Chapter1

Introduction

Inventorshavelongdreamedofcreatingmachinesthatthink.Thisdesiredates backtoatleastthetimeofancientGreece.ThemythicalfiguresPygmalion, Daedalus,andHephaestusmayallbeinterpretedaslegendaryinventors,and Galatea,Talos,andPandoramayallberegardedasartificiallife (OvidandMartin , 2004Sparkes; , 1996Tandy; , 1997).

Whenprogrammablecomputerswerefirstconceived,peoplewonderedwhether suchmachinesmightbecomeintelligent,overahundredyearsbeforeonewas built (Lovelace, 1842).Today, **artificialintelligence** (AI)isathrivingfieldwith manypracticalapplicationsandactiveresearchtopics.Welooktointelligent softwaretoautomateroutinelabor,understandspeechorimages,makediagnoses inmedicineandsupportbasicscientificresearch.

Intheearlydaysofartificialintelligence,thefieldrapidlytackledandsolved problemsthatareintellectuallydifficultforhumanbeingsbutrelativelystraight-forwardforcomputers—problemsthatcanbedescribedbyalistofformal,math-ematicalrules. Thetruechallengetoartificialintelligenceprovedtobesolving thetasksthatareeasyforpeopletoperformbuthardforpeopletodescribe formally—problemsthatwesolveintuitively,thatfeelautomatic,likerecognizing spokenwordsorfacesinimages.

Thisbookisaboutasolutiontothesemoreintuitiveproblems.Thissolutionis toallowcomputerstolearnfromexperienceandunderstandtheworldintermsof ahierarchyofconcepts,witheachconceptdefinedthroughitsrelationtosimpler concepts.Bygatheringknowledgefromexperience,thisapproachavoidstheneed forhumanoperatorstoformallyspecifyalltheknowledgethatthecomputerneeds. Thehierarchyofconceptsenablesthecomputertolearncomplicatedconceptsby buildingthemoutofsimplerones.Ifwedrawagraphshowinghowtheseconcepts

第一章

介绍

发明家们一直梦想着创造出能够思考的机器，这一愿望至少可以追溯到古希腊时期，神话人物皮格马利翁、代达罗斯和赫菲斯托斯都可能被视为传奇发明家，而伽拉忒亚、塔洛斯和潘多拉都可能被视为人工生命（Ovid and Martin 2004 Sparkes 1996 Tandy 1997； ， ； ， ）。

当可编程计算机首次被设想出来时，人们想知道这种机器是否会变得智能，这比实际制造出一台机器早了一百多年（Lovelace 1842）。今天，人工智能（AI）是一个蓬勃发展的领域，拥有许多实际应用和活跃的研究课题。我们期待智能软件能够自动化日常劳动、理解语音或图像、进行医学诊断并支持基础科学研究。

在人工智能发展的早期，该领域迅速处理和解决了那些对人类来说在智力上很困难但对计算机来说相对简单的问题——这些问题可以用一系列正式的数学规则来描述。事实证明，人工智能面临的真正挑战是解决那些对人类来说很容易执行但难以正式描述的任务——我们可以直观地、自动地解决这些问题，比如识别口头单词或图像中的人脸。

本书旨在解决这些更直观的问题。该解决方案旨在让计算机从经验中学习，并根据概念的层级结构理解世界，每个概念都通过其与更简单概念的关系来定义。通过从经验中收集知识，这种方法避免了人类操作员正式指定计算机所需的所有知识。概念的层级结构使计算机能够通过从简单概念构建复杂概念来学习它们。如果我们画一个图表来展示这些概念如何

arebuiltontopofeachother,thegraphisdeep,withmanylayers.Forthisreason, wecallthisapproachtoAI **deeplearning** .

ManyoftheearlysuccessesofAItookplaceinrelativelysterileandformal environmentsanddidnotrequirecomputerstohavemuchknowledgeabout theworld.Forexample,IBM'sDeepBluechess-playingsystemdefeatedworld championGarryKasparovin1997 (Hsu, 2002).Chessisofcourseaverysimple world,containingonlysixty-fourlocationsandthirty-twopiecesthatcanmove inonlyrigidlycircumscribedways.Devisingasuccessfulchessstrategyis a tremendousaccomplishment, butthechallengeisnotduetothedifficultyof describingthesetofchesspiecesandallowablemovestothecomputer.Chess canbecompletelydescribedbyaverybrieflistofcompletelyformalrules,easily providedaheadoftimebytheprogrammer.

Ironically,abstractandformaltasksthatareamongthemostdifficultmental undertakingsforahumanbeingareamongtheeasiestforacomputer.Computers havelongbeenabletodefeateventhebesthumanchessplayerbutonlyrecently havebegunmatchingsomeoftheabilitiesofaveragehumanbeingstorecognize objects orspeech.A person's everyday life requires an immenseamount of knowledgeabouttheworld.Muchofthisknowledgeissubjectiveandintuitive, andthereforedifficulttoarticulateinaformalway.Computersneedtocapture thissameknowledgeinordertobehaveinanintelligentway.Oneofthekey challengesinartificialintelligenceishowtogetthisinformalknowledgeintoa computer.

Severalartificialintelligenceprojectshavesoughttohard-codeknowledge abouttheworldinformallanguages.Acomputercanreasonautomaticallyabout statementsintheseformallanguagesusinglogicalinferencerules.Thisisknownas the **knowledgebase** approachtoartificialintelligence.Noneoftheseprojectshas ledtoamajorsuccess.Oneofthemostfamoussuchprojectsis Cyc (Lenatand Guha, 1989).Cycisaninferenceengineandadatabaseofstatementsinalanguage calledCycL.Thesestatementsareenteredbyastaffofhumansupervisors.Itisan unwieldyprocess.Peoplestruggletodeviseformalruleswithenoughcomplexity toaccuratelydescribetheworld.Forexample,Cycfailedtounderstandastory aboutapersonnamedFredshavinginthemorning (Linde, 1992).Itsinference enginedetectedaninconsistencyinthestory: itknewthatpeopledonothave electricalparts,butbecauseFredwasholdinganelectricrazor,itbelievedthe entity"FredWhileShaving"containedelectricalparts.Itthereforeaskedwhether Fredwasstillapersonwhilehewasshaving.

Thedifficultiesfacedbysystemsrelyingonhard-codedknowledgesuggest thatAIsystemsneedtheabilitytoacquiretheirownknowledge,byextracting

彼此叠加，图形很深，有很多层。因此，我们将这种人工智能方法称为深度学习。人工智能的许多早期成功都发生在相对贫瘠和正式的环境中，并且不需要计算机对世界有太多了解。例如，IBM 的 DeepBlue 国际象棋系统在 1997 年击败了世界冠军加里卡斯帕罗夫（,）。国际象棋当然是一个非常简单的世界，只包含 64 个位置和 32 个棋子，只能以严格限定的方式移动。设计一个成功的国际象棋策略是一项巨大的成就，但挑战并不在于难以向计算机描述棋子集和允许的移动。国际象棋可以通过非常简短的完全正式规则列表轻松描述由程序员提前提供。

讽刺的是，对于人类来说最困难的脑力劳动中的抽象和正式任务对于计算机来说是最容易的。计算机早已能够击败最好的人类棋手，但直到最近才开始匹配普通人类的一些能力，例如识别物体或语音。一个人的日常生活需要大量关于世界的知识。这些知识大部分是主观和直觉的，因此很难以正式的方式表达。计算机需要捕捉这些相同的知识才能以智能的方式运行。人工智能的关键挑战之一是如何将这些非正式知识输入计算机。

一些人工智能项目试图将关于世界的知识硬编码到非正式语言中。计算机可以使用逻辑推理规则自动推理这些正式语言中的语句。这被称为人工智能的知识库方法。这些项目都没有取得重大成功。最著名的此类项目之一是 Cyc (Lenatand Guha 1989,)。Cyc 是一个推理引擎和名为 CycL 的语言语句数据库。这些语句由人类主管人员输入。这是一个笨拙的过程。人们努力设计足够复杂的正式规则来准确描述世界。例如，Cyc 未能理解一个名叫 Fred 的人早上刮胡子的故事（ , ）。它的推理 Linde 1992 引擎检测到故事中的不一致之处：它知道人没有电气部件，但是因为弗雷德拿着一把电动剃须刀，它认为实体“弗雷德在剃须时”包含电气部件。因此它询问弗雷德在剃须时是否仍然是一个人。

依赖硬编码知识的系统所面临的困难表明，人工智能系统需要能够通过提取

patternsfromrawdata.Thiscapabilityisknownas **machinelearning** .The introductionofmachinelearningenabledcomputerstotackleproblemsinvolving knowledgeoftherealworldandmakedecisionsthatappearsubjective.Asimple machinelearningalgorithmcalled **logisticregression** candeterminewhetherto recommendcesareandelivery (Mor-Yosef *etal.* , 1990).Asimplemachinelearning algorithmcalled **naiveBayes** canseparatelegitimatee-mailfromspame-mail.

Theperformanceofthesesimplemachinelearningalgorithmsdependsheavily onthe **representation** ofthedatatheyaregiven.Forexample,whenlogistic regressionisusedtorecommendcesareandelivery,theAIsystemdoesnotexamine thepatientdirectly.Instead,thedoctortellsthesystemseveralpiecesofrelevant information,suchasthepresenceorabsenceofauterinescar.Eachpieceof informationincludedintherepresentationofthepatientisknownasa **feature**. Logisticregressionlearnshoweachofthesefeaturesofthepatientcorrelateswith variousoutcomes. However,itcannotinfluencehowfeaturesaredefinedinany way.Iflogisticregressionweregivenan MRIscanofthepatient,ratherthan thedoctor'sformalizedreport,itwouldnotbeabletomakeusefulpredictions. IndividualpixelsinanMRIscanhavenegligiblecorrelationwithanycomplications thatmightoccurduringdelivery.

Thisdependenceonrepresentationsisageneralphenomenonthatappears throughoutcomputerscienceandevendailylife.Incomputerscience,operations suchassearchingacollectionofdatacanproceedexponentiallyfasterifthecollection isstructuredandindexedintelligently.Peoplecaneasilyperformarithmetic onArabicnumeralsbutfindarithmeticonRomannumeralsmuchmoretime consuming.Itisnotsurprisingthatthechoiceofrepresentationhasanenormous effectontheperformanceofmachinelearningalgorithms.Forasimplevisual example,seefigure 1.1.

Manyartificialintelligencetaskscanbesolvedbydesigningtherightsetof featurestoextractforthattask,thenprovidingthesefeaturestoasimplemachine learningalgorithm.Forexample,ausefulfeatureforspeakeridentificationfrom soundisanestimateofthesizeofthespeaker'svocaltract.Thisfeaturegivesa strongclueastowhetherthespeakerisaman,woman,orchild.

For manytasks, however, it isdifficult toknowwhat featuresshould be extracted.Forexample,supposethatwewouldliketowriteaprogramtodetect carsinphotographs.Weknowthatcarshavewheels,sowemightliketousethe presenceofawheelasafeature.Unfortunately,itisdifficulttodescribeexactly whatawheellookslikeintermsofpixelvalues.Awheelhasasimplegeometric shape,butitsimagemaybecomplicatedbyshadowsfallingonthewheel,thesun glaringoffthemetalpartsofthewheel,thefenderofthecaroranobjectinthe

模式来自原始数据。这种能力被称为机器学习。机器学习的引入使计算机能够解决涉及现实世界知识的问题并做出看似主观的决策。一种称为逻辑回归的简单机器学习算法可以确定是否建议剖腹产和分娩（Mor-Yosef 1990etal.,）。一种称为朴素贝叶斯的简单机器学习算法可以将合法电子邮件与垃圾邮件区分开来。

这些简单的机器学习算法的性能在很大程度上取决于它们所给出的数据的表示。例如，当使用逻辑回归来推荐剖腹产时，人工智能系统不会直接检查病人。相反，医生会告诉系统几条相关信息，例如是否存在子宫疤痕。病人表示中包含的每条信息都称为一个特征。

逻辑回归可以显示患者的各项特征与各种结果之间的关联。然而，它无法以任何方式影响这些特征的定义。如果使用患者的 MRI 扫描结果而非医生的正式报告进行逻辑回归分析，则无法做出有用的预测。MRI 图像中的单个像素与分娩过程中可能发生的任何并发症之间的关联微乎其微。

这种对表示的依赖是整个计算机科学甚至日常生活中出现的普遍现象。在计算机科学中，如果数据集合是结构化的并且索引是智能的，那么搜索数据集合等操作可以以指数级的速度进行。人们可以轻松地对阿拉伯数字进行算术运算，但发现对罗马数字进行算术运算要耗费得多。表示的选择对机器学习算法的性能有很大的影响，这并不奇怪。对于一个简单的视觉示例，请参见图 1.1
许多人工智能任务可以通过设计正确的一组特征来提取，然后将这些特征提供给简单的机器学习算法来解决。例如，从声音中识别说话人的有用特征是估计说话人声道的大小。这个特征提供了一个强有力的线索来表明说话人是男人、女人还是孩子。

然而，对于许多任务来说，很难知道应该提取哪些特征。例如，假设我们想编写一个程序来检测照片中的汽车。我们知道汽车有轮子，所以我们可能想使用轮子的存在作为安全特征。不幸的是，很难用像素值来准确描述轮子的样子。轮子有简单的几何形状，但它的图像可能会因为落在轮子上的阴影、轮子金属部分上的阳光、汽车的挡泥板或车内的物体而变得复杂。

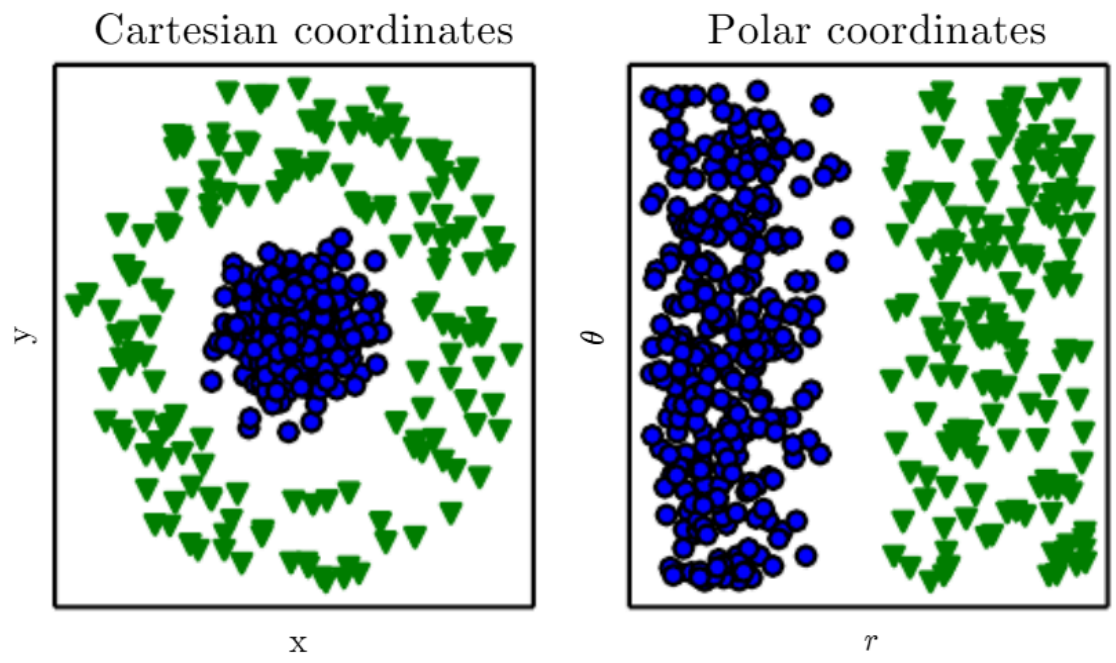

Figure1.1:Example ofdifferentrepresentations:supposewewanttoseparate two categoriesofdatabydrawingalinebetweentheminascatterplot.Intheplotontheleft, werepresentsomedatausingCartesiancoordinates,andthetaskisimpossible.Intheplot ontheright,werepresentthedatawithpolarcoordinatesandthetaskbecomessimpleto solvewithaverticalline.(Figureproducedincollaborationwith DavidWarde-Farley.)

foregroundobscuringpartofthewheel,andsoon.

Onesolutiontothisproblemistousemachinelearningtodiscovernotonly themappingfromrepresentationtooutputbutalsotherepresentationitself. Thisapproachisknownas **representationlearning** . Learnedrepresentations oftenresultinmuchbetterperformancethancanbeobtainedwithhand-designed representations.TheyalsoenableAIsystemstorapidlyadapttonewtasks,with minimalhumanintervention.Arepresentationlearningalgorithmcandiscovera goodsetoffeaturesforasimpletaskinminutes,orforacomplextaskinhoursto months.Manuallydesigningfeaturesforacomplextaskrequiresagreatdealof humantimeandeffort;itcantakedecadesforanentirecommunityofresearchers.

Thequintessentialexampleofarepresentationlearningalgorithmisthe **autoencoder** .Anautoencoderisthecombinationofan **encoder** function,which convertstheinputdataintoadifferentrepresentation,anda **decoder** function, whichconvertsthenewrepresentationbackintotheoriginalformat.Autoencoders aretrainedtopreserveasmuchinformationaspossiblewhenaninputisrun throughtheencoderandthenthedecoder,buttheyarealsotrainedtomakethe newrepresentationhavevariousniceproperties.Differentkindsofautoencoders aimtoachievedifferentkindsofproperties.

Whendesigningfeaturesoralgorithmsforlearningfeatures,ourgoalisusually toseparatethe **factorsofvariation** thatexplaintheobserveddata.Inthis

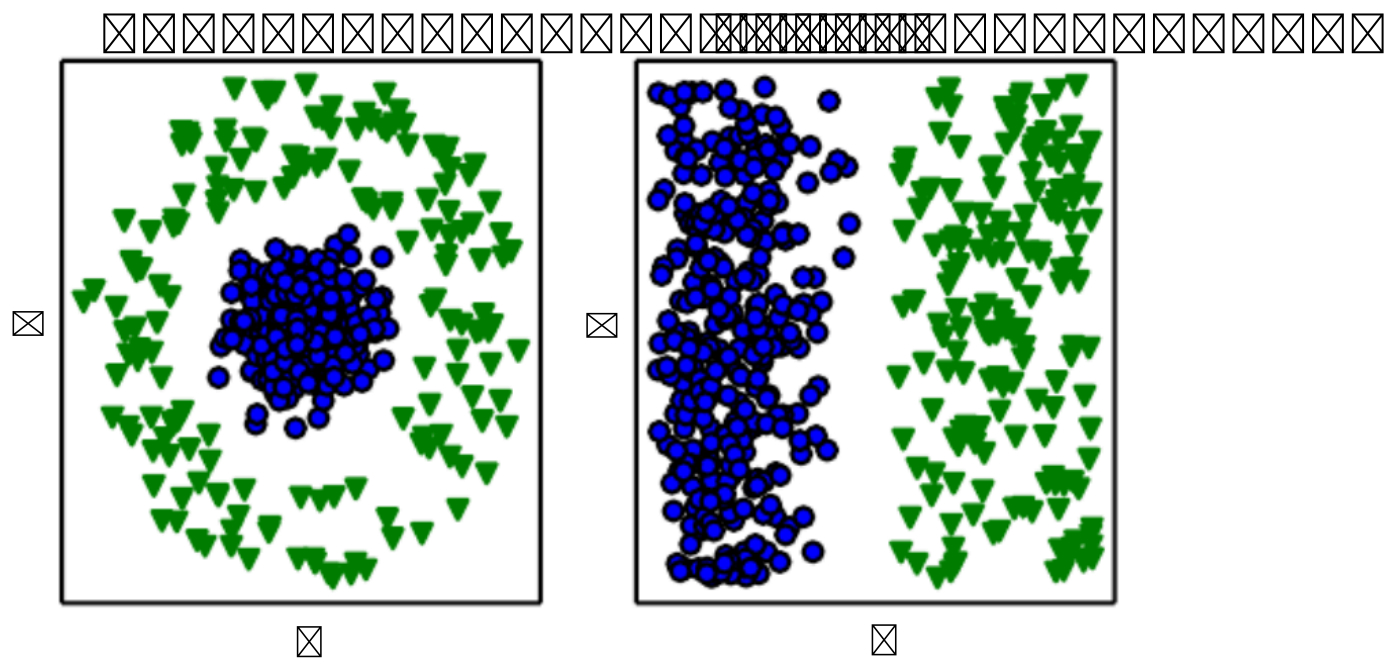

图 1.1：不同表示的示例：假设我们要通过在散点图之间画一条线来分离两类数据。在左侧的图中，我们使用笛卡尔坐标表示一些数据，这个任务是不可能完成的。在右侧的图中，我们用极坐标表示数据，用垂直线就可以很容易地解决这个任务。（该图是与 DavidWarde-Farley 合作制作的。）

前景遮挡了部分车轮，很快。

解决这个问题的一个方法是利用机器学习，不仅发现从表征到输出的映射，还发现表征本身。这种方法被称为表征学习。学习到的表征通常比手工设计的表征性能更好。它们还能使人工智能系统快速适应新任务，并最大限度地减少人工干预。表征学习算法可以在几分钟内为简单任务发现一组良好的特征，或者在几小时到几个月内为复杂任务发现一组良好的特征。手动设计复杂任务的特征需要大量的人力和精力；这可能需要整个研究群体数十年的时间。

演示学习算法的典型例子是自动编码器。自动编码器是编码器功能和解码器功能的组合，编码器功能将输入数据转换为不同的表示，解码器功能将新的表示转换回原始格式。自动编码器经过训练，在输入通过编码器和解码器时保留尽可能多的信息，但它们也经过训练，使新的表示具有各种良好的属性。不同类型的自动编码器旨在实现不同类型的属性。

在设计特征或学习特征的算法时，我们的目标通常是分离变异因素
解释观察到的数据。在此

context,weusetheword“factors”simplytorefertoseparatesourcesofinfluence; thefactorsareusuallynotcombinedbymultiplication.Suchfactorsareoftennot quantitiesthataredirectlyobserved.Instead,theymayexistaseitherunobserved objectsorunobservedforcesinthephysicalworldthataffectobservablequantities. Theymayalsoexistasconstructsinthehumanmindthatprovideusefulsimplifying explanationsorinferredcausesoftheobserveddata.Theycanbethoughtofas conceptsorabstractionsthathelpusmakesenseoftherichvariabilityinthedata. Whenanalyzingaspeechrecording,thefactorsofvariationincludethespeaker's age,theirsex,theiraccentandthewordstheyarespeaking.Whenanalyzingan imageofacar,thefactorsofvariationincludethepositionofthecar,itscolor, andtheangleandbrightnessofthesun.

Amajorsourceofdifficultyinmanyreal-worldartificialintelligenceapplications isthatmanyofthefactorsofvariationinfluenceeverysinglepieceofdataweare abletoobserve.Theindividualpixelsinanimageofaredcarmightbeveryclose toblackatnight.Theshapeofthecar'ssilhouettedependsontheviewingangle. Mostapplicationsrequireusto *disentangle* thefactorsofvariationanddiscardthe onesthatwedonotcareabout.

Ofcourse,itcanbeverydifficulttoextractsuchhigh-level,abstractfeatures fromrawdata.Manyofthesefactorsofvariation,suchasaspeaker'saccent, canbeidentifiedonlyusingsophisticated,nearlyhuman-levelunderstandingof thedata.Whenitisnearlyasdifficulttoobtainarepresentationastosolvethe originalproblem,representationlearningdoesnot,atfirstglance,seemtohelpus.

Deeplearning solvesthiscentralprobleminrepresentationlearningbyintroducingrepresentationsthatareexpressedintermsofother,simplerrepresentations. Deeplearningenablesthecomputertobuildcomplexconceptsoutofsimplerconcepts.Figure 1.2 showshowadeeplearningsystemcanrepresenttheconceptof animageofapersonbycombiningsimplerconcepts,suchascornersandcontours, whichareinturndefinedintermsofedges.

Thequintessentialexampleofadeeplearningmodelisthefeedforwarddeep network,or **multilayerperceptron** (MLP).Amultilayerperceptronisjusta mathematicalfunctionmappingsomesetofinputvaluestooutputvalues.The functionisformedbycomposingmanysimplerfunctions.Wecanthinkofeach applicationofadifferentmathematicalfunctionasprovidinganewrepresentation oftheinput.

Theideaoflearningtherightrepresentationforthedataprovidesoneperspectiveondeeplearning.Anotherperspectiveondeeplearningisthatdepth enablesthecomputertolearnamultistepcomputerprogram.Eachlayeroftherepresentationcanbethoughtofasthestateofthecomputer'smemoryafter

在本文中，我们使用“因素”一词仅指独立的影响源；这些因素通常不会通过乘法组合而成。这些因素通常不是直接观察到的量，而是存在于物理世界中影响可观察量的未观察到的物体或未观察到的力。它们也可能存在于人类思维中，为观察到的数据提供有用的简化解释或推断原因。它们可以被认为是概念或抽象概念，帮助我们理解数据中丰富的可变性。分析语音录音时，变化因素包括说话者的年龄、性别、口音和他们所说的词语。分析汽车图像时，变化因素包括汽车的位置、颜色以及太阳的角度和亮度。

许多现实世界人工智能应用中的一个主要困难来源是，许多变化因素会影响我们能够观察到的每一条数据。汽车图像中的各个像素在夜间可能非常接近黑色。汽车轮廓的形状取决于视角。

大多数应用程序都要求我们考虑变化的因素，
并丢弃那些我们不关心的因素。

当然，从原始数据中提取这种高级抽象特征是非常困难的。许多这样的变化因素，比如说话者的口音，只有使用复杂的、接近人类水平的数据理解才能识别。当获得一个表征几乎和解决原始问题一样困难时，表征学习乍一看似乎对我们没有帮助。

深度学习通过引入用其他更简单的表示形式来表达的表示形式，解决了表示学习中的这一核心问题。深度学习使计算机能够从简单的概念构建复杂的概念。如图所示，深度学习系统可以通过组合更简单的概念（例如角和轮廓，这些概念又用边来定义）来表示人物图像的概念。

深度学习模型的典型例子是前馈深度网络，或多层感知器（MLP）。多层感知器只是一个将一些输入值映射到输出值的数学函数。该函数由许多更简单的函数组合而成。我们可以将不同数学函数的每个应用视为提供输入的新表示。

学习数据的正确表示的想法为深度学习提供了一个视角。深度学习的另一个视角是深度使计算机能够学习多步骤的计算机程序。表示的每一层都可以被认为是计算机内存状态的快速变化。

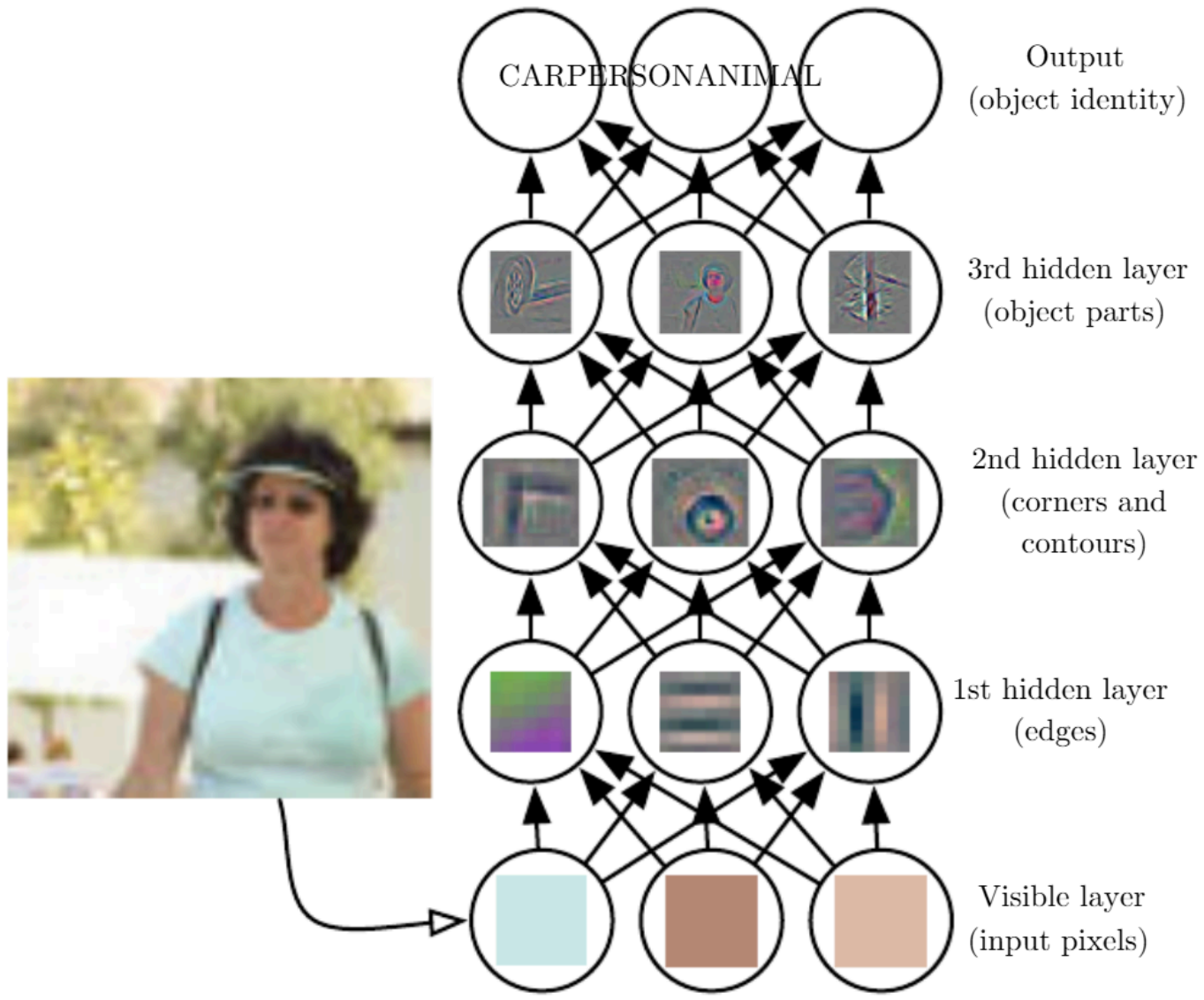

Figure1.2:Illustrationofadeeplearningmodel.Itisdifficultforacomputertounderstand themeaningofrawsensoryinputdata,suchasthisimagerepresentedasacollection ofpixelvalues.Thefunctionmappingfromasetofpixelstoanobjectidentityisvery complicated.Learningorevaluatingthismappingseemsinsurmountableiftackleddirectly. Deeplearningresolvesthisdifficultybybreakingthedesiredcomplicatedmappingintoa seriesofnestedsimplemappings,eachdescribedbyadifferentlayerofthemodel.The inputispresentedatthe **visiblelayer** ,sonamedbecauseitcontainsthevariablesthat weareabletoobserve.Thenaseriesof **hiddenlayers** extractsincreasinglyabstract featuresfromtheimage.Theselayersarecalled"hidden"becausetheirvaluesarenotgiven inthedata;insteadthemodelmustdeterminewhichconceptsareusefulforexplaining therelationshipsintheobserveddata.Theimagesherearevisualizationsofthekind offeaturerepresentedbyeachhiddenunit.Giventhepixels,thefirstlayercaneasily identifyedges,bycomparingthebrightnessofneighboringpixels.Giventhefirsthidden layer'sdescriptionoftheedges,thesecondhiddenlayercaneasilysearchforcornersand extendedcontours,whicharerecognizableascollectionsofedges.Giventhesecondhidden layer'sdescriptionoftheimageintermsofcornersandcontours,thethirdhiddenlayer candetectentirepartsofspecificobjects,byfindingspecificcollectionsofcontoursand corners.Finally,thisdescriptionoftheimageintermsoftheobjectpartsitcontainscan beusedtorecognizetheobjectspresentintheimage.Imagesreproducedwithpermission from ZeilerandFergus2014 ().

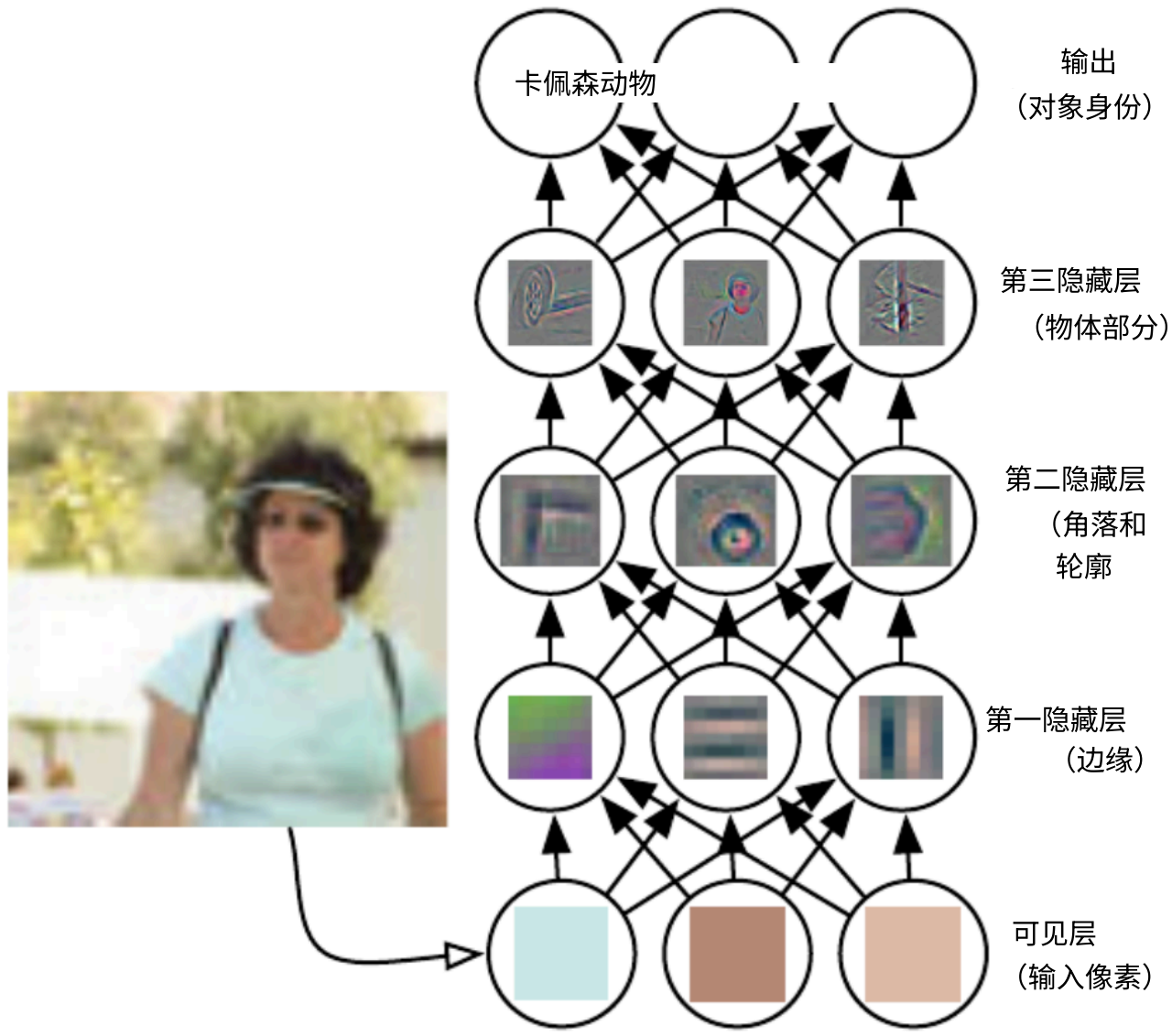

图 1.2：深度学习模型图解。计算机很难理解原始感官输入数据的含义，例如这幅以像素值集合表示的图像。从一组像素到对象身份的功能映射非常复杂。如果直接处理，学习或评估这种映射似乎是无法克服的。深度学习通过将所需的复杂映射分解为一系列嵌套的简单映射来解决这一难题，每个映射由模型的不同层描述。输入呈现在可见层，之所以这样命名，是因为它包含我们能够观察到的变量。一系列隐藏层从图像中提取越来越抽象的特征。这些层被称为“隐藏”，因为它们的值未在数据中给出；相反，模型必须确定哪些概念对于解释观察到的数据中的关系是有用的。这里的图像是每个隐藏单元所代表的特征类型的可视化。给定像素，第一层可以通过比较相邻像素的亮度轻松识别边缘。给定第一个隐藏层的边缘描述，第二个隐藏层可以轻松搜索可识别为边缘集合的角和扩展轮廓。给定第二个隐藏层对图像的角和轮廓的描述，第三个隐藏层可以通过查找特定的轮廓和角的集合来检测特定对象的整个部分。最后，此图像对其所包含的对象部分的描述可用于识别图像中存在的对象。图像经 ZeilerandFergus2014（）许可复制。

executinganothersetofinstructionsinparallel.Networkswithgreaterdepthcan executemoreinstructionsinsequence.Sequentialinstructionsoffergreatpower becauselaterinstructionscanreferbacktotheresultsofearlierinstructions.Accordingtothisviewofdeeplearning,notalltheinformationinalayer'sactivations necessarilyencodesfactorsofvariationthatexplaintheinput.Therepresentation alsostoresstateinformationthathelpstoexecuteaprogramthatcanmakesense oftheinput. Thisstateinformationcouldbeanalogoustoacounterorpointer inatraditionalcomputerprogram.Ithasnothingtodowiththecontentofthe inputspecifically,butithelpsthemodeltoorganizeitsprocessing.

Therearetwomainwaysofmeasuringthedepthofamodel.Thefirstviewis basedonthenumberofsequentialinstructionsthatmustbeexecutedtoevaluate thearchitecture.Wecanthinkofthisasthelengthofthelongestpaththrough aflowchartthatdescribeshowtocomputeeachofthemodel'soutputsgiven itsinputs.Justastwoequivalentcomputerprogramswillhavedifferentlengths dependingonwhichlanguagetheprogramiswrittenin,thesamefunctionmay bedrawnasaflowchartwithdifferentdepthsdependingonwhichfunctionswe allowtobeusedasindividualstepsintheflowchart.Figure 1.3 illustrateshowthis choiceoflanguagecangivetwodifferentmeasurementsforthesamearchitecture.

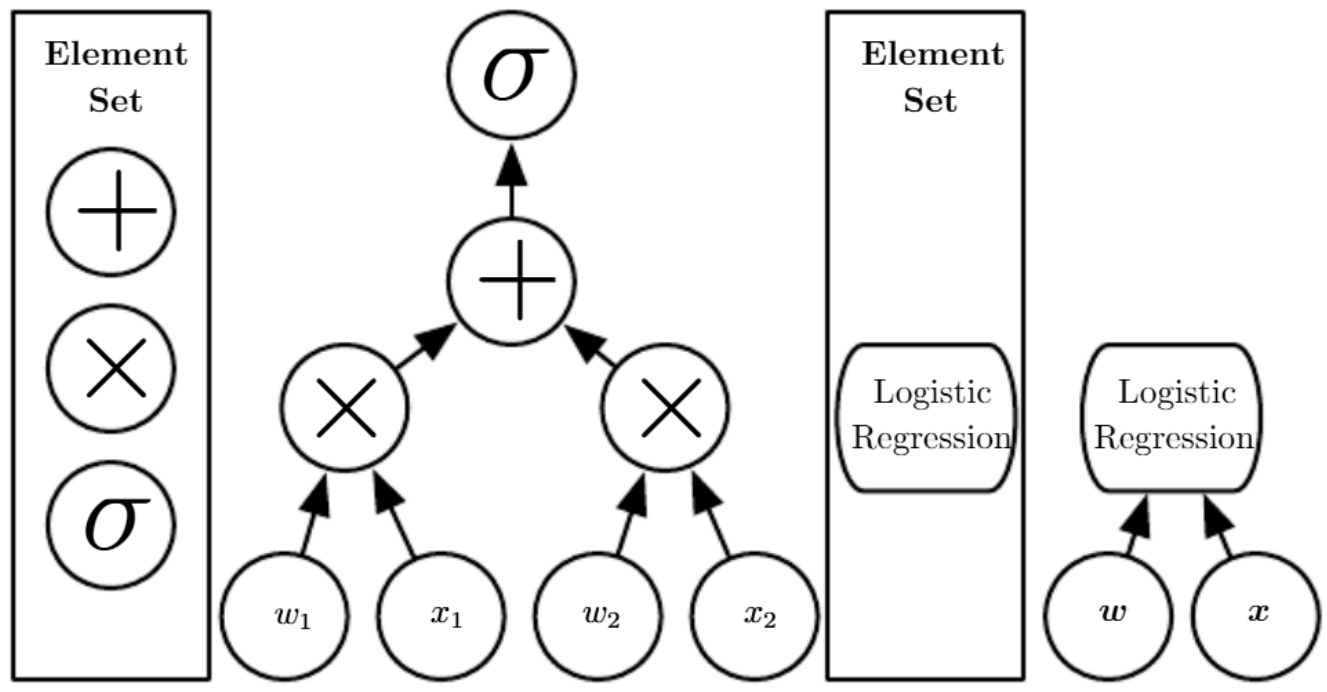

Figure1.3:Illustrationofcomputationalgraphsmappinganinputtoanoutputwhere eachnodeperformsanoperation.Depthisthelengthofthelongestpathfrominputto outputbutdependsonthedefinitionofwhatconstitutesapossiblecomputationalstep. Thecomputationdepictedinthesegraphsistheoutputofalogisticregressionmodel, $\sigma(\boldsymbol{w}^T\boldsymbol{x})$,where σ isthelogisticsigmoidfunction.Ifweuseaddition,multiplicationand logisticsigmoidsastheelementsofourcomputerlanguage,thenthismodelhasdepth three.Ifweviewlogisticregressionasanelementitself,thenthismodelhasdepthone.

并行执行另一组指令。深度更大的网络可以按顺序执行更多指令。顺序指令具有强大的功能，因为后续指令可以参考先前指令的其他结果。根据这种深度学习观点，层中激活的信息并非一定都编码了解释输入的变化因素。这种表示还存储了有助于执行能够理解输入的程序的状态信息。这种状态信息可以类似于传统计算机程序中的计数器或指针。它与输入的内容无关，但它有助于模型组织其处理。

有两种主要方法可以测量模型的深度。第一种观点基于评估架构所必须执行的顺序指令的数量。我们可以将其视为流程图中最长路径的长度，该流程图描述了如何根据输入计算每个模型的输出。正如两个等效的计算机程序将具有不同的长度（取决于程序编写的语言）一样，相同的功能可以绘制为具有不同深度的流程图，具体取决于我们允许将哪些功能用作流程图中的单独步骤。该图说明了这种1.3 语言选择如何为相同的架构提供两种不同的测量结果。

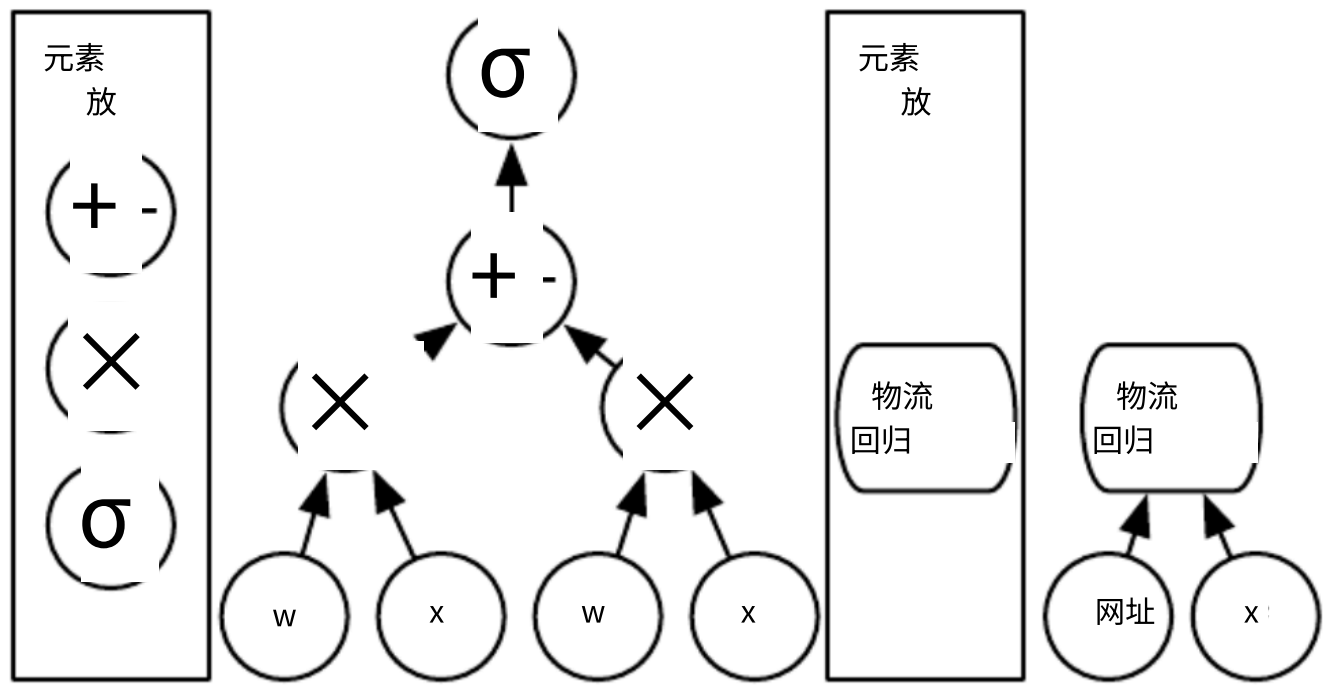

图 1.3：计算图的示意图，将输入映射到输出，其中每个节点执行一项操作。深度是从输入到输出的最长路径的长度，但取决于构成可能计算步骤的定义。这些图中描述的计算是逻辑回归模型σ(wx)的输出，其中σ是逻辑 S 形函数。如果我们使用加法、乘法和逻辑 S 形函数作为计算机语言的元素，则该模型的深度为三。如果我们将逻辑回归本身视为一个元素，则该模型的深度为一。

Anotherapproach,usedbydeepprobabilisticmodels,regardsthedepthofa modelasbeingnotthedepthofthecomputationalgraphbutthedepthofthe graphdescribinghowconceptsarerelatedtoeachother.Inthiscase,thedepth oftheflowchartofthecomputationsneededtocomputetherepresentationof eachconceptmaybemuchdeeperthanthegraphoftheconceptsthemselves. Thisisbecausethesystem'sunderstandingofthesimplerconceptscanberefined giveninformationaboutthemorecomplexconcepts.Forexample,anAIsystem observinganimageofafacewithoneeyeinshadowmayinitiallyseeonlyone eye.Afterdetectingthatafaceispresent,thesystemcantheninferthatasecond eyeisprobablypresentaswell.Inthiscase,thegraphofconceptsincludesonly twolayers—alayerforeyesandalayerforfaces—butthegraphofcomputations includes $2n$ layersifwerefineourestimateofeachconceptgiventheother n times.

Becauseitisnotalwaysclearwhichofthesetwoviews—thedepthofthe computationalgraph,orthedepthoftheprobabilisticmodelinggraph—ismost relevant,andbecausedifferentpeoplechoosedifferentsetsofsmallestelements fromwhichtoconstructtheirgraphs,thereisnosinglecorrectvalueforthe depthofanarchitecture,justasthereisnosinglecorrectvalueforthelengthof acomputerprogram. Noristhereaconsensusabouthowmuchdepthamodel requirestoqualifyas"deep."However,deeplearningcanbesafelyregardedasthe studyofmodelsthatinvolveagreateramountofcompositionofeitherlearned functionsorlearnedconceptsthantraditionalmachinelearningdoes.

Tosummarize,deeplearning,thesubjectofthisbook,isanapproachtoAI. Specifically,itisatypeofmachinelearning,atechniquethatenablescomputer systemstoimprovewithexperienceanddata.Wecontendthatmachinelearning istheonlyviableapproachtobuildingAIsystemsthatcanoperateincomplicated real-worldenvironments.Deeplearningisaparticularkindofmachinelearning thatachievesgreatpowerandflexibilitybyrepresentingtheworldasanested hierarchyofconcepts,witheachconceptdefinedinrelationtosimplerconcepts,and moreabstractrepresentationscomputedintermsoflessabstractones.Figure 1.4 illustratestherelationshipbetweenthesedifferentAIdisciplines.Figure 1.5 gives ahigh-levelschematicofhoweachworks.

1.1WhoShouldReadThisBook?

Thisbookcanbeusefulforavarietyofreaders,butwewroteitwithtwotarget audiencesinmind. Oneofthesetargetaudiencesisuniversitystudents(under-graduateorgraduate)learningaboutmachinelearning,includingthosewhoare beginningacareerindeeplearningandartificialintelligenceresearch.Theother

深度概率模型使用的另一种方法认为模型的深度不是计算图的深度，而是描述概念如何相互关联的图的深度。在这种情况下，计算每个概念的表示所需的计算流程图的深度可能比概念本身的图深得多。

这是因为系统对简单概念的理解可以通过提供有关更复杂概念的信息来改进。例如，一个人工智能系统在阴影中观察一张只有一只眼睛的脸的图像时，最初可能只看到一只眼睛。在检测到一张脸后，系统可以推断第二只眼睛也可能存在。在这种情况下，概念图只包括两层——眼睛层和脸层——但如果我们对每个概念的估计都考虑到另一个概念，那么计算图包括 2n 层

n 次。

由于计算图的深度和概率建模图的深度这两种观点中哪一种最相关并不总是很清楚，而且不同的人会选择不同的最小元素集来构建他们的图，因此对于架构的深度没有单一的正确值，就像对于计算机程序的长度没有单一的正确值一样。对于模型需要多深才能被称为“深度”，目前也没有达成共识。然而，深度学习可以安全地视为研究比传统机器学习涉及更多学习函数或学习概念组合的模型。

总而言之，本书的主题是深度学习，是一种人工智能方法。具体来说，它是一种机器学习，一种使计算机系统能够通过经验和数据进行改进的技术。我们认为，机器学习是构建可以在复杂的现实世界环境中运行的人工智能系统的唯一可行方法。深度学习是一种特殊的机器学习，它通过将世界表示为概念的嵌套层次结构来实现强大的功能和灵活性，每个概念都相对于更简单的概念进行定义，并且更抽象的表示是根据不太抽象的概念计算出来的。图 1.4 说明了这些不同人工智能学科之间的关系。图 1.5 给出了每个工作原理的高级示意图。

1.1 谁应该读这本书?

本书适合各类读者，但我们在撰写时主要针对两类目标读者。一类目标读者是正在学习机器学习的大学生（本科生或研究生），包括那些刚开始从事深度学习和人工智能研究的人士。另一类目标读者是……

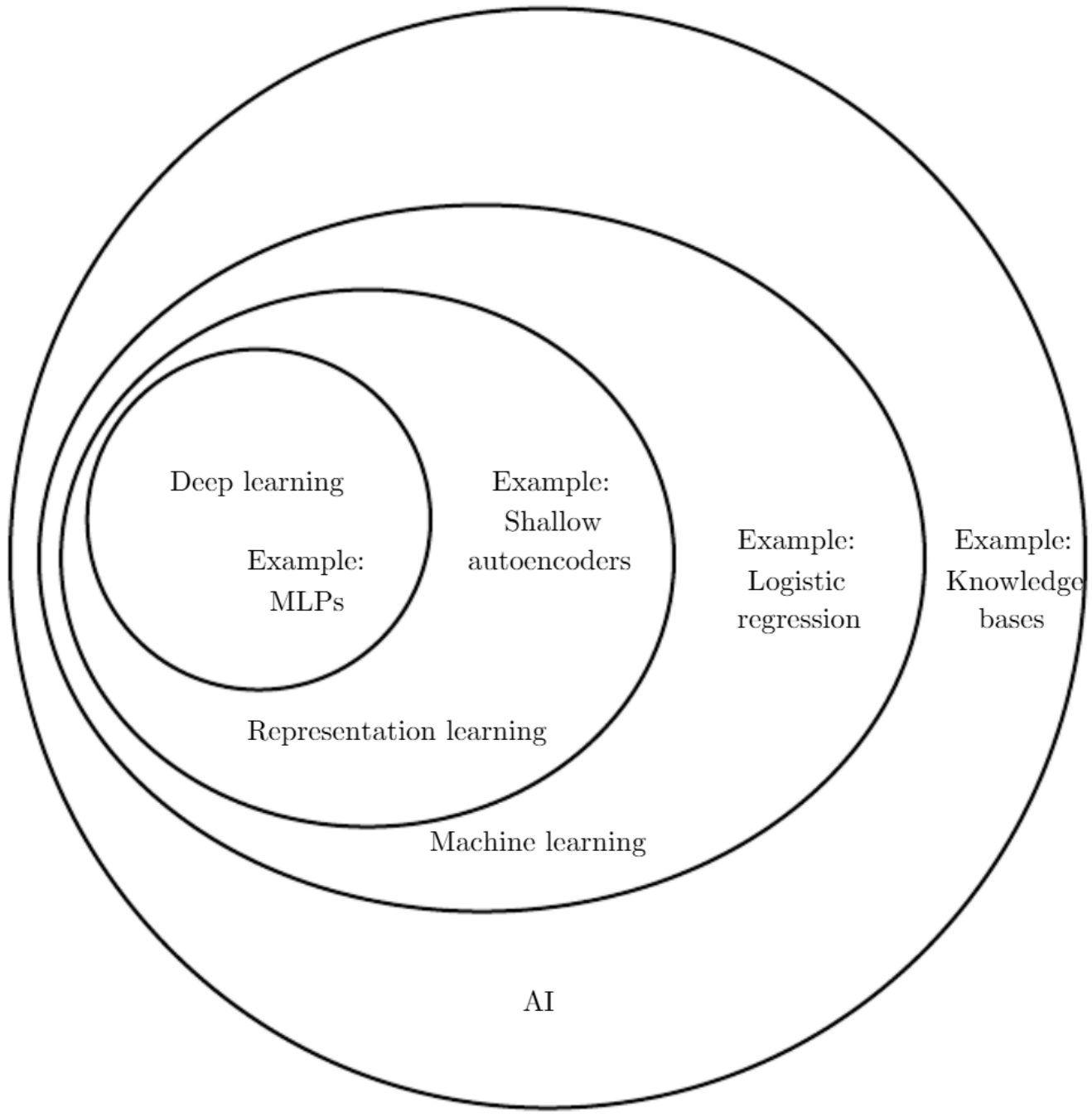

Figure1.4:AVenndiagramshowinghowdeeplearningisakindofrepresentationlearning, whichisinturnakindofmachinelearning,whichisusedformanybutnotallapproaches toAI.EachsectionoftheVenndiagramincludesanexampleofanAItechnology.

targetaudienceissoftwareengineerswhodonothaveamachinelearningorstatis-ticsbackgroundbutwanttorapidlyacquireoneandbeginusingdeeplearningin theirproductorplatform.Deeplearninghasalreadyprovedusefulinmanysoft-waredisciplines,includingcomputervision,speechandaudioprocessing,natural languageprocessing,robotics,bioinformaticsandchemistry,videogames,search engines,onlineadvertisingandfinance.

Thisbookhasbeenorganizedintothreepartstobestaccommodateavariety ofreaders.Part I introducesbasicmathematicaltoolsandmachinelearning concepts.Part II describesthemostestablisheddeeplearningalgorithms,which areessentiallysolvedtechnologies.Part III describesmorespeculativeideasthat arewidelybelievedtobeimportantforfutureresearchindeeplearning.

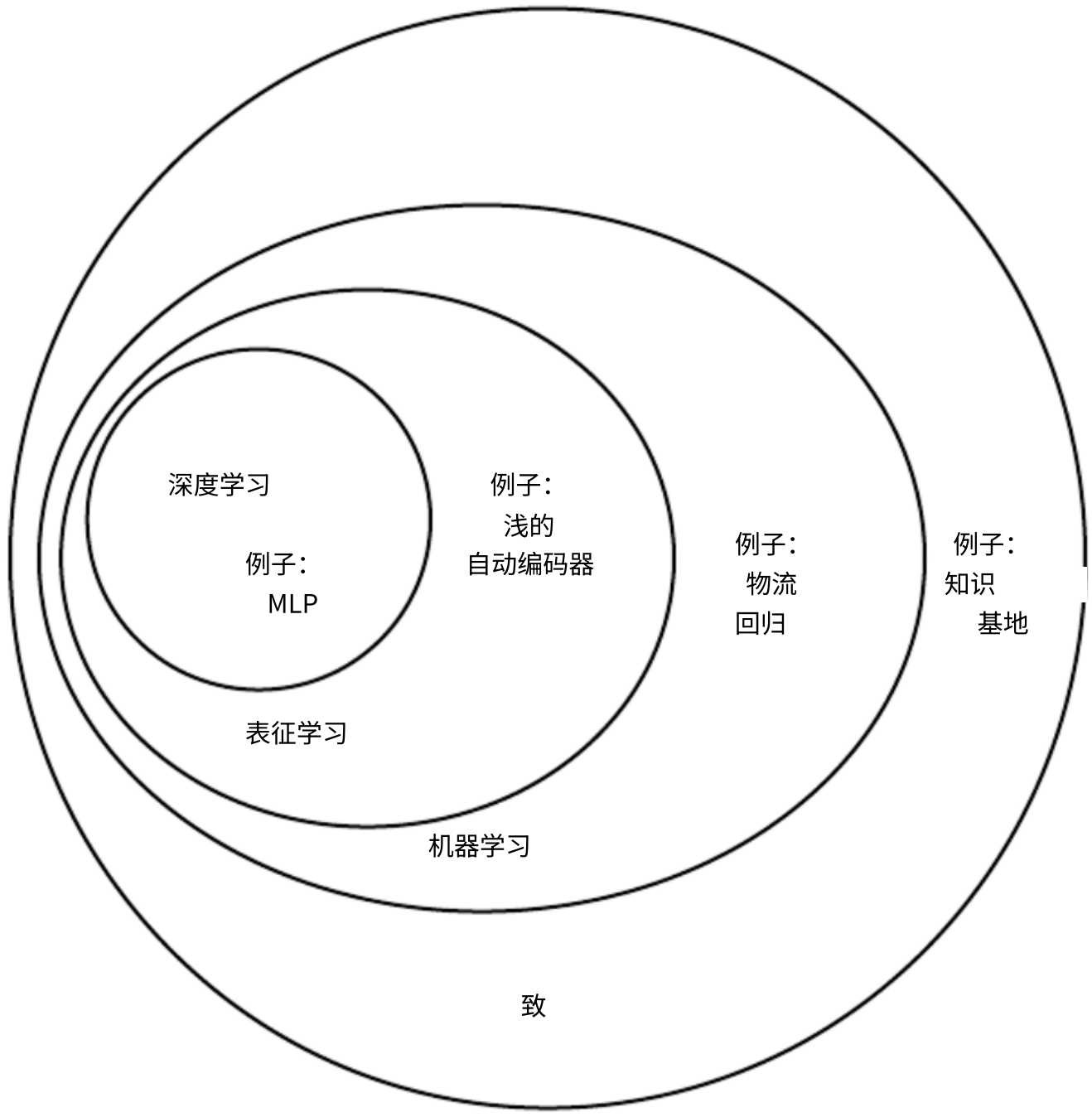

图 1.4：维恩图展示了深度学习是一种表征学习，而表征学习又是一种机器学习，用于许多但并非所有人工智能方法。维恩图的每一部分都包含一个人工智能技术的示例。

目标受众是那些没有机器学习或统计学背景，但想要快速掌握这些背景并开始在其产品或平台中使用深度学习的软件工程师。深度学习已经被证明在许多软件学科中有用，包括计算机视觉、语音和音频处理、自然语言处理、机器人技术、生物信息学和化学、视频游戏、搜索引擎、在线广告和金融。

本书分为三部分，以最好地适应各种读者。第一部分介绍基本的数学工具和机器学习概念。第二部分描述最成熟的深度学习算法，这些算法本质上是已解决的技术。第三部分描述更多推测性的想法，人们普遍认为这些想法对未来的深度学习研究很重要。

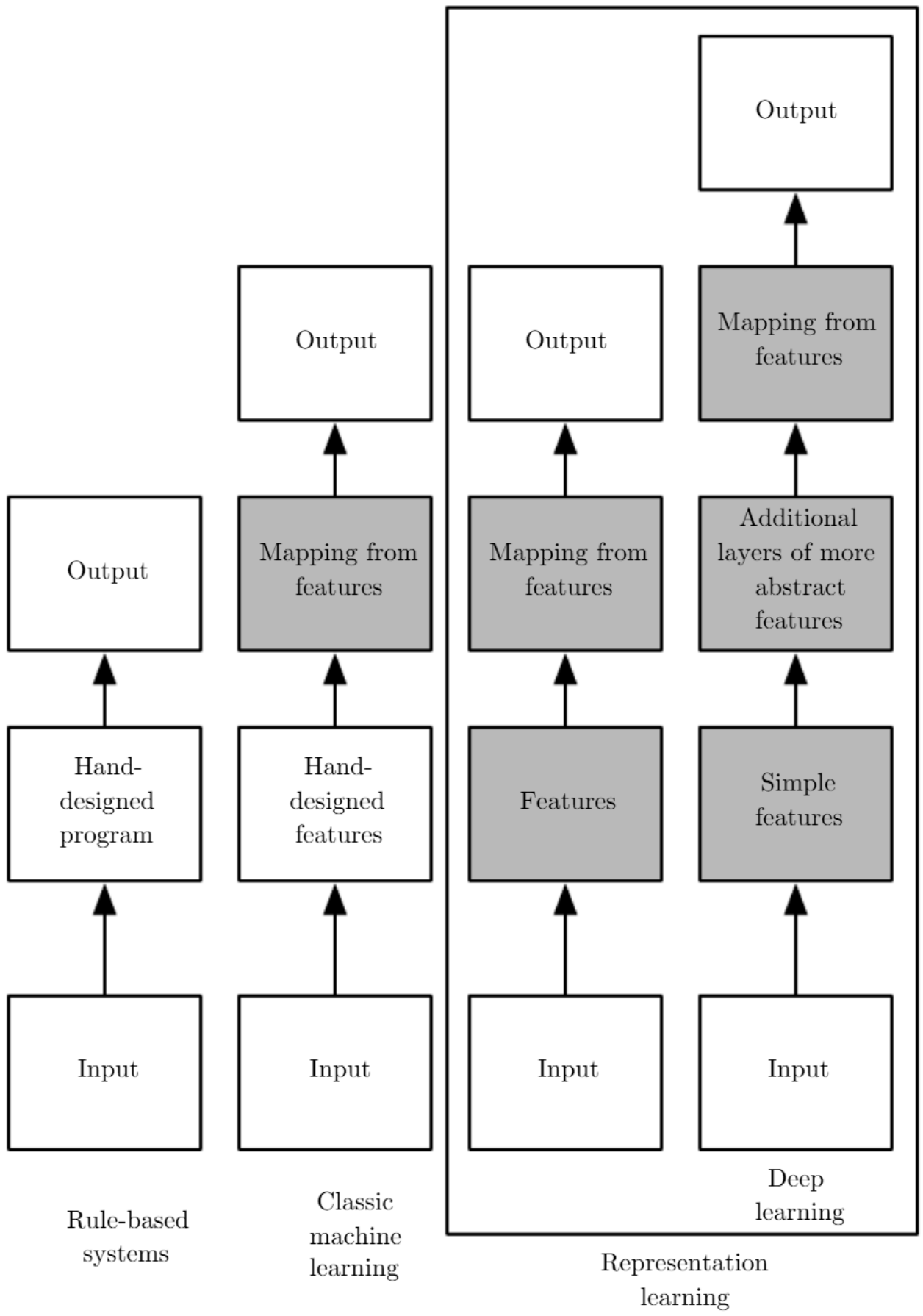

Figure1.5: FlowchartsshowinghowthedifferentpartsofanAIsystemrelatetoeach otherwithindifferentAIdisciplines.Shadedboxesindicatecomponentsthatareableto learnfromdata.

Readersshouldfeelfreetoskippartsthatarenotrelevantgiventheirinterests orbackground.Readersfamiliarwithlinearalgebra,probability,andfundamental machinelearningconceptscanskippart I,forexample,whilethosewhojustwant toimplementaworkingsystemneednotreadbeyondpart II.Tohelpchoosewhich

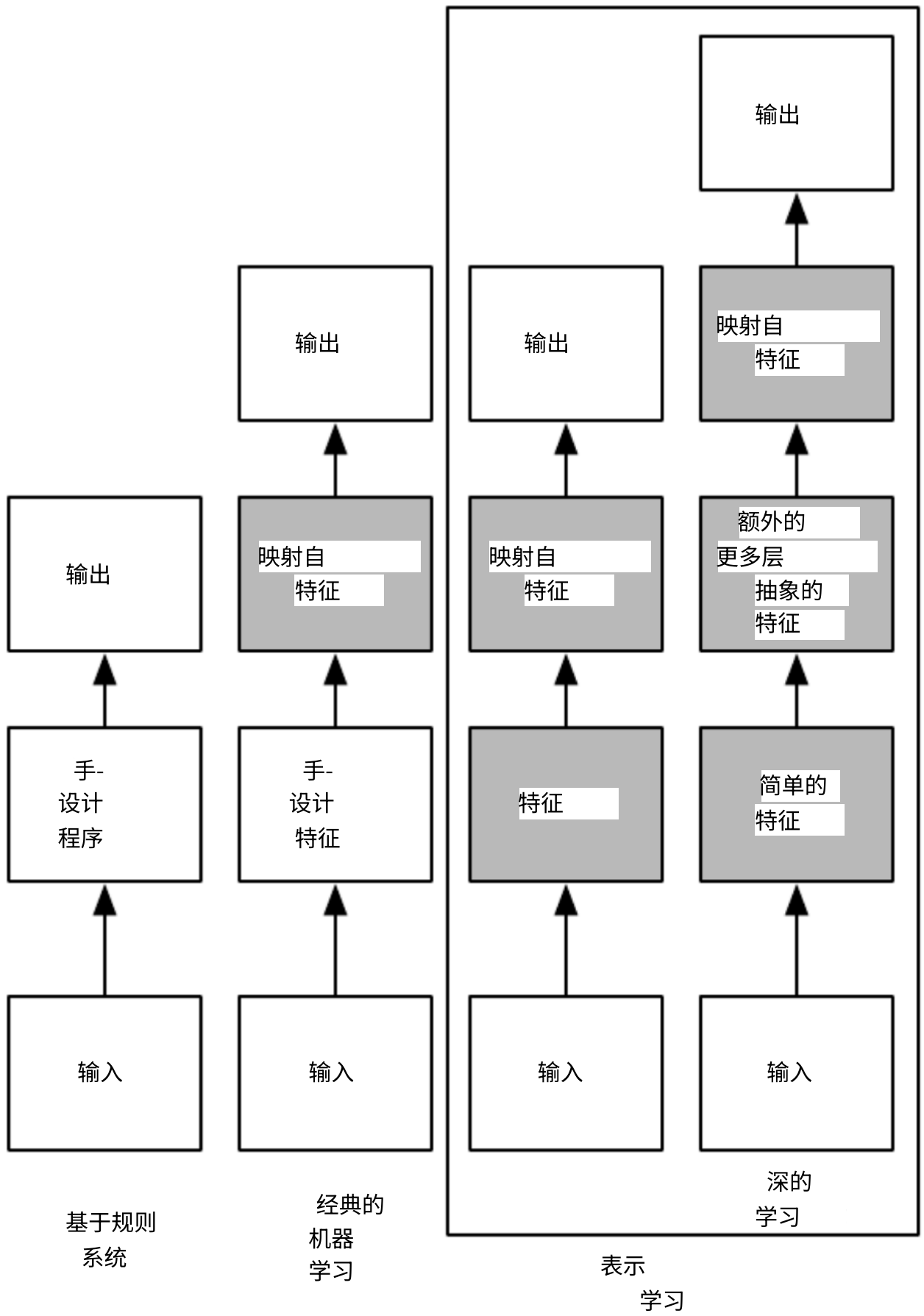

图 1.5：流程图展示了人工智能系统的不同部分在不同的人工智能学科中如何相互关联。阴影框表示能够从数据中学习的组件。

读者可以根据自己的兴趣或背景跳过不相关的部分。例如，熟悉线性代数、概率和基本机器学习概念的读者可以跳过此部分，而只想实现一个工作系统的读者则无需阅读此部分之后的内容。

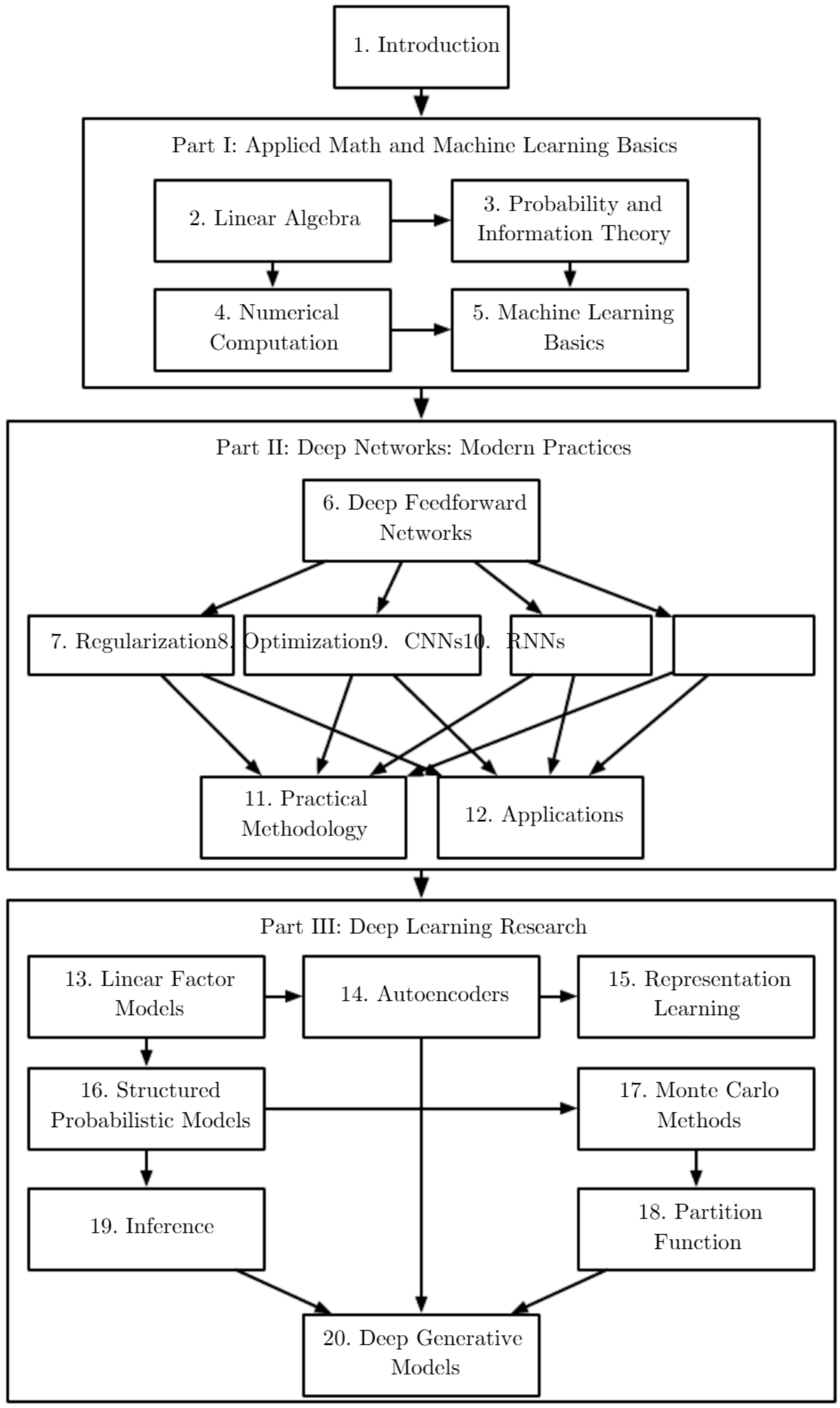

Figure1.6:Thehigh-levelorganizationofthebook.Anarrowfromonechaptertoanother indicatesthattheformerchapterisprerequisitematerialforunderstandingthelatter.

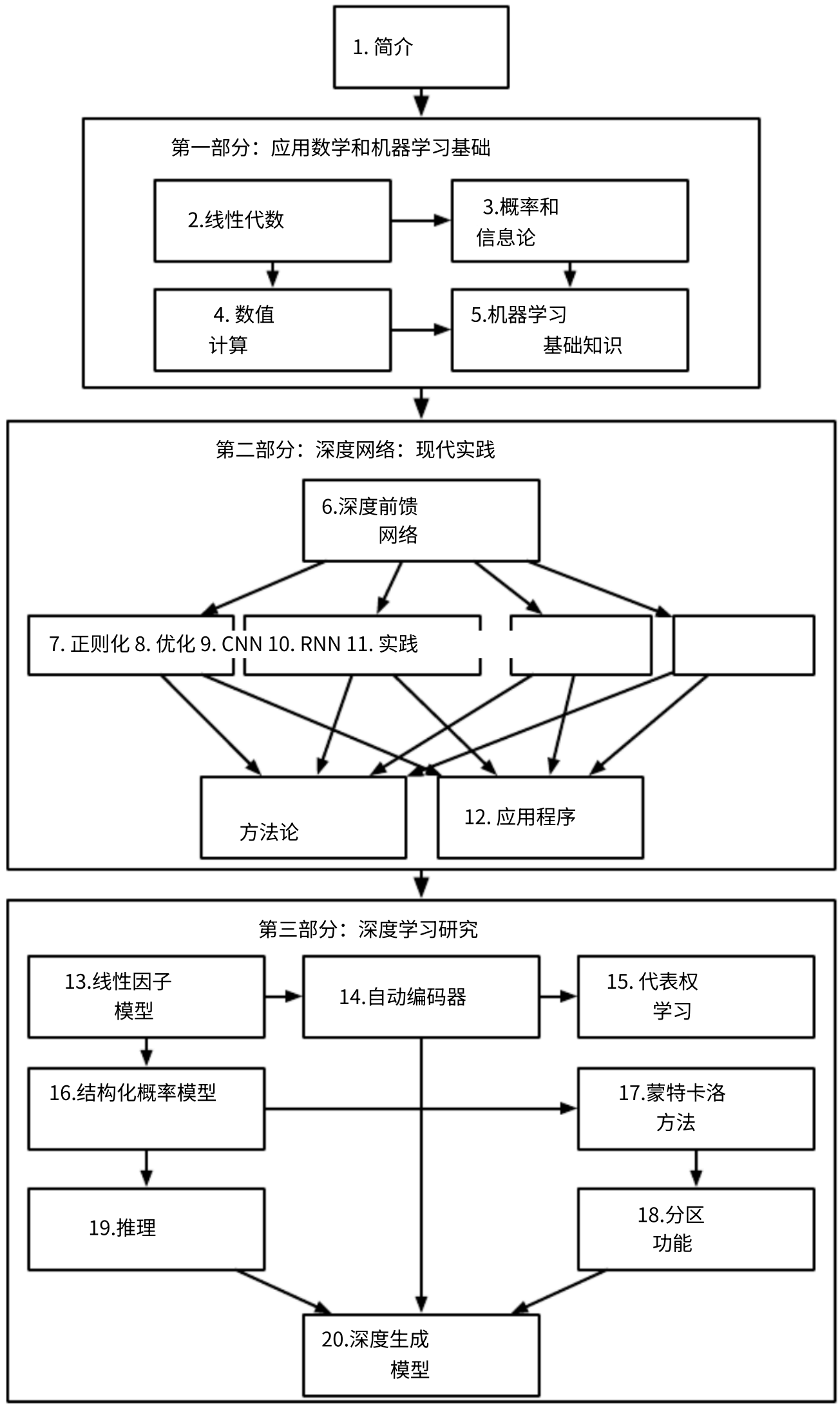

图 1.6：本书的高层组织结构。从一章到另一章的箭头表示前一章是理解后一章的先决条件材料。

chapterstoread,figure 1.6 providesaflowchartshowingthehigh-levelorganization ofthebook.

Wedoassumethatallreaderscomefromacomputersciencebackground.We assumefamiliaritywithprogramming,abasicunderstandingofcomputational performanceissues,complexitytheory,introductorylevelcalculusandsomeofthe terminologyofgraphtheory.

1.2HistoricalTrendsinDeepLearning

Itiseasiesttounderstanddeeplearningwithsomehistoricalcontext.Ratherthan providingadetailedhistoryofdeeplearning,weidentifyafewkeytrends:

- Deeplearninghashadalongandrichhistory,buthasgonebymanynames, reflectingdifferentphilosophicalviewpoints,andhaswaxedandwanedin popularity.
- Deeplearninghasbecomemoreusefulastheamountofavailabletraining datahasincreased.
- Deeplearningmodelshavegrowninsizeovertimeascomputerinfrastructure (bothhardwareandsoftware)fordeeplearninghasimproved.
- Deeplearninghassolvedincreasinglycomplicatedapplicationswithincreasing accuracyovertime.

1.2.1TheManyNamesandChangingFortunesofNeuralNetworks

Weexpectthatmanyreadersofthisbookhaveheardofdeeplearningasanexciting newtechnology,andaresurprisedtoseeamentionof“history”inabookaboutan emergingfield.Infact,deeplearningdatesbacktothe1940s.Deeplearningonly *appears* tobenew,becauseitwasrelativelyunpopularforseveralyearspreceding itscurrentpopularity,andbecauseithasgonethroughmanydifferentnames,only recentlybeingcalled“deeplearning.”Thefieldhasbeenrebrandedmanytimes, reflectingtheinfluenceofdifferentresearchersanddifferentperspectives.

Acomprehensivehistoryofdeeplearningisbeyondthescopeofthistextbook. Somebasiccontext,however,isusefulforunderstandingdeeplearning.Broadly speaking,therehavebeenthreewavesofdevelopment:deeplearningknownas **cybernetics** inthe1940s–1960s,deeplearningknownas **connectionism** inthe

章节阅读，图提供了一个流
程图，展示了本书的高级组
织结构。

我们确实假设所有读者都来自计算机科学背景。我们假设他们熟悉编程、对计算性能问题、复杂性理论、入门级微积分和一些图论术语有基本的了解。

1.2 深度学习的历史趋势

通过一些历史背景来理解深度学习是最容易的。我们不会提供深度学习的详细历史，而是确定一些关键趋势：

- 深度学习有着悠久而丰富的历史，但其名称众多，反映了不同的哲学观点，而且受欢迎程度时高时低。
- 随着可用训练数据量的增加，深度学习变得更加有用。
- 随着深度学习的计算机基础设施（硬件和软件）的改进，深度学习模型的规模不断扩大。
- 随着时间的推移，深度学习已经解决了越来越复杂的应用问题，并且准确性也越来越高。

1.2.1 神经网络的众多名称和变迁

我们预计本书的许多读者都听说过深度学习这一令人兴奋的新技术，并且很惊讶地看到在关于新兴领域的书中提到“历史”。事实上，深度学习可以追溯到 20 世纪 40 年代。深度学习只是看起来很新，因为在它现在流行之前的几年里它相对不受欢迎，并且因为它经历了许多不同的名称，直到最近才被称为“深度学习”。该领域已被多次重新命名，反映了不同研究人员和不同观点的影响。

深度学习的全面历史超出了本教科书的范围。
然而，了解一些基本背景对理解深度学习还是有用的。广义上讲，深度学习经历了三次发展浪潮：20 世纪 40 年代至 60 年代被称为控制论的深度学习，20 世纪 60 年代被称为 联结主义

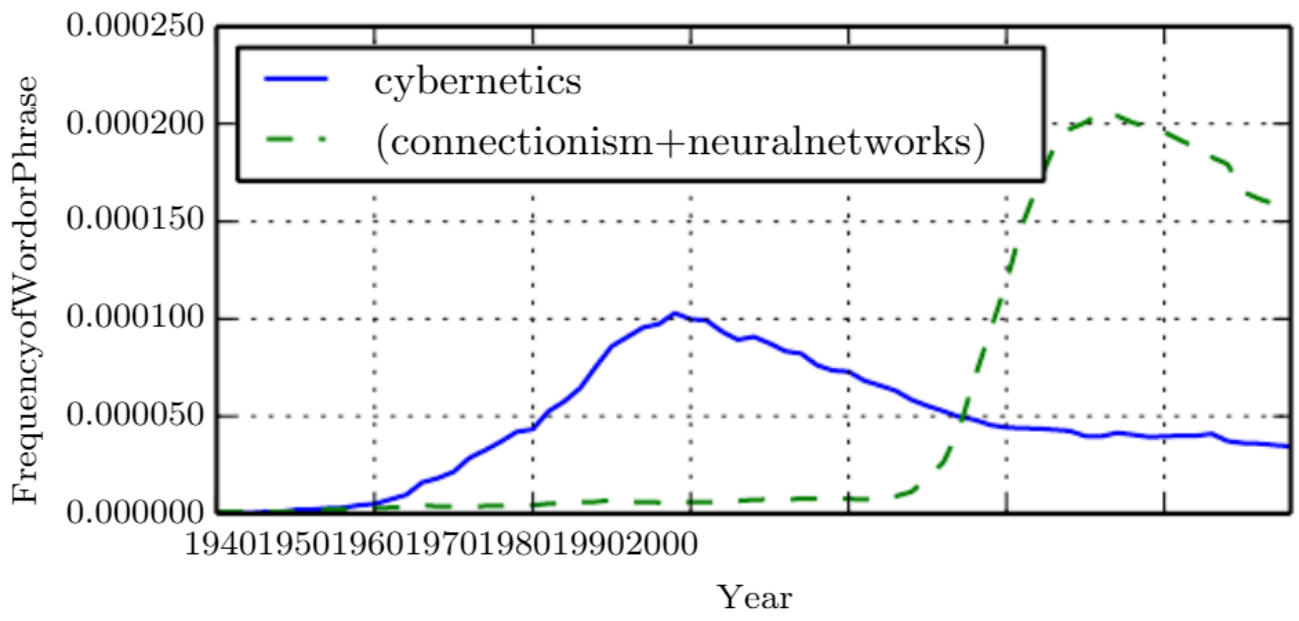

Figure1.7:Twofthethreehistoricalwavesofartificialneuralnetsresearch,asmeasured bythefrequencyofthephrases“cybernetics”and“connectionism”or“neuralnetworks,” accordingtoGoogleBooks(thethirdwaveistoorecenttoappear).Thefirstwave startedwithcyberneticsinthe1940s–1960s,withthedevelopmentoftheoriesofbiological learning (McCullochandPitts , 1943Hebb; , 1949) andimplementationsofthefirstmodels, suchastheperceptron (Rosenblatt, 1958),enablingthetrainingofasingleneuron.The secondwavestartedwiththeconnectionistapproachofthe1980–1995period,withback-propagation (Rumelhart *etal.* , 1986a) totrainaneuralnetworkwithoneortwohidden layers. Thecurrentandthirdwave,deeplearning,startedaround2006 (Hinton *etal.* , 2006Bengio; *etal.* , 2007Ranzato; *etal.* , 2007a) andisjustnowappearinginbookform asof2016. Theothertwowavessimilarlyappearedinbookformmuchlaterthanthe correspondingscientificactivityoccurred.

1980s–1990s,andthecurrentresurgenceunderthenamedeeplearningbeginning in2006.Thisisquantitativelyillustratedinfigure 1.7.

Someoftheearliestlearningalgorithmswerecognizetodaywereintendedto becomputationalmodelsofbiologicallearning,thatis,modelsofhowlearning happensorcouldhappeninthebrain. Asaresult,oneofthenamesthatdeep learninghasgonebyis **artificialneuralnetworks** (ANNs).Thecorresponding perspectiveondeeplearningmodelsisthattheyareengineeredsystemsinspired bythebiologicalbrain(whetherthehumanbrainorthebrainofanotheranimal). Whilethekindsofneuralnetworksusedformachinelearninghavesometimes beenusedtounderstandbrainfunction (HintonandShallice , 1991),theyare generallynotdesignedtoberealisticmodelsofbiologicalfunction.Theneural perspectiveondeeplearningismotivatedbytwomainideas.Oneideaisthat thebrainprovidesaproofbyexamplethatintelligentbehaviorispossible,anda conceptuallystraightforwardpathtobuildingintelligenceistoreverseengineerthe computationalprinciplesbehindthebrainandduplicateitsfunctionality.Another

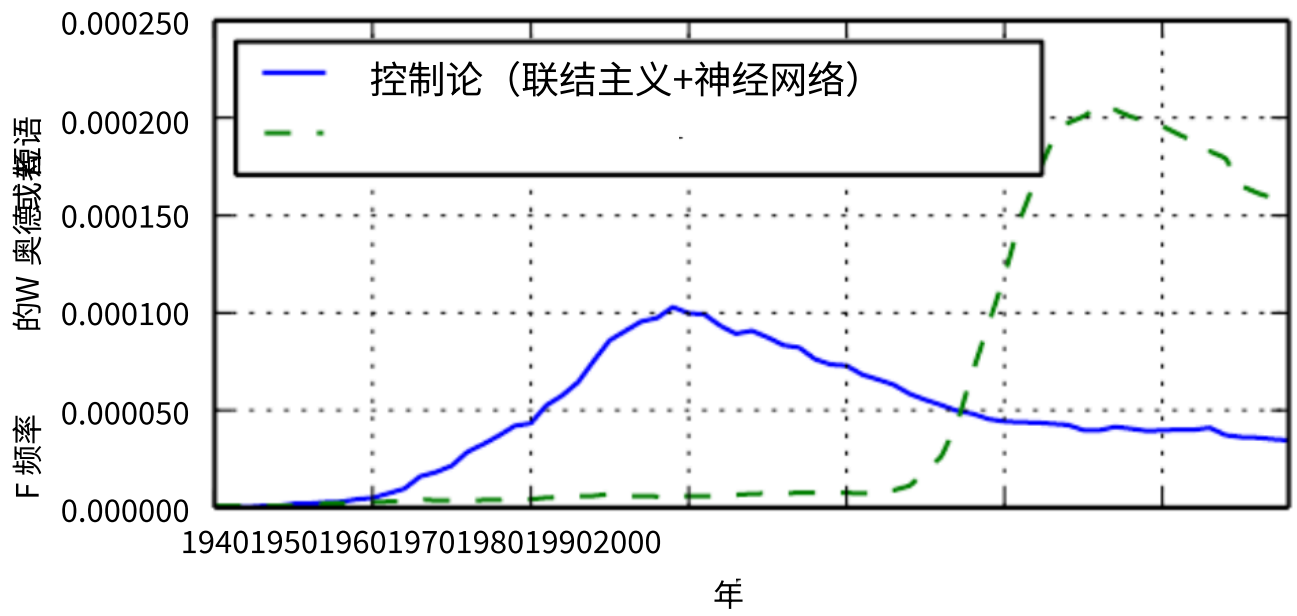

图 1.7：根据 Google 图书的数据，以“控制论”和“联结主义”或“神经网络”等词语出现的频率来衡量，人工神经网络研究经历了三次历史浪潮中的两次（第三次浪潮出现得较晚）。第一次浪潮始于 20 世纪 40 年代至 60 年代的控制论，当时出现了生物学习理论（McCulloc 和 Pitts 1943；Hebb 1949，1951；1952），并实现了第一批模型，如感知器（，），使单个神经元的训练成为可能。第二次浪潮始于 1958 年的 Rosenblatt 浪潮，始于 1980 年至 1995 年期间的联结主义方法，使用反向传播（，）来训练具有一层或两层隐藏层的神经网络（Rumelhart et al. 1986）。当前的第三次浪潮，深度学习，始于 2006 年左右（Hinton 等人）。

2006Bengio 2007Ranzato 2007a; etal.,; etal.,) 并且刚刚于 2016 年以书籍形式出现。另外两波浪潮同样以书籍形式出现，但比相应的科学活动发生的时间要晚得多。

深度学习最初在 20 世纪 80 年代至 90 年代兴起，并从 2006 年开始以深度学习的名义重新兴起。图 1.7 定量地说明了这一点。今天公认的一些最早的学习算法旨在成为生物学习的计算模型，即学习如何在大脑中发生或可能发生的模型。因此，深度学习的一个名称是人工神经网络 (ANN)。深度学习模型的相应观点是，它们是受生物大脑（无论是人类大脑还是其他动物的大脑）启发的工程系统。虽然用于机器学习的各种神经网络有时被用来理解大脑功能（，），但它们通常不是 Hinton 和 Shallice 1991 设计为生物功能的现实模型。深度学习的神经视角受到两个主要思想的推动。一个思想是，大脑通过例子证明了智能行为是可能的，而构建智能的概念上直接的途径是逆向工程大脑背后的计算原理并复制其功能。另一个

perspectiveisthatitwouldbedeeplyinterestingtounderstandthebrainandthe principlesthatunderliehumanintelligence,somachinelearningmodelsthatshed lightonthesebasicscientificquestionsareusefulapartfromtheirabilitytosolve engineeringapplications.

Themodernterm"deeplearning"goesbeyondtheneuroscientificperspective onthecurrentbreedofmachinelearningmodels.Itappealstoamoregeneral principleoflearning *multiplelevelsofcomposition* ,whichcanbeappliedinmachine learningframeworksthatarenotnecessarilyneurallyinspired.

Theearliestpredecessorsofmoderndeeplearningweresimplelinearmodels motivatedfromaneuroscientificperspective.Thesemodelsweredesignedto takeasetof n inputvalues $x_1,...,x_n$ andassociatethemwithanoutput y. Thesemodelswouldlearnasetofweights $w_1,...,w_n$ andcomputetheiroutput $f(\boldsymbol{x}, \boldsymbol{w}) = x_1 w_1 + \cdots + x_n w_n$.Thisfirstwaveofneuralnetworksresearchwas knownascybernetics,asillustratedinfigure 1.7.

TheMcCulloch-Pittsneuron (McCullochandPitts , 1943) wasanearlymodel ofbrainfunction.Thislinearmodelcouldrecognizetwodifferentcategoriesof inputsbytestingwhether $f(\boldsymbol{x}, \boldsymbol{w})$ ispositiveornegative.Ofcourse,forthemodel tocorrespondtothedesireddefinitionofthecategories,theweightsneededtobe setcorrectly.Theseweightscouldbesetbythehumanoperator.Inthe1950s,the perceptron (Rosenblatt, 1958, 1962) becamethefirstmodelthatcouldlearnthe weightsthatdefinedthecategoriesgivenexamplesofinputsfromeachcategory. The **adaptivelinearelement** (ADALINE),whichdatesfromaboutthesame time,simplyreturnedthevalueof $f(\boldsymbol{x})$ itselftopredictarealnumber (Widrow andHoff , 1960) andcouldalsolearntopredictthesenumbersfromdata.

Thesesimplelearningalgorithmsgreatlyaffectedthemodernlandscapeofma- chinelearning.Thetrainingalgorithmusedtoadapttheweightsoftheadaline wasaspecialcaseofanalgorithmcalled **stochasticgradientdescent** .Slightly modifiedversionsofthestochasticgradientdescentalgorithmremainthedominant trainingalgorithmsfordeeplearningmodelstoday.

Modelsbasedonthe $f(\boldsymbol{x}, \boldsymbol{w})$ usedbytheperceptronandadalinearecalled **linearmodels** .Thesemodelsremainsomeofthemostwidelyusedmachine learningmodels,thoughinmanycasestheyare *trained* indifferentwaysthanthe originalmodelsweretrained.

Linearmodelshavemanylimitations.Mostfamously,theycannotlearnthe XORfunction,where $f([0,1], \boldsymbol{w}) = 1$ and $f([1,0], \boldsymbol{w}) = 1$ but $f([1,1], \boldsymbol{w}) = 0$ and $f([0,0], \boldsymbol{w}) = 0$.Criticswhoobservedtheseflawsinlinearmodelscaused abacklashagainstbiologicallyinspiredlearningingeneral (MinskyandPapert , 1969).Thiswasthefirstmajordipinthepopularityofneuralnetworks.

我的观点是，了解大脑和人类智能背后的原理将会非常有趣，因此，阐明这些基本科学问题的机器学习模型除了能够解决工程应用之外，还非常有用。

现代术语“深度学习”超越了神经科学对当前机器学习模型的视角。它诉诸于学习多层次组合的更普遍的原理，该原理可应用于不一定受神经启发的机器学习框架中。

现代深度学习的早期前身是从神经科学角度出发的简单线性模型。这些模型被设计用来接受一组输入值

x1，...，xnandas 将它们与输出关联起来 y.

深度学习模型的相应观点是，它们是受生物大脑启发而设计的系统（即使是输出 f(x脑，)=x这些模型将学习 w组权重第一波神经网络研究被称为控制论，如图 1.7 所示。McCulloch-Pitts 神经元（,）是大脑功能的早期模型。该线性模型通过测试 f(xw,) 是正数还是负数，可以识别两种不同的输入类别。当然，为了使模型与所需的类别定义相对应，需要正确设置权重。这些权重可以由人类操作员设置。20 世纪 50 年代，感知器（Rosenblatt 1958 1962,,）成为第一个可以通过给定每个类别的输入示例来学习定义类别的权重的模型。

自适应线性元素（ADALINE）大约在同一时期出现，它只是返回 f(x) 本身的值来预测一个实数（Widrow 和 Hoff 1960,），并且还可以学习从数据中预测这些数字。

这些简单的学习算法极大地影响了机器学习的现代格局。用于调整 ADALINE 权重的训练算法是称为随机梯度下降的算法的一个特例。稍微修改后的随机梯度下降算法版本仍然是当今深度学习模型的主要训练算法。

感知器和 ADALINE 使用的基于 f(xw,) 的模型称为线性模型。这些模型仍然是一些最广泛使用的机器学习模型，尽管在许多情况下，它们的训练方式与原始模型不同。

线性模型有许多局限性。最著名的是，它们不能学习 XOR 函数，其中 f([0,1], w) = 1 和 f([1,0], w) = 1，但 f([1,1], w) = 0 和 f([0,0], w) = 0。批评者观察到线性模型中的这些缺陷，引起了对生物启发式学习的强烈反对（MinskyandPapert，1969）。这是神经网络流行的第一个主要原因。

Today,neuroscienceisregardedasanimportantsourceofinspirationfordeep learningresearchers,butitisnolongerthepredominantguideforthefield.

Themainreasonforthediminishedrole ofneuroscienceindeeplearning researchtodayisthatwesimplydonothaveenoughinformationaboutthebrain touseitasaguide.Toobtainadeepunderstandingoftheactualalgorithmsused bythebrain,wewouldneedtobeabletomonitortheactivityof(atthevery least)thousandsofinterconnectedneuronssimultaneously.Becausewearenot abletodothis,wearefarfromunderstandingevensomeofthemostsimpleand well-studiedpartsofthebrain (OlshausenandField , 2005).

Neurosciencehasgivenusareasontohopethatasingledeeplearningalgorithm cansolvemanydifferenttasks.Neuroscientistshavefoundthatferretscanlearnto "see"withtheauditoryprocessingregionoftheirbrainiftheirbrainsarerewired tosendvisualsignalstothatarea (VonMelchner *etal.* , 2000).Thissuggeststhat muchofthemammalianbrainmightuseasinglealgorithmtosolvemostofthe differenttasksthatthebrainsolves.Beforethishypothesis,machinelearning researchwasmorefragmented,withdifferentcommunitiesofresearchersstudying naturallanguageprocessing,vision,motionplanningandspeechrecognition.Today, theseapplicationcommunitiesarestillseparate,butitiscommonfordeeplearning researchgroupstostudymanyorevenalltheseapplicationareassimultaneously.

Weareabletodrawsomeroughguidelinesfromneuroscience.Thebasic ideaofhavingmanycomputationalunitsthatbecomeintelligentonlyviatheir interactionswitheachotherisinspiredbythebrain.Theneocognitron (Fukushima, 1980) introducedapowerfulmodelarchitectureforprocessingimagesthatwas inspiredbythestructureofthemammalianvisualsystemandlaterbecamethe basisforthemodernconvolutionalnetwork (LeCun *etal.* , 1998b),aswewillsee insection 9.10.Mostneuralnetworkstodayarebasedonamodelneuroncalled the **rectifiedlinearunit** .Theoriginalcognitron (Fukushima, 1975) introduced amorecomplicatedversionthatwashighlyinspiredbyourknowledgeofbrain function.Thesimplifiedmodernversionwasdevelopedincorporatingideasfrom manyviewpoints,with NairandHinton (2010) and Glorot *etal.* (2011a) citing neuroscienceasaninfluence,and Jarrett *etal.* (2009) citingmoreengineering-orientedinfluences.Whileneuroscienceisanimportantsourceofinspiration,it neednotbetakenasarigidguide.Weknowthatactualneuronscomputevery differentfunctionsthanmodernrectifiedlinearunits,butgreaterneuralrealism hasnotyetledtoanimprovementinmachinelearningperformance.Also,while neurosciencehassuccessfullyinspiredseveralneuralnetwork *architectures*,we donotyetknowenoughaboutbiologicallearningforneurosciencetooffermuch guidanceforthe *learningalgorithms* weusetotrainthesearchitectures.

如今，神经科学被视为深度学习研究人员的重要灵感来源，但它不再是该领域的主要指南。

如今，神经科学在深度学习研究中的作用减弱的主要原因是我们没有足够的大脑信息来指导它。为了深入了解大脑使用的实际算法，我们需要能够同时监控（至少）数千个相互连接的神经元的活动。因为我们无法做到这一点，所以我们甚至无法理解大脑中一些最简单、研究最透彻的部分（ , ）。OlshausenandField 2005

神经科学让我们有理由希望单一的深度学习算法可以解决许多不同的任务。神经科学家发现，如果雪貂的大脑重新连接以向该区域发送视觉信号，它们就可以学会用大脑的听觉处理区域“看”（VonMelchner 2000etal。, ）。这表明大部分哺乳动物的大脑可能使用单一算法来解决大脑解决的大多数不同任务。在此假设之前，机器学习研究更加分散，不同的研究人员社区研究自然语言处理，视觉，运动规划和语音识别。今天，这些应用社区仍然是分开的，但深度学习研究小组同时研究许多甚至所有这些应用领域是很常见的。

我们可以从神经科学中得出一些粗略的指导方针。许多计算单元仅通过相互作用才能变得智能的基本思想是受到大脑的启发。新认知机（Fukushima，1980）引入了一种强大的图像处理模型架构，其灵感来自哺乳动物视觉系统的结构，后来成为现代卷积网络（ ， ）的基础，正如我们将在 1998b 节中看到的 LeCun 等人。当今大多数神经网络都基于称为 9.10 的模型神经元

整流线性单元。原始认知机（Fukushima 1975, ）引入了一个更复杂的版本，极大地启发了您对大脑功能的知识。简化的现代版本结合了许多观点的想法而开发，其中（）和（）引用了 NairandHinton 2010 Glorot et al. 2011a 神经科学作为影响，并且（）引用了更多工程 - Jarrett et al. 2009 导向的影响。虽然神经科学是重要的灵感来源，但不必将其作为僵化的指南。我们知道实际的神经元计算的功能与现代整流线性单元完全不同，但更大的神经现实主义尚未导致机器学习性能的改进。此外，虽然神经科学已经成功启发了几种神经网络架构，但我们对生物学习的了解还不足以让神经科学为我们用来训练搜索结构的学习算法提供足够的指导。

Mediaaccountsoftenemphasizethesimilarityofdeeplearningtothebrain. Whileitistruethatdeeplearningresearchersaremorelikelytocitethebrainasan influencethanresearchersworkinginothermachinelearningfields,suchaskernel machinesorBayesianstatistics,oneshouldnotviewdeeplearningasanattempt tosimulatethebrain.Moderndeeplearningdrawsinspirationfrommanyfields, especiallyappliedmathfundamentalslikelinearalgebra,probability,information theory,andnumericaloptimization.Whilesomedeeplearningresearcherscite neuroscienceasanimportantsourceofinspiration,othersarenotconcernedwith neuroscienceatall.

Itis worth notingthat theefforttounderstandhowthe brainworkson an algorithmic level is alive andwell.This endeavor isprimarily knownas "computationalneuroscience"andisaseparatefieldofstudyfromdeeplearning. Itiscommonforresearcherstomovebackandforthbetweenbothfields.The fieldofdeeplearningisprimarilyconcernedwithhowtobuildcomputersystems thatareabletosuccessfullysolvetasksrequiringintelligence,whilethefieldof computationalneuroscienceisprimarilyconcernedwithbuildingmoreaccurate modelsofhowthebrainactuallyworks.

Inthe1980s,thesecondwaveofneuralnetworkresearchemergedingreat partviaamovementcalled **connectionism**,or **paralleldistributedprocess-ing** (Rumelhart *etal.* , 1986c; McClelland *etal.* , 1995). Connectionismarosein thecontextofcognitivescience.Cognitivescienceisaninterdisciplinaryapproach tounderstandingthemind,combiningmultipledifferentlevelsofanalysis.During theearly1980s,mostcognitivescientistsstudiedmodelsofsymbolicreasoning. Despitetheirpopularity,symbolicmodelsweredifficulttoexplainintermsof howthebraincouldactuallyimplementthemusingneurons.Theconnectionists begantostudymodelsofcognitionthatcouldactuallybegroundedinneural implementations (TouretzkyandMinton , 1985),revivingmanyideasdatingback totheworkofpsychologistDonaldHebbinthe1940s (Hebb, 1949).

Thecentralideainconnectionismisthatalargenumberofsimplecomputational unitscanachieveintelligentbehaviorwhennetworkedtogether.Thisinsightapplies equallytoneuronsinbiologicalnervoussystemsasitdoestohiddenunitsin computationalmodels.

Severalkeyconceptsaroseduringtheconnectionismmovementofthe1980s thatremaincentraltotoday'sdeeplearning.

Oneoftheseconceptsisthatof **distributedrepresentation** (Hinton *etal.* , 1986).Thisistheideathateachinputtoasystemshouldberepresentedby manyfeatures,andeachfeatureshouldbeinvolvedintherepresentationofmany possibleinputs.Forexample,supposewehaveavisionsystemthatcanrecognize

媒体报道常常强调深度学习与大脑的相似性。虽然深度学习研究人员比其他机器学习领域（例如核机器或贝叶斯统计）的研究人员更有可能引用大脑作为参考，但不应将深度学习视为模拟大脑的尝试。现代深度学习从许多领域汲取灵感，

尤其是应用数学基础知识，如线性代数、概率、信息论和数值优化。虽然一些深度学习研究人员将神经科学视为重要的灵感来源，但其他人根本不关心神经科学。

值得注意的是，理解大脑在算法层面如何运作的努力依然活跃。这项努力主要被称为“计算神经科学”，是一个独立于深度学习的研究领域。研究人员在两个领域之间来回切换是很常见的。深度学习领域主要关注如何构建能够成功解决需要智能的任务的计算机系统，而计算神经科学领域主要关注如何构建更精确的大脑实际运作模型。

20 世纪 80 年代，神经网络研究的第二波浪潮兴起，很大程度上是通过一场被称为联结主义或并行分布式处理的运动（ ， ； ， ）出现的。联结主义出现于 Rumelhart 等人 1986c McClelland 等人 1995 认知科学的背景下。认知科学是一种理解思维的跨学科方法，结合了多个不同层次的分析。在 20 世纪 80 年代早期，大多数认知科学家研究符号推理模型。尽管符号模型很受欢迎，但它们很难解释大脑如何真正实现那些令人沉思的神经元。联结主义者开始研究能够真正建立在神经实现基础上的认知模型（TouretzkyandMinton 1985, ），复兴了许多可以追溯到心理学家唐纳德・赫布在 20 世纪 40 年代的工作的想法（ , ）。赫布 1949

联结主义的核心思想是，大量简单的计算单元在联网时可以实现智能行为。这一见解同样适用于生物神经系统中的神经元，就像它适用于计算模型中的隐藏单元一样。

20 世纪 80 年代的联结主义运动中出现了几个关键概念，这些概念至今仍是深度学习的核心。

其中一个概念是分布式表示（Hinton 等人，1986）。

cars,trucks,andbirds,andtheseobjectscaneachbered,green,orblue.Oneway ofrepresentingtheseinputswouldbetohaveaseparateneuronorhiddenunit thatactivatesforeachoftheninepossiblecombinations:redtruck,redcar,red bird,greentruck,andsoon.Thisrequiresninedifferentneurons,andeachneuron mustindependentlylearntheconceptofcolorandobjectidentity.Onewayto improveonthissituationistouseadistributedrepresentation,withthreeneurons describingthecolorandthreeneuronsdescribingtheobjectidentity.Thisrequires onlysixneuronstotalinsteadofnine,andtheneurondescribingrednessisableto learnaboutrednessfromimagesofcars,trucksandbirds,notjustfromimages ofonespecificcategoryofobjects. Theconceptofdistributedrepresentationis centraltothisbookandisdescribedingreaterdetailinchapter 15.

Anothermajoraccomplishmentoftheconnectionistmovementwasthesuccessfuluseofback-propagationtotraindeepneuralnetworkswithinternalrepresentationsandthepopularizationoftheback-propagationalgorithm (Rumelhart *etal.* , 1986a; LeCun, 1987).Thisalgorithmhaswaxedandwanedinpopularity but,asofthiswriting,isthedominantapproachtotrainingdeepmodels.

Duringthe1990s,researchersmadeimportantadvancesinmodelingsequences withneuralnetworks. Hochreiter (1991) and Bengio *etal.* (1994) identifiedsomeof thefundamentalmathematicaldifficultiesinmodelinglongsequences,describedin section 10.7. HochreiterandSchmidhuber1997 () introducedthelongshort-term memory(LSTM)networktoresolvesomeofthesedifficulties.Today,theLSTMis widelyusedformanysequencemodelingtasks,includingmanynaturallanguage processingtasksatGoogle.

Thesecondwaveofneuralnetworksresearchlasteduntilthemid-1990s.Venturesbasedonneuralnetworksandother AItechnologiesbegantomakeunrealistically ambitiousclaimswhileseekinginvestments.WhenAIresearchdidnotfulfill theseunreasonableexpectations,investorsweredisappointed.Simultaneously, otherfieldsofmachinelearningmadeadvances.Kernelmachines (Boser *etal.* , 1992CortesandVapnik; , 1995Schölkopf; *etal.* , 1999) andgraphicalmodels (Jordan, 1998) bothachievedgoodresultsonmanyimportanttasks.Thesetwofactors ledtoadeclineinthepopularityofneuralnetworksthatlasteduntil2007.

Duringthistime,neuralnetworkscontinuedtoobtainimpressiveperformance onsometasks (LeCun *etal.* , 1998b; Bengio *etal.* , 2001).TheCanadianInstitute forAdvancedResearch(CIFAR)helpedtokeepneuralnetworksresearchalive viaitsNeuralComputationandAdaptivePerception(NCAP)researchinitiative. Thisprogramunitedmachinelearningresearchgroupsledby GeoffreyHintonat UniversityofToronto,YoshuaBengioatUniversityofMontreal,andYannLeCun atNewYorkUniversity.ThemultidisciplinaryCIFARNCAPresearchinitiative

汽车、卡车和鸟类，这些物体可以是红色、绿色或蓝色。表示这些输入的一种方法是拥有一个单独的神经元或隐藏单元，可以激活九种可能的组合中的每一种：红色卡车、红色汽车、红色鸟、绿色卡车等等。这需要九个不同的神经元，每个神经元都必须独立学习颜色和物体身份的概念。改善这种情况的一种方法是使用分布式表示，其中三个神经元描述颜色，三个神经元描述物体身份。这只需要六个神经元而不是九个，并且描述红色的神经元能够从汽车、卡车和鸟类的图像中学习红色，而不仅仅是从某一特定类别的物体的图像中学习。分布式表示的概念是本书的核心，并将在第 15 章中详细描述。联结主义运动的另一项主要成就是成功使用反向传播来训练具有内部表示的深度神经网络，以及反向传播算法的普及（Rumelhart 等人，；，）。这种算法在 1986 年 LeCun 1987 年流行起来并逐渐衰落，但截至撰写本文时，它仍是训练深度模型的主要方法。

在 20 世纪 90 年代，研究人员在利用神经网络对序列进行建模方面取得了重要进展。() 和() 发现了长序列建模的一些基本数学难题，如第 10.7 节所述。Hochreiter 和 Schmidhuber1997() 引入了长短期记忆 (LSTM) 网络来解决其中的一些难题。如今，LSTM 被广泛用于许多序列建模任务，包括谷歌的许多自然语言处理任务。

神经网络研究的第二波浪潮一直持续到 20 世纪 90 年代中期。基于神经网络和其他人工智能技术的企业在寻求投资时开始做出不切实际的雄心勃勃的主张。当人工智能研究未能满足这些不合理的期望时，投资者感到失望。与此同时，机器学习的其他领域取得了进展。内核机器（Boser 等人，1992；CortesandVapnik，1995；Schölkopf，1999；Jor-；等人，）和图形模型（dan 1998，）都在许多重要任务上取得了良好的结果。这两个因素导致神经网络的流行度下降，直到 2007 年才有所下降。

在此期间，神经网络在某些任务上继续取得令人印象深刻的表现（，；，）。加拿大高等研究院（CIFAR）通过其神经计算与自适应感知（NCAP）研究计划，帮助神经网络研究保持活力。该计划联合了由多伦多大学的 Geoffrey Hinton、蒙特利尔大学的 Yoshua Bengio 和纽约大学的 Yann LeCun 领导的机器学习研究小组。这项多学科的 CIFAR 研究计划

alsoincludedneuroscientistsandexpertsinhumanandcomputervision.

Atthispoint,deepnetworksweregenerallybelievedtobeverydifficultto train.Wenowknowthatalgorithmsthathaveexistedsincethe1980sworkquite well,butthiswasnotapparentcirca2006.Theissueisperhapssimplythatthese algorithmsweretoocomputationallycostlytoallowmuchexperimentationwith thehardwareavailableatthetime.

Thethirdwaveofneuralnetworksresearchbeganwithabreakthroughin 2006.GeoffreyHintonshowedthatakindofneuralnetworkcalledadeepbelief networkcould be efficientlytrainedusinga strategycalled greedylayer-wise pretraining (Hinton *etal.* , 2006),whichwedescribeinmoredetailinsection 15.1. TheotherCIFAR-affiliatedresearchgroupsquicklyshowedthatthesamestrategy couldbeusedtotrainmanyotherkindsofdeepnetworks (Bengio *etal.* , 2007; Ranzato *etal.* , 2007a) andsystematicallyhelpedtoimprovegeneralizationon testexamples.Thiswaveofneuralnetworksresearchpopularizedtheuseofthe term“deeplearning”toemphasizethatresearcherswerenowabletotraindeeper neuralnetworksthanhadbeenpossiblebefore,andtofocusattentiononthe theoreticalimportanceofdepth (BengioandLeCun , 2007; DelalleauandBengio , 2011Pascanu; *etal.* , 2014aMontufar; *etal.* , 2014).Atthistime,deepneural networksoutperformedcompetingAIsystemsbasedonothermachinelearning technologiesaswellashand-designedfunctionality.Thisthirdwaveofpopularity ofneuralnetworkscontinuestothetimeofthiswriting,thoughthefocusofdeep learningresearchhaschangeddramaticallywithinthetimeofthiswave.The thirdwavebeganwithafocusonnewunsupervisedlearningtechniquesandthe abilityofdeepmodelstogeneralizewellfromsmalldatasets,buttodaythereis moreinterestinmucholdersupervisedlearningalgorithmsandtheabilityofdeep modelstoleveragelargelabeleddatasets.

1.2.2IncreasingDatasetSizes

Onemaywonderwhydeeplearninghasonlyrecentlybecomerecognizedasacrucial technologyeventhoughthefirstexperimentswithartificialneuralnetworkswere conductedinthe1950s.Deeplearninghasbeensuccessfullyusedincommercial applicationssincethe1990sbutwasoftenregardedasbeingmoreofanartthana technologyandsomethingthatonlyanexpertcoulduse,untilrecently.Itistrue thatsomeskillisrequiredtogetgoodperformancefromadeeplearningalgorithm. Fortunately,theamountofskillrequiredreducesastheamountoftrainingdata increases.Thelearningalgorithmsreachinghumanperformanceoncomplextasks todayarenearlyidenticaltothelearningalgorithmsthatstruggledtosolvetoy problemsinthe1980s,thoughthemodelswetrainwiththesealgorithmshave

还包括神经科学家以及人类和计算机视觉专家。

此时，人们普遍认为深度网络很难训练。我们现在知道，自 20 世纪 80 年代以来存在的算法已经运行得很好，但这在 2006 年左右并不明显。问题可能只是这些算法的计算成本太高，无法利用当时可用的硬件进行大量实验。

第三次神经网络研究浪潮始于 2006 年的一项突破。Geoffrey Hinton 表明，一种称为深度信念网络的神经网络可以通过一种称为贪婪逐层预训练的策略进行有效训练（，），我们将在第 3 节中对此进行更详细的描述。Hinton 等人。 2006 15.1 其他 CIFAR 附属研究小组很快证明，同样的策略可以用于训练许多其他类型的深度网络（，；Bengio 等人 2007；Ranzato 等人 2007，），并系统地帮助提高测试样本的泛化能力。这波神经网络研究浪潮普及了“深度学习”一词的使用，强调研究人员现在能够训练比以前更深的神经网络，并将注意力集中在

深度的理论重要性（BengioandLeCun 2007；DelalleauandBengio 2011；Pascanu 2014a；Montufar 2014；etal.，；etal.，）。目前，深度神经网络
网络的表现优于基于其他机器学习技术以及手工设计功能的竞争人工智能系统。神经网络的第三次流行浪潮一直持续到本文撰写之时，尽管深度学习研究的重点在这一浪潮中发生了巨大变化。第三次浪潮开始时关注的是新的无监督学习技术和深度模型从小数据集很好地概括的能力，但今天人们对多持股监督学习算法和深度模型利用大型标记数据集的能力更感兴趣。

1.2.2 增加数据集大小

人们或许会疑惑，尽管深度学习的首次实验早在 20 世纪 50 年代就已开展，但为什么它直到最近才被公认为一项关键技术。深度学习自 20 世纪 90 年代以来就已成功应用于商业领域，但直到最近，它通常被认为更像是一门艺术而非技术，而且只有专家才能使用。的确，要使深度学习算法获得良好的性能，需要一些技能。幸运的是，随着训练数据量的增加，所需的技能数量会减少。如今，在复杂任务上达到人类水平的学习算法，与 20 世纪 80 年代难以解决玩具问题的学习算法几乎完全相同，尽管我们用这些算法训练的模型已经……

undergonechangesthatsimplifythetrainingofverydeeparchitectures.Themost importantnewdevelopmentisthattodaywecanprovidethesealgorithmswith theresourcestheyneedtosucceed.Figure 1.8 showshowthesizeofbenchmark datasetshasexpandedremarkablyovertime.Thistrendisdrivenbytheincreasing digitizationofsociety.Asmoreandmoreofouractivitiestakeplaceoncomputers, moreandmoreofwhatwedoisrecorded.Asourcomputersareincreasingly networkedtogether,itbecomeseasiertocentralizetheserecordsandcuratethem intoadatasetappropriateformachinelearningapplications.Theageof"BigData"

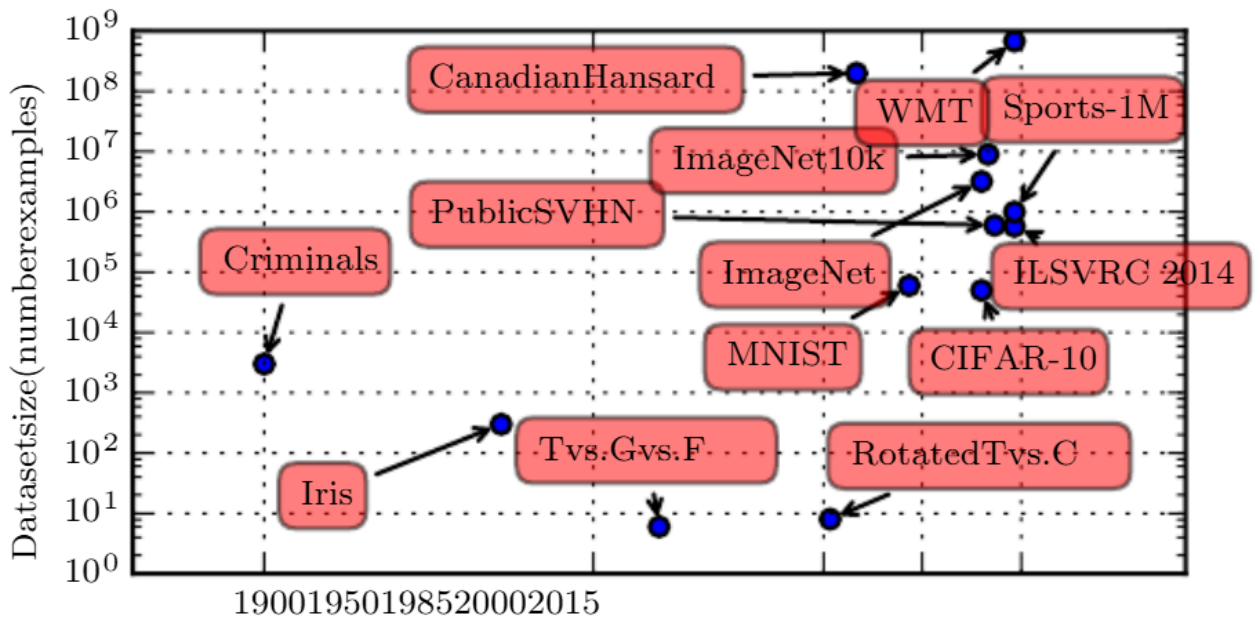

Figure1.8:Increasingdatasetsizeovertime.Intheearly1900s,statisticiansstudied datasetsusinghundredsorthousandsofmanuallycompiledmeasurements (Garson, 1900; Gosset, 1908Anderson; , 1935Fisher; , 1936).Inthe1950sthroughthe1980s,thepioneers ofbiologicallyinspiredmachinelearningoftenworkedwithsmallsyntheticdatasets,such aslow-resolutionbitmapsofletters,thatweredesignedtoincurlowcomputationalcostand demonstratethatneuralnetworkswereabletolearnspecifickindsoffunctions (Widrow andHoff , 1960Rumelhart; *etal.* , 1986b).Inthe1980sand1990s,machinelearning becamemorestatisticalandbegantoleveragelargerdatasetscontainingtensofthousands ofexamples,suchastheMNISTdataset(showninfigure 1.9) ofscansofhandwritten numbers (LeCun *etal.* , 1998b).Inthefirstdecadeofthe2000s,moresophisticated datasetsofthissamesize,suchastheCIFAR-10dataset (KrizhevskyandHinton , 2009), continuedtobeproduced.Towardtheendofthatdecadeandthroughoutthefirsthalfof the2010s,significantlylargerdatasets,containinghundredsofthousandstotensofmillions ofexamples,completelychangedwhatwaspossiblewithdeeplearning.Thesedatasets includedthepublicStreetViewHouseNumbersdataset (Netzer *etal.* , 2011),various versionsoftheImageNetdataset (Deng *etal.* , 2009, 2010a; Russakovsky *etal.* , 2014a), andtheSports-1Mdataset (Karpathy *etal.* , 2014).Atthetopofthegraph,weseethat datasetsoftranslatedsentences,suchasIBM'sdatasetconstructedfromtheCanadian Hansard (Brown *etal.* , 1990) andtheWMT2014EnglishtoFrenchdataset (Schwenk, 2014),aretypicallyfaraheadofotherdatasetsizes.

经历了简化非常深层架构的训练的改变。最重要的新发展是，今天我们可以为这些算法提供它们成功所需的资源。图显示，基准 1.8 数据集的大小随着时间的推移显着扩大。这一趋势是由社会日益数字化所驱动的。随着我们越来越多的活动在计算机上进行，我们所做的也越来越多地被记录下来。随着我们的计算机越来越多地联网在一起，将这些记录集中起来并将它们整理成适用于机器学习应用程序的数据集变得更加容易。“大数据”时代

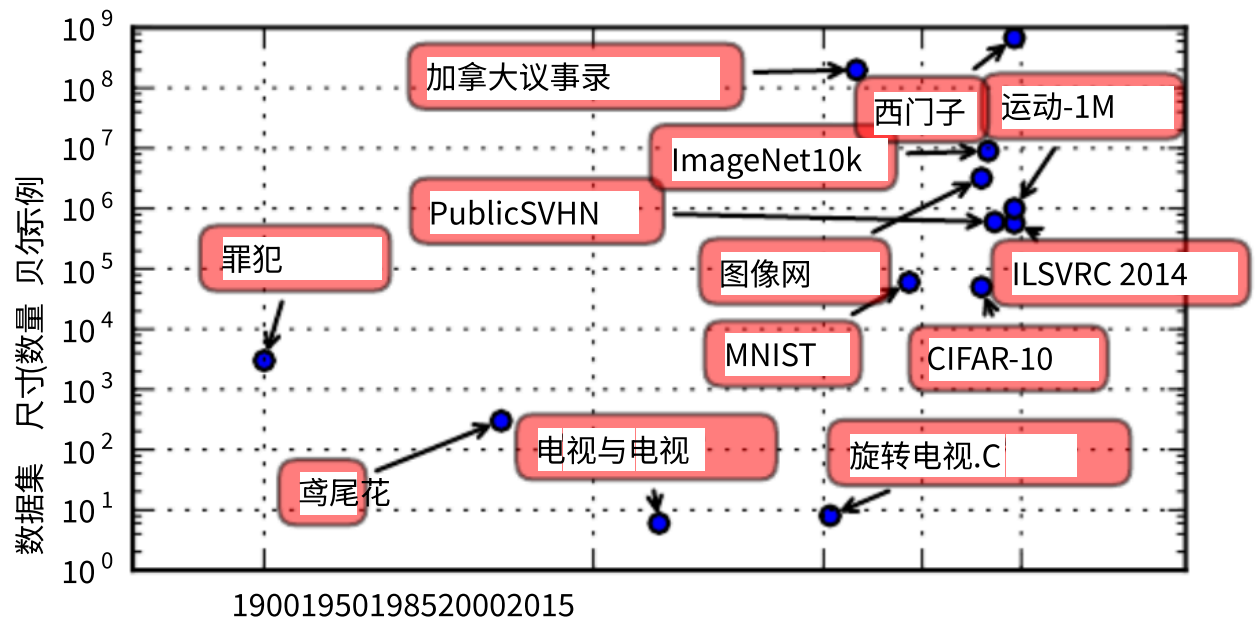

图 1.8：数据集大小随时间推移不断增加。20 世纪初，统计学家使用数百上千个手动编制的测量数据来研究数据集（Garson 1900、Gosset 1908、Anderson 1935、Fisher 1936，1991；2001；2002；2003；2004，2005）。20 世纪 50 年代到 80 年代，受生物启发的机器学习的先驱者经常使用小型合成数据集，例如低分辨率字母位图，旨在降低计算成本并证明神经网络能够学习特定类型的函数（Widrow and Hoff 1960、Rumelhart 1986b，1992；2003 年）。20 世纪 80 年代和 90 年代，机器学习变得更加具有统计意义，并开始利用包含数万个示例的更大数据集，例如包含 1.9 手写数字扫描图的 MNIST 数据集（如图所示）(,)。在 21 世纪的第一个十年，同样大小的更复杂的 LeCun et al. 1998b 数据集，例如 CIFAR-10 数据集（KrizhevskyandHinton 2009, ）不断涌现。在 21 世纪第一个十年的末期以及整个 21 世纪第一个十年的前半段，包含数亿到数千万个示例的更大的数据集彻底改变了深度学习的可能性。这些数据集包括公共 StreetViewHouseNumbers 数据集（ , ）以及各种 Netzer et al. 2011 版 ImageNet 数据集（ , ， ；Deng et al. 2009 2010a Russakovsky 2014aetal. , ）和 Sports-1M 数据集（ , ）。在图的顶部，我们看到 Karpathy et al. 2014

翻译句子的数据集，例如 IBM 根据加拿大 Hansard 构建的数据集（Brown 1990 等,）和 WMT2014 英语到法语数据集（Schwenk, 2014），通常远远领先于其他数据集的规模。

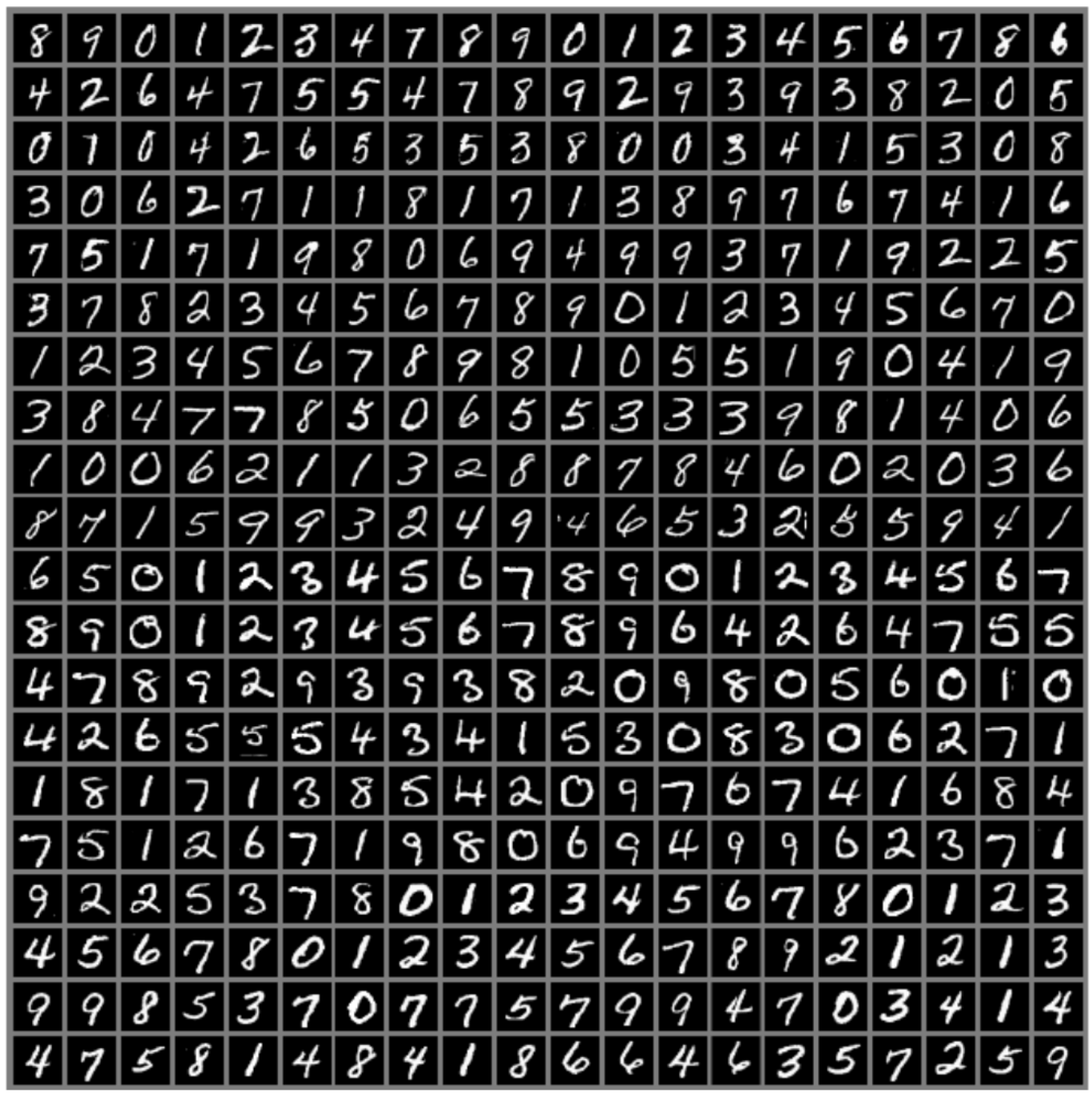

Figure1.9:ExampleinputsfromtheMNISTdataset.The“NIST”standsforNational InstituteofStandardsandTechnology,theagencythatoriginallycollectedthisdata. The“M”standsfor“modified,”sincethedatahasbeenpreprocessedforeasierusewith machinelearningalgorithms.TheMNISTdatasetconsistsofscansofhandwrittendigits andassociatedlabelsdescribingwhichdigit0–9iscontainedineachimage.Thissimple classificationproblemisoneofthesimplestandmostwidelyusedtestsindeeplearning research.Itremainspopulardespitebeingquiteeasyformoderntechniquestosolve. GeoffreyHintonhasdescribeditas“the *drosophila* ofmachinelearning,”meaningthatit enablesmachinelearningresearcherstostudytheiralgorithmsincontrolledlaboratory conditions,muchasbiologistsoftenstudyfruitflies.

hasmademachinelearningmucheasierbecausethekeyburdenofstatistical estimation—generalizingwelltonewdataafterobservingonlyasmallamount ofdata—hasbeenconsiderablylightened.Asof2016,aroughruleofthumb isthatasupervisedeeplearningalgorithmwillgenerallyachieveacceptable performancewitharound5,000labeledexamplespercategoryandwillmatchor

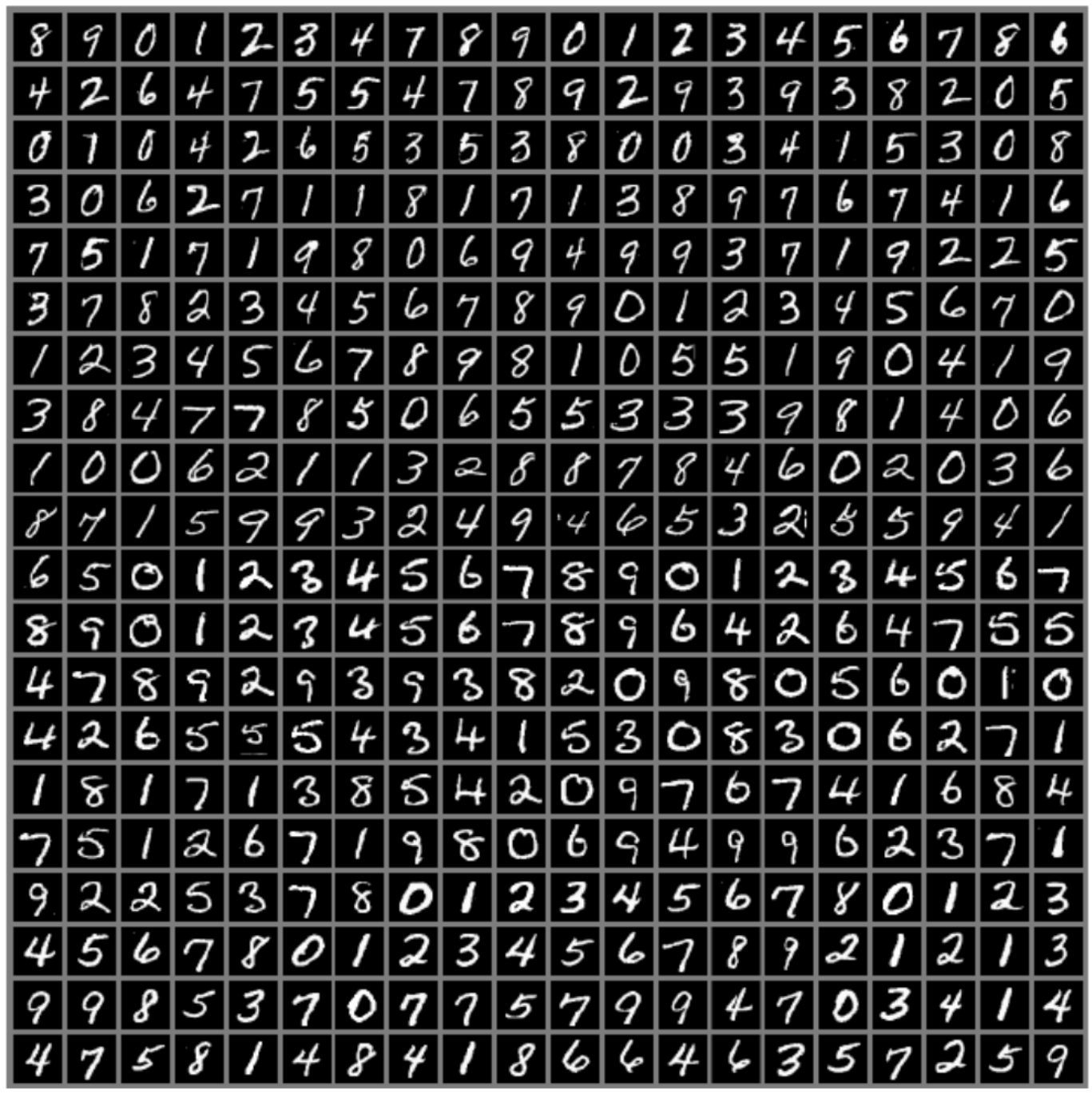

图 1.9：MNIST 数据集的示例输入。“NIST”代表美国国家标准与技术研究院 (National Institute of Standards and Technology)，该机构最初收集了这些数据。“M”代表“已修改”，因为数据已进行预处理，以便于机器学习算法使用。MNIST 数据集包含手写数字扫描图以及描述每幅图像中包含哪个数字 0-9 的相关标签。这个简单的分类问题是深度学习研究中简单且最广泛的测试之一。尽管对于现代技术来说，它很容易解决，但仍然很受欢迎。

杰弗里·辛顿 (Geoffrey Hinton) 将其描述为“机器学习中的果蝇”，这意味着它使机器学习研究人员能够在受控的实验室条件下研究他们的算法，就像生物学家经常研究果蝇一样。

使得机器学习变得更容易，因为统计估计的关键负担——在仅观察少量数据后很好地推广到新数据——已经大大减轻。截至 2016 年，一个粗略的经验法则是，监督深度学习算法通常会在每个类别中大约 5,000 个带标签的示例中获得可接受的性能，并且将匹配或

exceedhumanperformancewhentrainedwithadatasetcontainingatleast10 millionlabeledexamples.Workingsuccessfullywithdatasetssmallerthanthisis animportantresearcharea,focusinginparticularonhowwecantakeadvantage oflargequantitiesofunlabeledexamples,withunsupervisedorsemi-supervised learning.

1.2.3IncreasingModelSizes

Anotherkeyreasonthatneuralnetworksarewildlysuccessfultodayafterenjoying comparativelylittlesuccesssincethe1980sisthatwehavethecomputational resourcestorunmuchlargermodelstoday.Oneofthemaininsightsofconnection- ismisthatanimalsbecomeintelligentwhenmanyoftheirneuronsworktogether. Anindividualneuronorsmallcollectionofneuronsisnotparticularlyuseful.

Biologicalneuronsarenotespeciallydenselyconnected.Asseeninfigure 1.10, ourmachinelearningmodelshavehadanumberofconnectionsperneuronwithin anorderofmagnitudeofevenmammalianbrainsfordecades.

Intermsofthetotalnumberofneurons,neuralnetworkshavebeenastonishingly smalluntilquiterecently,asshowninfigure 1.11.Sincetheintroductionofhidden units,artificialneuralnetworkshavedoubledinsizeroughlyevery2.4years.This growthisdrivenbyfastercomputerswithlargermemoryandbytheavailability oflargerdatasets.Largernetworksareabletoachievehigheraccuracyonmore complextasks.Thistrendlookssettocontinuefordecades.Unlessnewtechnologies enablefasterscaling,artificialneuralnetworkswillnothavethesamenumber ofneuronsasthehumanbrainuntilatleastthe2050s. Biologicalneuronsmay representmorecomplicatedfunctionsthancurrentartificialneurons,sobiological neuralnetworksmaybeevenlargerthanthisplotportrays.

Inretrospect,itisnotparticularlysurprisingthatneuralnetworkswithfewer neuronsthanaleechwereunabletosolvesophisticatedartificialintelligenceprob- lems.Eventoday'snetworks,whichweconsiderquitelargefromacomputational systemspointofview,aresmallerthanthenervoussystemofevenrelatively primitivevertebrateanimalslikefrogs.

Theincreaseinmodelsizeovertime,duetotheavailabilityoffasterCPUs, theadventofgeneralpurposeGPUs(describedinsection 12.1.2),fasternetwork connectivityandbettersoftwareinfrastructurefordistributedcomputing,isoneof themostimportanttrendsinthehistoryofdeeplearning.Thistrendisgenerally expectedtocontinuewellintothefuture.

在使用至少包含 1000 万个标记示例的数据集进行训练时，其表现将超越人类。成功处理小于此的数据集是一个重要的研究领域，尤其关注如何利用无监督或半监督学习来利用大量未标记的示例。

1.2.3 增加模型尺寸

神经网络自 1980 年代以来一直没有取得太大成功，而如今却异常成功的另一个关键原因是，我们如今拥有了处理更大模型的计算资源。联结主义的主要见解之一是，当许多神经元协同工作时，动物就会变得聪明。

单个神经元或小群神经元并不是特别有用。

生物神经元的连接并不是特别密集。如图 1 所示，几十年来，我们的机器学习模型每个神经元的连接数量甚至与哺乳动物大脑的连接数量相当。

就神经元的总数而言，神经网络直到最近才变得非常小，如图所示。自从引入 1.11 个隐藏单元以来，人工神经网络的规模大约每 2.4 年翻一番。这种增长是由速度更快、内存更大的计算机以及更大数据集的可用性驱动的。更大的网络能够在更复杂的任务上实现更高的准确率。这种趋势看起来将持续几十年。除非新技术能够实现更快的扩展，否则至少在 2050 年代之前，人工神经网络的神经元数量都不会与人脑相同。生物神经元可能比当前的人工神经元代表更复杂的功能，因此生物神经网络可能比该图描绘的还要大。

回想起来，神经元比水蛭少的神经网络无法解决复杂的人工智能问题，这并不特别令人惊讶。即使从计算系统的角度来看，今天的网络已经相当大了，但它甚至比青蛙等相对原始的脊椎动物的神经系统还要小。

由于更快的 CPU 的出现、通用 GPU 的出现（如第节所述）、更快的网络连接以及更好的分布式计算软件基础设施，模型大小随着时间的推移而增加，这是深度学习历史上最重要的趋势之一。人们普遍预计这种趋势将持续到未来。

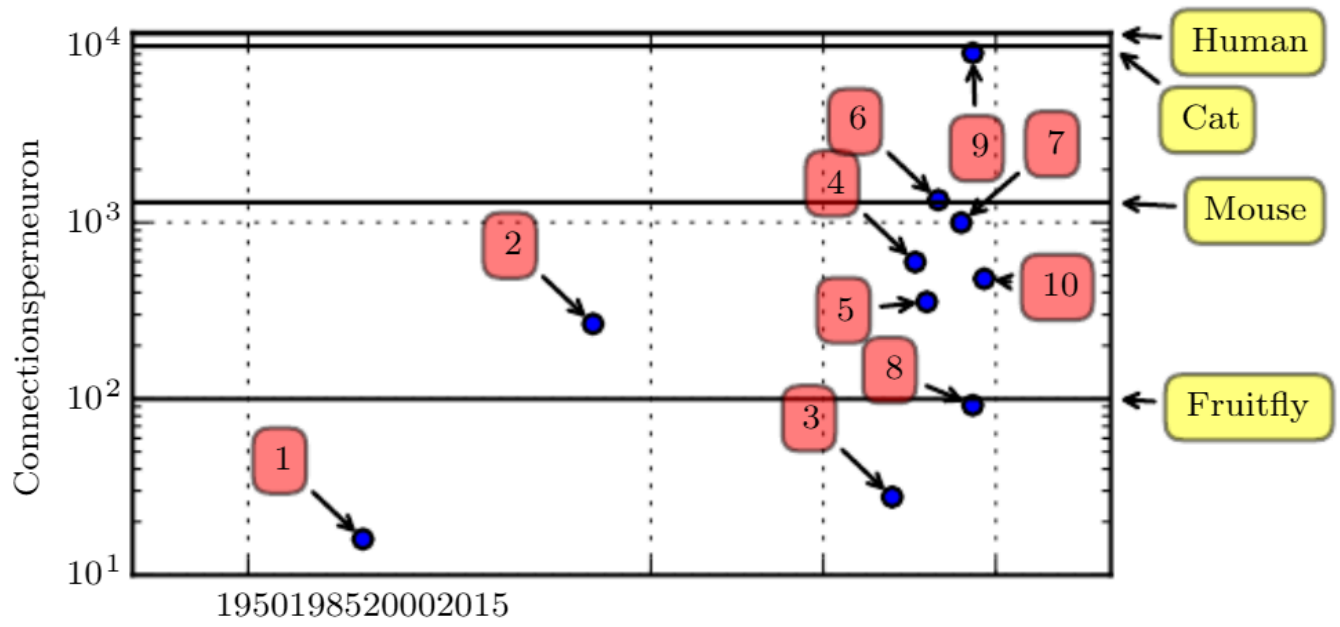

Figure1.10:Numberofconnectionsperneuronovertime.Initially,thenumberofconnectionsbetweenneuronsinartificialneuralnetworkswaslimitedbyhardwarecapabilities. Today,thenumberofconnectionsbetweenneuronsismostlyadesignconsideration.Some artificialneuralnetworkshavenearlyasmanyconnectionsperneuronasacat,andit isquitecommonforotherneuralnetworkstohaveasmanyconnectionsperneuronas smallermammalslikemice.Eventhehumanbraindoesnothaveanexorbitantamount ofconnectionsperneuron.Biologicalneuralnetworksizesfrom Wikipedia (2015).

1.Adaptivelinearelement (WidrowandHoff , 1960)
2.Neocognitron (Fukushima, 1980)
3.GPU-acceleratedconvolutionalnetwork (Chellapilla *etal.* , 2006)
4.DeepBoltzmannmachine (SalakhutdinovandHinton , 2009a)
5.Unsupervisedconvolutionalnetwork (Jarrett *etal.* , 2009)
6.GPU-acceleratedmultilayerperceptron (Ciresan *etal.* , 2010)
7.Distributedautoencoder (Le *etal.* , 2012)
8.Multi-GPUconvolutionalnetwork (Krizhevsky *etal.* , 2012)
9.COTSHPCunsupervisedconvolutionalnetwork (Coates *etal.* , 2013)
10.GoogLeNet (Szegedy *etal.* , 2014a)

1.2.4IncreasingAccuracy,ComplexityandReal-WorldImpact

Sincethe1980s,deeplearninghasconsistentlyimprovedinitsabilitytoprovide accuraterecognitionandprediction.Moreover,deeplearninghasconsistentlybeen appliedwithsuccesstobroaderandbroadersetsofapplications.

Theearliestdeepmodelswereusedtorecognizeindividualobjectsintightly cropped,extremelysmallimages (Rumelhart *etal.* , 1986a).Sincethentherehas beenagradualincreaseinthesizeofimagesneuralnetworkscouldprocess.Modern objectrecognitionnetworksprocessrichhigh-resolutionphotographsanddonot

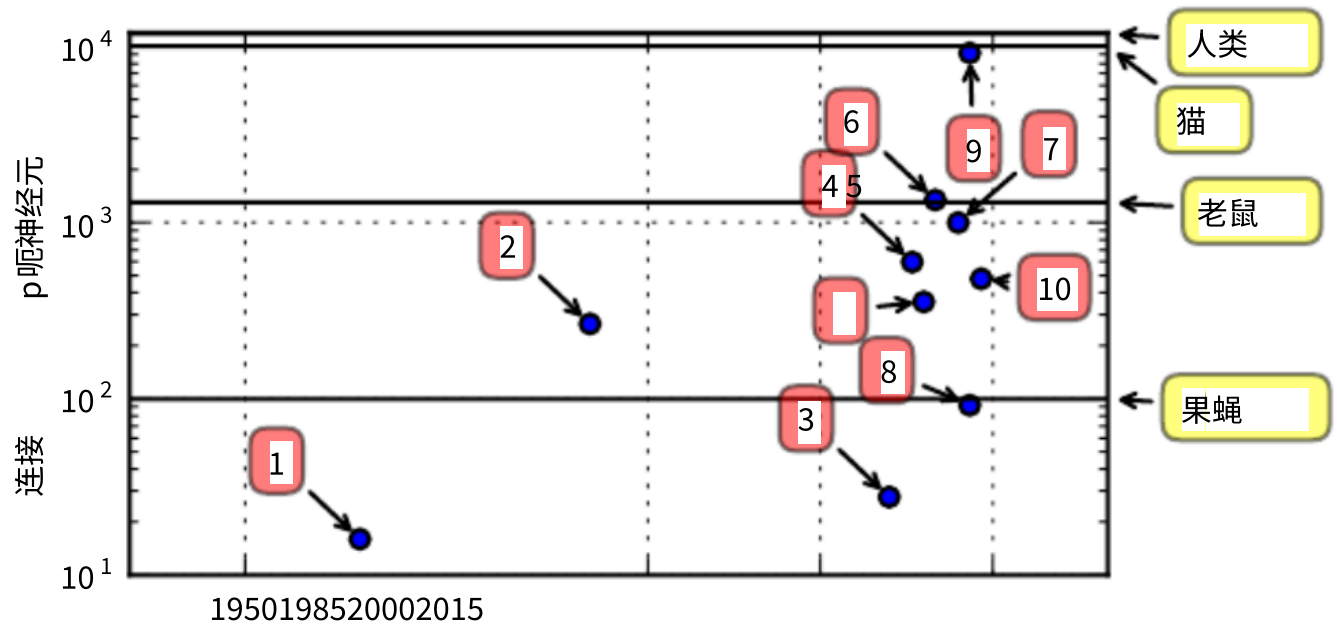

图 1.10：每个神经元的连接数量随时间变化。最初，人工神经网络中神经元之间的连接数量受限于硬件能力。如今，神经元之间的连接数量主要取决于设计。一些人工神经网络早期每个神经元的连接数量与猫一样多，其他神经网络每个神经元的连接数量也相当常见。体型较小的哺乳动物，如老鼠。即使人类大脑，每个神经元的连接数量也远不及人类。生物神经网络的大小来自（）。维基百科 2015

1.自适应线性单元（，）WidrowandHoff 1960

2.Neocognitron（福岛 1980，）

3.GPU 加速卷积网络（，）Chellapilla 等人，2006

4.深玻尔兹曼机（SalakhutdinovandHinton 2009a，）

5.无监督卷积网络（，）Jarrett 等人，2009

6.GPU 加速多层感知器（，）Ciresan 等人，2010

7.分布式自动编码器（Le et al. 2012）

8.多 GPU 卷积网络（，）Krizhevsky 等人，2012

9.COTSHPCunsupervisedconvolutionalnetwork（，）Coates 等人，2013

10.GoogLeNet (,)Szegedy 等人。 2014a

1.2.4 提高准确性、复杂性和现实世界影响

自 20 世纪 80 年代以来，深度学习不断提高其提供准确识别和预测的能力。此外，深度学习已成功应用于越来越广泛的应用领域。

最早的深度模型用于识别紧密裁剪、极小图像中的单个物体（，）。自 1986 年 Rumelhart 等人提出以来，神经网络可以处理的图像大小逐渐增加。现代物体识别网络可以处理丰富的高分辨率照片，并且不会

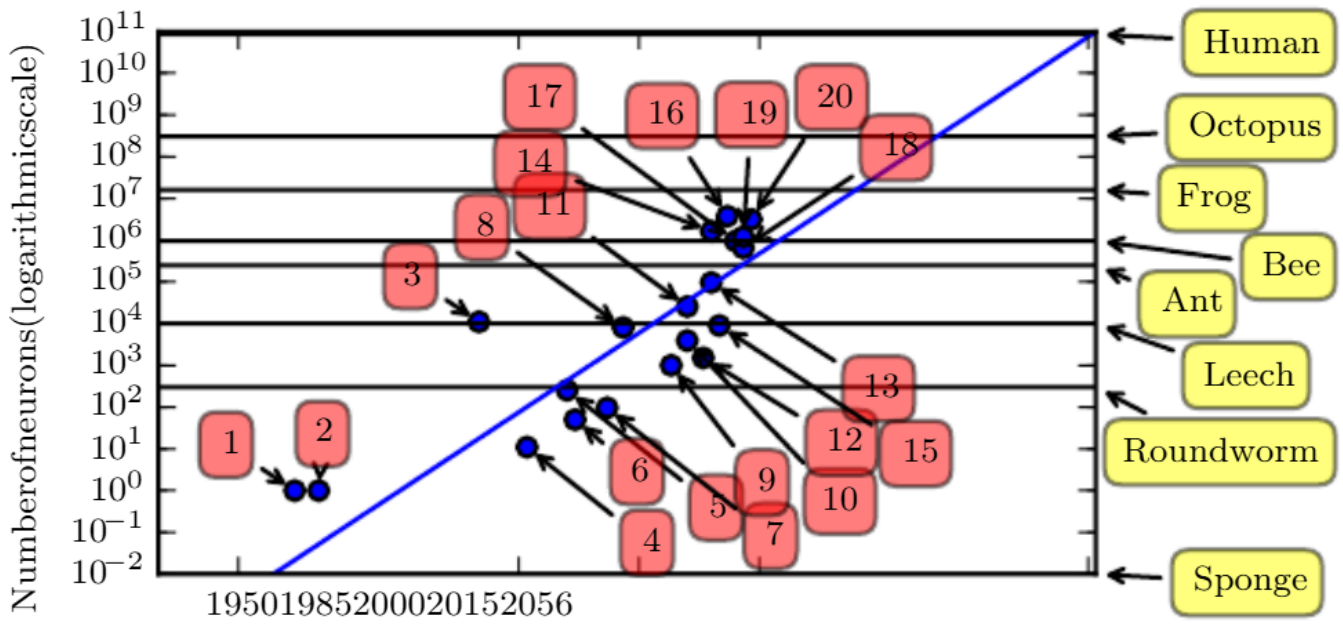

Figure1.11:Increasingneuralnetworksizeovertime.Sincetheintroductionofhidden units,artificialneuralnetworkshavedoubledinsizeroughlyevery2.4years.Biological neuralnetworksizesfrom Wikipedia (2015).

1.Perceptron (Rosenblatt, 1958, 1962)

2.Adaptivelinearelement (WidrowandHoff , 1960)

3.Neocognitron (Fukushima, 1980)

4.Earlyback-propagationnetwork (Rumelhart *etal.* , 1986b)

5.Recurrentneuralnetworkforspeechrecognition (RobinsonandFallside , 1991)

6.Multilayerperceptronforspeechrecognition (Bengio *etal.* , 1991)

7.Meanfieldsigmoidbeliefnetwork (Saul *etal.* , 1996)

8.LeNet-5 (LeCun *etal.* , 1998b)

9.Echostatenetwork (JaegerandHaas , 2004)

10.Deepbeliefnetwork (Hinton *etal.* , 2006)

11.GPU-acceleratedconvolutionalnetwork (Chellapilla *etal.* , 2006)

12.DeepBoltzmannmachine (SalakhutdinovandHinton , 2009a)

13.GPU-accelerateddeepbeliefnetwork (Raina *etal.* , 2009)

14.Unsupervisedconvolutionalnetwork (Jarrett *etal.* , 2009)

15.GPU-acceleratedmultilayerperceptron (Ciresan *etal.* , 2010)

16.OMP-1network (CoatesandNg , 2011)

17.Distributedautoencoder (Le *etal.* , 2012)

18.Multi-GPUconvolutionalnetwork (Krizhevsky *etal.* , 2012)

19.COTSHPCunsupervisedconvolutionalnetwork (Coates *etal.* , 2013)

20.GoogLeNet (Szegedy *etal.* , 2014a)

havearequirementthatthephotobecroppedneartheobjecttoberecognized (Krizhevsky *etal.* , 2012).Similarly,theearliestnetworkscouldrecognizeonly twokindsofobjects(orinsomecases,theabsenceorpresenceofasinglekindof object),whilethesemodernnetworkstypicallyrecognizeatleast1,000different categoriesofobjects. ThelargestcontestinobjectrecognitionistheimageNet

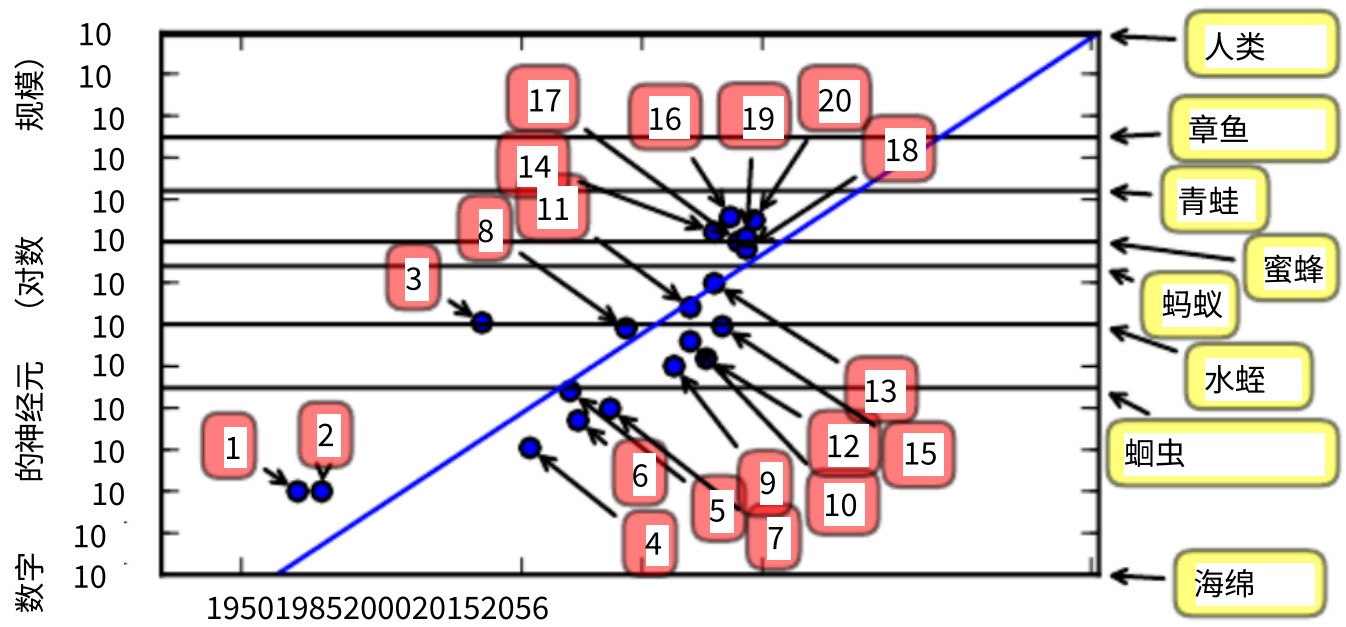

图 1.11：神经网络规模随时间推移而增加。自引入隐藏单元以来，人工神经网络的规模大约每 2.4 年翻一番。生物神经网络的规模从（）。维基百科 2015

1.感知器（,,）Rosenblatt 1958 1962

2.自适应线性单元（，）WidrowandHoff 1960

3.Neocognitron（福岛 1980,）

4.早期反向传播网络（，）Rumelhart 等人，1986b

5.Recurrentneuralnetworkforspeechrecognition（RobinsonandFallside 1991,）

6.Multilayerperceptronforspeechrecognition（，）Bengio 等人，1991

7.Meanfieldsigmoidbeliefnetwork（,）Saul 等人，1996 年

8.LeNet-5 (,)LeCun 等人。 1998b

9.Echostatenetwork（,）JaegerandHaas 2004

10.深度信念网络（Hinton 2006 等,）

11.GPU 加速卷积网络（,）Chellapilla 等人，2006 年

12.深玻尔兹曼机（SalakhutdinovandHinton 2009a,）

13.GPU 加速的深度信念网络（Raina 等人，2009 年）

14.无监督卷积网络（，）Jarrett 等人，2009

15.GPU 加速多层感知器（，）Ciresan 等人，2010

16.OMP-1 网络（,）CoatesandNg 2011

17.分布式自动编码器（Le et al. 2012）

18.多 GPU 卷积网络（，）Krizhevsky 等人，2012

19.COTSHPCunsupervisedconvolutionalnetwork（,）Coates 等人，2013

20.GoogLeNet (,)Szegedy 等人。 2014a

要求将照片裁剪到物体附近才能识别（，）。同样，最早的网络只能识别 Krizhevsky etal. 2012 两种物体（或者在某些情况下，识别单一类型物体的存在或不存在），而这些现代网络通常可以识别至少 1,000 种不同类别的物体。物体识别领域最大的竞赛是 ImageNet

LargeScaleVisualRecognitionChallenge(ILSVRC)heldeachyear.Adramatic momentinthemeteoricriseofdeeplearningcamewhenaconvolutionalnetwork wonthischallengeforthefirsttimeandbyawidemargin,bringingdownthe state-of-the-arttop-5errorratefrom26.1percentto15.3percent (Krizhevsky *etal.* , 2012),meaningthattheconvolutionalnetworkproducesarankedlistof possiblecategoriesforeachimage,andthecorrectcategoryappearedinthefirst fiveentriesofthislistforallbut15.3percentofthetestexamples.Sincethen, thesecompetitionsareconsistentlywonbydeepconvolutionalnets,andasofthis writing,advancesindeeplearninghavebroughtthelatesttop-5errorrateinthis contestdownto3.6percent,asshowninfigure 1.12.

Deeplearninghasalsohadadramaticimpactonspeechrecognition.After improvingthroughoutthe1990s,theerrorratesforspeechrecognitionstagnated startinginabout2000.Theintroductionofdeeplearning (Dahl *etal.* , 2010; Deng *etal.* , 2010b; Seide *etal.* , 2011; Hinton *etal.* , 2012a) tospeechrecognitionresulted inasuddendropinerrorrates,withsomeerrorratescutinhalf.Weexplorethis historyinmoredetailinsection 12.3.

Deepnetworkshavealsohadspectacularsuccessesforpedestriandetectionand imagesegmentation (Sermanet *etal.* , 2013; Farabet *etal.* , 2013Couprie; *etal.* , 2013) andyieldedsuperhumanperformanceintrafficsignclassification (Ciresan *etal.* , 2012).

Atthesametimethatthescaleandaccuracyofdeepnetworkshaveincreased,

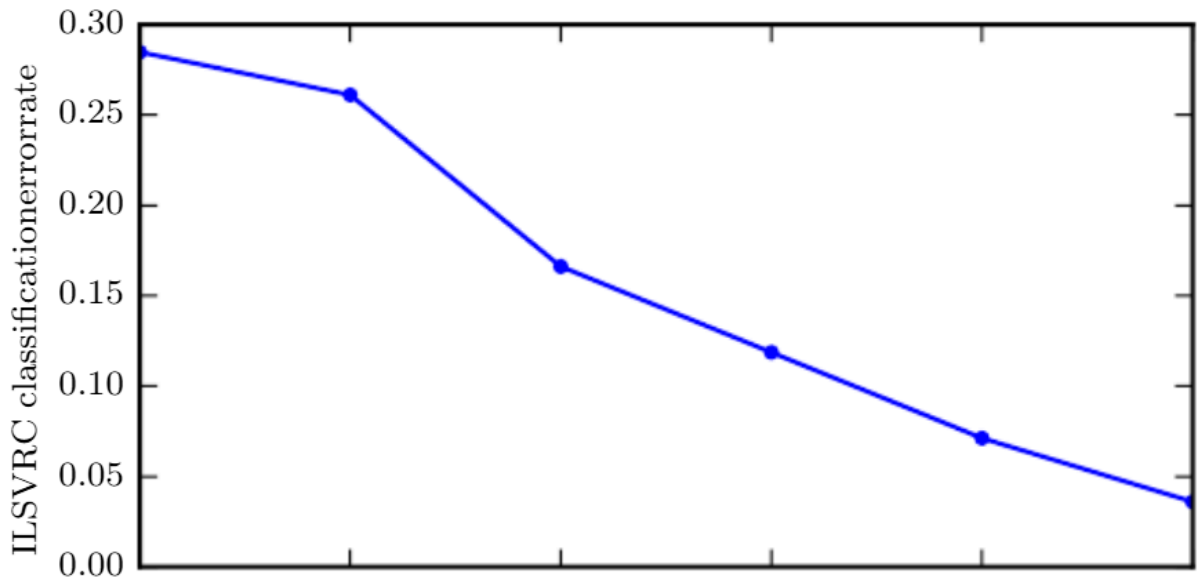

Figure1.12:Decreasingerrorrateovertime.Sincedeepnetworksreachedthescale necessarytocompeteintheImageNetLargeScaleVisualRecognitionChallenge,they haveconsistentlywonthecompetitioneveryyear,yieldinglowerandlowererrorrates eachtime.Datafrom Russakovsky *etal.* (2014b) and He *etal.* (2015).

大规模视觉识别挑战赛 (ILSVRC) 每年举办一次。深度学习飞速发展的戏剧性时刻是，一个卷积网络首次以大幅优势赢得该挑战，将当时最先进的 top-5 错误率从 26.1% 降至 15.3% (Krizhevsky et al.,)，这意味着卷积网络为每个图像生成一个包含 2012 个可能类别的排序列表，并且除了 15.3% 的测试示例外，所有测试示例的正确类别都出现在此列表的前五个条目中。自那以后，这些比赛一直由深度卷积网络赢得，截至本文撰写时，深度学习的进步已将本次竞赛的最新 top-5 错误率降至 3.6%，如图 1.12 所示深度学习也对语音识别产生了巨大的影响。在经历了整个 20 世纪 90 年代的改进之后，语音识别的错误率从 2000 年左右开始停滞不前。深度学习（ , ; Dahl et al. 2010 Deng et al. et al. 2010b Seide , ; 2011 Hinton , ）引入语音识别导致错误率突然下降，有些错误率减少了一半。我们将在第 12.3 节中更详细地探讨这段历史。深度网络在行人检测和图像分割（ , ; Sermanet et al. 2013 Farabet 2013Couprieetal. , ; et al. , 2013）并在交通标志分类中取得了超越人类的表现（ Ciresan et al. , ）。2012 在深度网络规模和准确性不断提高的同时，

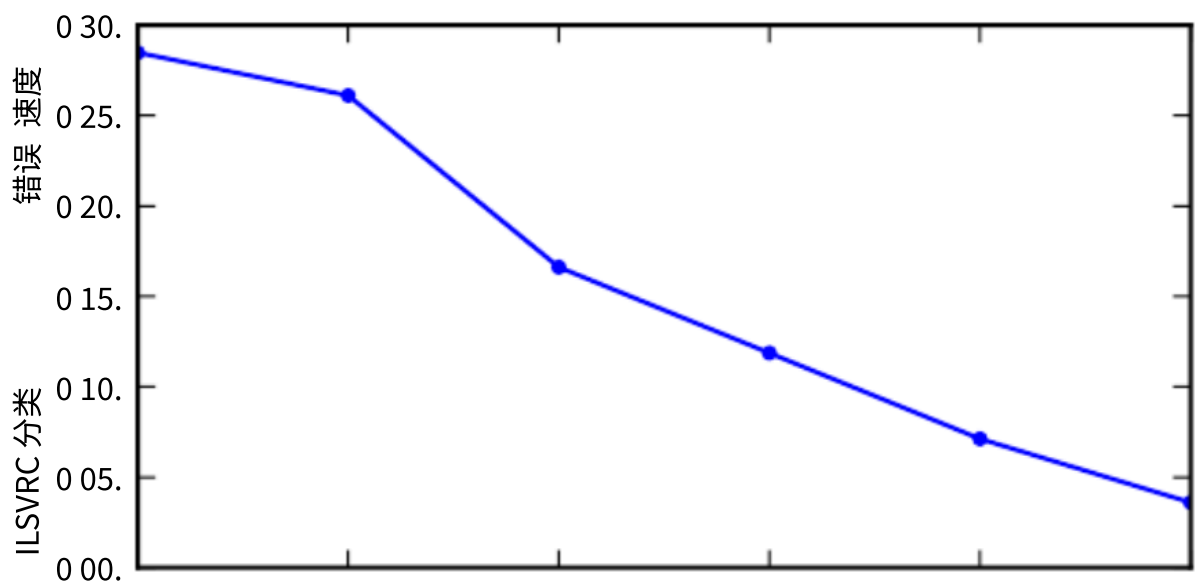

图 1.12：随着时间的推移，错误率不断降低。自从深度网络达到参加 ImageNet 大规模视觉识别挑战赛所需的规模以来，它们每年都连续赢得比赛，并且每次的错误率都越来越低。数据来自 Russakovsky 2014b、He 2015 等（ ）和等（ ）。

sohasthecomplexityofthetasksthattheycansolve. Goodfellow *etal.* (2014d) showedthatneuralnetworkscouldlearntooutputanentiresequenceofcharacters transcribedfromanimage,ratherthanjustidentifyingasingleobject.Previously, itwaswidelybelievedthatthiskindoflearningrequiredlabelingoftheindividual elementsofthesequence (GülçehreandBengio , 2013).Recurrentneuralnetworks, suchastheLSTMsequencemodelmentionedabove,arenowusedtomodel relationshipsbetween *sequences* andother *sequences* ratherthanjustfixedinputs. Thissequence-to-sequencelearningseemstobeonthecuspofrevolutionizing anotherapplication:machinetranslation (Sutskever *etal.* , 2014Bahdanau; *etal.* , 2015).

Thistrendofincreasingcomplexityhasbeenpushedtoitslogicalconclusion withtheintroductionofneuralTuringmachines (Graves *etal.* , 2014) thatlearn toreadfrommemorycellsandwritearbitrarycontenttomemorycells.Such neuralnetworkscanlearnsimpleprogramsfromexamplesofdesiredbehavior.For example,theycanlearntosortlistsofnumbersgivenexamplesofscrambledand sortedsequences.Thisself-programmingtechnologyisinitsinfancy,butinthe futureitcouldinprinciplebeappliedtonearlyanytask.

Anothercrowningachievementofdeeplearningisitsextensiontothedomainof **reinforcementlearning** .Inthecontextofreinforcementlearning,anautonomous agentmustlearntoperformataskbytrialanderror,withoutanyguidancefrom thehumanoperator.DeepMinddemonstratedthatareinforcementlearningsystem basedondeeplearningiscapableoflearningtoplayAtarivideogames,reaching human-levelperformanceonmanytasks (Mnih *etal.* , 2015).Deeplearninghas alsosignificantlyimprovedtheperformanceofreinforcementlearningforrobotics (Finn *etal.* , 2015).

Manyoftheseapplicationsofdeeplearningarehighlyprofitable.Deeplearning isnowusedbymanytoptechnologycompanies, includingGoogle,Microsoft, Facebook,IBM,Baidu,Apple,Adobe,Netflix,NVIDIA,andNEC.

Advancesindeeplearninghavealsodependedheavilyonadvancesinsoftware infrastructure.SoftwarelibrariessuchasTheano (Bergstra *etal.* , 2010; Bastien *etal.* , 2012),PyLearn2 (Goodfellow *etal.* , 2013c),Torch (Collobert *etal.* , 2011b), DistBelief (Dean *etal.* , 2012),Caffe (Jia, 2013),MXNet (Chen *etal.* , 2015),and TensorFlow (Abadi *etal.* , 2015) haveallsupportedimportantresearchprojectsor commercialproducts.

Deeplearninghasalsomadecontributionstootherssciences.Modernconvolutionalnetworksforobjectrecognitionprovideamodelofvisualprocessingthat neuroscientistscanstudy (DiCarlo, 2013).Deeplearningalsoprovidesusefultools forprocessingmassiveamountsofdataandmakingusefulpredictionsinscientific

因此，它们所能解决的任务也很复杂。(）Goodfellow 等人，2014 年发现，神经网络可以学习输出从图像转录的整个字符序列，而不仅仅是识别单个对象。此前，人们普遍认为这种学习需要标记序列中的各个元素（ ， ）。循环神经网络，例如上面提到的 LSTM 序列模型，现在用于模拟序列之间的关系，而不仅仅是固定的输入。这种序列到序列的学习似乎正处于另一个应用革命的风口浪尖：机器翻译（Sutskever，2014；Bahdanauetal.,；等人，2015)。

随着神经图灵机（Graves 2014 等）的引入，这种复杂性不断增加的趋势被推向了逻辑结论，神经图灵机可以学习从记忆单元读取数据并向记忆单元写入任意内容。这种神经网络可以从所需行为的示例中学习简单的程序。例如，它们可以学习根据给出的混乱和排序序列的示例对数字列表进行排序。这种自编程技术尚处于起步阶段，但在未来，它原则上可以应用于任何任务。

深度学习的另一项伟大成就是其向强化学习领域的扩展。在强化学习的背景下，自主智能体必须通过反复试验来学习特定格式的任务，而无需人类操作员的任何指导。DeepMind 证明，基于深度学习的强化学习系统能够学习玩 Atari 电子游戏，在许多任务上达到人类水平的表现（ ，)。深度学习还显著提高了机器人强化学习的性能（ ，)。Finn 等人，2015

深度学习的许多应用都利润丰厚。目前，许多顶尖科技公司都在使用深度学习，包括谷歌、微软、Facebook、IBM、百度、苹果、Adobe、Netflix、NVIDIA 和 NEC。

深度学习的进步也在很大程度上依赖于软件基础设施的进步。诸如 Theano（2011，2012；Bergstra 等人，2010 Bastien 等人，2011)、PyLearn2（2012 Goodfellow 等人，2012)、Torch（2013、2014)、DistBelief（2013、2014)、Caffe（2013、2014)、MXNet（2013、2014）和 TensorFlow（2013、2014）等软件库都支持重要的研究项目或商业产品。

深度学习也对其他科学做出了贡献。用于物体识别的现代卷积网络为神经科学家提供了研究视觉处理的理想模型（ ，)。深度学习还提供了有用的工具 DiCarlo 2013，用于处理大量数据并在科学研究中做出有用的预测

fields.Ithasbeensuccessfullyusedtopredicthowmoleculeswillinteractinorder tohelppharmaceuticalcompaniesdesignnewdrugs (Dahl *etal.* , 2014),tosearch forsubatomicparticles (Baldi *etal.* , 2014),andtoautomaticallyparsemicroscope imagesusedtoconstructa3-Dmapofthehumanbrain (Knowles-Barley *etal.* , 2014).Weexpectdeeplearningtoappearinmoreandmorescientificfieldsinthe future.

Insummary,deeplearningisanapproachtomachinelearningthathasdrawn heavilyonourknowledgeofthehumanbrain,statisticsandappliedmathasit developedoverthepastseveraldecades.Inrecentyears,deeplearninghasseen tremendousgrowthinitspopularityandusefulness,largelyastheresultofmore powerfulcomputers,largerdatasetsandtechniquestotraindeepernetworks.The yearsaheadarefullofchallengesandopportunitiestoimprovedeeplearningeven furtherandtobringittonewfrontiers.

它已被成功用于预测分子如何相互作用，以帮助制药公司设计新药（ ， ），搜索 Dahl et al. 2014 中的亚原子粒子（ ， ），以及自动解析用于构建人脑 3-D 地图的显微镜 Baldi et al. 2014 图像（ ，Knowles-Barley et al. 2014）。我们期待深度学习在未来出现在越来越多的科学领域。

总而言之，深度学习是一种机器学习方法，它大量借鉴了我们对人脑、统计学和应用数学的知识，并在过去几十年中得到了发展。近年来，深度学习的普及度和实用性得到了巨大增长，这主要是由于更强大的计算机、更大的数据集和训练更深层网络的技术。未来几年充满了挑战和机遇，可以进一步改进深度学习并将其带入新的领域。