This last and biggest part of the book consists of thirteen contributions from leading experts in machine learning. They range from reviews of wide areas of research to more specialized papers oriented mainly toward obtaining new results.

The opening chapter, Chap. 13, is by Alexey Chervonenkis. It follows closely the slides of his talk at the symposium “Measures of Complexity” (see Fig. IV.1). The author both summarizes classical results of VC theory and reports his recent results on learnability in the case of infinite VC dimension.

The following chapter, Chap. 14, is by R.M. Dudley. In part it is based on his talk at “Measures of Complexity” (see Fig. IV.2). The main topic of this chapter is the extension of VC theory from uniform laws of large numbers to central limit theorems that hold uniformly over suitable classes of functions. The chapter also contains a brief review of different ways of generalizing the notion of VC dimension to function classes.

Chapter 15, by Alexander Rakhlin and Karthik Sridharan, deals with non-i.i.d. (namely, martingale) generalizations of VC theory. Similarly to Chap. 13, this

Fig. IV.1 Alexey Chervonenkis delivering his 50-minute talk at the “Measures of Complexity” symposium
Chapter again concentrates on uniform laws of large numbers. The main applications of this chapter’s results are in the theory of on-line learning.

Chapter 16 by Ingo Steinwart is a review of different techniques used in the analysis of the ERM principle. It pays special attention to the measures of complexity of function classes that have proved to be useful in this analysis.

In Chap. 17 Nikolay Vereshchagin and Alexander Shen review algorithmic statistics, an approach to the theory of statistics based on Kolmogorov complexity.

Bernhard Schölkopf’s talk at the “Measures of Complexity” symposium (“Causal inference and statistical learning”; see Fig. IV.3) introduced the main ideas of causal inference from the point of view of machine learning and discussed machine-learning implications of causal knowledge. Chapter 18 of the book, by Dominik Janzing, Bastian Steudel, Naji Shajarisales, and Bernhard Schölkopf, is devoted to a new approach to causal inference, Information-Geometric Causal Inference, that can be used to distinguish between cause and effect for two variables.
The following chapter, Chap. 19, by Vladimir Cherkassky and Sauptik Dhar, discusses ways of making the output of an SVM more interpretable. After reviewing existing methods and pointing out their potential weaknesses, the authors propose a graphical technique for understanding decisions made by SVM classifiers.

Chapter 20 by Olivier Catoni is devoted to the PAC-Bayesian approach to statistical learning theory. It synthesises two existing PAC-Bayesian methods (by Seeger and by Catoni) and applies the synthesis to deriving performance guarantees for support vector machines.

A concept class is maximum if Sauer’s lemma holds for it with equality. Chapter 21 by Hyam Rubinstein, Benjamin Rubinstein, and Peter Bartlett gives a new characterization of maximum classes. Motivated by the Sample Compression Hypothesis, the authors apply this characterization to study the possibility of embedding concept classes of finite VC dimension into maximum classes without a significant increase in their VC dimension.

Chapter 22 by László Györfi and Harro Walk is devoted to nonparametric hypothesis testing. We are given two probability measures $\nu_1$ and $\nu_2$ and know that the true data-generating distribution $\mu$ is closer to one of these than to the other. It turns out that, almost surely, we can figure out whether $\mu$ is closer to $\nu_1$ or $\nu_2$ making only finitely many errors in the usual i.i.d. scenario.

In Chap. 23, Ran El-Yaniv and Yair Wiener study Version Space Compression Set Size, which can be regarded as a measure of complexity of a data set given a concept class. The chapter reviews known properties of this measure of complexity and applies it to selective prediction and active learning.

Chapter 24 by Andreas Maurer, Massimiliano Pontil, and Luca Baldassare gives negative results for the problem of sparse coding, i.e., the problem of approximating a random vector in a high-dimensional linear space by a sparse linear combination of dictionary vectors. It turns out that the quality of approximation will be poor unless the data-generating distribution is concentrated in a low-dimensional subspace; the authors quantify this phenomenon and discuss its implications.

Chapter 25 by Asaf Noy and Koby Crammer is another chapter devoted to the PAC-Bayesian approach. It combines the PAC-Bayesian approach with robust methods based on the Laplace distribution, applies these ideas to boosting, and reports very encouraging experimental results.