ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Naudé, Wim; Bray, Amy; Lee, Celina

Working Paper

Crowdsourcing Artificial Intelligence in Africa: Findings from a Machine Learning Contest

IZA Discussion Papers, No. 14545

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Naudé, Wim; Bray, Amy; Lee, Celina (2021) : Crowdsourcing Artificial Intelligence in Africa: Findings from a Machine Learning Contest, IZA Discussion Papers, No. 14545, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/245596

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14545

Crowdsourcing Artificial Intelligence in Africa: Findings from a Machine Learning Contest

Wim Naudé
Amy Bray
Celina Lee

JULY 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14545

Crowdsourcing Artificial Intelligence in Africa: Findings from a Machine Learning Contest

Wim Naudé
University College Cork, RWTH Aachen University and University of Johannesburg and IZA

Amy Bray
Zindi

Celina Lee
Zindi

JULY 2021

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Crowdsourcing Artificial Intelligence in Africa: Findings from a Machine Learning Contest

In this paper, we study the crowdsourcing of innovation in Africa through a data science contest on an intermediated digital platform. We ran a Machine Learning (ML) contest on the continent's largest data science contest platform, Zindi. Contestants were surveyed on their motivations to take part and their perceptions about AI in Africa. In total, 614 contestants submitted 15,832 entries, and 559 responded to the accompanying survey. From the findings, we answered several questions: who take part in these contests and why? Who is most likely to win? What are contestants' entrepreneurial aspirations in deploying AI? What are the obstacles they perceive to the greater diffusion of AI in Africa? We conclude that crowdsourcing of AI via data contest platforms offers a potential mechanism to alleviate some of the constraints in the adoption and diffusion of AI in Africa. Recommendations for further research are made.

JEL Classification: O31, O33, O36, O55

Keywords: crowdsourcing, innovation, data science, artificial intelligence, Africa

Corresponding author:
Wim Naudé
Technology and Innovation Management (TIM)
RWTH Aachen University
Kackertstraße 7
52072 Aachen
Germany
E-mail: naude@time.rwth-aachen.de

1 Introduction

Artificial Intelligence (AI) has been described as a general-purpose technology with potentially far-reaching societal impacts (Goldfarb et al., 2021; Trajtenberg, 2018). Many developing countries, including many in Africa, are pinning hopes on AI to help accelerate growth, productivity, and competitiveness, and help to achieve the Sustainable Development Goals (SDGs) (Ndung'u & Signé, 2020; Vinuesa et al., 2020). However, as Brynjolfsson et al. (2017) and Crafts (2020) warned, the potential of AI to bring about rapid increases in productivity and GDP growth depend on supporting conditions and investments. With these lacking, and taking time to establish, the diffusion and adoption of AI will be slow, and the potential benefits out of reach. Recent research confirms that, despite the hype, adoption rates of AI remain modest - even in advanced economies[1] - and lagging in Africa (European Commission, 2020; Cisse, 2018). The slow diffusion of AI to Africa may contribute to further widening digital (and 'algorithmic') divides and data gaps between Africa and the advanced economies[2] (Naudé & Vinuesa, 2021; UNCTAD, 2021).

A voluminous literature has studied the determinants of the diffusion and adoption of new technologies, recognised the importance of intangible aspects, such learning, interaction, acceptance and experimentation, and tangible resources such as ICT infrastructure and financial capital – altogether key 'absorptive' capacities (Abramovitz, 1986; Keller, 2004; Kumar & Singh, 2019; Jovanovic & Rob, 1989). There is also agreement that, as Griliches (1957) pointed out more than half a century ago in the case of hybrid corn seed varieties, the demand for a technology can facilitate or hinder its diffusion. If, as with hybrid corn seed varieties in early 20th century America, a new technology is seen as not yet cost effective or profitable, it will not be used, or its adoption postponed until there is more certainty (Farzin et al., 1998). Both sets of determinants, respectively from the supply and demand side, are relevant to understand the slow diffusion and adoption of AI in Africa.

Overcoming these will moreover necessitate policies and investments not only on strengthening and building (tangible) ICT infrastructure, where most current digital economy development policies are focusing on (Ojanperä et al., 2019), but also to speed up learning,

[1] A U.S. Census Bureau Survey of over 800,000 firms in the USA found that only 2,9% were using Machine Learning (ML) in 2018 (Zolas et al., 2020). And a 2020 survey by the European Commission found that amongst EU firms who indicate to use AI, ' at the level of each technology, adoption in the EU is still relatively low. It ranges from merely 3% of enterprises currently having adopted sentiment analysis to 13% for anomaly detection and process/equipment optimisation' (European Commission, 2020, p.9).

[2] In UNCTAD's *Readiness for Frontier Technologies Index*, the world average score is 0,44, and the average score for Sub-Saharan Africa 0,17 (UNCTAD, 2021).

interaction and experimentation in data science and bring products and services based on AI faster to the market (Radhakrishnan & Chattopadhyay, 2020). It is not enough, as Graham et al. (2017) stress, to merely provide ICT infrastructure, referring to many instances where access to the internet did not result in it being used in Africa. What is important is to also encourage the participation of African-based agents in online content creation – using digital technologies for value creation, learning and interaction. For example, in 2017 the African share in collaborative coding on GitHub, domain registrations, and academic articles were respectively only 0,5%, 0,7% and 1,1% of all online content in 2017 (Graham et al., 2017, p.352).

Although still facing a digital divide and data gaps, and although rates of participation in the global knowledge economy is still relatively low, recent years have seen positive signs of progress and catch-up in Africa as far as digital technologies are concerned – beyond the use of mobile phones. Some of these are discussed in Cisse (2018) Naudé (2018) and Ndung'u & Signé (2020) and include the rise of indigenous digital platforms, the expansion of tech hubs and the growing amount of venture capital funding for tech start-ups, many of whom adopt AI. One notable area of progress has been in the crowdsourcing of innovation. In crowdsourcing innovation through intermediary digital platforms, specifically contest-based platforms, African-based initiatives have become internationally notable. In 2020, out of 107 data science contests held on 24 platforms worldwide, the third most (13%) was on an African-based platform (ML Contests, 2021; Olaleye, 2021).

In this paper, we study the crowdsourcing of innovation in Africa through data science contests on an intermediated digital platform. Specifically, we are interested in whether such data science contests can facilitate the diffusion and adoption of AI in Africa. This necessitates answering several questions: who takes part in these contests and why? Who is most likely to win? What are contestants' entrepreneurial aspirations and what are the obstacles they perceive to the diffusion of AI in Africa? To answer these questions, we designed and issued a ML challenge on Africa's largest data science contest platform, *Zindi*, in 2020 – contestants were challenged to submit a ML (*recommender)* model to predict sales of insurance products for *Zimnat*, the largest insurance company in Zimbabwe. The challenge was sponsored by the *Volkswagen Foundation* as part of a research project to understand the diffusion of AI better. In tandem with the ML challenge, contestants were surveyed on their motivations to take part and their perceptions about AI in Africa. In total, 614 contestants from across the continent and further afield submitted 15,832 entries and 559 responded to the accompanying survey. Our central hypothesis is that, given the low rates of diffusion of AI in Africa, and the relative lack of ICT skills, that the growing popularity of data science

contests partly reflect weaknesses in labor markets and educational facilities. In essence, intermediated data science contests may be a mechanism to overcome shortcomings in the supportive institutions for the diffusion of AI.

The results from the contest and survey are described in this paper, after a discussion of the relevant literature on the crowdsourcing of innovation, and the benefits and downsides of digital platforms as intermediaries. In summary, we found that data science contests via an intermediary digital platform in Africa may provide a mechanism to overcome labor market and educational obstacles to AI adoption in Africa. Several recommendations for further research are provided.

This paper contributes to several strands of recent literature on the digital economy, innovation, and artificial intelligence, and how this is evolving in Africa. Firstly, to the best of our knowledge, this is the first paper to undertake a systematic analysis of a crowdsourcing contest held on Africa's largest data science contest platform. This complements the rising tide of recent work on Africa's participation in the digital economy, which includes extensive surveys of the digital economy and digital work in Africa by Graham (2019), Anwar & Graham (2020), Friederici et al. (2017) and Ojanperä et al. (2019) amongst others. These authors studied amongst others the measurement of Africa's knowledge economy, the participation of African workers in digital online creation, the role and extent of integration of African digital workers into global labor markets as freelancers through digital work platforms, and the use of hackathons. They did not study the role and extent of data science competitions however, even though Africa is quite prominent in international data science competitions, and contests are used with increased frequency.

Second, our paper contributes to the knowledge base on crowdsourcing contests, more specifically on the crowdsourcing contests in the data science community. According to Tauchert et al. (2020, p.1) 'While there generally has been a lot of research done for crowdsourcing, there is, after an extensive investigation, almost no research available addressing the combination of both, crowdsourcing and data science.' Our paper explicitly addresses both crowdsourcing and data science.

Third, whilst there is a large literature on the motivations for taking part in crowdsourcing challenges in general (see e.g., Acar (2019); Ghezzi et al. (2018)), there is a gap in the literature on the motivations to take part in of data science contests. Our paper attempts to also address this gap, as motivations to take part in data science contests may provide us with information with which to probe indirectly the determinants of the (slow) diffusion of

AI.

Finally, given that Africa's knowledge economy is 'surprisingly understudied' (Ojanperä et al., 2019) and given that there is ' a dearth of data on all aspects of artificial intelligence (AI) in Africa' (Gwagwa et al., 2020, p.2), our paper contributes to expanding our perspectives and understanding of Africa's knowledge economy and specifically the acceptance, adoption and perspectives of AI.

The rest of the paper is structured as follows. In section 2 we place the study within the context of the scholarly literature on crowdsourcing of innovation. We also survey the current landscape of data science contests as a method of crowdsourcing innovation in the digital economy, and list notable data science contests platforms currently in operation, noting the predominance of African-based contests. Section 3 contains the empirical analyses, which are twofold: we present an analysis of a survey of contestants, as well as an analysis of the outcomes of a ML contest that we designed and held on *Zindi*. Section 4 concludes with a summary and recommendations for further research.

2 Relevant Literature

While the crowdsourcing[3] of AI in Africa is a very recent phenomenon, using the crowd to source a solution to a problem, or find an innovation, can according to Afuah & Tucci (2012, p.355) be traced 'as far back as 1714, when the British government offered a cash prize -the Longitude Prize- to anyone who would come up with an elegant way to determine the position of ships in the sea.' What has changed since then is that the availability of the internet and increased computing power have in recent years enabled the scaling up of the process of crowdsourcing (Mao et al., 2017; Piller & Walcher, 2006).

Crowdsourcing can more formally be described as 'the process of a crowd-seeker tapping the intellect of a large pool of independent individuals (usually referred to as crowd) to either collaboratively or competitively fulfil the requirements of a project normally delegated to employees in an organization or a task that requires human judgement' (Ayaburi et al., 2020, p.1227). It can take place through crowdsourcing communities - essentially a collaborative approach[4] - or through crowdsourcing challenges or contests, also referred to as tournament-

[3]The term crowdsourcing is ascribed to Jeff Howe who used it in an article in Wired magazine in 2006 (Acar, 2019).

[4]Examples include *InnoCentive's* Open Innovation Marketplace or *crowdSPRING*.

based crowdsourcing (Acar, 2019; Afuah & Tucci, 2012).

Contests or tournaments can in turn be hosted either directly by the crowd-seeker (organizer / firm) (see e.g., Lüttgens et al. (2014)) or indirectly via an intermediary platform through issuing a contest to find the best solution. An example of a directly issued crowdsourcing contest is the case of Netflix, which in 2006 offered a prize of US$ 1 million for software that 'could better predict which movies customers might like than its own in-house recommendation software, Cinematch [...] thousands of teams made submissions until one claimed the prize in 2009 by showing that its software was 10% better than Cinematch' (Carpenter, 2011, p.698). This contest resulted in around 45,000 entries, with the winning solution containing more than 100 individual algorithms (Acar, 2019). And an example of a notable crowdsourcing contest via an intermediary digital platform, in this case *Kaggle*,[5] was the 2019 Deepfake Detection Challenge (DFDC),[6]which offered a prize of US$ 1 million, and which was sponsored by the crowd-seekers Amazon Web Services (AWS), Facebook, Microsoft, and the Partnership on AI.

In a crowdsourcing contest via a platform intermediary - the main focus of interest in this paper - a crowd-seeker 'posts a task (e.g., logo design, programming task) on a platform and announces a monetary reward that he is willing to pay for a winning solution. Contestants (e.g., designers or programmers) submit solutions on the platform and the requester chooses the best solution (possibly more than one) and awards the prize'(Segev, 2020, p.241). Crowdsourcing platforms 'act as intermediaries between organisations and the crowd' (Pavlidou et al., 2020, p.2052).

The literature on crowdsourcing more broadly, and crowdsourcing contests more specifically, is voluminous and expanding, and covers various scholarly fields, in particular strategy, organization, economics, and innovation (Felin et al., 2017). In the latter it is given much attention in the areas of open innovation and co-creation studies (Ghezzi et al., 2018; Lüttgens et al., 2014). As far as crowdsourcing contest or tournaments are concerned, the literature has been concerned to provide an theoretical underpinning (mostly from game theory analyses of auctions); to determine for which projects crowdsourcing contests are best suited; and to better understand the motivation and behaviour of contestants (Segev, 2020). Reviews of these strands of literature are contained in Segev (2020), Mao et al. (2017), Ayaburi et al. (2020), Ghezzi et al. (2018), Zhao & Zhu (2014) and Corchón (2007).

[5]Kaggle, a platform established in 2010 in Australia, has become one of the most prominent of global data competition platforms. In 2017 it was bought by Google.

[6]See `https://www.kaggle.com/c/deepfake-detection-challenge/`. More than 2000 teams submitted over 8500 entries.

Some of the core findings of this literature is pertinent for the present study. For instance, are crowdsourcing always better than more standard procurement to obtain solutions? who takes part in contests, who wins, and how does the incentives to take part affects participation and performance? and how should a crowdsourcing contest best be designed to result in an optimal winning solution? Some of the answers to these questions, on which the literature seems to be converging, are that first, that crowdsourcing contests are not necessarily more efficient than traditional 'idealized' methods of procurement,[7] one reason being that 'the effort of losing contestants is wasted' (Chawla et al., 2015, p.80). However, idealised procurement is not always available, and there are many instances where practical realities depart significantly from the simplified theoretical world (Segev, 2020).

The literature has also established that individuals who takes part in contests are motivated by both intrinsic (e.g., entrepreneurial mindset, enjoyment), extrinsic motivations (monetary rewards, career objectives, reputation) and learning and prosocial motivations (Ghezzi et al., 2018). Acar (2019), using data from *InnoCentive*, found that the openness of contests can give rise to large numbers of inappropriate solutions and that moreover the motivation to enter a contest affects the appropriateness. Specifically, he found that contestants who were motivated by intrinsic and extrinsic motivation submitted more appropriate bids, whilst those motivated by learning and prosocial interaction submnitted less appropriate solutions.

Second, who wins and the quality of the winning solution depends on the effort, skills, and creativity of contestants (Moldovanu & Sela, 2001; Gross, 2020). In this respect, Körpeoglu & Cho (2018) showed that a free-entry open contest, with larger number of contestants, will elicit better performance amongst contestants due to increased competition. Chawla et al. (2015), modelling crowdsourcing as an all-pay auction, argued that the best contest design is an 'ironed virtual value optimizer' - in the parlance of game theory, meaning that the reward should be 'divided evenly among all contestants whose submissions are above a minimum quality threshold, and are tied under a weakly monotone transformation (via an ironed virtual value function) of the submission quality' (Chawla et al., 2015, p.81). However, this is complex and impractical and thus in practice what is often found are winner-takes-all or rank-based-rewards (e.g., decrease prizes for the best three solutions). Relatedly, Moldovanu & Sela (2006) studied the conditions under which it will be optimal to divide the prize amongst the finalists in the contest, and not follow a winner-takes-all approach.

So far, only a relatively small part of this literature has specifically been devoted to the

[7]They establish that 'Idealized procurement leads to a better quality outcome in comparison to crowdsourcing' (Chawla et al., 2015, p.82).

special case of data science crowdsourcing contests - which is most relevant for the case of crowdsourcing AI in Africa. Tauchert et al. (2020, p.1) recently concluded that 'While there generally has been a lot of research done for crowdsourcing, there is, after an extensive investigation, almost no research available addressing the combination of both, crowdsourcing and data science.' There is still a lack of scholarly attention to the process, participation in, and overall efficiency and contribution of data science contest platforms in general. There are no studies yet, to the best of our knowledge, that focuses on the case of data science crowdsourcing contests in Africa.

The extant literature on data science crowdsourcing contests have none the less being able to establish a number of findings that are relevant for present purposes. The first is that data science crowdsourcing contests can be efficient, and can contribute to the crowd-seekers' business. In this respect Humphries et al. (2018) hosted a competition on the data science contest platform *DrivenData* challenging participants to submit a time-series model that can predict the population of Antarctic penguins. Four of the submitted entries were able to provide better predictions that an earlier model developed by a team of domain experts. And Bojer & Meldgaard (2021) analysed and reviewed six *Kaggle* forecasting competitions, concluding that these competitions can indeed contribute to improve forecasting of daily and weekly business time series. Other studies that similarly reported on the efficiency of data science contests include Taieb & Hyndman (2014) and Makridakis et al. (2020). Data science contests have also helped in steering advances in artificial intelligence (AI). For example, the *ImageNet* classification competition pitched data scientists against each other to build models to classify the millions of images on *ImageNet*. This resulted, as Marconi et al. (2019) discuss, in an increase in classification accuracy from 70% in 2010 to 97% in 2017.

Second, although data science contests can be efficient, they are also subject to shortcomings, which can negatively affect their efficiency. Some of the general concerns about crowdsourcing remain valid (see above). Tauchert et al. (2020) contain one of the few studies so far to attempt to evaluate why firms may want to use data science contest via a platform, and how they evaluate the success or performance of the solution that result from it. Interviewing ten users of *Kaggle* contests, they concluded that although firms do perceive value from the solutions thus obtained, solutions were often not ideal, because not all the tasks characterizing the internal data processes and needs of firms can be crowdsourced. Brackbill & Centola (2020) studied the process of solution discovery among distributed groups of data scientists across 16 independent data science competitions, finding, counter-intuitively, that more efficient communication networks amongst contestants can reduce the likelihood of the

discovery of more novel solutions. This is because 'the faster solutions of moderate quality diffuse through an organization, the more likely groups will abandon novel and unproven ideas and settle for an existing solution rather than working to discover groundbreaking innovations' (Brackbill & Centola, 2020, p.1).

In addition to the peer-reviewed literature, less formal and rigorous assessments of data science contests have been provided by industry participants and others. For instance, ML Contests (2021) and Olaleye (2021) analysed the nature of participation in 127 data science contests globally over 24 platforms in 2020. According to ML Contests (2021) the total prize money in these competitions amounted to US$3,5 million, the most competitions were issued by *Kaggle*, and most competitions required solutions to problems from the perspective of supervised learning, computer vision, and reinforcement learning. Olaleye (2021) reports that 61% of winners were individuals (teams of 1), that 94% of these winners used the Python programming language, and that for Deep Learning they used only the PyTorch and TensorFlow frameworks.

There are many potential advantages to using a data science contest through an intermediary platform format to promote the diffusion of AI - and this explains the extent and growth in the number of data science contests in recent years as noted also in the previous paragraph. Table 1 provides a list of leading data science competition platforms circa 2021.

Table 1 lists 26 active platforms circa 2021. Most of these are in the USA, followed by India, Europe and China. There is one platform in Africa - Zindi - headquartered in South Africa, but with a regional office in Ghana. In addition to these 26 active platforms, mention can also be made of the annual *International Data Analysis Olympiad* (IDAO)[8] which is held in Russia and organised by the Higher School of Economics (HSE) and Yandex. This contest attracted 2756 participants from 83 countries in 2020.

Amongst the advantages that data science contest platforms offers, are opportunities for learning and skills development, for recruitment and better labor market participation, and for efficiently crowdsourcing innovations (as shown in the preceding paragraphs). While the actual provision of a solution is of importance to the crowd-seeker, given that not all the tasks characterizing the internal data processes and needs of firms can be crowdsourced as was mentioned (see (Tauchert et al., 2020)), data science contest platforms' value rests perhaps more in the learning/skills development and labor-job matching areas - i.e. matching scarce talent and skills with the employment demands of firms. Indeed, most of the data

[8]See `https://idao.world`.

Table 1: Notable data science competition platforms circa 2021

Platform	**Location**	**Website**
AICrowd	Switzerland	`www.aicrowd.com`
Alibaba Cloud Tianchi	China	`www.alibabacloud.com`
Bitgrit	Japan	`bitgrit.net/competition/`
Challenge Data	France	`challengedata.ens.fr`
CodaLab	France	`competitions.codalab.org`
CrowdANALYTIX	USA, India	`www.crowdanalytix.com`
Data mining cup	Germany	`www.data-mining-cup.com`
DataFountain	China	`www.datafountain.cn`
DataHack	India	`datahack.analyticsvidhya.com`
Datasource.ai	USA	`www.datasource.ai/en`
DPhi	Belgium	`https://dphi.tech`
DrivenData	USA	`www.drivendata.org`
EvalAI	USA	`https://eval.ai`
HackerEarth	USA, India	`www.hackerearth.com`
IEEE DataPort	USA	`ieee-dataport.org`
Kaggle	USA	`www.kaggle.com`
MachineHack	India	`machinehack.com`
Numerai	USA	`https://numer.ai`
Omdena	USA	`omdena.com/projects/`
RAMP	France	`https://ramp.studio`
Signate	Japan	`https://signate.jp`
Topcoder	USA	`www.topcoder.com`
Unearthed	Australia	`unearthed.solutions`
Waymo	USA	`waymo.com/open/challenges/`
Xeek	UK	`https://xeek.ai`
Zindi	South Africa	`https://zindi.africa`

Source: Authors' own compilation.

science contest platforms listed in Table 1 aim to establish a community of data scientists and support them through provision of short courses, resources, links, hackathons,[9] and datasets. *DataHack*, based in India, for example emphasizes in their aims to improve the job market potential of contestants, by allowing them to signal their skills, but moreover also to test and build their skills. They offer 44 different educational courses online, some free and some at payment, for example on AI and ML, Introduction to Web Scraping, Tableau for Beginners, Getting Started with Neural Networks and on Writing Powerful Data Science Articles, amongst many others. It is thus fairly accurate, as put by Neo (2019) that data

[9]Hackathons are also crowdsourcing of innovation activities, but distinct from data science competitions in that it has a shorter focus, a stronger collaborative effort, and a greater concern with software development. Specifically, they are short events ' often lasting a day or two -where organizers invite people to imagine and prototype software applications' (Irani, 2019, p.224). For a critical discussion of Hackathons in the context of developing countries, see Irani (2019).

science competitions are 'the perfect place to learn best practices, accrue feedback on your work, and augment your skills.'

Other platforms focus more on the job-market aspect. For example *TopCoder,* a contest platform concerned with the 'future of work', describe their core business as essentially providing an 'on-demand talent platform.' They make available what they term a 'Talent-as-a-Service app' as a 'Freelancer on-demand offering.' Similarly, *MachineHack* describe their purpose as to help industry to 'discover and evaluate talented data scientists' and to crowdsource innovations.

To the extent that availability of data sciences skills, and the ability of firms to employ sufficiently skilled data scientists limit the diffusion of AI in Africa, data science contest platforms with their community approach seems much to recommend it.

While the majority of data science contest platforms tend to be general and broad in the challenges they offer, some tend to be more specialized by focusing on specific domain challenges –examples include the Australian platform *Unearthed*,[10] which is concerned with matching novel data solutions and talent to the needs of the energy and natural resources industry, *Xeek*,[11] which has as purpose to 'unite the data and geoscience communities around the shared goal of crowdsourcing innovative solutions' and *Numerai*,[12] which holds tournaments for picking the best ML models for predicting the stock market. While most data science contest platforms offer all the mentioned services to some degree or the other, most tend to put more weight on one of these aspects and differ in their governance. For example, some platforms will emphasize learning, while another will emphasize recruitment, and another the crowdsourcing of innovations. Some will be private owned for-profit companies or non-profit foundations, whilst others are open-source web platforms, such as *Codalab*[13] and *EvalAI.*[14]

In comparison to the literature on crowdsourcing contest platforms more generally, the literature on data science contests are not yet voluminous. While the studies surveyed in the preceding paragraphs have dealt with issues such as the efficiency of data science contests and the methods used and the nature of the competition problems, it has neglected the characteristics and motivations of contestants, including their skills, experience, and their perceptions of AI and data science. This literature has also not taken a regional view, e.g.,

[10]See `https://unearthed.solutions`. They describe themselves as a community of 'startups, developers, and data scientists making the energy and resources industry more efficient and sustainable.'

[11]See `https://xeek.ai`.

[12]See `https://numer.ai`.

[13]See `https://competitions.codalab.org`

[14]See `https://eval.ai`.

investigated whether the location of contestants' matter for their success.

In the rest of the paper we attempt to contribute to filling the gaps that we have noted in the current literature on data science contest platforms, focusing on the case of data science contests in Africa and their role in the diffusion of AI. Our central hypothesis is that, given the low rates of diffusion of AI in Africa and the evidence of a relative lack of ICT skills, as was discussed in the introduction, that the growth in data science contests on the continent reflect weaknesses in labor markets and educational facilities in data science generally, and AI specifically. In essence, the rise of data science contests may be a response to shortcomings in the supportive institutions for the diffusion of AI.

3 Empirical Analysis

3.1 Methodology

3.1.1 The Data Science Contest Platform

Zindi has more than 26,000 registered users across 45 African countries, making it the largest data science competition platform in Africa. Since its establishment in September 2018 it has run more than 130 contests and hackathons, eliciting around 51,000 enrollments and awarding more than US$ 250,000 in prize money. Between 2019 and 2020 its number of contests and hackathons more than tripled, from 25 in 2019 to 81 in 2020.

The running of contests and hackathons are central to the business model of *Zindi*. Various organizations ('crowd-seekers') approach *Zindi* with data and a problem that needs a solution. For a hosting fee, *Zindi* prepares the dataset, defines the challenge, sets the prize money, the competition duration, and other specifications of the competition. Its competitions are published on `www.zindi.africa/competitions`. *Zindi* promotes its contests on its social media platforms. These are also redistributed on websites of initiatives such as *ML Contests*,[15] which publishes details of data science competitions from more than 20 different data science competition platforms worldwide.

Over the course of a challenge, contestants download the data from the competition page, build ML models to solve the problem, and submit files (in *csv*-format) that hold answers

[15] See `https://blog.mlcontests.com`.

predicted by their models. Submissions are typically scored using one of several error metrics (in the present case a binary log loss function). The better the score, the better the model is performing against the public test data set, and contestants are ranked on a public leader-board. At the end of the contest, contestants can select their two best submissions, and these are then scored on a private test set (this prevents model over-fitting). The top three contestants are then asked to submit their full code for review and verification.

Once the winning code is verified and cleaned up, it is delivered to the organization that sponsored the competition, and IP rights transferred from the winning contestants to the organization.

3.1.2 The Challenge

The present challenge was organized on *Zindi* by Zimbabwe's largest insurance company, *Zimnat* (the crowd-seeker in this case). The challenge was very simply stated as 'Can you predict which insurance products existing clients will want next?' Hence, *Zimnat* desired a Recommender System (RSs) for their insurance products. "Recommender Systems (RSs) are tools and techniques for information retrieval and filtering, used to suggest items to be used or consumed [...] RSs either generate a set of personalized recommendations/suggestions of items that are expected to be useful for a certain user or try to predict whether a specific item will be of interest to a user or not, based on his/her previous preferences and those that are observed on similar users. In their simplest form, the recommendation is provided in form of a list of ranked items" (Gatzioura et al., 2019, p.4).

Three prizes were offered: A US$ 2,500 first prize, a second prize of US$ 1,500 (60% of the first prize) and a third prize of US$ 1,000 (40% of the first prize). Contestants had 75 days to provide a solution or bid[16] , were allowed to work in teams, and to submit multiple bids. A leader-board provided up-to-date process feedback on the quality submitted entries. Such feedback has been found to be useful in the empirical literature for galvanizing competition and improving the quality of the winning bids (Segev, 2020). In the present case the quality of feedback as reflected on the leader-board was based on an objective metric for determining the best solution, namely the log-loss function of the binary classifier used to model the answer to the question (see below for a further explanation).

Once contestants registered for the contest on *Zindi's* website and had been issued with a

[16]Entries or solutions are often referred to in the literature as 'bids' as these are analogous to an auction (Ayaburi et al., 2020).

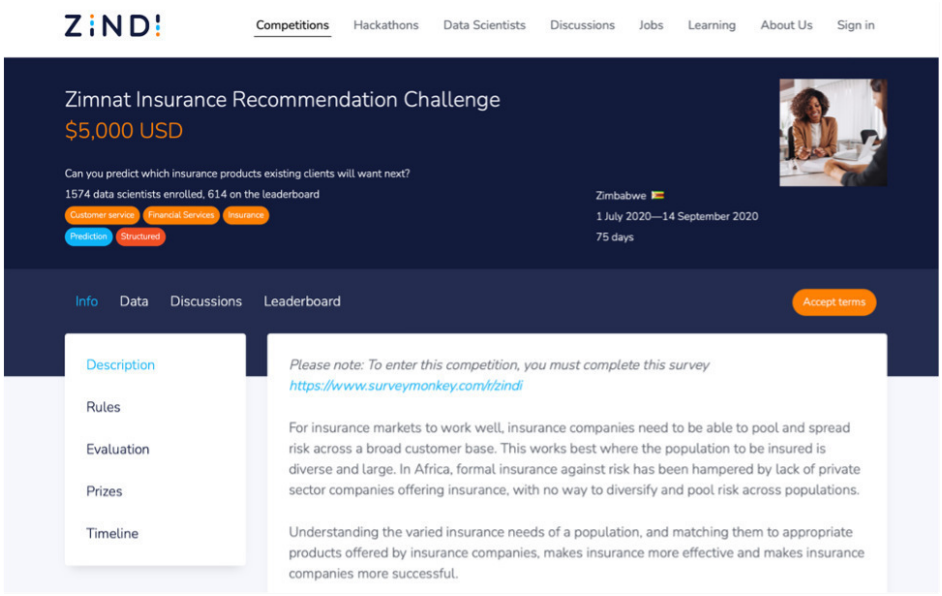
Zindi username, they gained partial access to a large dataset from *Zimnat*, covering more than 29,000 customers. Contestants could only access incomplete data on 10,000 customers of *Zimnat* who had purchased more than one insurance product. Specifically, contestants were given data on all but one of the insurance products of every *Zimnat* customer. To answer the challenge, they had to develop a Recommender System. There were 29 insurance products to choose from. The value of such a model is that it can then also be applied to any customer to identify further insurance products that might be useful to them given their current profile. Thus, the contestants face a (binary) classification task – classifying further insurance products into those that will be chosen and those that will not be chosen. We chose a Recommender System contests because Recommender System are amongst the most established and widely diffused uses of ML, and hence could be seen as a basic AI model building challenge.

Communication about the contest and its requirements and how it will be evaluated are important steps in crowdsourcing innovations. In this regard Pavlidou et al. (2020, p.2053) stated that 'Following planning, a project announces an open call for participation. A precise description with timeline, requirements and expected goal makes it easy for an individual to assess whether they are interested and suitable for the project.' The present contest was announced and open on 1 July 2020 and ran until 13 September 2020 - thus for 75 days. The announcement (See Figure 1) provided precise information, an important requirement as stressed by Pavlidou et al. (2020), and contained a description, the rules, the prizes, and the timeline. On the competition website contestants could also find the data, a discussion board, as well as a leader-board, which tracked progress as the competition progressed. An advantage in this contest, as in most data science competitions, is that the metric for deciding the winner (s) is clear and objective.

During the period 1 July to 13 September 2020, contestants could upload multiple entries, i.e., submit more than one model – with provisional results shown on the leader-board as mentioned. Although it is a competition, contestants could also collaborate in teams, and submit a solution on behalf of a team. Indeed, some of the best solutions came from teams, even though, as in data science competitions in general, most entries are by individuals (Olaleye, 2021).

By the close of the competition, 614 contestants submitted in total 15832 entries. Submissions were received from participants in 62 countries. Most were from Africa, but there were also a significant number of submissions (21% in fact) from India. From within Africa, most submissions were received from Nigeria (15%), South Africa (9%), Kenya (8%), Zimbabwe

Figure 1: Screenshot of the Zindi Webpage Announcing the Zimnat Insurance Recommendation Challenge



Source: `https://zindi.africa/competitions/zimnat-insurance-recommendation-challenge`

(8%), Tunisia (3%), and Ghana and Ethiopia (each 2%).

3.1.3 Model Assessment

In this type of crowdsourcing contest, selecting the best solution is relatively easy. The task here required is to develop a binary classifier model (Naik & Purohit, 2017) as Recommender System. That is, they need to predict, given clients of *Zimnat's* past insurance choices, which other insurance product they will also choose. For each possible additional insurance product, it needs to predict either yes (1), or no (0) with respect to whether the client will choose it. With such an aim, the best performing model will be the one with the highest accuracy in predicting the second insurance product that customers had bought. Such accuracy can be measured in terms of the cross-entropy log-loss (CE) score (S) (see e.g.,Godoy (2018)), which can be written as:

$$S = -\frac{1}{n}\Sigma[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \tag{1}$$

Where n is the number of insurance products being predicted; $\hat{y}_i$ the predicted probability of the insurance product being purchased by the client. $y_i = 1$ if the insurance product

had been purchased, = 0 if not. Use of the cross-entropy log-loss score is a well-established metric in data science competitions with which to rank proposed solutions. In the case of the present competition, the winning entry had a log-loss score of 0.0257, implying an approximate accuracy of 97%. Binary classifiers can be obtained using several algorithmic methods, including Logistic Regression, Decision Trees, K-Nearest Neighbors, Support Vector Machine, Random Forest, Neural Networks and Naïve Bayes (Ortner, 2020). In the present case the winning team used Decision-Trees (Catboost, LightGBM) and Neural Networks.

3.1.4 The Survey

Participation in the competition was made conditional on completion of a short online questionnaire. The questionnaire aimed to capture the perceptions of data scientists in Africa, including broadly all those participating in African data science contests - hence aspiring to showcase and match their talents in Africa. The questionnaire is contained in the Appendix.

The questionnaire was designed to collect, in addition to basic information on the contestants (e.g., location, education and other personal characteristics) also information on contestants' labor market participation, and on their perceptions regarding AI in Africa – their views on opportunities, obstacles and their entrepreneurial proclivities in this regard.

In total, 559 responses were collected, of which 83% were of males, and 15% of females. The median age of participants was 25 years. Only 10% of the participants were older than 35 years. This indicates that contestants participating in data science contests in Africa are largely (not exclusively) male and young. Their young age may reflect that they are participating to improve their skills and signal to prospective employers, as per our hypothesis. Below we will further investigate this.

3.2 Results

3.2.1 Who participates in data science contests?

The contest as described in the previous section required completion of a separate questionnaire. This allowed us to obtain information on who participates in data science contests for AI solutions, match this with who win in these contests, and furthermore solicit their views as knowledgeable participants in Africa's knowledge economy on how they perceive AI and

AI-entrepreneurship opportunities and constraints on the continent.

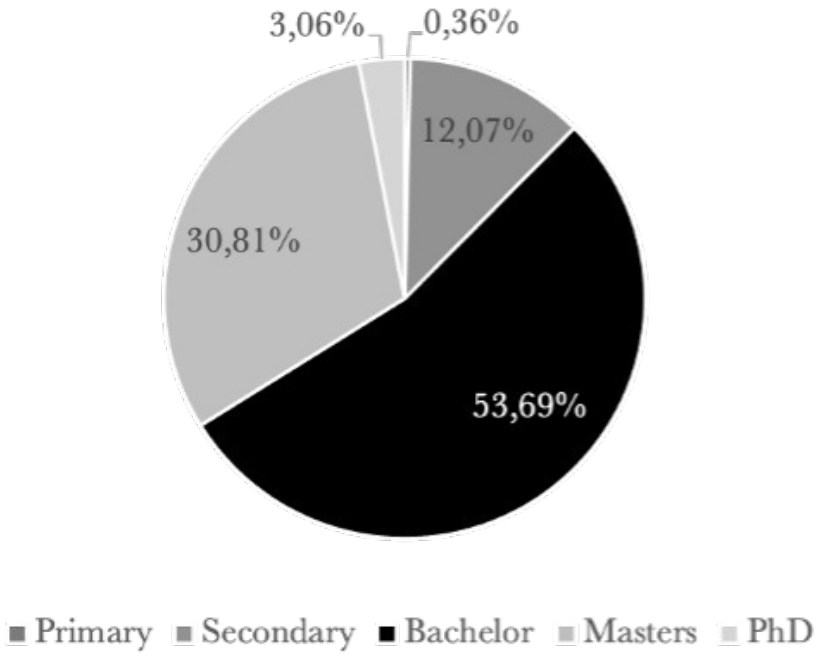
The first question is : who participates?

The answer is, mainly young people, and mainly men. The median age of participants were 25 years (27 years average) and the majority – 83% - were male. The gender share roughly corresponds to the share of males in the *Zindi* user base of 17,000 data scientists (79%).

Participants came from 62 countries, in Africa mainly from Nigeria, South Africa, Kenya, Zimbabwe, Ghana and Ethiopia. In African data science competitions, the participation of data scientists from India is noticeable – in the present case 21% of participants were from India. This is perhaps not surprising in view of the fact that no less than four of the major data science contest platforms in the world are based in India (see Table 1).

The broader literature on crowdsourcing has established that 'domain relevant skills' and education that is 'at least partially related to the project' were characteristics of contestants who submitted effective (winning) solutions (Pavlidou et al., 2020, p.2050). Therefore, the first section of our questionnaire collected data on the skills profile of participants. From the responses it was found that 53% of participants had a bachelor's degree as their highest level of education, 31% held a master's degree, and only 3% had a PhD – see Figure 2.

Figure 2: Highest level of education of contestants



Source: Authors

Only 35% of the participants indicated that they had a formal qualification in data science – most participants therefore seemed to have learned their data science abilities outside of the formal education sector. Consistent with this is that only 9% of participants rated their own AI-expertise as advanced whereas the majority (56%) considered it to be intermediate and 36% rated their expertise as beginner.

These findings are somewhat consistent, but with a subtle difference, with the findings of Anwar & Graham (2020, p.96) that 'as machine learning – the science of getting computers to make specific decisions – becomes ever more important in the contemporary era, there will be a concomitant rise in the need for people to train those machine learning systems [...] many of these new jobs require little training other than basic computer literacy on the part of the worker.'

Thus, the profile of who participated in our data science competition is of relatively young people, with intermediate to beginner level skills in AI and mostly not formally qualified in data science. This may suggest that a motivation to enter a data science competition may be job market signaling and/or gaining experience and feedback. In other words, a data science competition may partly fill in for gaps in the formal education and labor institutions in the participants' countries. This is preliminary confirmation of our hypothesis.

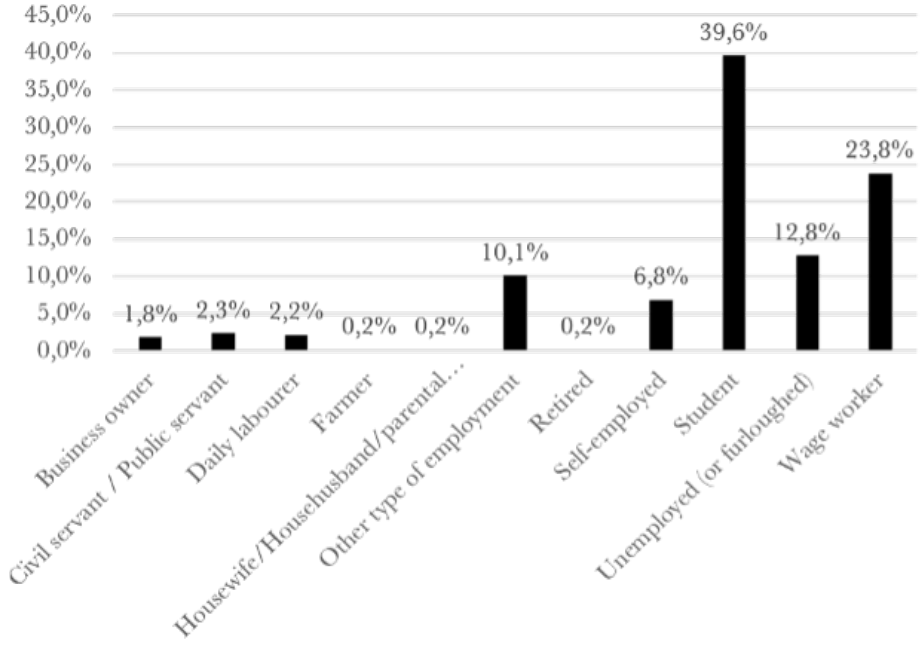
This conclusion is further support by the fact that the largest category of participants were students (40%). Figure 3 shows all the responses to this question.

As can be seen from Figure 3, altogether 53% of participants were either still studying, or unemployed. Only 24% were in some form of employment, and even less – around 7% indicated that they were in self-employment and 2% that they were a business owner. As one could expect from such a labor market profile, a significant share of participants was dependent on their income from others – for example 38% indicated dependence on their parents for an income.

Of those who are currently employed, or had been in employment in the recent past, the most (43%) were employed in the ICT industry – see Figure 4.

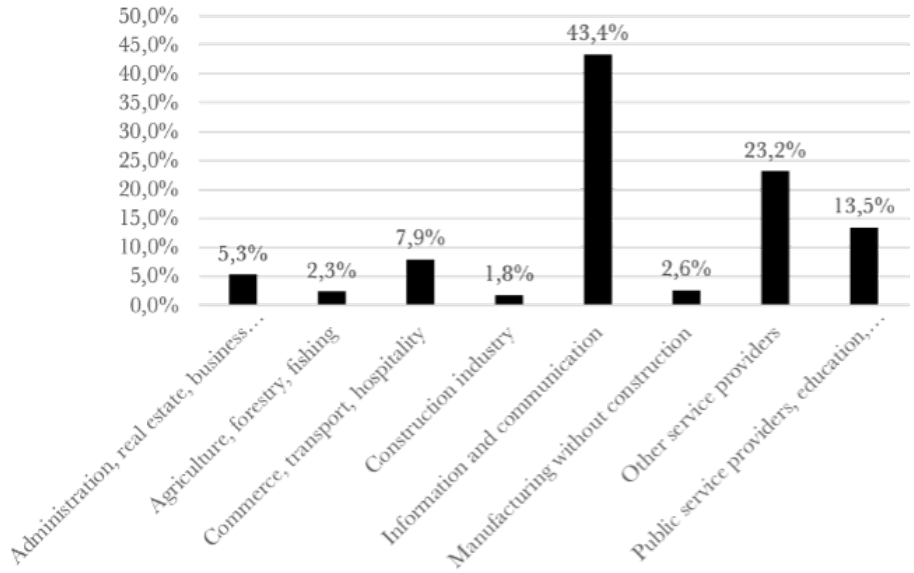
Only about half (48%) of those currently employed indicated that they are employed as a data scientist in their organizations. Why would these young and aspiring data scientists wish to take part in a data science competition? The scholarly literature has generally, in the case of crowdsourcing of innovation recognised both intrinsic and extrinsic motivations for people to take part (Tauchert et al., 2020). Pavlidou et al. (2020, p.2054-2056) surveys the literature

Figure 3: Labor market status



Source: Authors

Figure 4: Sector of current and past employment of participants

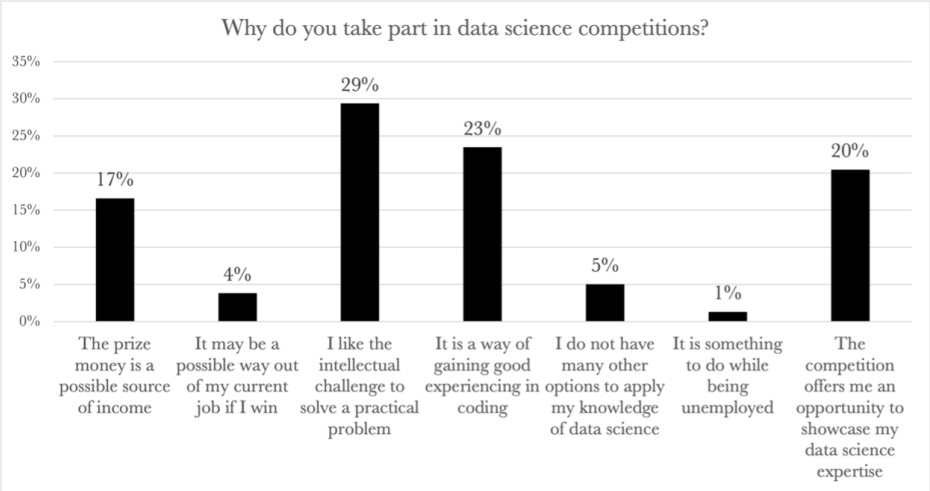


Source: Authors

on the motivation to participate in crowdsourcing of innovation activities, noting financial, career-related, individual-level, community-interaction, and altruistic motivations. In our own survey, we therefore included questions that relate to intrinsic motivations (e.g., it is an

intellectual challenge, or a good way of gaining experience) and extrinsic motivations (e.g., earning income and finding employment or promoting a career). Based on the contestants' educational and labour market status, one may suspect that extrinsic motivations, such as using the competition to improve skills and signal experience to potential employers, may dominate. Figure 5 shows their responses to the question.

Figure 5: Motivation to take part in a data science contest in Africa



Source: Authors

From Figure 5 the major reasons reported by the participants for taking part in data science competitions are due to the intellectual challenge it provided, the opportunity it provides to gain experience in coding, the opportunity to showcase their expertise, and earning income. This perhaps surprisingly then indicates that intrinsic motivations play a greater role than one could have expected *a priori*, but that indeed competitions offer an opportunity to gain both skills and employment. The latter three motivations are consistent with our expectations that data science contests are mechanisms to overcome labor market and educational obstacles in Africa – see also section 4.2 below where the perceptions on these constraints are further analyzed.

3.2.2 Who wins?

Our first aim was to capture some of the dynamics that goes into participating in and winning a data science competition involving the development of an AI model.

As was mentioned, contestants download the data for this challenge from the competition page and submit *csv* files with their models' predicted answers. After the closing date (13

September 2020) the submissions we automatically scored using a binary cross-entropy / log loss function. Based on the entries received, the following contestants were ranked in the top 10. Table 2 indicates that many contestants collaborated in teams. Four of the top 10 entries were in fact provided by teams collaborating across countries.

Table 2: Top 10 Bids for the Zimnat Contest

Rank	**Username**	**Gender**	**Country**	**Entries**	**Contests**
1	Team Super Kind Recommenders	male	Kenya	117	35
1	Team Super Kind Recommenders	male	India	44	22
1	Team Super Kind Recommenders	male	Nigeria	54	30
2	Pandas	female	Kenya	179	5
3	Team Sbernat	male	Russia	43	19
3	Team Sbernat	male	Russia	5	4
4	FC	male	UK	71	5
5	Team ARK	male	Ghana	286	0
5	Team ARK	male	Tunisia	29	27
5	Team ARK	male	Tunisia	47	38
6	johnpateha	male	Russia	43	4
7	Team DrSpark	male	Nigerian	35	30
7	Team DrSpark	male	South Africa	39	6
7	Team DrSpark	male	Nigeria	37	41
7	Team DrSpark	male	Nigeria	141	40
8	Guillaume$_{F}ilteau$	male	USA	236	1
9	KarmaML	male	Russia	231	3
10	Icfstat	male	Peru	79	4

Source: Authors' compilation based on contest data.

The top three contestants, in terms of their log loss function ranking, were (reporting only their usernames / team names) *Team Super Kind Recommenders*, with a log loss score of 0,0257 (approximately 97% accuracy) *Pandas* and *Team Sbernat*. The winner, *Team Super Kind Recommenders* was a team consisting of three data scientists from Kenya, Nigeria, and India. The runners up were respectively from Kenya (*Pandas*) and Russia (*Team Sbernat*).

Table 1 further indicates that of the top 10 entries, only one was by a female (*Pandas*, from Kenya) and that four came from Nigeria, four from Russia, and two from Tunisia, with entries from Peru. USA and UK as well. Amongst the top ten entries were one participant each from Kenya, Ghana and South Africa.

Furthermore, all the contestants in the top ten entries submitted multiple solutions – the

average for the top ten is 95 – and virtually all had taken part in previous *Zindi* data science competitions, with an average participation in 17 previous competitions.

We performed a simple OLS regression analysis to gain more insight into the determinants of winning submissions. The regression was performed on merging data from the competition submissions to the results from the survey. This resulted in data covering 251 participants who completed both the survey as well as submitted a competition entry. Note: not everyone who completed the survey submitted a competition entry, and not everyone who submitted an entry completed the questionnaire.

Of the 251 participants who completed both the survey as well as submitted a competition entry, 48% were from an African country (the majority from Nigeria, Kenya, and South Africa), 13% were female, the average age was 27 years, on average 33 submissions were entered per contestant, 19% were married, only 3% had a PhD degree, 91% had a tertiary education (bachelors, masters of PhD degree), 9% has secondary education as highest qualification. Moreover, only 36% indicated that they held a formal qualification in data science. Only 9% considered themselves to have advanced skills in the field of artificial intelligence (AI). Around 39% are seriously consideration starting up a new venture based on AI. 37% of these participants were students, and only 6% were self-employment.

To measure success or competence (performance) in the contest we used the lowest score in terms of the log loss function obtained in the *Zimnat* competition as dependent variable. The lower the score, the better the performance. As independent variables we used the number of submissions entered and the number of competitions that a contestant has taken part - as indicators respectively of learning and experience; we also used variables reflecting on the individual characteristics of contestants, such as gender, age, country of residence, education level, employment status, formal qualifications in data science, and intention to start a new venture using AI technology. The regression results (with robust standard errors in brackets) are shown in Table 3. A post-regression variance inflation factor (VIF) test indicated no significant multicollinearity between the independent variables.

In Table 3 four models are estimated: model 1 contains variables reflecting on contestants' experience, model 2 on their skills, model 3 on their labor market status. Model 4 combines all of these. Before discussing the results, it is worth pointing out that these models only explain a small proportion of the variation in scores, given the low adjusted-R^2 statistics. There are evidently unknown and unmeasured variables that these models omit.

Model 1 indicates that experience improves the performance of contestants in the *Zimnat*

Table 3: Determinants of Contest Performance – OLS Regression Results; Dependent variable : score on log loss function (robust standard errors in brackets)

Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Constant	0.26	**0.20**	**0.15**	0.09
	(0.101)	(0.04)**	(0.05)**	(0.13)
Submissions	**-0.002**			-0.002
	(0.00)**			(0.00)
Zindi contests	**-0.003**			-0.01
	(0.002)*			(0.002)
Age	0.0005			0.004
	(0.004)			(0.003)
PhD		**-0.18**		**-0.23**
		(0.04)**		(0.10)*
Formal qual		0.02		0.03
		(0.08)		(0.07)
Advanced AI		**-0.14**		-0.08
		(0.05)**		(0.05)
Student			0.04	0.03
			(0.06)	(0.06)
Self-employed			0.31	0.31
			(0.35)	(0.38)
Entrepreneurial			0.02	0.01
			(0.07)	(0.06)
Gender				-0.05
				(0.13)
Africa located				**0.16**
				(0.08)*
Diagnostics:				
Adjusted R^2	0.03	0.01	0.08	0.07
N	250	249	250	247

Source: Authors' compilation based on contest data. *Note: a ** and * indicates significance at respectively the 1% and 5% levels.*

contests. Thus, contestants with more submissions (reflecting learning by attempting more) as well as with more experience from participating in *Zindi* contests, had lower (better) scores. Age is not significant.

Model 2 indicates that skills also matter – contestants with a PhD achieved better scores, and those who rated themselves to have advanced skills in AI also performed better. Having a formal qualification in data science did not matter.

According to model 3 in Table 3, being self- employed is associated with poorer (higher)

scores. Being a student though or having plans to use AI for new venture creation, were insignificant.

Finally, in the combined model (model 4) in Table 3, we also control for gender, as well as whether a participant is resident in an African country. The results are very similar across models, only in the combined model AI expertise and self-employment status becomes insignificant in the presence of the other variables. Moreover, the gender indicator is insignificant, indicating that being female carries no penalty in terms of performance (one of the finalist in our contest is female). Finally, participants located in Africa tended to have on average poorer scores given the positive sign and significance of the indicator variable for their location.

In conclusion, what these very tentative findings show, is that experience and skills matter most in performing well in a data science contest. Having an advanced degree, and having taken part in previous contests contribute to better performance, as does submitting more entries (perhaps reflecting learning). The fact that participants in Africa seem, on average to obtain lower results may be due to relative lack of opportunities to build skills, and/or relative inexperience in data science contests. We have to stress that these are tentative results - the regressions models themselves only explain a very small proportion of the variance of the success measure, suffer from omitted variable bias, and only pertain to a single contest. Further research is needed to gather data and analyse the determinants of winning,learning and collaborating in data science contests.

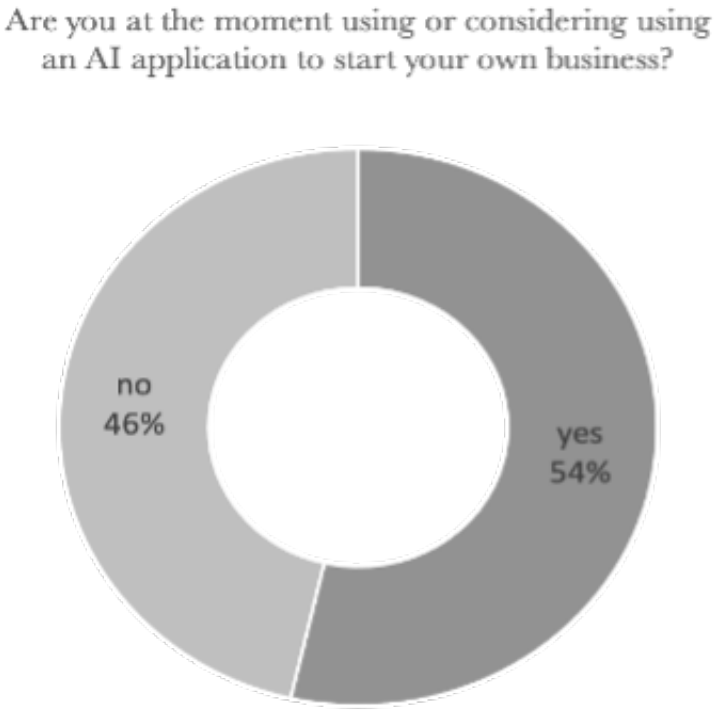
3.2.3 Perceptions on AI in Africa

Our second aim was, through the survey attached to the challenge, to obtain information on the perceptions of data scientists on AI diffusion and AI-entrepreneurship in Africa. For this we designed the short questionnaire contained in the Appendix and discussed in section 3.3.

From the responses we found the following. First, a slight majority (54%) of respondents indicated that they are now using or considering using an AI application to start their own business (see Figure 6). This suggest that there is a significant interest in the entrepreneurial potential of using AI amongst data scientists working in Africa.

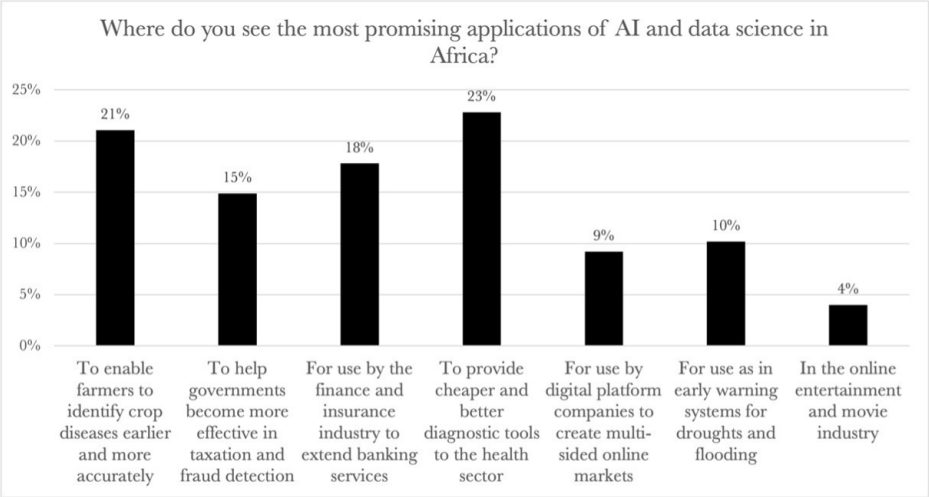
The most promising applications of AI in Africa, according to the participants, are in healthcare, farming and the finance industry (Figure 7).

Figure 6: Entrepreneurial intentions of participants



Source: Authors

Figure 7: Perceptions of Most Promising AI Applications in Africa

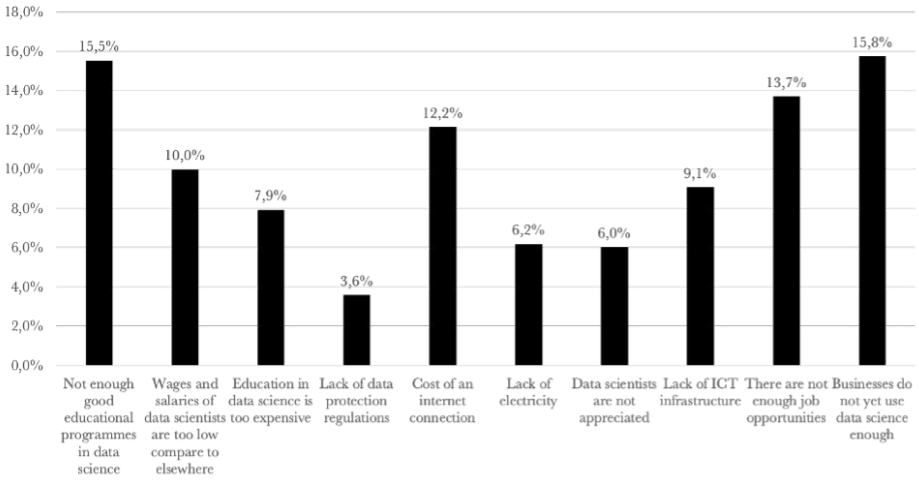


Source: Authors

Despite these intentions, and perceptions of promising applications the participants in our survey and competition do however perceive significant obstacles. These include both obstacles to be a data scientist in Africa (Figure 7), and obstacles in the adoption of AI by

business firms (Figure 8).

Figure 8: Perceived obstacles to be a data scientist in Africa



Source: Authors

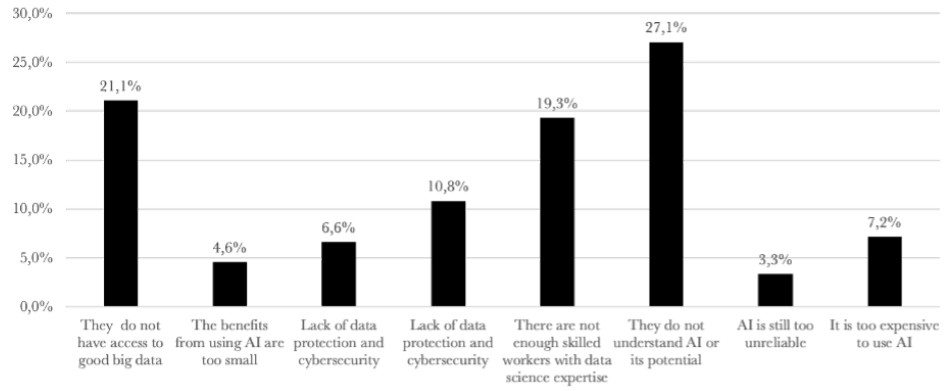
Figure 8 shows that the most serious obstacles to be a data scientist in Africa are perceived to be that businesses do not yet use data science enough (i.e., low adoption of data-driven decision making by businesses in Africa), as well as a lack of education programs in data science. Related to the first is that around 13% of participants rated the lack of sufficient job opportunities for data scientists as a problem, and 12% the cost of an internet connection. The lack of education is also compounded by the expense of education in data science – 10% of participants rated this is as the most serious constraint. It can also be seen that lack of data protection regulations and electricity are not seen as the most serious constraints.

Figure 9 indicates that participants in our data science contest regard the most significant obstacles in the adoption of AI by businesses in Africa as a lack of understanding of AI and its potential (27%), lack of access to good big data (21%) and lack of skilled workers with data science expertise (19%). The obstacles to the adoption of AI may explain why, in the perception of the contest participants, the extent of AI adoption in their countries were so low, as Figure 10 shows.

Figure 10 shows that 76% of respondents judged that fewer than 30% of businesses in their countries are currently using AI and/or data-driven technology to optimize business process and customer engagement.

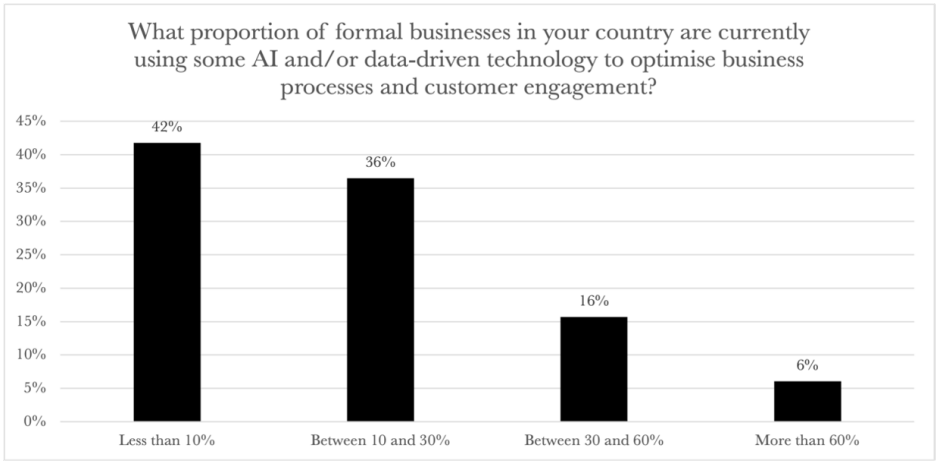
Given the importance of lack of education and skills as an obstacle in this slow diffusion of AI, and the fact that most of the participants did not have a formal data science qualification, it

Figure 9: Perceived obstacles to AI adoption by business firms



Source: Authors

Figure 10: Perceived extent of diffusion of AI in Africa

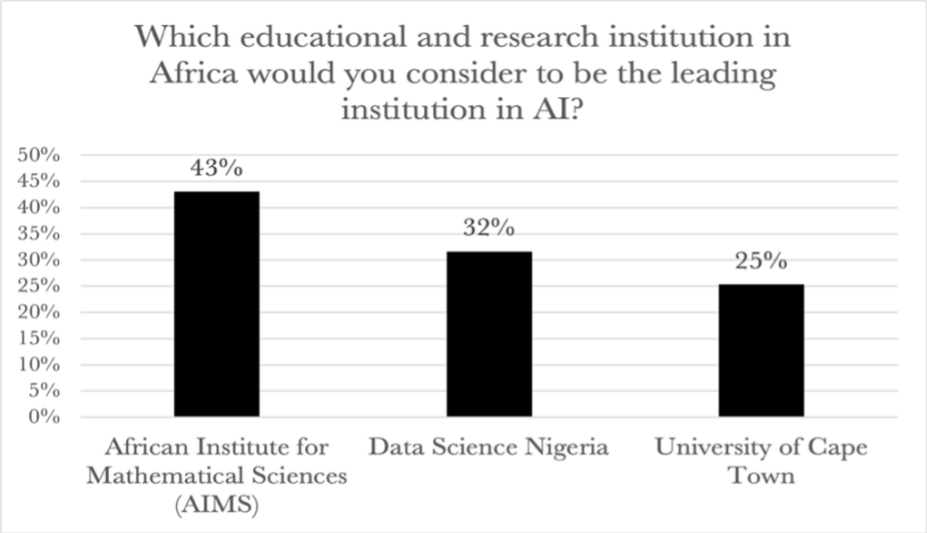


Source: Authors

is interesting to note that participants seemed very aware of providers of AI and data science education in Africa. However, only a handful of institutions were recognised or acknowledged my participants for their expertise in AI. The leading providers of AI education according to the responses are depicted in Figure 11.

Figure 11 indicates that the three leading AI education providers in Africa according to contest participants were the African Institute for Mathematical Sciences, Data Science Nigeria, and the University of Cape Town (UCT). Generally speaking, apart from the recognition of UCT (and some participants also mentioned Pretoria University and Stellenbosch University) what is noticeable is the absence of African universities from being prominent in

Figure 11: Perceptions of leading AI education providers in Africa

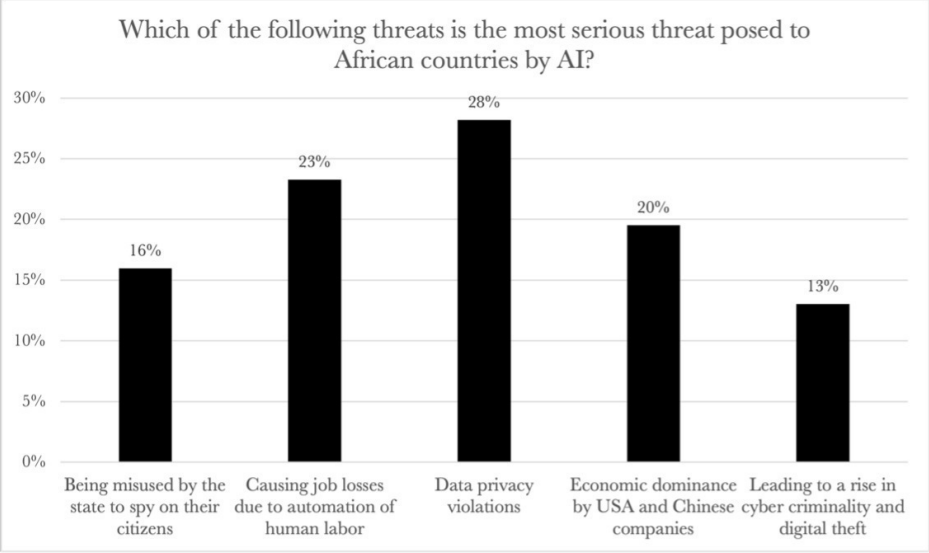


Source: Authors

providing AI education.

Finally, contest participants were asked which threat posed by AI they consider to be the most serious. The responses are shown in Figure 12.

Figure 12: Perceptions of Threats Posed by AI in Africa



Source: Authors

Thus, as shown in Figure 12, participants consider data privacy violations and job losses from automation to be the two most important or serious threats posed by AI. Around 20% of participants also feared the economic dominance of USA and Chinese digital platform

companies, and 16% feared that misuse by states to spy on their citizens posed the greatest threat.

4 Concluding Remarks

Although still facing a noticeable digital divide, and although rates of participation in the global knowledge economy is still relatively low, recent years have seen positive signs of progress and catch-up in Africa as far as digital technologies are concerned – beyond the use of mobile phones. These include the rise of indigenous digital platforms, the expansion of tech hubs and the growing amount of venture capital funding for tech start-ups, many of whom adopt AI. One notable area of progress has been in the crowdsourcing of innovation. In crowdsourcing innovation through intermediary digital platforms, specifically contest-based platforms, African-based initiatives have become internationally notable. In 2020, out of 107 data science contests held on 24 platforms worldwide, the third most (13%) was on an African-based platform (ML Contests, 2021; Olaleye, 2021).

There are no studies yet, to the best of our knowledge, that focuses on the case of data science crowdsourcing contests in Africa. This paper is an attempt to fill this gap. Specifically, we were interested in whether data science contests can facilitate the diffusion and adoption of AI in Africa through providing a conduit for learning, interaction, and experimentation in Machine Learning (ML).

This necessitated answering several questions: who takes part in these contests and why? Who are most likely to win? What are contestants' entrepreneurial aspirations in deploying AI and what are the obstacles they perceive to the greater diffusion of AI in Africa?

To answer these questions, we designed and issued a ML challenge on Africa's largest data science contest platform, *Zindi*, in 2020 – contestants were challenged to submit a ML (*recommender)* model to predict sales of insurance products for Zimnat, the largest insurance company in Zimbabwe. In total, 614 contestants from across the continent and further afield submitted 15,832 entries and 559 responded to the accompanying survey. From the survey and the contest, we found the following answers to the questioned posed.

Firstly, *who participates in data science contests?* The answer was that it is mainly young people, and mainly men. The median age of participants were 25 years and 83% were male. Only 35% of the participants indicated that they had a formal qualification in data science

and only 9% rated their own AI-expertise as advanced. The largest category of participants were students (40%). Thus, the profile of who participated in our data science competition is of relatively young people, with intermediate to beginner level skills in AI and mostly not formally qualified in data science.

Secondly, *why take part in these contests?* We found that participation was driven by the intellectual challenge it provided, the opportunity it provides to gain experience in coding, the opportunity to showcase expertise, and also to earn an income. This surprisingly indicates that intrinsic motivations play a greater role than one could have expected *a priori*, but that indeed contests offer an opportunity to gain both skills and employment.

Thirdly: *who wins data science contests?* We found that that experience and skills matter most in performing well in our contest. Having an advanced degree and having taken part in previous contests contribute to better performance, as does submitting more entries. We also found that participants based in Africa seems, on average, to obtain lower results. This may reflect relative lack of opportunities to build skills, or relative inexperience in taking part in data science contests.

Fourthly, we asked the contestants about their perceptions about AI in Africa, including their entrepreneurial aspirations and obstacles to AI adoption. A slight majority (54%) of respondents indicated that they are now using or considering using an AI application to start their own business. This suggest that there is a non-negligible interest in the entrepreneurial potential of using AI amongst data scientists working in Africa. The most promising applications of AI in Africa, according to the participants, are in healthcare, farming and the finance industry.

As far as the adoption of AI in Africa is concerned, 76% of respondents judged that fewer than 30% of businesses in their countries are currently using AI and/or data-driven technology to optimize business process and customer engagement. The contestants regarded the most significant obstacles in this adoption as a lack of understanding of AI and its potential (27%), lack of access to good big data (21%) and lack of skilled workers with data science expertise (19%). For data scientists working in Africa, they perceived the most serious obstacles to be that businesses do not yet use data science enough (i.e., low adoption of data-driven decision making by businesses in Africa), as well as a lack of education programs in data science.

Finally, participants consider data privacy violations and job losses from automation to be the two most important or serious threats posed by AI. Around 20% of participants also feared the economic dominance of USA and Chinese digital platform companies, and 16%

feared that misuse by states to spy on their citizens posed the greatest threat.

Overall, our empirical results suggest, in confirmation of our hypothesis, that data science contests via an intermediary digital platform in Africa may reflect labor market and educational obstacles in Africa, which results in low adoption of AI and relative low demand on an individual country level for the expensive skills of data scientists. As such, the crowdsourcing of AI via data contest platforms offers a way around these obstacles, and a potential mechanism to raise both the adoption of AI amongst businesses (supporting adoption from the demand side) as well as the building up of AI skills (supporting adoption from the supply side). As such the future of crowdsourcing platforms is pertinent. According toSegev (2020, p.253) the survival of platforms for crowdsourcing contests 'depends on the rules they use and their ability to attract good contestants' as well as to draw in many crowd-seekers. The research we reported in this paper is a point of departure for platforms to meet these requirements for survival. But, given that we studied but a single contest on a single platform, our results should be seen as provisional. Further research is needed not only to study more contests across multiple platforms and over time, but also to expand the scope of questions to investigate. In this respect we have three recommendations.

First, we agree with Pavlidou et al. (2020, p.2057) that 'the effect of earning on the crowd's personal and professional development has not been explored' and that hence more research is needed to determine whether and how participants' labor market participation and career prospects are affected by data science contests.

Second, future research could investigate whether and how firms based in Africa benefit from data science contests – in other words, does this platform for the crowdsourcing of data-related innovations help them significantly to perform better? How effective are the solutions that they obtain through data science competitions? What novel contributions to the expansion of data science and AI have data science contests stimulated in Africa? Data science contest platforms are a hybrid of contest-based and collaboration-based crowdsourcing models, in that there is competition, but teamwork (collaboration) is also allowed. The question with respect to the diffusion and adoption of AI in Africa is, whether this organizational model for intermediary crowdsourcing is optimal, and whether perhaps evolution of the model towards more collaboration will not be required, particularly considering the multi-disciplinary challenges that AI by its nature, requires.

Finally, it is not only the sustainability or survival of data science platform contests that are pertinent, but also their scalability. Kohler (2018) for instance pointed out that '*scale can*

enhance the effectiveness of crowdsourcing platforms because of the interdependencies among participants.' Achieving scale and sustainability are imperatives for data science platform business models in Africa if they are to play a continuing an enlarged role in the diffusion of new technologies, such as AI throughout the continent.

Acknowledgements

We are grateful to the participants of various workshops of the Research Unit for the Diffusion of Quality AI at Paderborn University for their comments on earlier versions of this paper. The financial assistance of the Volkswagen Stiftung, through planning grant AZ 97042 from their project on Artificial Intelligence and the Society of the Future is gratefully acknowledged. The usual disclaimer applies.

Appendix
Zimnat Insurance Recommendation Challenge: Questionnaire

PARTICIPANT QUESTIONNAIRE

As per the competition rules, all participants are required to complete the following questionnaire. Thank you for your time. This questionnaire should not take more than 10 minutes of your time.

Name: [Name]
Zindi Username: [username]

SECTION 1: Personal Information

Q1.1: Gender [Female / Male / Other]

Q1.2: Age (in completed years) [Number]

Q1.3: Which country do you currently live in? [Country]

Q1.4: Nationality [nationality]

Q1.5: Location: do you live in a rural, semi-/peri-urban or urban area? [Rural / peri-urban / urban]

Q1.6: Marital status

- Single (never married)
- Married or in a domestic partnership
- Divorced
- Separated
- Widowed

Q1.7: What is your highest educational qualification? [PhD, Masters, Bachelors, Secondary Education, Primary, None]

Q1.8: Do you hold a formal qualification in data science? [yes/ no]

Q1.9: Did you study data science outside your country of birth? [yes/ no]

Q1.10: How many previous Zindi data science competitions did you take part in? [number]

Q1.11: How many Kaggle data science competition have you taken part in? [number]

Q1.12: How did you find out about this competition? [Zindi web page / Social media / Friends/colleagues]

SECTION 2: Labor Market Information

Q2.1: Which of the following best describes your current situation:

- Wage worker
- Daily labourer

- Civil servant / public servant
- Self-employed
- Business owner
- Farmer
- Other type of employment
- Unemployed (or furloughed)
- Student
- Housewife/Househusband/parental leave
- Unable to work due to disability
- Retired

Q2.2: The main provider of income in my household is:

- Myself
- My spouse / partner
- Both me and my spouse / partner, equally
- My parent(s)
- My child(ren)
- Other
- I prefer not to answer

Q2.3: [if Q1 == Unemployed or furloughed]: Since when?

- In the last month
- In the last three months

- In the last six months
- In the last year
- Longer ago than one year

Q2.4: [if Q1 == working or unemployed]:
In which sector is your current / was your most recent, main employment (that is, the job that is / was responsible for most of your income):

- Agriculture, forestry, fishing
- Manufacturing without construction
- Construction industry
- Commerce, transport, hospitality
- Information and communication
- Administration, real estate, business service providers
- Public service providers, education, health
- Other service providers

Q2.5: [if Q1 == working or unemployed]: Do you currently have a job as a data scientist in the organization? [yes/ no]

Q2.6: [if Q1 == self-employed or business owner]: Is your current main business or self-employment activities the selling of data science services? [yes/ no]

Q2.7: Does your business sell a product or service online ? [yes/ no]

Q2.8: How much has your monthly net income changed since the start of the COVID-19 crisis?

- Drastically decreased
- Moderately decreased
- Did not change
- Moderately increased
- Drastically increased
- I prefer not to answer/doesn't apply

Q2.9: Why do you take part in data science competitions? Choose the two that best apply to you:

- The prize money is a possible source of income
- I like the intellectual challenge to solve a practical problem
- The competition offers me an opportunity to showcase my data science expertise
- It is a way of gaining good experiencing in coding
- I do not have many other options to apply my knowledge of data science
- It is something to do while being unemployed
- It may be a possible way out of my current job if I win

SECTION 3: Artificial Intelligence (AI) Perspectives

Q3.1: How would you rate your own expertise in Artificial Intelligence (AI):

- Beginner
- Intermediate
- Advanced

Q3.2: What are the three most serious obstacles for you as a data scientist working in Africa?

- There are not enough job opportunities
- Data scientists are not appreciated
- Lack of data protection regulations
- Lack of ICT infrastructure
- Lack of electricity
- Cost of an internet connection
- Not enough good educational programmes in data science
- Education in data science is too expensive
- Wages and salaries of data scientists are too low compare to elsewhere
- Businesses do not yet use data science enough

Q3.3: Which educational and research institution in Africa would you consider to be the leading institution in AI? [name of institution]

Q3.4: Are you at the moment using or considering using an AI application to start your own business? [yes/ no]

Q3.5: What proportion of formal businesses in your country are currently using some AI and/or data-driven technology to optimise business processes and customer engagement?

- Less than 10
- Between 10 and 30
- Between 30 and 60
- More than 60

Q3:6: What are the 2 most important reasons why businesses in your country do not use AI and other data-science methods more intensively?

- They don't understand AI or its potential
- They don't have access to good big data
- It is too expensive to use AI
- There are not enough skilled workers with data science expertise
- The benefits from using AI are too small
- Lack of data protection and cybersecurity
- Lack of ICT infrastructure and broadband
- AI is still too unreliable

Q3.7: Where do you see the most promising applications of AI and data science in Africa? Choose the two most significant:

- For use by digital platform companies to create multi-sided online markets
- For use by the finance and insurance industry to extend banking services
- To help governments become more effective in taxation and fraud detection
- To enable farmers to identify crop diseases earlier and more accurately
- To provide cheaper and better diagnostic tools to the health sector
- For use as in early warning systems for droughts and flooding
- In the online entertainment and movie industry

Q3.8: What are the most serious threats posed to African countries by AI?

- Being misused by the state to spy on their citizens

- Leading to a rise in cyber criminality and digital theft
- Causing job losses due to automation of human labor
- Data privacy violations
- Economic dominance by USA and Chinese companies

THE END

References

Abramovitz, M. (1986). Catching Up, Forging Ahead, and Falling Behind. *Journal of Economic History*, 46(2), 385–406.

Acar, O. (2019). Motivations and Solution Appropriateness in Crowdsourcing Challenges for Innovation. *Research Policy*, 48(8), 103716.

Afuah, A. & Tucci, C. (2012). Crowdsourcing as a Solution to Distant Search. *Academy of Management Review*, 37(3), 355–375.

Anwar, M. & Graham, M. (2020). Digital Labour at Economic Margins: African Workers and the Global Information Economy. *Review of African Political Economy*, 47 (163)(163), 95–105.

Ayaburi, E., Lee, J., & Maasberg, M. (2020). Understanding Crowdsourcing Contest Fitness Strategic Decision Factors and Performance: An Expectation-Confirmation Theory Perspective. *Information Systems Frontiers*, 22, 12271240.

Bojer, C. & Meldgaard, J. (2021). Kaggle Forecasting Competitions: An Overlooked Learning opportunity. *International Journal of Forecasting*, 37(2), 587–603.

Brackbill, D. & Centola, D. (2020). Impact of Network Structure on Collective Learning: An Experimental Study in a Data Science Competition. *PLoS ONE*, 15(9), e0237978.

Brynjolfsson, E., Rock, D., & Syverson, C. (2017). Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. *NBER Working Paper no. 24001. National Bureau for Economic Research.*

Carpenter, J. (2011). May the Best Analyst Win. *Science*, 331(6018), 698–699.

Chawla, S., Hartline, J., & Sivan, B. (2015). Optimal Crowdsourcing Contests. *Games and Economic Behavior*, 13, 80–96.

Cisse, M. (2018). Look to Africa to Advance Artificial Intelligence. *Nature*, 562, 461.

Corchón, L. (2007). The Theory of Contests: A Survey. *Review of Economic Design*, 11, 69–100.

Crafts, N. (2020). AI as a GPT: An Historical Perspective. *Paper presented at the Bank of England Conference on "The Impact of Machine Learning and AI on the UK Economy" March 25.*

European Commission (2020). European Enterprise Survey on the Use of Technologies Based on Artificial Intelligence. *European Commission DG Communications Networks, Content & Technology. Brussels*.

Farzin, Y., Huismann, K., & Kort, P. (1998). Optimal Timing of Technology Adoption. *Journal of Dynamics & Control*, 22, 779– 799.

Felin, T., Lakhani, K., & Tushman, M. (2017). Firms, Crowds, and Innovation. *Strategic Organization*, 15(2), 119–140.

Friederici, N., Ojanperä, S., & Graham, M. (2017). The Impact of Connectivity in Africa: Grand Visions and the Mirage of Inclusive Digital Development. *Electronic Journal of Information Systems in Developing Countries*, 79(2), 1–20.

Gatzioura, A., Sánchez-Marr, M., , & Gibert, K. (2019). A Hybrid Recommender System to Improve Circular Economy in Industrial Symbiotic Networks. *Energies*, 12, 3546.

Ghezzi, A., Gabelloni, D., Martini, A., & Natalicchio, A. (2018). Crowdsourcing: A Review and Suggestions for Future Research. *International Journal of Management Reviews*, 20(2), 343–363.

Godoy, D. (2018). Understanding Binary Cross-Entropy Log Loss: A Visual Explanation. *Towards Data Science*, 21 November.

Goldfarb, A., Taska, B., & Teodoridis, F. (2021). Could Machine Learning be a General Purpose Technology? A Comparison of Emerging Technologies Using Data from Online Job Postings. *Available at SSRN: http://dx.doi.org/10.2139/ssrn.3468822*.

Graham, M. (2019). Digital Economies at Global Margins. *Cambridge MA: MIT Press*.

Graham, M., Ojanperä, S., Anwar, M., & Friederici, N. (2017). Digital Connectivity and African Knowledge Economies. *Questions de Communication*, 32, 345–360.

Griliches, Z. (1957). Hybrid Corn: An Exploration in the Economics of Technological Change. *Econometrica*, 25(4), 501–522.

Gross, D. (2020). Creativity Under Fire: The Effects of Competition on Creative Production. *The Review of Economics and Statistics*, 102(3), 583–599.

Gwagwa, A., Kraemer-Mbula, E., Rizk, N., Rutenberg, I., & de Beer, J. (2020). Artificial Intelligence (AI) Deployments in Africa: Benefits, Challenges and Policy dimensions. *The African Journal of Information and Communication*, 26, 1–28.

Humphries, G., Che-Castaldo, C., Bull, J., Lipstein, G., Ravia, A., Carrión, B., Bolton, T., Ganguly, A., & Lynch, H. (2018). Predicting the Future is Hard and other Lessons from a Population Time Series Data Science Competition. *Ecological Informatics*, 48, 1–11.

Irani, L. (2019). Hackathons and the Cultivation of Platform Dependence. *In Graham, M. ed. 2019. Digital Economies at Global Margins. Cambridge MA: MIT Press. Pp. 223-249.*

Jovanovic, B. & Rob, R. (1989). The Growth and Diffusion of Knowledge. *The Review of Economic Studies*, 56(4), 569–582.

Keller, W. (2004). International Technology Diffusion. *Journal of Economic Literature*, (42), 752–782.

Kohler, T. (2018). How to Scale Crowdsourcing Platforms. *California Management Review*, 60(2), 98–121.

Körpeoglu, E. & Cho, S.-H. (2018). Incentives in Contests with Heterogeneous Solvers. *Management Science*, 64(6), 2709–2715.

Kumar, S. & Singh, B. (2019). Barriers to the International Diffusion of Technological Innovations. *Economic Modelling*, 82, 74–86.

Lüttgens, D., Pollok, P., Antons, D., & Piller, F. (2014). Wisdom of the Crowd and Capabilities of a Few: Internal Success Factors of Crowdsourcing for Innovation. *Journal of Business Economics*, 84(3), 339–374.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 Time Series and 61 Forecasting methods. *International Journal of Forecasting*, 36(1), 54–74.

Mao, K., Capra, L., Harman, M., & Jia, Y. (2017). A Survey of the Use of Crowdsourcing in Software Engineering. *Journal of Systems and Software*, 126, 57–84.

Marconi, S., Graves, S., Gong, D., Nia, M., Bras, M. L., Dorr, B., Fontana, P., Gearhart, J., Greenberg, C., Harris, D., Kumar, S., Nishant, A., Prarabdh, J., Rege, S., Bohlman, S., White, E., & Wang, D. (2019). A Data Science Challenge for Converting Airborne Remote Sensing Data into Ecological Information. *PeerJ*, 6.

ML Contests (2021). Winning at Competitive ML in 2021: An Analysis of Over 100 ML Contest Winners. *ML Contests Blog*, 13 April.

Moldovanu, B. & Sela, A. (2001). The Optimal Allocation of Prizes in Contests. *American Economic Review*, 91, 542–558.

Moldovanu, B. & Sela, A. (2006). Contest Architecture. *Journal of Economic Theory*, 126(1), 70–96.

Naik, N. & Purohit, S. (2017). Comparative Study of Binary Classification Methods to Analyze a Massive Dataset on Virtual Machine. *Procedia Computer Science*, 112, 1863–1870.

Naudé, W. (2018). Brilliant Technologies and Brave Entrepreneurs: A New Narrative for African Manufacturing. *Journal of International Affairs*, 72(1), 143–158.

Naudé, W. & Vinuesa, R. (2021). Data Deprivations, Data Gaps and Digital Divides: Lessons from the COVID-19 Pandemic. *Big Data & Society*, DOI: 10.1177/20539517211025545.

Ndung'u, N. & Signé, L. (2020). The Fourth Industrial Revolution and Digitization will Transform Africa into a Global Powerhouse. *In Coulibay, B.S. and Golubski, C. (eds). Foresight Africa : Top priorities for the continent 2020-2030. Washington DC: The Brookings Institution. Chapter 5, pp. 60-73.*

Neo, B. (2019). 11 Data Science Competitions for you to Hone your Skills for 2020. *Towards Data Science*, 2 December.

Ojanperä, S., Graham, M., & Zook, M. (2019). The Digital Knowledge Economy Index: Mapping Content Production. *The Journal of Development Studies*, 55(12), 2626– 2643.

Olaleye, E. (2021). How To Win Any ML Contest. *Towards Data Science*, 13 April.

Ortner, A. (2020). Top 10 Binary Classification Algorithms - a Beginners Guide. *Medium*, 28 May.

Pavlidou, I., Papagiannidis, S., & Tsui, E. (2020). Crowdsourcing: A Systematic Review of the Literature using Text Mining. *Industrial Management & Data Systems*, 120(11), 2041–2065.

Piller, F. & Walcher, D. (2006). Tookits for Idea Competitions: A Novel Method to Integrate Users in New Product Development. *R/&D Management*, 36(3), 307–318.

Radhakrishnan, J. & Chattopadhyay, M. (2020). Determinants and Barriers of Artificial Intelligence Adoption A Literature Review. *In: Sharma S.K., Dwivedi Y.K., Metri B., Rana N.P. (eds) Re-imagining Diffusion and Adoption of Information Technology and*

Systems: A Continuing Conversation. TDIT 2020. IFIP Advances in Information and Communication Technology, vol 617. Springer, Cham.

Segev, E. (2020). Crowdsourcing Contests. *European Journal of Operational Research*, 281(2), 241–255.

Taieb, S. & Hyndman, R. (2014). A Gradient Boosting Approach to the Kaggle Load Forecasting Competition. *International Journal of Forecasting*, 30, 382–394.

Tauchert, C., Buxmann, P., & Lambinus, J. (2020). Crowdsourcing Data Science: A Qualitative Analysis of Organizations' Usage of Kaggle Competitions. *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

Trajtenberg, M. (2018). AI as the Next GPT: A Political-Economy Perspective. *NBER Working Paper no. 24245. National Bureau for Economic Research*.

UNCTAD (2021). Technology and Innovation Report 2021. *United Nations Conference on Trade and Development. New York*.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S., Tegmark, M., & Nerini, F. F. (2020). The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature Communications*, 11(233).

Zhao, Y. & Zhu, Q. (2014). Evaluation on Crowdsourcing Research: Current Status and Future Direction. *Information System Frontiers*, 16, 417–434.

Zolas, N., Kroff, Z., Brynjolfsson, E., McElheran, K., Beede, D. N., Buffington, C., Goldschlag, N., Foste, L., & Dinlersoz, E. (2020). Advanced Technologies Adoption and Use by U.S. Firms: Evidence from the Annual Business Survey. *NBER Working Paper No. 28290, National Bureau of Economic Research*.