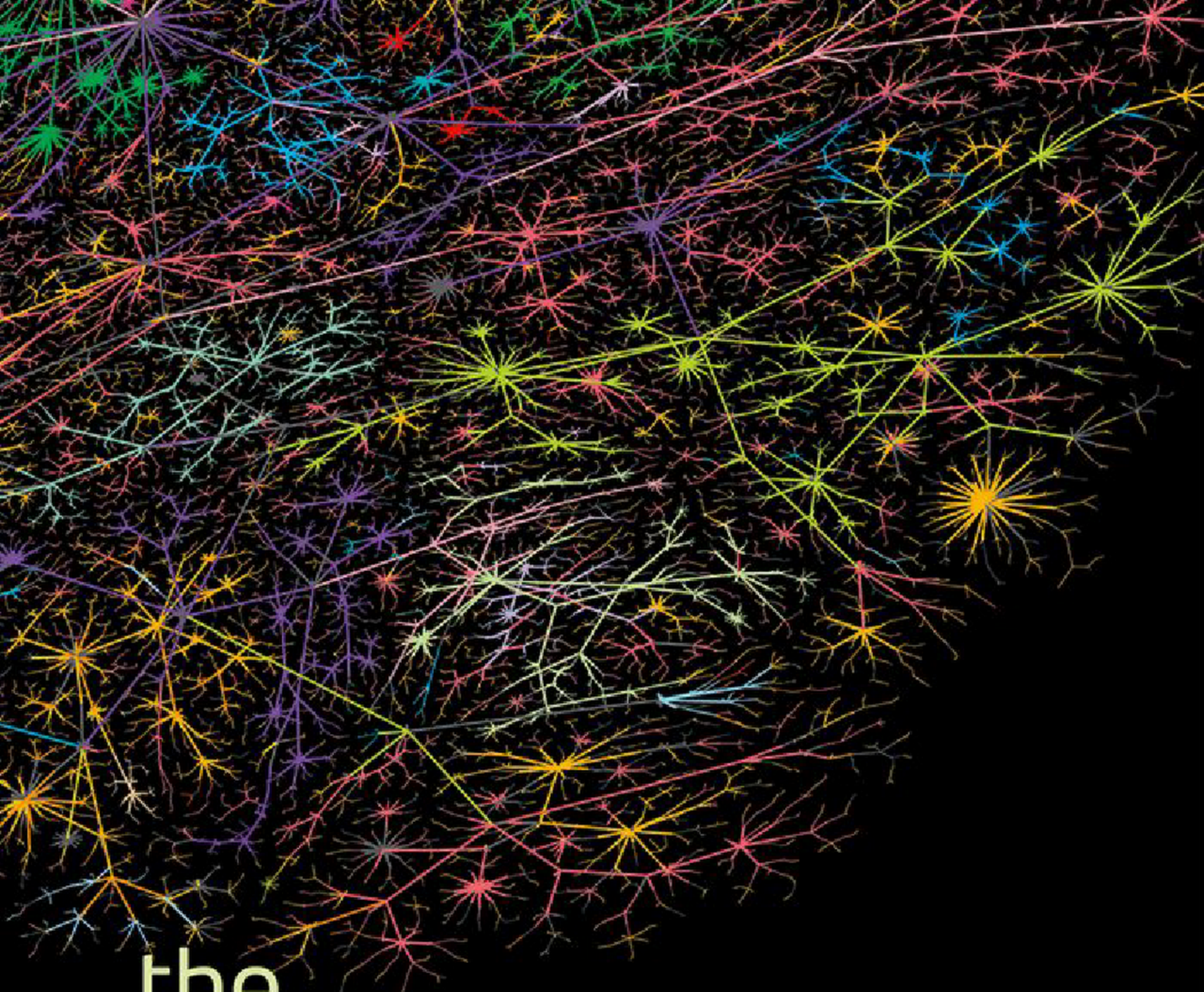

the Intelligent Web

Search, smart algorithms, and big data

GAUTAM SHROFF

the Intelligent Web

Search, Smart Algorithms, and Big Data

GAUTAM SHROFF

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2013
Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2013938816

ISBN 978–0–19–964671–5

Printed in Italy by
L.E.G.O. S.p.A.-Lavis TN

CONTENTS

LIST OF FIGURES

This page intentionally left blank

Prologue

POTENTIAL

I grew up reading and being deeply influenced by the popular science books of George Gamow on physics and mathematics. This book is my attempt at explaining a few important and exciting advances in computer science and artificial intelligence (AI) in a manner accessible to all. The incredible growth of the internet in recent years, along with the vast volumes of 'big data' it holds, has also resulted in a rather significant confluence of ideas from diverse fields of computing and AI. This new 'science of *web intelligence*', arising from the marriage of many AI techniques applied together on 'big data', is the stage on which I hope to entertain and elucidate, in the spirit of Gamow, and to the best of my abilities.

* * *

The computer science community around the world recently celebrated the centenary of the birth of the British scientist Alan Turing, widely regarded as the father of computer science. During his rather brief life Turing made fundamental contributions in mathematics as well as some in biology, alongside crucial practical feats such as breaking secret German codes during the Second World War.

Turing was the first to examine very closely the meaning of what it means to 'compute', and thereby lay the foundations of computer science. Additionally, he was also the first to ask whether the capacity of intelligent thought could, in principle, be achieved by a machine that 'computed'. Thus, he is also regarded as the father of the field of enquiry now known as 'artificial intelligence'.

In fact, Turing begins his classic 1950 article[1] with, 'I propose to consider the question, "Can machines think?" ' He then goes on to describe the famous 'Turing Test', which he referred to as the 'imitation game', as a way to think about the problem of machines thinking. According to the Turing Test, if a computer can converse with any of us humans in so convincing a manner as to fool us into believing that it, too, is a human, then we should consider that machine to be 'intelligent' and able to 'think'.

Recently, in February 2011, IBM's Watson computer managed to beat champion human players in the popular TV show *Jeopardy!*. Watson was able to answer fairly complex queries such as 'Which New Yorker who fought at the Battle of Gettysburg was once considered the inventor of baseball?'. Figuring out that the answer is actually Abner Doubleday, and not Alexander Cartwright who actually wrote the rules of the game, certainly requires non-trivial natural language processing as well as probabilistic reasoning; Watson got it right, as well as many similar fairly difficult questions.

During this widely viewed *Jeopardy!* contest, Watson's place on stage was occupied by a computer panel while the human participants were visible in flesh and blood. However, imagine if instead the human participants were also hidden behind similar panels, and communicated via the same mechanized voice as Watson. Would we be able to tell them apart from the machine? Has the Turing Test then been 'passed', at least in this particular case?

There are more recent examples of apparently 'successful' displays of artificial intelligence: in 2007 Takeo Kanade, the well-known Japanese expert in computer vision, spoke about his early research in face recognition, another task normally associated with humans and at best a few higher-animals: 'it was with pride that I tested the program on 1000 faces, a rare case at the time when testing with 10 images was considered a "large-scale experiment".'[2] Today, both Facebook and Google's Picasa regularly recognize faces from among the hundreds of

millions contained amongst the billions of images uploaded by users around the world.

Language is another arena where similar progress is visible for all to see and experience. In 1965 a committee of the US National Academy of Sciences concluded its review of the progress in automated translation between human natural languages with, 'there is no immediate or predicable prospect of useful machine translation'.[2] Today, web users around the world use Google's translation technology on a daily basis; even if the results are far from perfect, they are certainly good enough to be very useful.

Progress in spoken language, i.e., the ability to recognize speech, is also not far behind: Apple's Siri feature on the iPhone 4S brings usable and fairly powerful speech recognition to millions of cellphone users worldwide.

As succinctly put by one of the stalwarts of AI, Patrick Winston: 'AI is becoming more important while it becomes more inconspicuous', as 'AI technologies are becoming an integral part of mainstream computing'.[3]

* * *

What, if anything, has changed in the past decade that might have contributed to such significant progress in many traditionally 'hard' problems of artificial intelligence, be they machine translation, face recognition, natural language understanding, or speech recognition, all of which have been the focus of researchers for decades?

As I would like to convince you during the remainder of this book, many of the recent successes in each of these arenas have come through the deployment of many known but disparate techniques working together, and most importantly their deployment at *scale*, on large volumes of 'big data'; all of which has been made possible, and indeed driven, by the internet and the world wide web. In other words, rather than 'traditional' artificial intelligence, the successes we are witnessing are better described as those of '*web intelligence*'

arising from 'big data'. Let us first consider what makes big data so 'big', i.e., its *scale*.

* * *

The web is believed to have well over a trillion web pages, of which at least 50 billion have been catalogued and *indexed* by search engines such as Google, making them searchable by all of us. This massive web content spans well over 100 million domains (i.e., locations where we point our browsers, such as <http://www.wikipedia.org>). These are themselves growing at a rate of more than 20,000 net domain additions daily. Facebook and Twitter each have over 900 million users, who between them generate over 300 million posts a day (roughly 250 million tweets and over 60 million Facebook updates). Added to this are the over 10,000 credit-card payments made per *second*,* the well-over 30 billion point-of-sale transactions per year (via dial-up POS devices†), and finally the over 6 billion mobile phones, of which almost 1 billion are smartphones, many of which are GPS-enabled, and which access the internet for e-commerce, tweets, and post updates on Facebook.‡ Finally, and last but not least, there are the images and videos on YouTube and other sites, which by themselves outstrip all these put together in terms of the sheer volume of data they represent.

This deluge of data, along with emerging techniques and technologies used to handle it, is commonly referred to today as 'big data'. Such big data is both valuable and challenging, because of its sheer volume. So much so that the volume of data being created in the current five years from 2010 to 2015 will far exceed all the data generated in human history (which was estimated to be under 300 exabytes as of 2007§). The web, where all this data is being produced and resides, consists of millions of servers, with data storage soon to be measured in zetabytes.¶

* <http://www.creditcards.com>.
† <http://www.gaoresearch.com/POS/pos.php>.
‡ <http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats>.
§ <http://www.bbc.co.uk/news/technology-12419672>.
¶ petabyte = 1,000 GB, exabyte = 1,000 petabytes, and a zetabyte = 1,000 petabytes.

On the other hand, let us consider the volume of data an average human being is exposed to in a lifetime. Our sense of vision provides the most voluminous input, perhaps the equivalent of half a million hours of video or so, assuming a fairly a long lifespan. In sharp contrast, YouTube alone witnesses 15 million hours of *fresh* video uploaded every year.

Clearly, the volume of data available to the millions of machines that power the web far exceeds that available to any human. Further, as we shall argue later on, the millions of servers that power the web at least match if not exceed the raw computing capacity of the 100 billion or so neurons in a single human brain. Moreover, each of these servers are certainly much much faster at computing than neurons, which by comparison are really quite slow.

Lastly, the advancement of computing technology remains relentless: the well-known Moore's Law documents the fact that computing power per dollar appears to double every 18 months; the lesser known but equally important Kryder's Law states that storage capacity per dollar is growing even faster. So, for the first time in history, we have available to us both the computing power as well as the raw data that matches and shall very soon far exceed that available to the average human.

Thus, we have the *potential* to address Turing's question 'Can machines think?', at least from the perspective of raw computational power and data of the same order as that available to the human brain. How far have we come, why, and where are we headed? One of the contributing factors might be that, only recently after many years, does 'artificial intelligence' appear to be regaining a semblance of its initial ambition and unity.

* * *

In the early days of artificial intelligence research following Turing's seminal article, the diverse capabilities that might be construed to comprise intelligent behaviour, such as vision, language, or logical

reasoning, were often discussed, debated, and shared at common forums. The goals exposed by the now famous Dartmouth conference of 1956, considered to be a landmark event in the history of AI, exemplified both a unified approach to all problems related to machine intelligence as well as a marked overconfidence:

> We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.[4]

These were clearly heady times, and such gatherings continued for some years. Soon the realization began to dawn that the 'problem of AI' had been grossly underestimated. Many sub-fields began to develop, both in reaction to the growing number of researchers trying their hand at these difficult challenges, and because of conflicting goals. The original aim of actually answering the question posed by Turing was soon found to be too challenging a task to tackle all at once, or, for that matter, attempt at all. The proponents of 'strong AI', i.e., those who felt that true 'thinking machines' were actually possible, with their pursuit being a worthy goal, began to dwindle. Instead, the practical applications of AI techniques, first developed as possible answers to the strong-AI puzzle, began to lead the discourse, and it was this 'weak AI' that eventually came to dominate the field.

Simultaneously, the field split into many sub-fields: image processing, computer vision, natural language processing, speech recognition, machine learning, data mining, computational reasoning, planning, etc. Each became a large area of research in its own right. And rightly so, as the practical applications of specific techniques necessarily appeared to lie within disparate

areas: recognizing faces versus translating between two languages; answering questions in natural language versus recognizing spoken words; discovering knowledge from volumes of documents versus logical reasoning; and the list goes on. Each of these were so clearly separate application domains that it made eminent sense to study them separately and solve such obviously different practical problems in purpose-specific ways.

Over the years the AI research community became increasingly fragmented. Along the way, as Pat Winston recalled, one would hear comments such as 'what are all these vision people doing here'[3] at a conference dedicated to say, 'reasoning'. No one would say, 'well, because we think with our eyes',[3] i.e., our perceptual systems are intimately involved in thought. And so fewer and fewer opportunities came along to discuss and debate the 'big picture'.

* * *

Then the web began to change everything. Suddenly, the practical problem faced by the web companies became larger and more holistic: initially there were the search engines such as Google, and later came the social-networking platforms such as Facebook. The problem, however, remained the same: how to make more money from advertising?

The answer turned out to be surprisingly similar to the Turing Test: Instead of merely fooling us into believing it was human, the 'machine', i.e., the millions of servers powering the web, needed to *learn* about each of us, individually, just as we all learn about each other in casual conversation. Why? Just so that better, i.e., more closely targeted, advertisements could be shown to us, thereby leading to better 'bang for the buck' of every advertising dollar. This then became the holy grail: not intelligence per se, just doing better and better at this 'reverse' Turing Test, where instead of us being observer and 'judge', it is the machines in the web that observe and seek to 'understand' us better for their own selfish needs, if only to 'judge' whether or not we are likely

buyers of some of the goods they are paid to advertise. As we shall see soon, even these more pedestrian goals required weak-AI techniques that could mimic many of *capabilities* required for intelligent thought.

Of course, it is also important to realize that none of these efforts made any strong-AI claims. The manner in which seemingly intelligent capabilities are computationally realized in the web does not, for the most part, even attempt to mirror the mechanisms nature has evolved to bring intelligence to life in real brains. Even so, the results are quite surprising indeed, as we shall see throughout the remainder of this book.

At the same time, this new holy grail could not be grasped with disparate weak-AI techniques operating in isolation: our queries as we searched the web or conversed with our friends were *words*; our actions as we surfed and navigated the web were *clicks*. Naturally we wanted to *speak* to our phones rather than type, and the videos that we uploaded and shared so freely were, well, videos.

Harnessing the vast trails of data that we leave behind during our web existences was essential, which required expertise from different fields of AI, be they language processing, learning, reasoning, or vision, to come together and connect the dots so as to even come close to understanding *us*.

First and foremost the web gave us a different way to *look* for information, i.e., web search. At the same time, the web itself would *listen* in, and *learn*, not only about us, but also from our collective knowledge that we have so well digitized and made available to all. As our actions are observed, the web-intelligence programs charged with pinpointing advertisements for us would need to *connect* all the dots and *predict* exactly which ones we should be most interested in.

Strangely, but perhaps not surprisingly, the very synthesis of techniques that the web-intelligence programs needed in order to connect the dots in their practical enterprise of online advertising appears, in many respects, similar to how we ourselves integrate our different

perceptual and cognitive abilities. We consciously *look* around us to gather information about our environment as well as *listen* to the ambient sea of information continuously bombarding us all. Miraculously, we *learn* from our experiences, and *reason* in order to *connect* the dots and make sense of the world. All this so as to *predict* what is most likely to happen next, be it in the next instant, or eventually in the course of our lives. Finally, we *correct* our actions so as to better achieve our goals.

* * *

I hope to show how the cumulative use of artificial intelligence techniques at web scale, on hundreds of thousands or even millions of computers, can result in behaviour that exhibits a very basic feature of human intelligence, i.e., to colloquially speaking 'put two and two together' or 'connect the dots'. It is this ability that allows us to make sense of the world around us, make intelligent guesses about what is most likely to happen in the future, and plan our own actions accordingly.

Applying web-scale computing power on the vast volume of 'big data' now available because of the internet, offers the *potential* to create far more intelligent systems than ever before: this defines the new science of *web intelligence*, and forms the subject of this book.

At the same time, this remains primarily a book about weak AI: however powerful this web-based synthesis of multiple AI techniques might appear to be, we do not tread too deeply in the philosophical waters of strong-AI, i.e., whether or not machines can ever be 'truly intelligent', whether consciousness, thought, self, or even 'soul' have reductionist roots, or not. We shall neither speculate much on these matters nor attempt to describe the diverse philosophical debates and arguments on this subject. For those interested in a comprehensive history of the confluence of philosophy, psychology, neurology, and artificial intelligence often referred to as 'cognitive science', Margaret

Boden's recent volume *Mind as Machine: A History of Cognitive Science*[5] is an excellent reference.

Equally important are Turing's own views as elaborately explained in his seminal paper[1] describing the 'Turing test'. Even as he clearly makes his own philosophical position clear, he prefaces his own beliefs and arguments for them by first clarifying that 'the original question, "Can machines think?" I believe to be too meaningless to deserve discussion'.[1] He then rephrases his 'imitation game', i.e., the Turing Test that we are all familiar with, by a *statistical* variant: 'in about fifty years' time it will be possible to program computers . . . so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning'.[1] Most modern-day machine-learning researchers might find this formulation quite familiar indeed. Turing goes on to speculate that 'at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted'.[1] It is the premise of this book that such a time has perhaps arrived.

As to the 'machines' for whom it might be colloquially acceptable to use the word 'thinking', we look to the web-based engines developed for entirely commercial pecuniary purposes, be they search, advertising, or social networking. We explore how the computer programs underlying these engines sift through and make sense of the vast volumes of 'big data' that we continuously produce during our online lives—our collective 'data exhaust', so to speak.

In this book we shall quite often use Google as an example and examine its innards in greater detail than others. However, when we speak of Google we are also using it as a metaphor: other search engines, such as Yahoo! and Bing, or even the social networking world of Facebook and Twitter, all share many of the same processes and purposes.

The purpose of all these web-intelligence programs is simple: 'all the better to understand us', paraphrasing Red Riding Hood's wolf in grandmother's clothing. Nevertheless, as we delve deeper into what these vast syntheses of weak-AI techniques manage to achieve in practice, we do find ourselves wondering whether these web-intelligence systems might end up serving us a dinner far closer to strong AI than we have ever imagined for decades.

That hope is, at least, one of the reasons for this book.

* * *

In the chapters that follow we dissect the ability to connect the dots, be it in the context of web-intelligence programs trying to understand us, or our own ability to understand and make sense of the world. In doing so we shall find some surprising parallels, even though the two contexts and purposes are so very different. It is these connections that offer the potential for increasingly capable web-intelligence systems in the future, as well as possibly deeper understanding and appreciation of our own remarkable abilities.

Connecting the dots requires us to *look* at and experience the world around us; similarly, a web-intelligence program looks at the data stored in or streaming across the internet. In each case information needs to be stored, as well as retrieved, be it in the form of memories and their recollection in the former, or our daily experience of web search in the latter.

Next comes the ability to *listen*, to focus on the important and discard the irrelevant. To recognize the familiar, discern between alternatives or identify similar things. Listening is also about 'sensing' a momentary experience, be it a personal feeling, individual decision, or the collective sentiment expressed by the online masses. Listening is followed eventually by deeper understanding: the ability to *learn* about the structure of the world, in terms of facts, rules, and relationships. Just as we learn common-sense knowledge about the world around us, web-intelligence systems learn about our preferences and

behaviour. In each case the essential underlying processes appear quite similar: detecting the regularities and patterns that emerge from large volumes of data, whether derived from our personal experiences while growing up, or via the vast data trails left by our collective online activities.

Having learned something about the structure of the world, real or its online rendition, we are able to *connect* different facts and derive new conclusions giving rise to reasoning, logic, and the ability to deal with uncertainty. Reasoning is what we normally regard as unique to our species, distinguishing us from animals. Similar reasoning by machines, achieved through smart engineering as well as by crunching vast volumes of data, gives rise to surprising engineering successes such as Watson's victory at *Jeopardy!*.

Putting everything together leads to the ability to make *predictions* about the future, albeit tempered with different degrees of belief. Just as we predict and speculate on the course of our lives, both immediate and long-term, machines are able to predict as well—be it the supply and demand for products, or the possibility of crime in particular neighbourhoods. Of course, predictions are then put to good use for *correcting* and controlling our own actions, for supporting our own decisions in marketing or law enforcement, as well as controlling complex, autonomous web-intelligence systems such as self-driving cars.

In the process of describing each of the elements: *looking*, *listening*, *learning*, *connecting*, *predicting*, and *correcting*, I hope to lead you through the computer science of semantic search, natural language understanding, text mining, machine learning, reasoning and the semantic web, AI planning, and even swarm computing, among others. In each case we shall go through the principles involved virtually from scratch, and in the process cover rather vast tracts of computer science even if at a very basic level.

Along the way, we shall also take a closer look at many examples of web intelligence at work: AI-driven online advertising for sure, as well

as many other applications such as tracking terrorists, detecting disease outbreaks, and self-driving cars. The promise of self-driving cars, as illustrated in Chapter 6, points to a future where the web will not only provide us with information and serve as a communication platform, but where the computers that power the web could also help us *control* our world through complex web-intelligence systems; another example of which promises to be the energy-efficient 'smart grid'.

* * *

By the end of our journey we shall begin to suspect that what began with the simple goal of optimizing advertising might soon evolve to serve other purposes, such as safe driving or clean energy. Therefore the book concludes with a note on *purpose*, speculating on the nature and evolution of large-scale web-intelligence systems in the future. By asking where goals come from, we are led to a conclusion that surprisingly runs contrary to the strong-AI thesis: instead of ever mimicking human intelligence, I shall argue that web-intelligence systems are more likely to evolve synergistically with our own evolving collective social intelligence, driven in turn by our use of the web itself.

In summary, this book is at one level an elucidation of artificial intelligence and related areas of computing, targeted for the lay but patient and diligent reader. At the same time, there remains a constant and not so hidden agenda: we shall mostly concern ourselves with exploring how today's web-intelligence applications are able to mimic some aspects of intelligent behaviour. Additionally however, we shall also compare and contrast these immense engineering feats to the wondrous complexities that the human brain is able to grasp with such surprising ease, enabling each of us to so effortlessly 'connect the dots' and make sense of the world every single day.

This page intentionally left blank

1

LOOK

In 'A Scandal in Bohemia'[6] the legendary fictional detective Sherlock Holmes deduces that his companion Watson had got very wet lately, as well as that he had 'a most clumsy and careless servant girl'. When Watson, in amazement, asks how Holmes knows this, Holmes answers:

> 'It is simplicity itself . . . My eyes tell me that on the inside of your left shoe, just where the firelight strikes it, the leather is scored by six almost parallel cuts. Obviously they have been caused by someone who has very carelessly scraped round the edges of the sole in order to remove crusted mud from it. Hence, you see, my double deduction that you had been out in vile weather, and that you had a particularly malignant boot-slitting specimen of the London slavery.'

Most of us do not share the inductive prowess of the legendary detective. Nevertheless, we all continuously *look* at the the world around us and, in our small way, draw inferences so as to make sense of what is going on. Even the simplest of observations, such as whether Watson's shoe is in fact dirty, requires us to first look at his shoe. Our skill and intent drive what we look at, and look *for*. Those of us that may share some of Holmes's skill look for far greater detail than the rest of us. Further, more information is better: 'Data! Data! Data! I can't make bricks without clay', says Holmes in another episode.[7] No inference is

possible in the absence of input data, and, more importantly, the *right* data for the task at hand.

How does Holmes connect the observation of 'leather . . . scored by six almost parallel cuts' to the cause of 'someone . . . very carelessly scraped round the edges of the sole in order to remove crusted mud from it'? Perhaps, somewhere deep in the Holmesian brain lies a memory of a similar boot having been so damaged by another 'specimen of the London slavery'? Or, more likely, many different 'facts', such as the potential causes of damage to boots, including clumsy scraping; that scraping is often prompted by boots having been dirtied by mud; that cleaning boots is usually the job of a servant; as well as the knowledge that bad weather results in mud.

In later chapters we shall delve deeper into the process by which such 'logical inferences' might be automatically conducted by machines, as well as how such knowledge might be learned from experience. For now we focus on the fact that, in order to make his logical inferences, Holmes not only needs to look *at* data from the world without, but also needs to look *up* 'facts' learned from his past experiences. Each of us perform a myriad of such 'lookups' in our everyday lives, enabling us to recognize our friends, recall a name, or discern a car from a horse. Further, as some researchers have argued, our ability to converse, and the very foundations of all human language, are but an extension of the ability to correctly look up and classify past experiences from *memory*. 'Looking at' the world around us, relegating our experiences to memory, so as to later 'look them up' so effortlessly, are most certainly essential and fundamental elements of our ability to connect the dots and make sense of our surroundings.

The MEMEX Reloaded

Way back in 1945 Vannevar Bush, then the director of the US Office of Scientific Research and Development (OSRD), suggested that scientific

effort should be directed towards emulating and augmenting human memory. He imagined the possibility of creating a 'MEMEX': a device

> which is a sort of mechanised private file and library . . . in which an individual stores all his books, records, and communications, and which is mechanised so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.[8]

A remarkably prescient thought indeed, considering the world wide web of today. In fact, Bush imagined that the MEMEX would be modelled on human memory, which

> operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain. It has other characteristics, of course; trails that are not frequently followed are prone to fade, items are not fully permanent, memory is transitory. Yet the speed of action, the intricacy of trails, the detail of mental pictures, is awe-inspiring beyond all else in nature.[8]

At the same time, Bush was equally aware that the wonders of human memory were far from easy to mimic: 'One cannot hope thus to equal the speed and flexibility with which the mind follows an associative trail, but it should be possible to beat the mind decisively in regard to the permanence and clarity of the items resurrected from storage.'[8]

Today's world wide web certainly does 'beat the mind' in at least these latter respects. As already recounted in the Prologue, the volume of information stored in the internet is vast indeed, leading to the coining of the phrase 'big data' to describe it. The seemingly intelligent 'web-intelligence' applications that form the subject of this book all exploit this big data, just as our own thought processes, including Holmes's inductive prowess, are reliant on the 'speed and flexibility' of human memory.

How is this big data stored in the web, so as to be so easily accessible to all of us as we surf the web every day? To what extent does it resemble, as well as differ from, how our own memories are stored

and recalled? And last but not least, what does it portend as far as augmenting our own abilities, much as Vannevar Bush imagined over 50 years ago? These are the questions we now focus on as we examine what it means to remember and recall, i.e., to 'look up things', on the web, or in our minds.

* * *

When was the last time you were to meet someone you had never met before in person, even though the two of you may have corresponded earlier on email? How often have you been surprised that the person you saw looked different than what you had expected, perhaps older, younger, or built differently? This experience is becoming rarer by the day. Today you can Google persons you are about to meet and usually find half a dozen photos of them, in addition to much more, such as their Facebook page, publications or speaking appearances, and snippets of their employment history. In a certain sense, it appears that we can simply 'look up' the global, collective memory-bank of mankind, as collated and managed by Google, much as we internally look up our own personal memories as associated with a person's name.

Very recently Google introduced Google Glass, looking through which you merely need to look at a popular landmark, such as the Eiffel Tower in Paris, and instantly retrieve information about it, just as if you had typed in the query 'Eiffel Tower' in the Google search box. You can do this with books, restaurant frontages, and even paintings. In the latter case, you may not even know the name of the painting; still Glass will 'look it up', using the image itself to drive its search. We know for a fact that Google (and others, such as Facebook) are able to perform the same kind of 'image-based' lookup on human faces as well as images of inanimate objects. They too can 'recognize' people from their faces. Clearly, there is a scary side to such a capability being available in such tools: for example, it could be easily misused by stalkers, identity thieves, or extortionists. Google has deliberately not yet released a face recognition feature in Glass, and maintains that

'we will not add facial recognition to Glass unless we have strong privacy protections in place'.[9] Nevertheless, the ability to recognize faces is now within the power of technology, and we can experience it every day: for example, Facebook automatically matches similar faces in your photo album and attempts to name the people using whatever information it finds in its own copious memory-bank, while also tapping Google's when needed. The fact is that technology has now progressed to the point where we can, in principle, 'look up' the global collective memory of mankind, to recognize a face or a name, much as we recognize faces and names every day from our own personal memories.

* * *

Google handles over 4 billion search queries a day. How did I get that number? By issuing a few searches myself, of course; by the time you read this book the number would have gone up, and you can look it up yourself. Everybody who has access to the internet uses search, from office workers to college students to the youngest of children. If you have ever introduced a computer novice (albeit a rare commodity these days) to the internet, you might have witnessed the 'aha' experience: it appears that every piece of information known to mankind is at one's fingertips. It is truly difficult to remember the world before search, and realize that this was the world of merely a decade ago.

Ubiquitous search is, some believe, more than merely a useful tool. It may be changing the way we connect the dots and make sense of our world in fundamental ways. Most of us use Google search several times a day; after all, the entire collective memory-bank of mankind is just a click away. Thus, sometimes we no longer even bother to remember facts, such as when Napoleon was defeated at Waterloo, or when the East India Company established its reign in the Indian subcontinent. Even if we do remember our history lessons, our brains often compartmentalize the two events differently as both of them pertain to different geographies; so ask us which preceded the other, and we are

usually stumped. Google comes to the rescue immediately, though, and we quickly learn that India was well under foreign rule when Napoleon met his nemesis in 1815, since the East India Company had been in charge since the Battle of Plassey in 1757. Connecting disparate facts so as to, in this instance, put them in chronological sequence, needs extra details that our brains do not automatically connect across compartments, such as European vs Indian history; however, within any one such context we are usually able to arrange events in historical sequence much more easily. In such cases the ubiquity of Google search provides instant satisfaction and serves to augment our cognitive abilities, even as it also reduces our need to memorize facts.

Recently some studies, as recounted in Nicholas Carr's *The Shallows: What the internet is Doing to Our Brains,*[10] have argued that the internet is 'changing the way we think' and, in particular, diminishing our capacity to read deeply and absorb content. The instant availability of hyperlinks on the web seduces us into 'a form of skimming activity, hopping from one source to another and rarely returning to any source we might have already visited'.[11] Consequently, it is argued, our motivation as well as ability to stay focused and absorb the thoughts of an author are gradually getting curtailed.

Be that as it may, I also suspect that there is perhaps another complementary capability that is probably being enhanced rather than diminished. We are, of course, talking about the ability to connect the dots and make sense of our world. Think about our individual memories: each of these is, as compared to the actual event, rather sparse in detail, at least at first glance. We usually remember only certain aspects of each experience. Nevertheless, when we need to connect the dots, such as recall where and when we might have met a stranger in the past, we seemingly need only 'skim through' our memories without delving into each in detail, so as to correlate some of them and use these to make deeper inferences. In much the same manner, searching and surfing the web while trying to connect the dots is probably a

boon rather than a bane, at least for the purpose of correlating disparate pieces of information. The MEMEX imagined by Vannevar Bush is now with us, in the form of web search. Perhaps, more often than not, we regularly discover previously unknown connections between people, ideas, and events every time we indulge in the same 'skimming activity' of surfing that Carr argues is harmful in some ways. We have, in many ways, already created Vannevar Bush's MEMEX-powered world where

> the lawyer has at his touch the associated opinions and decisions of his whole experience, and of the experience of friends and authorities. The patent attorney has on call the millions of issued patents, with familiar trails to every point of his client's interest. The physician, puzzled by its patient's reactions, strikes the trail established in studying an earlier similar case, and runs rapidly through analogous case histories, with side references to the classics for the pertinent anatomy and histology. The chemist, struggling with the synthesis of an organic compound, has all the chemical literature before him in his laboratory, with trails following the analogies of compounds, and side trails to their physical and chemical behaviour. The historian, with a vast chronological account of a people, parallels it with a skip trail which stops only at the salient items, and can follow at any time contemporary trails which lead him all over civilisation at a particular epoch. There is a new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record. The inheritance from the master becomes, not only his additions to the world's record, but for his disciples the entire scaffolding by which they were erected.[8]

In many ways therefore, web search is in fact able to augment our own powers of recall in highly synergistic ways. Yes, along the way we do forget many things we earlier used to remember. But perhaps the things we forget are in fact irrelevant, given that we now have access to search? Taking this further, our brains are poor at indexing, so we search the web instead. Less often are we called upon to traverse our memory-to-memory links just to recall facts. We use those links only when making connections or correlations that augment mere search, such as while inferring patterns, making predictions, or hypothesizing conjectures, and we shall return to all these elements later in the

book. So, even if by repeatedly choosing to use search engines over our own powers of recall, it is indeed the case that certain connections in our brains are in fact getting weaker, as submitted by Nicholas Carr.[11] At the same time, it might also be the case that many other connections, such as those used for deeper reasoning, may be getting strengthened.

Apart from being a tremendously useful tool, web search also appears to be important in a very fundamental sense. As related by Carr, the Google founder Larry Page is said to have remarked that 'The ultimate search engine is something as smart as people, or smarter . . . working on search is a way to work on artificial intelligence.'[11] In a 2004 interview with *Newsweek*, his co-founder Sergey Brin remarks, 'Certainly if you had all the world's information directly attached to your brain, or an artificial brain that was smarter than your brain, you would be better off.'

In particular, as I have already argued above, our ability to connect the dots may be significantly enhanced using web search. Even more interestingly, what happens when search and the collective memories of mankind are automatically tapped by computers, such as the millions that power Google? Could these computers themselves acquire the ability to 'connect the dots', like us, but at a far grander scale and infinitely faster? We shall return to this thought later and, indeed, throughout this book as we explore how today's machines are able to 'learn' millions of facts from even larger volumes of big data, as well as how such facts are already being used for automated 'reasoning'. For the moment, however, let us turn our attention to the computer science of web search, from the inside.

Inside a Search Engine

'Any sufficiently advanced technology is indistinguishable from magic'; this often-quoted 'law' penned by Arthur C. Clarke also applies

to internet search. Powering the innocent 'Google search box' lies a vast network of over a million servers. By contrast, the largest banks in the world have at most 50,000 servers each, and often less. It is interesting to reflect on the fact that it is within the computers of these banks that your money, and for that matter most of the world's wealth, lies encoded as bits of ones and zeros. The magical Google-like search is made possible by a computing behemoth two orders of magnitude more powerful than the largest of banks. So, how does it all work?

Searching for data is probably the most fundamental exercise in computer science; the first data processing machines did exactly this, i.e., store data that could be searched and retrieved in the future. The basic idea is fairly simple: think about how you might want to search for a word, say the name 'Brin', in this very book. Naturally you would turn to the index pages towards the end of the book. The index entries are sorted in alphabetical order, so you know that 'Brin' should appear near the beginning of the index. In particular, searching the index for the word 'Brin' is clearly much easier than trawling through the entire book to figure out where the word 'Brin' appears. This simple observation forms the basis of the computer science of 'indexing', using which all computers, including the millions powering Google, perform their magical searches.

Google's million servers continuously crawl and index over 50 billion web pages, which is the estimated size of the *indexed** world wide web as of January 2011. Just as in the index of this book, against each word or phrase in the massive web index is recorded the web address (or URL†) of *all* the web pages that contain that word or phrase. For common words, such as 'the', this would probably be the entire English-language web. Just try it; searching for 'the' in Google yields

* Only a small fraction of the web is indexed by search engines such as Google; as we see later, the complete web is actually far larger.

† 'Universal record locater', or URL for short, is the technical term for a web address, such as <http://www.google.com>.

over 25 billion results, as of this writing. Assuming that about half of the 50 billion web pages are in English, the 50 billion estimate for the size of the *indexed* web certainly appears reasonable.

Each web page is regularly scanned by Google's millions of servers, and added as an entry in a huge web index. This web index is truly massive as compared to the few index pages of this book. Just imagine how big this web index is: it contains every word ever mentioned in any of the billions of web pages, in any possible language. The English language itself contains just over a million words. Other languages are smaller, as well as less prevalent on the web, but not by much. Additionally there are proper nouns, naming everything from people, both real (such as 'Brin') or imaginary ('Sherlock Holmes'), to places, companies, rivers, mountains, oceans, as well as every name ever given to a product, film, or book. Clearly there are many millions of words in the web index. Going further, common phrases and names, such as 'White House' or 'Sergey Brin' are also included as separate entries, so as to improve search results. An early (1998) paper[12] by Brin and Page, the now famous founders of Google, on the inner workings of their search engine, reported using a dictionary of 14 million unique words. Since then Google has expanded to cover many languages, as well as index common phrases in addition to individual words. Further, as the size of the web has grown, so have the number of unique proper nouns it contains. What is important to remember, therefore, is that today's web index probably contains hundreds of millions of entries, each a word, phrase, or proper noun, using which it indexes many billions of web pages.

What is involved in searching for a word, say 'Brin', in an index as large as the massive web index? In computer science terms, we need to explicitly define the steps required to 'search a sorted index', regardless of whether it is a small index for a book or the index of the entire web. Once we have such a prescription, which computer scientists call an 'algorithm', we can program an adequately powerful computer to

search any index, even the web index. A very simple program might proceed by checking each word in the index one by one, starting from the beginning of the index and continuing to its end. Computers are fast, and it might seem that a reasonably powerful computer could perform such a procedure quickly enough. However, size is a funny thing; as soon as one starts adding a lot of zeros numbers can get very big very fast. Recall that unlike a book index, which may contain at most a few thousand words, the web index contains millions of words and hundreds of millions of phrases. So even a reasonably fast computer that might perform a million checks per second would still take many hours to search for just one word in this index. If our query had a few more words, we would need to let the program work for months before getting an answer.

Clearly this is not how web search works. If one thinks about it, neither is it how we ourselves search a book index. For starters, our very simple program completely ignores that fact that index words were already sorted in alphabetical order. Let's try to imagine how a smarter algorithm might search a sorted index faster than the naive one just described. We still have to assume that our computer itself is rather dumb, and, unlike us, it does *not* understand that since 'B' is the second letter in the alphabet, the entry for 'Brin' would lie roughly in the first tenth of all the index pages (there are 26 letters, so 'A' and 'B' together constitute just under a tenth of all letters). It is probably good to assume that our computer is ignorant about such things, because in case we need to search the web index, we have no idea how many unique letters the index entries begin with, or how they are ordered, since all languages are included, even words with Chinese and Indian characters.

Nevertheless, we do know that there is *some* ordering of letters that includes all languages, using which the index itself has been sorted. So, ignorant of anything but the size of the complete index, our smarter search program begins, not at the beginning, but at the very middle