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

Manyoftheearlysuccessesof AItookplaceinrelativelysterileandformal 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.Ittherefore askedwhether Fredwasstillapersonwhilehewasshaving.

Thedifficultiesfacedbysystemsrelyingonhard-codedknowledgesuggest thatAIsystemsneedtheabilitytoacquiretheirownknowledge,byextracting 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.IflogisticregressionweregivenanMRIscanofthepatient,ratherthan thedoctor'sformalizedreport,itwouldnotbeabletomakeusefulpredictions. IndividualpixelsinanMRIscanhavenegligiblecorrelationwithanycomplications thatmightoccurduringdelivery.

Thisdependenceonrepresentationsisageneralphenomenonthatappears throughoutcomputerscienceandevendailylife.Incomputerscience,operations suchassearchingacollectionofdatacanproceedexponentiallyfasterifthecollectionisstructuredandindexedintelligently.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

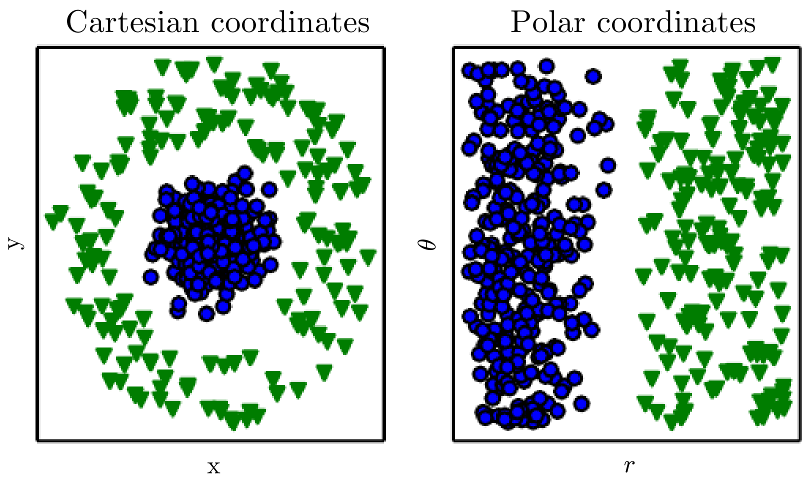

Figure1.1:Example ofdifferentrepresentations:supposewewanttoseparate two categoriesofdatabydrawingalinebetweentheminascatterplot.Intheplotontheleft, werepresentsomedatausingCartesiancoordinates,andthetaskisimpossible.Intheplot ontheright,werepresentthedatawithpolarcoordinatesandthetaskbecomessimpleto solvewithaverticalline.(FigureproducedincollaborationwithDavidWarde-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 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 solvesthiscentralprobleminrepresentationlearningbyintro-ducingrepresentationsthatareexpressedintermsofother,simplerrepresentations. Deeplearningenablesthecomputertobuildcomplexconceptsoutofsimplercon-cepts.Figure 1.2 showshowadeeplearningsystemcanrepresenttheconceptof animageofapersonbycombiningsimplerconcepts,suchascornersandcontours, whichareinturndefinedintermsofedges.

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

Theideaoflearningtherightrepresentationforthedataprovidesoneper-spectiveondeeplearning.Anotherperspectiveondeeplearningisthatdepth enablesthecomputertolearnamultistepcomputerprogram.Eachlayerofthe representationcanbethoughtofasthestateofthecomputer’smemoryafter

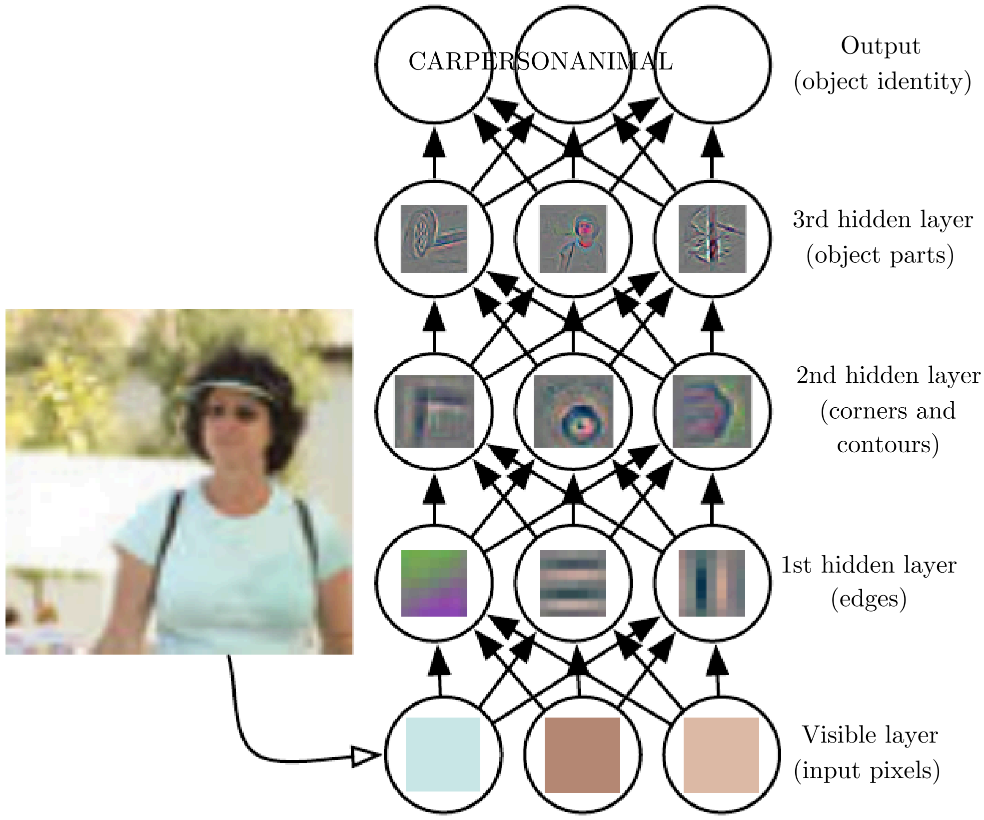

Figure1.2:Illustrationofadeeplearningmodel.Itisdifficultforacomputertounderstand themeaningofrawsensoryinputdata,suchasthisimagerepresentedasacollection ofpixelvalues.Thefunctionmappingfromasetofpixelstoanobjectidentityisvery complicated.Learningorevaluatingthismappingseemsinsurmountableiftackleddirectly.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

executinganothersetofinstructionsinparallel.Networkswithgreaterdepthcan executemoreinstructionsinsequence.Sequentialinstructionsoffergreatpower becauselaterinstructionscanreferbacktotheresultsofearlierinstructions.According tothisviewofdeeplearning,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.

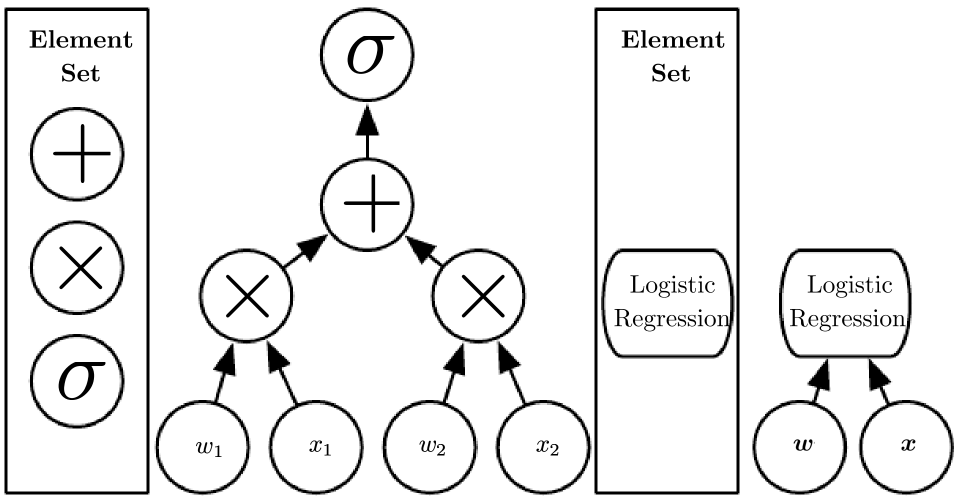

Figure1.3:Illustrationofcomputationalgraphsmappinganinputtoanoutputwhere eachnodeperformsanoperation.Depthisthelengthofthelongestpathfrominputto outputbutdependsonthedefinitionofwhatconstitutesapossiblecomputationalstep. Thecomputationdepictedinthesegraphsistheoutputofalogisticregressionmodel, $\sigma(\boldsymbol{w}^T\boldsymbol{x})$,where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

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

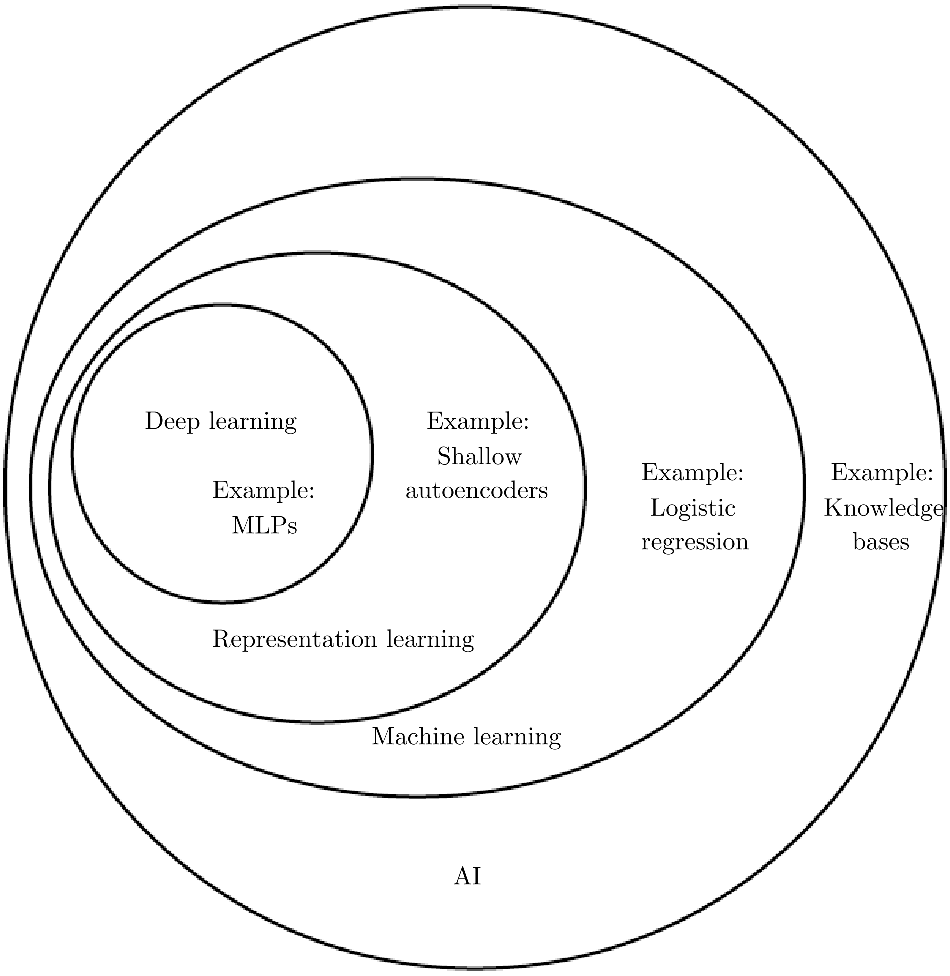

Figure1.4:AVenndiagramshowinghowdeeplearningisakindofrepresentationlearning, whichisinturnakindofmachinelearning,whichisusedformanybutnotallapproaches toAI.EachsectionoftheVenndiagramincludesanexampleofanAItechnology.

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.

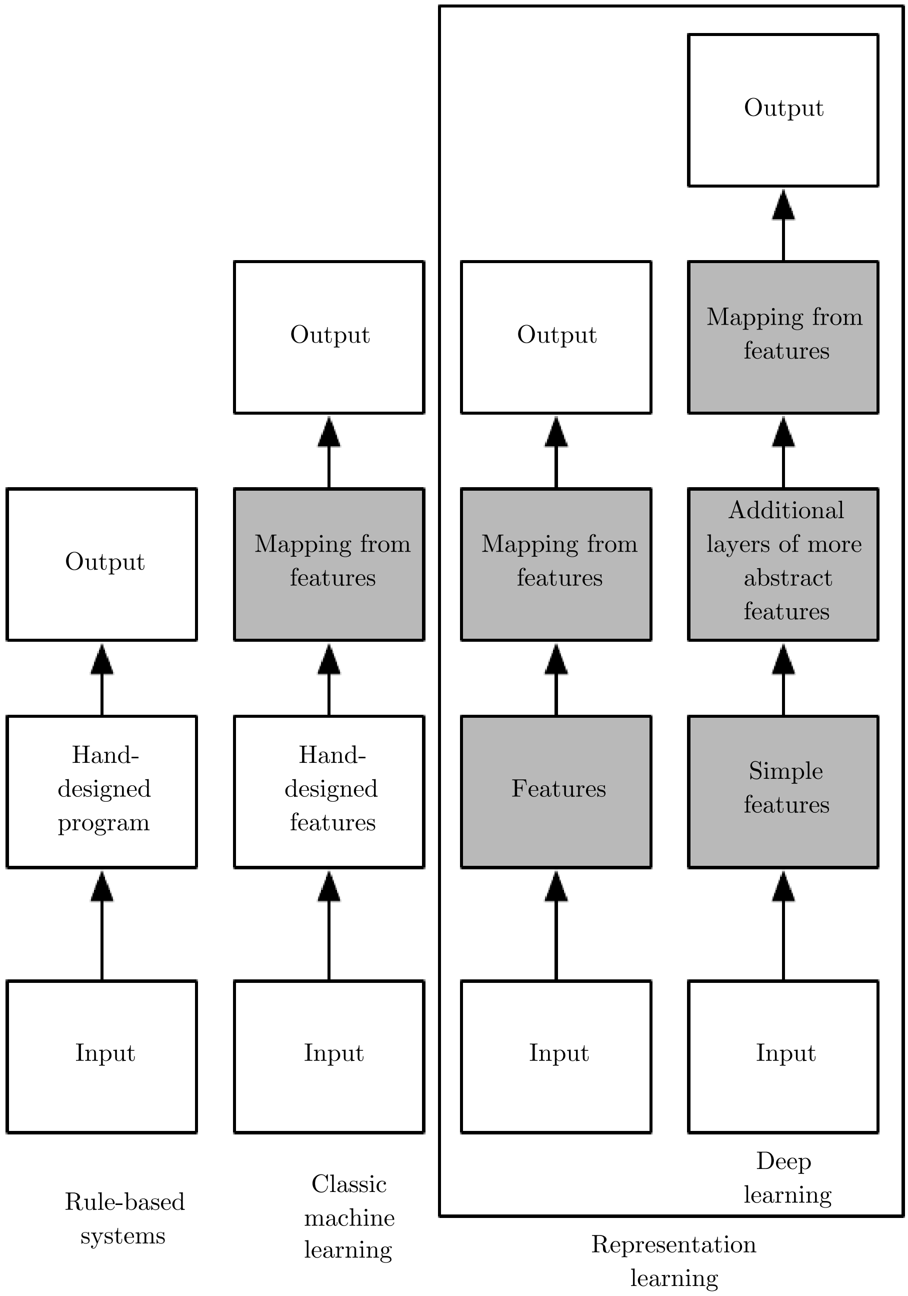

Figure1.5: FlowchartsshowinghowthedifferentpartsofanAIsystemrelatetoeach otherwithindifferentAIdisciplines.Shadedboxesindicatecomponentsthatareableto learnfromdata.

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

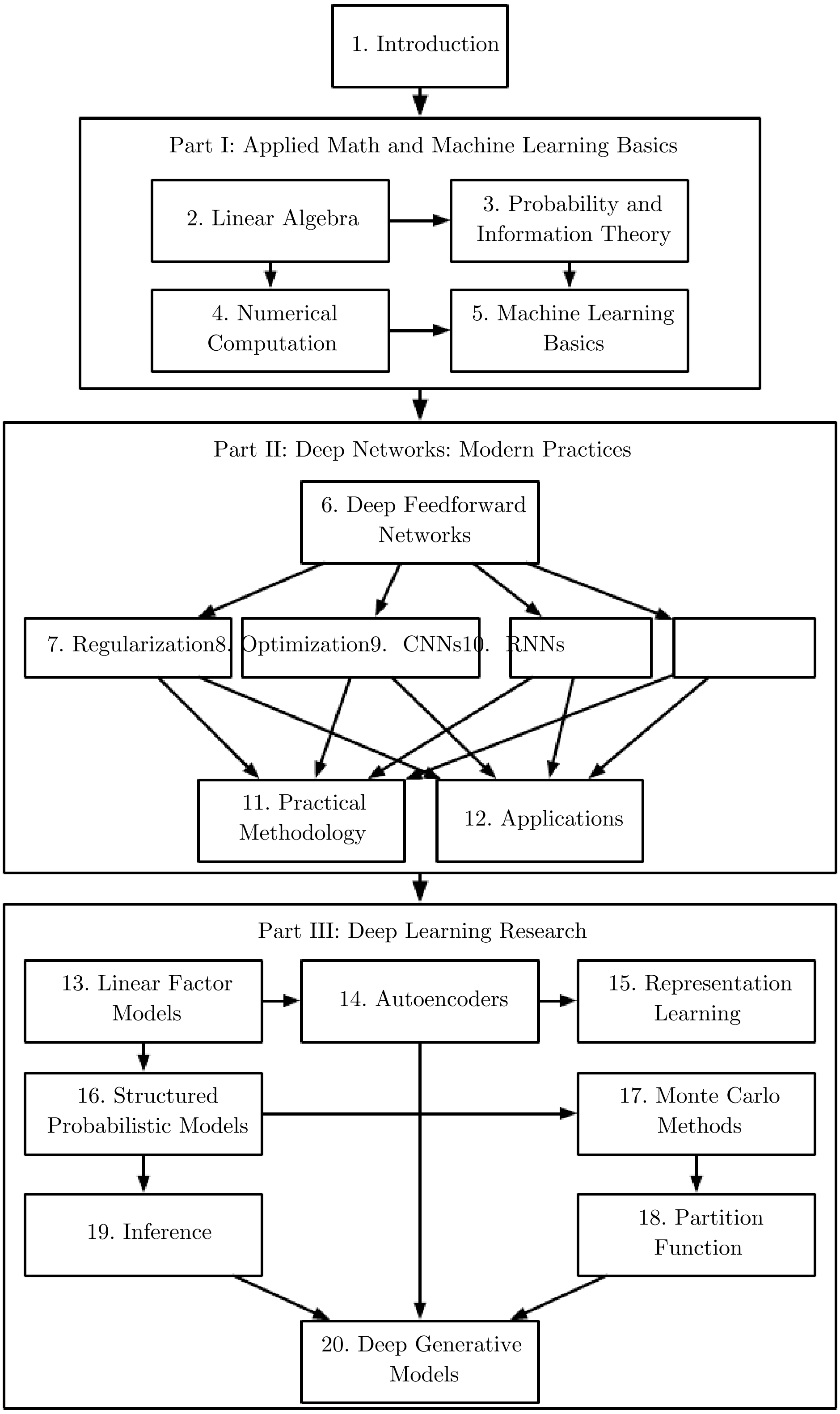

Figure1.6:Thehigh-levelorganizationofthebook.Anarrowfromonechaptertoanother indicatesthattheformerchapterisprerequisitematerialforunderstandingthelatter.

chapters to read, figure 1.6 provides a flowchart showing the high-level organization of the book.

We do assume that all readers come from a computer science background. We assume familiarity with programming, a basic understanding of computational performance issues, complexity theory, introductory level calculus and some of the terminology of graph theory.

1.2 Historical Trends in Deep Learning

It is easiest to understand deep learning with some historical context. Rather than providing a detailed history of deep learning, we identify a few key trends:

- Deep learning has had a long and rich history, but has gone by many names, reflecting different philosophical viewpoints, and has waxed and waned in popularity.
- Deep learning has become more useful as the amount of available training data has increased.
- Deep learning models have grown in size over time as computer infrastructure (both hardware and software) for deep learning has improved.
- Deep learning has solved increasingly complicated applications with increasing accuracy over time.

1.2.1 The Many Names and Changing Fortunes of Neural Networks

We expect that many readers of this book have heard of deep learning as an exciting new technology, and are surprised to see a mention of “history” in a book about an emerging field. In fact, deep learning dates back to the 1940s. Deep learning only *appears* to be new, because it was relatively unpopular for several years preceding its current popularity, and because it has gone through many different names, only recently being called “deep learning.” The field has been rebranded many times, reflecting the influence of different researchers and different perspectives.

A comprehensive history of deep learning is beyond the scope of this textbook. Some basic context, however, is useful for understanding deep learning. Broadly speaking, there have been three waves of development: deep learning known as **cybernetics** in the 1940s–1960s, deep learning known as **connectionism** in the

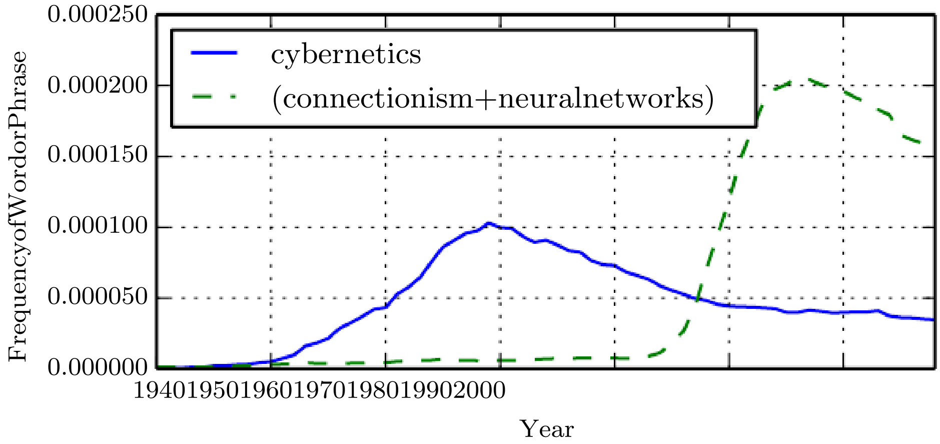

Figure1.7:Twoofthethreehistoricalwavesofartificialneuralnetsresearch,asmeasured bythefrequencyofthephrases“cybernetics”and“connectionism”or“neuralnetworks,” accordingtoGoogleBooks(thethirdwaveistoorecenttoappear).Thefirstwave startedwithcyberneticsinthe1940s–1960s,withthedevelopmentoftheoriesofbiological 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 perspectiveisthatitwouldbedeeplyinterestingtounderstandthebrainandthe principlesthatunderliehumanintelligence,somachinelearningmodelsthatshed lightonthesebasicscientificquestionsareusefulapartfromtheirabilitytosolve engineeringapplications.

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

Theearliestpredecessorsofmoderndeeplearningweresimplelinearmodels motivatedfromaneuroscientificperspective.Thesemodelsweredesignedto takeasetof n inputvalues $x_1,\ldots,x_n$ andassociatethemwithanoutput y. Thesemodelswouldlearnasetofweights $w_1,\ldots,w_n$ 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.

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.

Media accounts often emphasize the similarity of deep learning to the brain. While it is true that deep learning researchers are more likely to cite the brain as an influence than researchers working in other machine learning fields, such as kernel machines or Bayesian statistics, one should not view deep learning as an attempt to simulate the brain. Modern deep learning draws inspiration from many fields, especially applied math fundamentals like linear algebra, probability, information theory, and numerical optimization. While some deep learning researchers cite neuroscience as an important source of inspiration, others are not concerned with neuroscience at all.

It is worth noting that the effort to understand how the brain works on an algorithmic level is alive and well. This endeavor is primarily known as "computational neuroscience" and is a separate field of study from deep learning. It is common for researchers to move back and forth between both fields. The field of deep learning is primarily concerned with how to build computer systems that are able to successfully solve tasks requiring intelligence, while the field of computational neuroscience is primarily concerned with building more accurate models of how the brain actually works.

In the 1980s, the second wave of neural network research emerged in great part via a movement called **connectionism**, or **parallel distributed processing** (Rumelhart *et al.*, 1986c; McClelland *et al.*, 1995). Connectionism arose in the context of cognitive science. Cognitive science is an interdisciplinary approach to understanding the mind, combining multiple different levels of analysis. During the early 1980s, most cognitive scientists studied models of symbolic reasoning. Despite their popularity, symbolic models were difficult to explain in terms of how the brain could actually implement them using neurons. The connectionists began to study models of cognition that could actually be grounded in neural implementations (Touretzky and Minton, 1985), reviving many ideas dating back to the work of psychologist Donald Hebb in the 1940s (Hebb, 1949).

The central idea in connectionism is that a large number of simple computational units can achieve intelligent behavior when networked together. This insight applies equally to neurons in biological nervous systems as it does to hidden units in computational models.

Several key concepts arose during the connectionism movement of the 1980s that remain central to today's deep learning.

One of these concepts is that of **distributed representation** (Hinton *et al.*, 1986). This is the idea that each input to a system should be represented by many features, and each feature should be involved in the representation of many possible inputs. For example, suppose we have a vision system that can recognize 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.

Anothermajoraccomplishmentoftheconnectionistmovementwasthesuc-cessfuluseofback-propagationtotraindeepneuralnetworkswithinternalrepre-sentationsandthepopularizationoftheback-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.Ven-turesbasedonneuralnetworksandotherAItechnologiesbegantomakeunrealisti-callyambitiousclaimswhileseekinginvestments.WhenAIresearchdidnotfulfill theseunreasonableexpectations,investorsweredisappointed.Simultaneously, otherfieldsofmachinelearningmadeadvances.Kernelmachines (Boser *etal.* , 1992CortesandVapnik; , 1995Schölkopf; *etal.* , 1999) andgraphicalmodels (Jor-dan, 1998) bothachievedgoodresultsonmanyimportanttasks.Thesetwofactors ledtoadeclineinthepopularityofneuralnetworksthatlasteduntil2007.

Duringthistime,neuralnetworkscontinuedtoobtainimpressiveperformance onsometasks (LeCun *etal.* , 1998b; Bengio *etal.* , 2001).TheCanadianInstitute forAdvancedResearch(CIFAR)helpedtokeepneuralnetworksresearchalive viaitsNeuralComputationandAdaptivePerception(NCAP)researchinitiative. ThisprogramunitedmachinelearningresearchgroupsledbyGeoffreyHintonat UniversityofToronto,YoshuaBengioatUniversityofMontreal,andYannLeCun atNewYorkUniversity.ThemultidisciplinaryCIFARNCAPresearchinitiative 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 undergonechangesthatsimplifythetrainingofverydeeparchitectures.Themost importantnewdevelopmentisthattodaywecanprovidethesealgorithmswith theresourcestheyneedtosucceed.Figure 1.8 showshowthesizeofbenchmark datasetshasexpandedremarkablyovertime.Thistrendisdrivenbytheincreasing digitizationofsociety.Asmoreandmoreofouractivitiestakeplaceoncomputers, moreandmoreofwhatwedoisrecorded.Asourcomputersareincreasingly networkedtogether,itbecomeseasiertocentralizetheserecordsandcuratethem intoadatasetappropriateformachinelearningapplications.Theageof“BigData”

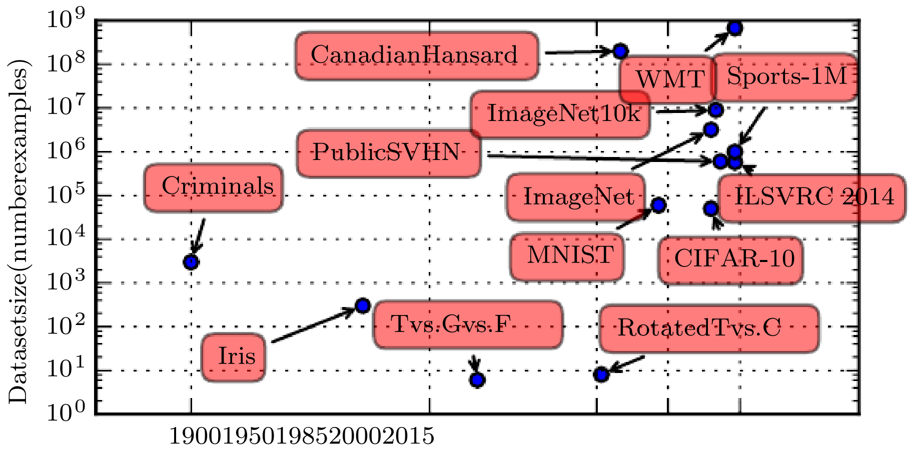

Figure1.8:Increasingdatasetsizeovertime.Intheearly1900s,statisticiansstudied datasetsusinghundredsorthousandsofmanuallycompiledmeasurements (Garson, 1900; Gosset, 1908Anderson; , 1935Fisher; , 1936).Inthe1950sthroughthe1980s,thepioneers ofbiologicallyinspiredmachinelearningoftenworkedwithsmallsyntheticdatasets,such 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.

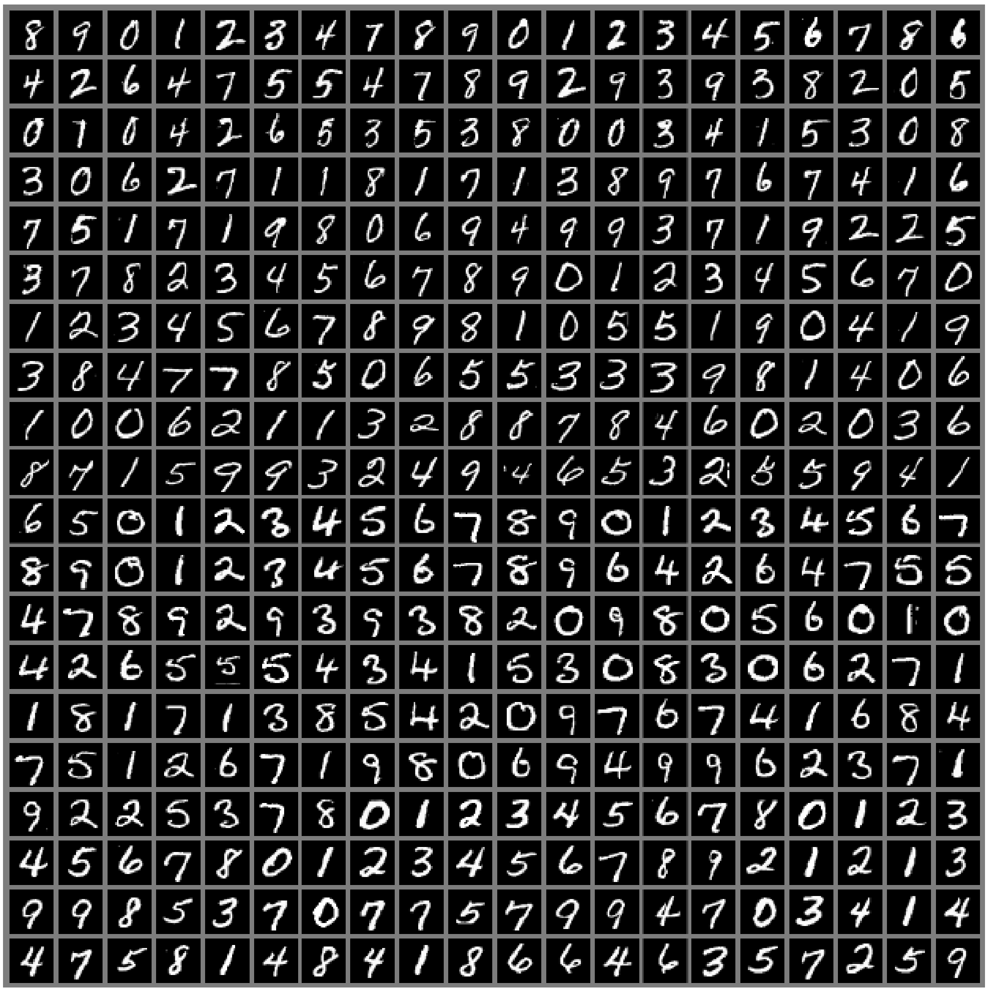

Figure1.9:ExampleinputsfromtheMNISTdataset.The“NIST”standsforNational InstituteofStandardsandTechnology,theagencythatoriginallycollectedthisdata. The“M”standsfor“modified,”sincethedatahasbeenpreprocessedforeasierusewith machinelearningalgorithms.TheMNISTdatasetconsistsofscansofhandwrittendigits andassociatedlabelsdescribingwhichdigit0–9iscontainedineachimage.Thi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

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

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

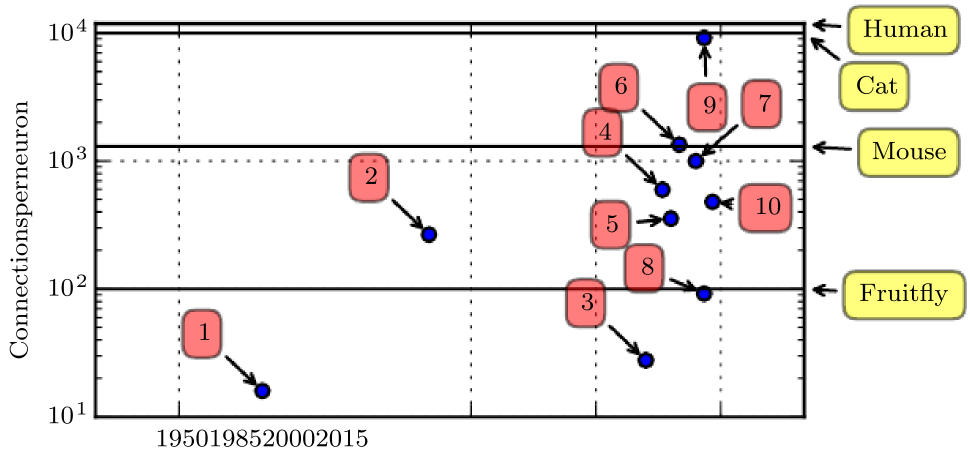

Figure1.10:Numberofconnectionsperneuronovertime.Initially,thenumberofconnectionsbetweenneuronsinartificialneuralnetworkswaslimitedbyhardwarecapabilities. Today,thenumberofconnectionsbetweenneuronsismostlyadesignconsideration.Some artificialneuralnetworkshavenearlyasmanyconnectionsperneuronasacat,andit 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

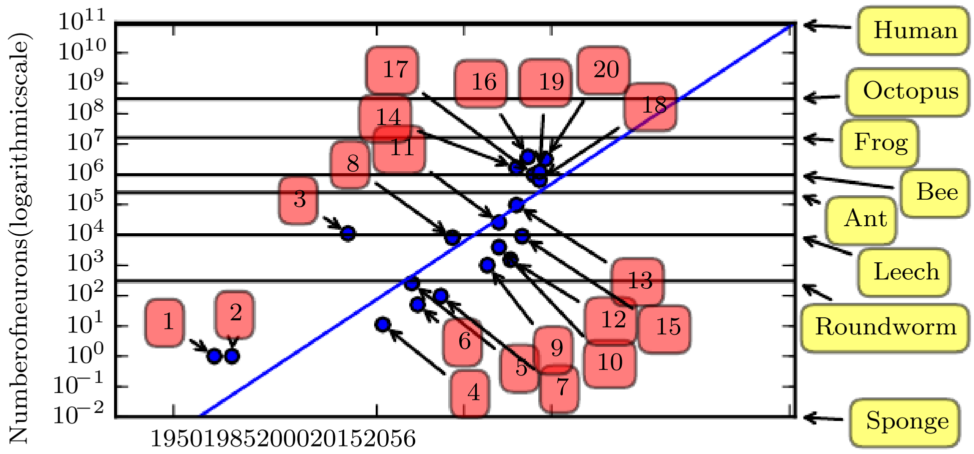

Figure1.11:Increasingneuralnetworksizeovertime.Sincetheintroductionofhidden units,artificialneuralnetworkshavedoubledinsizeroughlyevery2.4years.Biological neuralnetworksizesfrom Wikipedia (2015).

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

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,

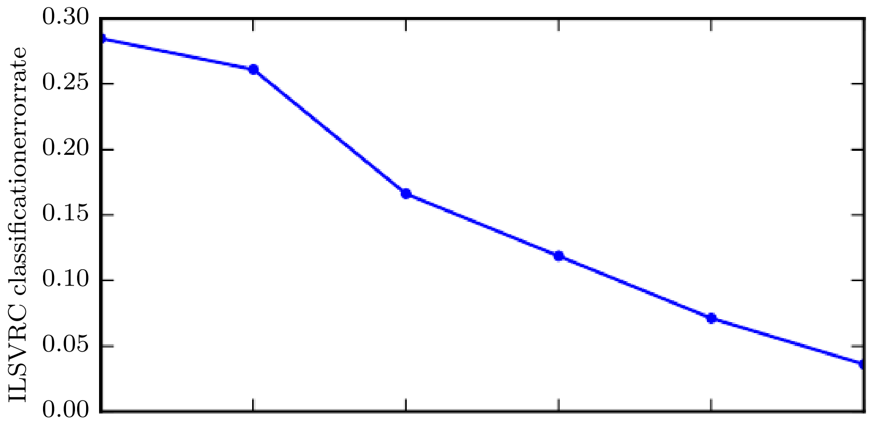

Figure1.12:Decreasingerrorrateovertime.Sincedeepnetworksreachedthescale necessarytocompeteintheImageNetLargeScaleVisualRecognitionChallenge,they haveconsistentlywonthecompetitioneveryyear,yieldinglowerandlowererrorrates eachtime.Datafrom Russakovsky *etal.* (2014b) and He *etal.* (2015).

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.

Deeplearninghasalsomadecontributionstoothersciences.Modernconvolu- tionalnetworksforobjectrecognitionprovideamodelofvisualprocessingthat neuroscientistscanstudy (DiCarlo, 2013).Deeplearningalsoprovidesusefultools forprocessingmassiveamountsofdataandmakingusefulpredictionsinscientific 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.