Working with Big Data

Scaling Data Discovery

Abdallah Bari

Working with Big Data

Scaling Data Discovery

Abdallah Bari

Math Coding & Analytics

ASIN: B0786QHJL7

Math Coding and Analytics
Westmount, Quebec, Canada
ASIN: B0786QHJL7

PREFACE

This book is part of an international collaborative research journey that I have started, with colleagues, some years ago to address the challenges of ambiguity and uncertainty in large datasets of mostly unstructured (image) raw data. Fractal geometry in combination of neural networks were used then to capture subtle genetic variation among individual genotypes. The results attracted the attention of both public and private sector as well as the International Research Board Members and published in journals as well as books such as the 2006 book "Complexus Mundi: Emergent Patterns In Nature".

In recent years and as result of Big Data' s volume, variety and velocity, I run into scale challenges, whether for storing, processing or analysing data. Addressing scale challenges and combining Big Data with machine learning helped to discover new and rare traits (genes), some of these traits and genes have been sought-for in vain. The results have been also considered as breakthroughs in 2015 and they have attracted the attention of both private and public sector as well as international media and renowned publishers.

Big Data has definitely created radical shifts with new opportunities, leading organisations and companies to shift their activities towards more data-driven decisions with the help of machine learning (ML) techniques and artificial intelligence (AI). A shift from stand-alone traditional desktop computing to embrace a more comprehensive strategy such as Mesh App and Service Architecture (MASA) strategy banking on Big Data to allow a more dynamic connection of people, processes, things and services that are supporting today's increasingly intelligent (AI) digital ecosystems, according to Garner (2017).

Along with Big Data' shifts and trends, the processing power has also increased dramatically defying Moor's Law with Quantum computers already in the market. Today's mobile phone processor (CPU) can run at speed of more than 1 GHz allowing millions of calculations per second. The dramatic increase of speed of a mobile is tens of thousands of times more faster than the processing rate of instructions of the computer used to guide spacecraft to the moon 45 years ago.

The processing power and the availability of Big Data offer opportunities for scaling analytics and discovery. The book Working with Big Data to scale Big Data' s Discovery refers to the different techniques and tools to address the ambiguity and uncertainty as well as the scaling challenges with the arrival of Big Data, spanning data integration, data preparation and data analytics. The tools presented in this book lend themselves to scale as the technology and the needs evolve.

Abdallah Bari - Math Coding and Analytics
Westmount, Canada - December 2017

CONTENTS

Chapter 1

INTRODUCTION

Big Data, as its name implies, involves enormous amount of data. The name first appeared in academic publications in the 1990s and as of 2008 it has been widely used with the spread of cloud infrastructure[1].

Today, there is even a more "rush to compute unlike anything we've ever seen before" wrote Matt Day, Technology Reporter of the Seattle Times[2]. Big Data is increasingly sought-for to detect hidden patterns in data since the presence of patterns in data is an indication for possibility for prediction and for discovery. The patterns can be used to predict when for example a customer is ready to make a purchase, what sample is likely to yield better results or when an aircraft jet-engine needs servicing.

The predictions in healthcare are helping to discover new drugs and their effectiveness, prior to their release. The Food and Drug Administration announced July 7, 2017 that its scaling Big Data analytics using high-performance computing (HPC) to make drug

[1] Tom Boellstorff (2013) Making big data, in theory. *First Monday* 18(10)

[2] Amazon cloud unit's grand plan: data centers in every major country worldwide. By Matt Day , Seattle Times technology reporter. Seattle Times (2017)

development and testing more effective[3]. In business, according to McKinsey, retailers who leverage the full power of big data are able to improve their operational margins by as much as 60%[4].

The poll conducted on April 2017 by Big Data Zone[5] listed several real-world problems solved based on Big Data involving different sectors spanning from retail, healthcare, media, and telecommunications to finance, government, IT, and to fleet management. The April 2017 Big Data Zone survey involved executives from 22 companies who are working with Big Data or providing Big Data solutions to clients today.

Big Data - Opportunity for discovery

The patterns identified using Big Data based on *in-silico* evaluation (math-based virtual screening) are helping to discover new genes in crops. Some of the genes discovered, in recent years, have been searched for in vain in the past. In-silico evaluation as virtual screening approach not only helped to tap on Big Data but also shortened the time to discovery, which is crucial today for business as well as for research and development to keep pace of rapid global changes including climate change.

The 2016 OECD[6] Report considered Big Data as a driving force for knowledge acquisition and value creation, fostering research and innovation with a potential to transform most if not all sectors[7]. The report referred to Big Data as the new research and development (R&D) for 21st century innovation systems, highlighting Big Data and its analytics as fundamental inputs to innovation, akin to R&D. The report revealed, based on available evidence, that companies using data-driven innovation (DDI) have raised productivity faster, by approximately 5-10% when compared

[3] Avanade's survey (2017) - https://www.avanade.com/en-us/technologies/data-analytics

[4] Scott Gottlieb, M.D. (July 7, 2017) How FDA Plans to Help Consumers Capitalize on Advances in Science. FDA Voice

[5] Tom Smith (2017) Executive Insights on the State of Big Data. Big Data Zone

[6] The Organisation for Economic Cooperation and Development (2016)

[7] The Organisation for Economic Cooperation and Development (2016)

to non-users.

Big Data's global market is expected to grow with an average of 25% per year by 2020[8]. Industrial companies such as GE and Siemens are turning to Big Data and promoting their respective corporations as Big Data firms[9].

In banking industry, banks are also shifting towards applications oriented around customer experience using Big Data and ML moving from solely more defensive applications involving security and risk. The BMO Canada 's shift has saved over $100CAD million in data re-use and data warehouse rationalization according to a recent article by Tom Davenport & Randy Bean (2017). The Bank has also established a data science platform including analytics sandboxes and open source software for machine learning, as well as software for robotic process automation (RPA). On overall the bank has already achieved several times more value in additional revenues over what it has saved in data rationalization[10].

According to Garner, during 2017 the world's most successful companies use technology to scale and outcompete traditional organisations [11]. Big Data moved from the "peak of inflated expectations" in 2012 towards "Trough of Disillusionment" in 2014 and now entering the "plateau of productivity". As was anticipated in 2012, Big Data was then in the 2 to 5 years horizon to reach the plateau of productivity (Figure 1.1).

[8] Frost & Sullivan (2014).

[9] (Economist May 6th 2017)

[10] Tom Davenport & Randy Bean (2017) Setting The Table For Data Science And AI At Bank Of Montreal, Forbes.

[11] Gartner, Inc - 2017 Hype Cycles Highlight Enterprise and Ecosystem Digital Disruptions