DATA SCIENCE ECOSYSTEM IN PAKISTAN

A REPORT BY ATOMCAMP

atomcamp

JANUARY 2023

About Us

atomcamp is a continuous learning platform that is helping the youth and organizations unlock opportunities with Data Science.

atomcamp hosts various courses and programs centered on tech education to upskill the Pakistani workforce and to create the awareness that continuous education is critical to keep up with the fast-paced world.

Technology Bootcamps

At atomcamp, we offer a variety of multidisciplinary courses, but our main focus is Data Science, Artificial Intelligence and Cloud Computing, which are relatively new and emerging fields, especially in Pakistan. Our goal is to make careers in these fields accessible to everyone in Pakistan - regardless of the educational or professional background.

atomcamp's 6-month Data Science Bootcamp enables participants to learn relevant data skills and launch their careers. The program is meant for those who are aiming to switch into a data science career as well as those who want to incorporate data science training into their current jobs/careers to remain competitive.

Our 3-month AI bootcamp is designed to train you to launch your career in AI, NLP and Computer Vision, regardless of the educational background. This program is designed for everyone as the first two months of AI bootcamp focus on building a foundation in Python, Math, and Machine Learning.

About the Author

This report has been authored by Mahnoor Imran Sayyed, a research analyst at atomcamp and Hussain Shahbaz Khawaja, a freelance Data Scientist.

atomcamp

Our Mission

promote a culture of continuous learning

provide skill development for youth

encourage interdisciplinary learning

build learning communities

provide contextual & accessible knowledge

Table of Contents

Data Science is here

Years ago, data science was an obscure and unknown field of study but today it is one of the fastest growing spaces in the world. From its humble beginnings, it has grown to dominate the way that we humans understand the world around us and make decisions about our day to day lives.

Today, data science has application in each and every field and data is one of the most valuable resources held by organizations.

Data and data science have also made a huge impact in the way that organizations and governments regulate and navigate issues related to data quantity, quality, security and privacy. With billions of terabytes of data being generated by the human species, data science holds the key to changing the way we interact with technology forever.

Defining Data Science

Data Science is an interdisciplinary subject that combines mathematics, statistics, advanced analytics and programming to derive actionable insights from subject specific data. Insights derived from data science can be utilized further to inform strategic decisions in organizations.[1]

Data Science Lifecycle

The typical data science life cycle is divided into four steps:

Step 1

Data Ingestion: Data ingestion involves collecting raw data and converting it into a format that can be analyzed. This can include structured data from databases or unstructured data from social media and other sources. Data ingestion is crucial for accurate and effective analysis, as the quality and completeness of the data directly impact the insights that can be drawn from it.

Step 2

Data storage and data processing: It involve storing the ingested data in a data warehouse or data lake and transforming it into a more usable format for analysis. This step also includes data cleaning, normalization, and identifying relevant subsets of data for analysis. The goal is to make the data easily accessible and usable for analysis while ensuring its quality and integrity.

Step 3

Data analysis: Data analysis involves examining, cleaning, transforming, and modeling data to extract meaningful insights and conclusions. It uses statistical and machine learning techniques to uncover patterns and relationships within the data for informed decision-making.

Step 4

Storytelling: Insights-once uncovered-need to be presented in visual formats that are easy to understand and communicate the findings accurately. It involves translating complex data into meaningful stories using visualizations, dashboards, reports, or presentations. The goal is to inform decision-making and drive action. Effective data storytelling is critical in the data science life cycle to ensure that data insights translate into real-world impact.

The Beginner's Guide: Key Definitions related to Data Science

Artificial Intelligence

Artificial intelligence is the ability of a digital computer or robot to perform tasks commonly associated with intelligent beings[1º].

Big Data

Big data refers to large and complex datasets that are difficult to process using traditional data processing techniques. It includes both structured and unstructured data from various sources such as social media, sensors, and machines.

Machine Learning

It is a subset of artificial intelligence (AI) that involves developing algorithms that can learn from data and make predictions or decisions based on that data. It enables computers to automatically improve their performance on a specific task by learning from experience.

Data Mining

Data mining is the process of analyzing large datasets to extract useful information and insights. It involves using statistical techniques, machine learning algorithms, and data visualization tools to identify patterns, relationships, and anomalies in the data.

Data Visualization

It is the process of presenting data in a graphical or visual format to help people understand complex data and identify patterns and trends. It involves using charts, graphs, and other visual aids to convey information in an easy-to-understand way.

Data Governance

Data governance refers to the management, policies, procedures, and standards for the effective and secure use of data in an organization. It involves ensuring data quality, integrity, and security, as well as compliance with relevant laws and regulations.

What can Data Science actually do?

In healthcare....

- Google's tool, LYNA, can be used to identify breast cancer tumors that metastasize into nearby lymph nodes. In one trial, LYNA — short for Lymph Node Assistant — accurately identified metastatic cancer 99 percent of the time using its machine-learning algorithm.

In transport....

- StreetLight has utilized data science to model traffic patterns for cars, bikes and pedestrians on North American streets. Using trillions of data points, Streetlight's traffic maps stay updated in real time.The company's maps also inform various city planning enterprises, including commuter transit design

In sport....

- Used by NBA and college teams, RSPCT's shooting analysis system uses a sensor on a basketball hoop's rim, to the exact when and where the ball strikes on each basket attempt. This data is then sent to a device that shows shot details in real time and provides predictive analytics.

In e-commerce....

- Social media giant, Instagram uses data science to tailor its marketing for sponsored posts. By tracking user's age, location and preferences, data scientists use data points from Instagram and Meta to craft a curated experience for each user.

Data Science: Changing the Global Landscape

Data science is a growing practice in technology. McKinsey predicts that by 2025, all workers in all roles will leverage data to better inform their decisions. With automation taking over the more routine and mundane tasks, human capital is bound to be directed to more innovation and strategy related tasks[3]. Moreover, they also predict that availability of data analytics tools and the volume of data will both increase exponentially-bringing about a digital revolution.

Today data science is a verified game changer for organizations across the globe. COVID 19 in particular caused more and more users to spend time online and the e-commerce market grew as a result [6]. This means that organizations now have access to much more data than ever before. Big Data, which refers to large datasets that can be used to glean insights, is becoming a critical driver for business success in industries all over the globe. And as the importance of data and data analysis skyrockets, the role being played by data scientists in organizations is becoming key to ensuring success in the industry. In a survey of 400 companies, Bain & Co found that companies with competent and well integrated data analytics teams were twice as likely to be in the top quartile of financial performance within their industries, three times more likely to execute decisions as intended and five times more likely to make decisions faster [5].

The size of the big data analytics market worldwide was 240 billion USD in 2021 and it is projected to grow to 655 billion USD by 2029 [12]. One EY study found that 93% of the companies surveyed were looking to increase their investments in data and data analytics [13]. The US Bureau of Statistics estimated that data science will see more growth than any other field between now and 2029 and already the trends indicate this.

240 billion USD

market size of the big data

655 billion USD

projected market size of the big data

93%

companies looking to increase investment in data analytics

By 2019, the job postings for data science on a popular job listing platform, Indeed, had increased by 256%.[14]

It is not just that the potential of data science is being realized globally but it is also that new developments in the field are making it more and more powerful. One important trend for example is the amount of data being created. According to Forbes, every day we generate 2.5 quintillion bytes of data and 90% of the data that exists in the world was only created in the last two years [15].

It is not just in business that data science is making a huge impact but in global and local governance too, data and its analytics can help governments make informed decisions. The United Nations has performed an in-depth analysis to understand how Big Data can be harnessed to improve progress towards the Sustainable Development Goals. Poverty levels for example can be better tracked by using mobile phones and e-commerce to determine spending patterns and income [7].

Data Scientist: a Profile

Roles in Data Science

Data Architect:
An IT professional that reviews and analyzes data infrastructure & storage.

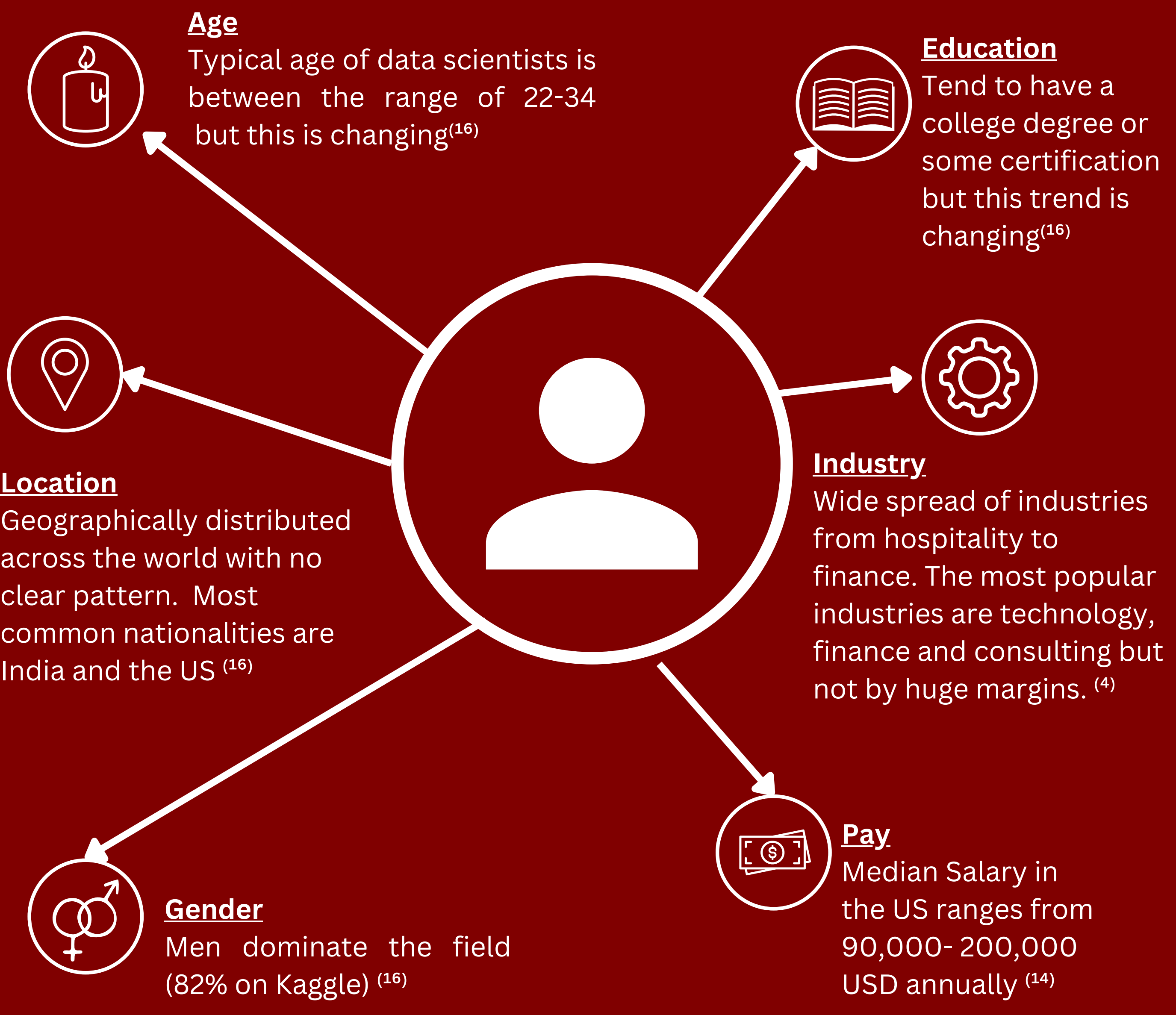
Data Engineer:
Expert in cloud skills and convert raw data into usable formats for data scientists and analysts.

Data Analyst:
A domain expert who gather and interpret data for organizations.

Machine Learning Engineer:
Research and build statistical as well as machine learning models using training data.

Data Scientist:
Work end to end to clean and interpret data into actionable insights.

Who works in Data?



Age
Typical age of data scientists is between the range of 22-34 but this is changing[16]

Education
Tend to have a college degree or some certification but this trend is changing[16]

Location
Geographically distributed across the world with no clear pattern. Most common nationalities are India and the US [16]

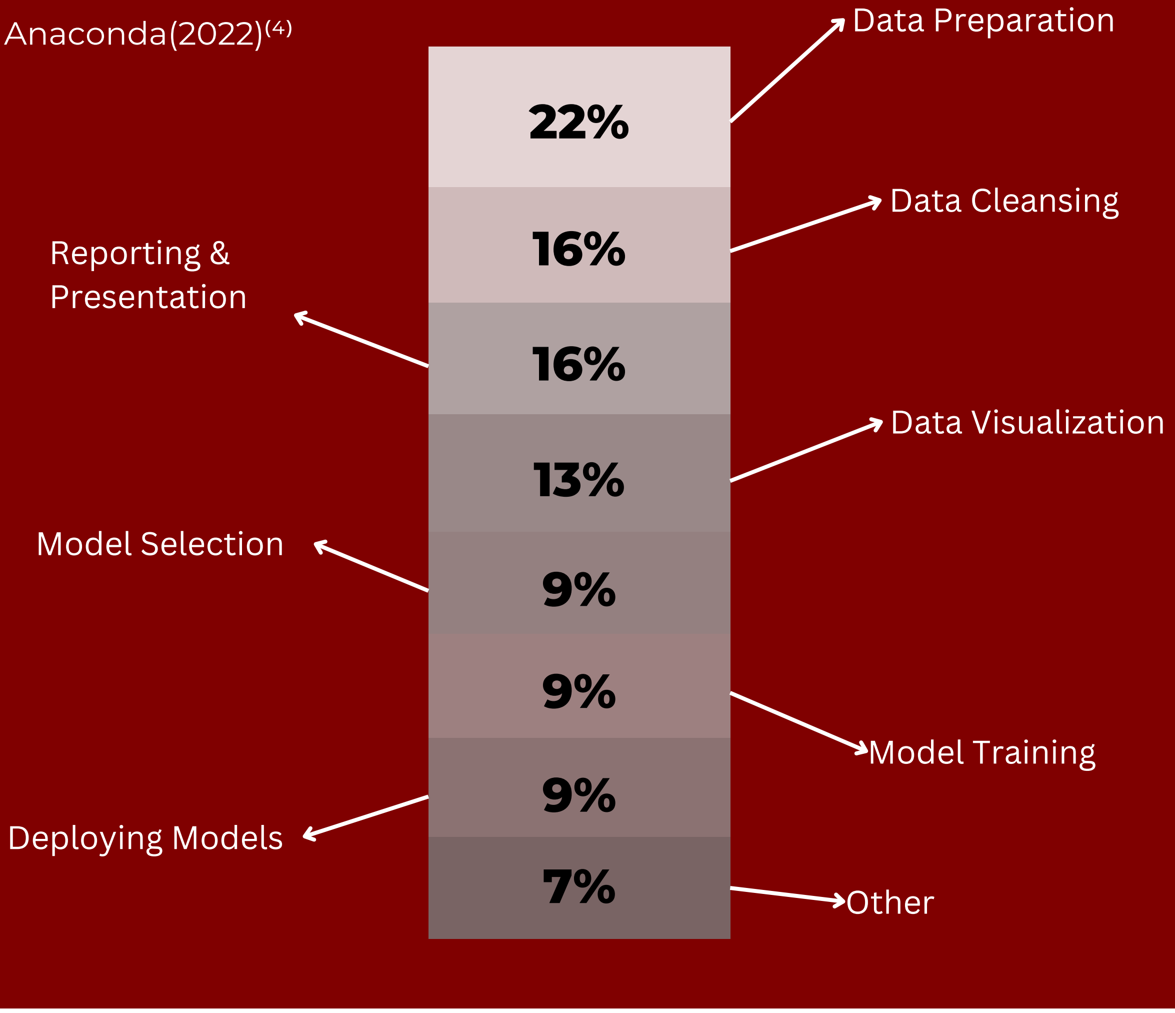
Industry
Wide spread of industries from hospitality to finance. The most popular industries are technology, finance and consulting but not by huge margins. [4]

Gender
Men dominate the field (82% on Kaggle) [16]

Pay
Median Salary in the US ranges from 90,000- 200,000 USD annually [14]

How are data scientists spending their time?

Source: Anaconda(2022)[4]



Global Data Science Governance & Policy

The onslaught of data science and data driven decision making has also been an important development for governments and governing bodies globally. Governments have responded to developments in data science by improving its adoption in their internal processes, by providing incentives to bolster data science centric industries in their regions and by establishing protocols for data security and governance.

Most countries in the developed world have now adopted a data or data governance strategy. The United States, for example, has an Office of Data Governance which outlines a comprehensive data strategy for the US government. This strategy involves the government into the management of the data ecosystem in various ways such as through ensuring data quality, maintaining a talent pipeline and assessing the existing landscape. Strategies were also devised for more specific verticals such as geo-spaitial data and enterprise strategy [17].

In China, there are several laws that lay down the basis for data governance and data science. Laws such as Data Security Law (2021) and Internet Information Service Algorithmic Recommendation Management Provisions (2022) lay down provisions to protect and organize data with an emphasized focus on security of government assets and citizen information [18].

The government of the UK has shared responsibility for data management and regulation across several departments with a focus on private public partnerships and collaboration to understand key risks and how these can be mitigated. Research is also being supported extensively through institutions such as the Alan Turing Institute and the UK Data Service [19].

India- a fast growing tech hub-has also adopted a multi-tiered approach to its data strategy. It has provided a unique identification number to each citizen through the Adhaar program which has allowed for greater authentication and tracking in the government system. Following this, the government has also focused on growing its digital payments system. The Data Empowerment and Protection Architecture has also been designed to push for consent in data sharing [20].

International bodies such as the United Nations (UN) and the Organization for Economic Co-operation and Development (OECD) have also taken several initiatives to adopt, support and regulate data science. The UN has several initiatives in place to support policy making in data such as the UN Commission on Science and Technology for Development (CSTD) which focuses on policy development in the area of technology, including data and information. It also supports research through the UN Global Pulse which is a research and innovation initiative that explores how big data can be used for development and humanitarian action. For data governance, initiatives such as the High-Level Panel on Digital Cooperation established by the Secretary-General promotes international cooperation on digital challenges and opportunities, including data governance [20]. Similarly, the OECD has put together Guidelines on the Protection of Privacy and Transborder Flows of Personal Data-these guidelines provide a framework for the protection of personal data in the context of international data flows [21].

Data governance is an evolving subject across the globe. As organizations and individuals create and work with a greater volume of data, governments and governing bodies must balance the interests of multiple stakeholders to best align domestic and international goals and priorities.

Data Science: the Pakistani Context

Though lagging behind in terms of development, Pakistan currently houses the world's fifth largest population. Moreover, this is an exceedingly young population with a median age of 22 [23]. The country's internet penetration of 54% in 2021 [24] is nearly a 40 percentage point jump from 15% in 2017 [25]. Internet usage in Pakistan has surged as a result of COVID 19. Moreover, almost one third of all online users in Pakistan have made a purchase online [24]. Pakistan's expanding digital footprint is an encouraging trend for two key reasons. Firstly, they indicate that Pakistan is a treasure trove of data and there is significant potential to be unlocked through data science in the country. Secondly, it indicates that the country hosts a young and digitally literate population that can be leveraged to create data science products for both local and global use. Already, there are signs that Pakistan's freelance and IT export markets are thriving.

Pakistan has a flourishing IT export sector which brings in USD 2.1 billion in revenue each year[26]. There are currently 12,000 IT companies in Pakistan and these employ nearly 600,000 professionals. The year on year growth of the industry is 18% and remittances have increased by 137% in the past five years [26]. As such, the IT industry has been recognized by the Government of Pakistan itself. According to Pakistan Vision 2025 and the Digital Policy of Pakistan 2018, the ICT industry size is targeted to reach $20 billion by 2025 [27].

The freelancing community is also thriving, bringing in USD 150 million in the financial year 2019-2020 alone. While a majority of these freelancers provide relatively low value services such as web development or graphic designing, there is significant promise that this market can be leveraged to build a dynamic data science community[28].

22

median age of the Pakistani population

2.1 billion

USD worth of IT exports

1/3

of all Pakistanti users have made a purchase online

54%

internet penetration in Pakistan

Pakistan's Market Landscape

In Pakistan, organizations and individuals working with data science are part of the larger IT and associated industries. Those working with data science can be divided into one of three categories or segments that operate in Pakistan's data and IT industry. It is difficult to get exact information on how large each segment is because no comprehensive primary study has been done to understand this market as it currently stands.

1 Freelancers

Freelancers are typically individual agents that operate through word of mouth or platforms such as UpWork, Freelancer and Fiverr. Freelancers working in data science typically take up projects from individuals and organizations in Pakistan and abroad. On these platforms, these individuals typically list a rate anywhere from about $30-50 per hour to $100 per hour depending on their expertise. There is a sizable community of Pakistanis working as freelance data scientists on such platforms. On Fiverr alone, 3158 Pakistani users have registered themselves as competent Data Scientists[29].

2 Service Providers

Service providers can...

implement specific solutions

OR

sell pre-made products

Arguably, the largest segment working directly with data science is that of software developers and service providers. These are large and mid-size organizations that work with local and international clients to sell particular products or services in IT consulting and development. Some of these organizations such as Emumba or Affiniti specialized in a particular product or practice. But for the most part, these organizations provide consulting and analytical services at large.

Some service providers working in Pakistan are described below along with their services and products.

Affiniti

A unicorn and valued at USD 1.6 billion[30], Affiniti is a major player in Pakistan's budding AI market. Based in the US, Affiniti provides an artificial intelligence and data science oriented product that uses data points on customer needs and trends to enhance the quality of their client's customer service. Using a "Pair Better" philosophy, Affiniti's AI can match customer service agents to customers and ensure that their client organization benefits from an improved brand perception.

Teradata Pakistan

Teradata is a global company with operations in Pakistan. It is a service provider that works with businesses to harness the power of data in various ways. It provides a host of solutions but has particular expertise in cloud and cloud management across a variety of sectors from automotive to finance [34].

Emumba

Emumba is a Pakistan-based organization that aims to provide agility and convenience to its clients. It has two major solutions. The first is K2 which is a dashboard framework that allows developers to quickly analyze data and create dashboards. The second is Axlerated Software which uses open source technologies to remove latency in high performing applications [31].

Systems Limited

Systems Limited is a renowned technology and IT consulting firm in Pakistan that works on solutions in-amongst others- Data Management & Analytics, Security and Cloud. It has worked with a variety of clients from Khaadi to Agha Khan University Hospital.Systems Limited is also Pakistan's only IT company to win Forbes Asia's Under a Billion Award three times in a row [32].

Bilytica

Another services provider, Bitlytica works with clients to provide solutions in data management, cloud, analytics and robotics. The company works in a variety of different industries across the country and the globe. Its work has been recognized by the Government of Pakistan with the President IT Award in 2021[33].

3 Mainstream Organizations

Apart from freelancers and service providers that work directly with data and data science, larger and more mainstream organizations in Pakistan are also adopting data science into their operations. Data & Analytics is often relegated to entire departments and teams in large scale organizations such as Unilever and P&G in Pakistan. In Telecom, Telenor and Jazz both have dedicated teams that track user data and conduct the relevant analytics to understand key trends in usage and preferences. Finally, start-up players such as Bazaar, Careem and Foodpanda also have dedicated protocols and teams to conduct data analytics.

Data Governance & Policy in Pakistan

As interest around the globe has risen in data and its various uses, governments have raced to put provisions in place to regulate this use of data by organizations and individuals. The Government of Pakistan passed a Personal Data Protection Bill in 2022 which establishes a National Commission for Personal Data Protection and protects users rights to their data. Accountability mechanisms are also set up to ensure that misuse of data is appropriately handled by the relevant authorities (35). The government has also sought to take advantage of the global growth in the industry by incentivising business to set up operations in the country and encouraging Pakistani to adopt tech-related skills. For example, the Special Technology Zones Authority has opened up seven technology zones all across Pakistan to promote tech-centric businesses. Similarly, the government has sponsored training programs such as the President's Initiative for Artificial Intelligence and the Skills for All program to provide the youth with training in technical skills.

It is not just the public sector that is making an effort to regulate and grow the data science industry in Pakistan but also other bodies such as P@sha and Pakistan Software Export Board. These associations promote IT exports in Pakistan and provide training to individuals and organizations in the country. Institutions such as the National Center for Cyber Security are also at the forefront of research and regulation in Pakistan from the private sector.

The Way Forward for Pakistan

Pakistan is one of the largest countries-in terms of population- in the world. As such, the onslaught of data science provides a host of opportunities and challenges for the country. In order to ensure that the country is able to take advantage of these opportunities and avoid the pitfalls that may come its way, the government of Pakistan needs to work in tandem with various stakeholders and devise a comprehensive strategy.

Key Challenges & Considerations

Data Protection

While Pakistan has made significant strides with its new bill, there is still room for improvement. Data protection and security is quickly evolving to be a more pressing concern than ever. To begin with, the language of the existing laws is ambiguous and leaves much to be desired. Moreover, there needs to emphasis on compliance with more steps being taken by the government to ensure that companies protect consumer rights to privacy.

Human Capital

For organizations and freelancers in Pakistan to truly take advantage of the growing data science market, it is necessary that human capital is digitally literate and skilled in the latest technologies. Pakistan's government needs to collaborate with universities and schools to update the curriculum being taught as well as support and facilliate programs that teach tech skills informally.

Opportunity

With Pakistan's IT exports offering promising growth, the government needs to capitalize on the global data science market and encourage Pakistani professionals and organizations to upskill with data science and produce high value data science products. This is a huge opportunity for a country like Pakistan which can take advantage of the difference in labour pay scales to improve its economy.

Data Quality

Data is a critical resource to harness the power of the latest and most cutting edge technologies. With a growing and increasingly online population, Pakistan generates a wealth of data but maintaining and bolstering this data so that it can be used by organizations and the government to uncover key trends is an important responsibility. The government needs to put in place mechanisms to ensure that data quality is high.

Adoption

Data science is a powerful tool and can-as shown in this report-bring about huge changes in the way that organizations operate. However in Pakistan, market conditions themselves may not be enough to encourage adoption of data science amongst organizations. Moreover, uplifting the existing broadband system to a high speed fiber optic networks is critical. To realize the demands of Pakistan's growing digital economy, this is a critical next step.

Collaboration

Not only is data science a growing field but also a fast changing one and as such, the government needs to work with other countries as well as international organizations to stay up to date on the latest protocols in data governance and regulation and ensure that it is able to make the right decisions when necessary.

References

1. IBM.(n.d) *What is Data Science?* https://www.ibm.com/topics/data-science#:~:text=Data%20science%20combines%20math%20and,decision%20making%20and%20strategic%20planning.
2. BuiltIn. (2023) *22 Data Science Applications and Examples* https://builtin.com/data-science/data-science-applications-examples
3. McKinsey.(2022)*The data driven enterprise of 2025*, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025
4. Anaconda (2022) *State of Data Science 2022* https://www.anaconda.com/state-of-data-science-report-2022
5. Bain & Co. (2013) *The Value of Big Data: How analytics differentiates winners.* https://www.bain.com/contentassets/5672af3b82f84aa2a80ca732fa8ea06c/bain20_brief_the_value_of_big_data.pdf
6. UNCTAD (2020), *COVID-19 has changed online shopping forever, survey shows,*https://unctad.org/news/covid-19-has-changed-online-shopping-forever-survey-shows
7. United Nations (n.d), *Big Data for Sustainable Development*, https://www.un.org/en/global-issues/big-data-for-sustainable-development
8. Venture Beat,(2022) *What is data analytics? Definition, models, life cycle and application best practice*s https://venturebeat.com/data-infrastructure/what-is-data-analytics-definition-models-life-cycle-and-application-best-practices/
9. IBM (n.d). *Big data analytics*, https://www.ibm.com/analytics/big-data-analytics
10. Copeland, B. (2022, November 11). *artificial intelligence*. *Encyclopedia Britannica*. https://www.britannica.com/technology/artificial-intelligence
11. IBM.(n.d) *Structured vs. Unstructured Data: What's the Difference?*, https://www.ibm.com/cloud/blog/structured-vs-unstructured-data
12. Statista (2022) *Global big data analytics market size 2021-2029*, https://www.statista.com/statistics/1336002/big-data-analytics-market-size/
13. EY (2022)*How companies are investing in data and analytics* https://www.ey.com/en_us/consulting/how-companies-are-investing-in-data-and-analytics
14. Harvard Business Review, *Is data scientist still the sexiest job of the 21st century?*https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century
15. Forbes (2018) *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read* https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/
16. Kaggle (2021) *State of Data Science and Machine Learning 2021* https://www.kaggle.com/kaggle-survey-2021
17. US Department of Labor (n.d) *Data Strategy* https://www.dol.gov/agencies/odg/strategy
18. U.S. China Economic And Security Review Comission (2022). *Chinas Evolving Data Governance Regime.* https://www.uscc.gov/sites/default/files/2022-07/Chinas_Evolving_Data_Governance_Regime.pdf
19. The Royal Society (2022)*The UK data governance landscape* https://royalsociety.org/-/media/policy/projects/data-governance/uk-data-governance-explainer.pdf
20. Feigenbaum, E. A., & Nelson, M. R. (2022, August 31). Data Governance, Asian Alternatives: How India and Korea Are Creating New Models and Policies. Carnegie Endowment for International Peace; Carnegie Endowment for International Peace. https://carnegieendowment.org/2022/08/31/data-governance-asian-alternatives-how-india-and-korea-are-creating-new-models-and-policies-pub-87765
21. United Nations (n.d), https://www.un.org/en/
22. OECD (2002), OECD *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data,* OECD Publishing, Paris, https://doi.org/10.1787/9789264196391-en.
23. Government of Pakistan, Pakistan Economic Survey 2012-2013 Link
24. Tribune. (2021, July 30). Country's internet penetration stands at 54%. The Express Tribune; Tribune. https://tribune.com.pk/story/2312994/countrys-internet-penetration-stands-at-54
25. Statista. (2017).*Pakistan: internet penetration rate 2017* https://www.statista.com/statistics/765487/internet-penetration-rate-pakistan/
26.
27. Government of Pakistan (n.d), Sector Profile: IT and Tech enabled industries https://invest.gov.pk/sites/default/files/inline-files/IT.pd
28. Government of Pakistan.(2020). Pakistan's IT Industry Overview http://www.moit.gov.pk/SiteImage/Misc/files/Pakistan%27s%20IT%20Industry%20Report-Printer.pdf
29. Fiverr. (n.d).https://www.fiverr.com/
30. Venture Beat, "Sales AI company Afiniti valued at $1.6 billion, files for IPO" Link
31. Emumba.(n.d) https://www.emumba.com/
32. Systems Limited (n.d) https://www.systemsltd.com/insights/case-studies
33. Bilytica (n.d) https://www.bilytica.com/
34. Teradata (n.d) https://www.teradata.com
35. Halim, W., Upadhyay, A., & Coflan, C. (2022). Data Access and Protection Laws in Pakistan: A technical review (Helpdesk Response No. 118). EdTech Hub. https://doi.org/10.53832/edtechhub.0098