics, one has to solve this equation. The only difference is how the set of functions in which one is looking for the solution is defined: in a "narrow set of parametric functions" or in a "wide set of non-parametric functions".[7]

However, this point of view was not clearly developed in the framework of classical statistics, since both theories (parametric and nonparametric) of density estimation were constructed *before* the theory of solving ill-posed problems was introduced.

The general setting of the density estimation problem was described for the first time in *EDBED*. Later in Chapter 2, Section 2.3 when we discuss the SVM method, we will consider a pattern recognition problem, and show the difference between the solutions obtained by nonparametric statistics (based on the philosophy of realism) and by an SVM solution (based on the philosophy of instrumentalism).

Regularization Techniques

The regularization theory as introduced by Tikhonov suggests minimizing the equation

$$R_\gamma(f) = ||Af - F_\varepsilon||^2_{E_2} + \gamma_\varepsilon \Omega(f). \tag{1.9}$$

Under very specific requirements on the set of functions defined both by the functional $\Omega(f)$ and the value $c > 0$

$$\Omega(f) \leq c \tag{1.10}$$

[7]The maximum likelihood method suggested by Fisher is valid just for a very narrow admissible set of functions. It is already invalid, for example, for the set of densities defined by the sum of two Gaussians with unknown parameters (see example [139], Section 1.7.4.)

(for any $c > 0$ the set should be *convex and compact*), and under the condition that the desired solution *belongs to the set with some fixed* c_0, it is possible to define a strategy of choosing the values of the parameter γ that asymptotically lead to the solution.

1.3.2 Structural Risk Minimization

The Structural Risk Minimization (SRM) principle generalizes the Ivanov scheme in two ways:

(1) It considers a structure on any sets of functions (not necessarily defined by inequality (1.5)).

(2) It does not require compactness or convexity on the set of functions that define the element of the structure. It also does not require the desired solution belonging to one of the elements of the structure.

The only requirement is that every element of the nested sets possesses a finite VC dimension (or other capacity factor).

Under these general conditions the risks provided by functions that minimize the VC bound converge to the smallest possible risk (even if the desired function belongs to the closure of the elements). Also, for any fixed number of observations it defines the smallest guaranteed risk.

In the early 1970s Chervonenkis and I introduced SRM for sets of indicator functions (used in solving pattern recognition problems) [13]. In *EDBED* the SRM principle was generalized for sets of real-valued functions (used in solving regression estimation problems).

Therefore the difference between regularization and structural risk minimization can be described as follows.

Regularization was introduced for solving ill-posed problems. It requires strong knowledge about the problem to be solved (the solution has to belong to the compact (1.10) defined by some constant c) and (generally speaking) does not have guaranteed bounds for a finite number of observations.

Structural risk minimization was introduced for solving predictive problems. It is more general (does not require strong restrictions of admissible set of functions) and has a guaranteed bound for a finite number of observations.

Therefore if the regularization method is the main instrument for solving ill-posed problems using the *philosophical realism* approach, then the structural risk minimization method is the main instrument for solving problems using the *philosophical instrumentalism* approach.

Remark. In the late 1990s the concept of regularization started to be used in the general framework of minimizing the functionals (1.9) to solve predictive generalization problems. The idea was that under any definition of the functional $\Omega(f)$ there exists a parameter γ which leads to convergence to the desired result. This is, however,

incorrect: first, it depends on the concept of convergence; second, there are functionals (for which the set of functions (1.5) can violate finiteness of capacity conditions) that do not lead to convergence in any sense.

1.4 The Beginning of the Split Between Classical Statistics and Statistical Learning Theory

The philosophy described above was more or less clear by the end of the 1960s.[8] By that time there was no doubt that in analyzing the pattern recognition problem we came up with a new direction in the theory of generalization. The only question that remained was how to describe this new direction. Is this a new branch of science or is it a further development in classical statistics? This question was the subject of discussions in the seminars at the Institute of Control Sciences of the Academy of Sciences of USSR (Moscow).

The formal decision, however, was made when it came time to publish these results in the *Reports of Academy of Sciences of USSR* [143]. The problem was in which section of *Reports* it should be published — in "Control Sciences (Cybernetics)" or in "Statistics". It was published as a contribution in the "Control Sciences" section.

This is how one of the leading statisticians of the time, Boris Gnedenco, explained why it should not be published in the "Statistics" section:

> It is true that this theory came from the same roots and uses the same formal tools as statistics. However, to belong to the statistical branch of science this is not enough. Much more important is to share the same belief in the models and to share the same philosophy. Whatever you are suggesting is not in the spirit of what I am doing or what A. Kolmogorov is doing. It is not what our students are doing nor will it be what the students of our students do. Therefore, you must have your own students, develop your own philosophy, and create your own community.

More than 35 years have passed since this conversation. The more time passed, the more impressed I became with Gnedenco's judgment. The next three decades (1970s, 1980s, and 1990s) were crucial for developments in statistics. After the shocking discovery that the classical approach suffers from the curse of dimensionality, statisticians tried to find methods that could replace classical methods in solving real-life problems. During this time statistics was split into two very different parts: theoretical statistics that continued to develop the classical paradigm of generative models, and applied statistics that suggested a compromise between theoretical justification of the algorithms and heuristic approaches to solving real-life problems. They tried to justify such a position by inventing special names for such activities (exploratory data analysis), where in fact the superiority of common sense over theoretical justification was declared. However, they never tried to construct or justify new algorithms using VC

[8] It was the content of my first book *Pattern Recognition Problem* published in 1971 (in Russian).

theory. Only after SVM technology became a dominant force in data mining methods did they start to use its technical ideas (but not its philosophy) to modify classical algorithms.[9]

Statistical learning theory found its home in computer science. In particular, one of the most advanced institutions where SLT was developing in the 1970s and 1980s was the Institute of Control Sciences of the Academy of Sciences of USSR. Three different groups, each with different points of view on the generalization problem, became involved in such research: the Aizerman–Braverman–Rozonoer's group, the Tsypkin group, and the Vapnik–Chervonenkis group.

Of these groups ours was the youngest: I just got my PhD (candidate of science) thesis, and Chervonenkis got his several years later. Even so, our research direction was considered one of the most promising. In order to create a VC community I was granted permission from the Academy of Sciences to have my own PhD students.[10]

From this beginning we developed a statistical learning community. I had several very strong students including Tamara Glaskov, Anatoli Mikhalsky, Anatoli Stehanuyk, Alexander Sterin, Fellix Aidu, Sergey Kulikov, Natalia Markovich, Ada Sorin, and Alla Juravel who developed both machine learning theory and effective machine learning algorithms applied to geology and medicine.

By the end of the 1960s my department head, Alexander Lerner, made an extremely important advance in the application of machine learning: he convinced the high-level bureaucrats to create a laboratory for the application of machine learning techniques in medicine.

In 1970 such a laboratory was created in the State Oncology Centre. The director of the laboratory was my former PhD student, Tamara Glaskov.

It is hard to overestimate how much this laboratory accomplished during this time. Only recently have the most advanced oncology hospitals in the West created groups to analyze clinical data. This was routine in USSR decades earlier.

In beginning of the 1970s I prepared my doctoral thesis.

1.5 The Story Behind This Book

Government control under the Soviet Communist regime was total. One of its main modus operandi was to control who was promoted into more or less prominent positions. From the government bureaucrat's perspective a scientific degree (and especially a doctoral degree) holder possessed influence, and therefore they wanted to control who obtained this degree.

The execution of such control was one of the obligations of the institution called

[9] Statisticians did not recognise conceptual aspects of VC theory. Their criticism of this theory before SVM was that the VC bounds were too loose to be useful. Therefore the theory is not practical and to create new methods it is better to use common sense than the results of this theory.

[10] In the Russian system there were two academic degrees: *candidate of science* (which is equivalent to the PhD degree in the United States) and *doctor of science* (which is equivalent to the *Habilitation a Diriger des Recherches (HDR)* in France). Normally only doctors of science could have PhD students. I was granted this privilege and had to defend my doctoral thesis soon.

the Supreme Certifying Commission[11] (SCC) closely related to the KGB. The rule was that any decision on any thesis defense made by any Scientific Councils anywhere in the country must be approved by this commission. If the SCC disapproved several decisions by a particular Scientific Council it could be dismissed. Therefore the normal policy of academic institutions was not to enter into conflict with the SCC.

From the KGB's point of view I was a wrong person to obtain the doctoral level: I was not a member of the Communist Party, I was Jewish, my PhD adviser, Alexander Lerner, had applied for immigration to Israel and became a "refusenik," some of my friends were dissidents, and so on.

In this situation everybody understood that the Institute would be in conflict with the SCC's mandate. Nevertheless the feeling was that the support of the scientific community would be so strong that the SCC would not start the battle.

The SCC, however, reacted with a trick that to my knowledge was never used before: it requested that the Scientific Council change one of the reviewers to their trusted man who did his job: wrote a negative review.

I had a long conversation with the Chairman of the Scientific Council, Yakov Tsypkin, after he discussed the situation with the members of the Council. He told me that everyone on the Scientific Council understood what was going on and if I decided to defend my thesis the Scientific Council would unanimously support me. However, I had no chance of being approved by the SCC since they would have a formal reason to reject my thesis. Also they would have a formal reason to express distrust of the Scientific Council of the Institute. In this situation the best solution was to withdraw my thesis and publish it as a book. However, since the names of the authors of books were also under the KGB's control (the authors should also be "good guys") I would only be able to publish the book if my name did not attract too much attention. This would allow the editor, Vladimir Levantovsky (who was familiar with this story), to successfully carry out all necessary procedures to obtain permission (from the institution that controls the press) to publish the book.

So, I withdrew my thesis, rewrote it as a book, and due to the strong support of many scientists (especially Tsypkin), the editor Levantovsky was able to publish it (in Russian) in 1979.

In 1982 the well known American statistician, S. Kotz, translated it into English under the title *Estimation of Dependencies Based on Empirical Data* which was published by Springer. The first part of this volume is its reprint.

The main message that I tried to deliver in the book was that classical statistics could not overcome the curse of dimensionality but the new approach could. I devoted three chapters of the book to different classical approaches and demonstrated that none of them could overcome the curse of dimensionality. Only after that did I describe the new theory.

[11] The Russian abbreviation is VAK.