LESSONS LEARNED

FROM LACUNA FUND'S FIRST YEAR OF FUNDING

October 2021

Lacuna Fund

Our voice on data

LACUNAFUND.ORG

ABOUT LACUNA FUND

Lacuna Fund is a collaborative effort dedicated to providing data scientists, researchers, and social entrepreneurs in low- and middle-income contexts globally with the resources they need to produce machine learning datasets that address urgent problems in their communities. Co-founded by The Rockefeller Foundation, Google.org, Canada's IDRC, and GIZ on behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ), Lacuna Fund launched in July of 2020 with a pooled fund of $4 million to support the creation, expansion, and maintenance of datasets used for training or evaluation of machine learning models.

ABOUT MERIDIAN INSTITUTE

Meridian Institute serves as the secretariat and fiscal sponsor for Lacuna Fund. Meridian is a mission-driven nonprofit. We help our partners build understanding, guide collaboration, and develop and implement solutions to complicated, often controversial problems—big and small, global and local.

AUTHORS AND REVIEWERS

Seth Blum, Jennifer Pratt Miles, Frances Burton (design)
Thank you to the Lacuna Fund Steering Committee, Technical Advisory Panel members, and the Open4Good Alliance for their thoughtful review of and comments on this report.

ACKNOWLEDGEMENTS AND DISCLAIMERS

This work was carried out with the aid of a grant from the International Development Research Centre, Ottawa, Canada. Lacuna Fund's activities are also supported by several philanthropic partners. The views expressed herein do not necessarily represent those of Lacuna Fund's contributors, Steering Committee members, or IDRC or its Board of Governors.

Read more about Lacuna Fund's funders.

PHOTO CREDITS

Cover image and page 17: USGS/Landsat
Page 5: courtesy of Machine Learning Datasets for Crop Pest and Disease Diagnosis based on Crop Imagery and Spectrometry Data Agriculture project team
Page 6: courtesy of Decision Support Tool for Community-led Land Use Plans Agriculture project
Page 10: courtesy of Building an Annotated Spoken Corpus for Igbo NLP Tasks Language project
All other images used under Unsplash License.

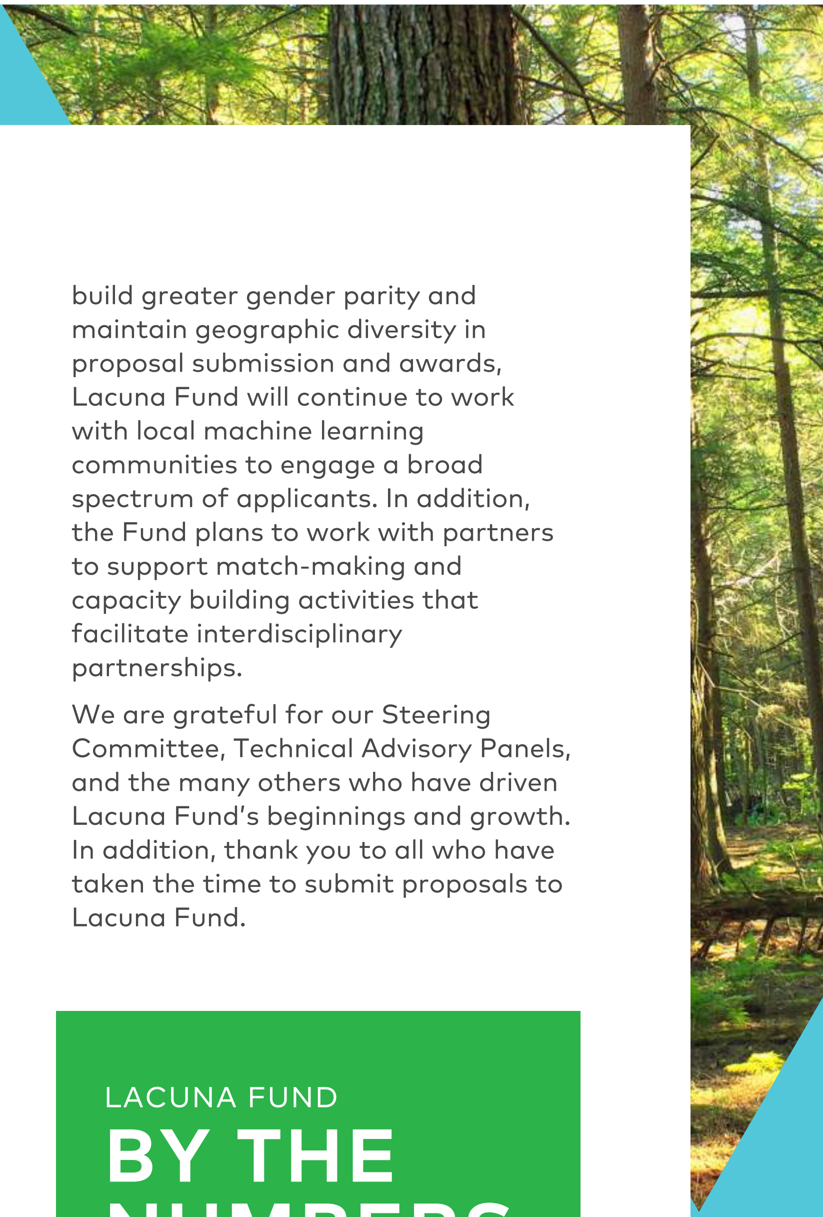
LESSONS LEARNED

EXECUTIVE SUMMARY

This report summarizes outcomes and learnings from Lacuna Fund's first two calls for proposals with an eye towards providing guidance to future applicants and insight to the field of machine learning and artificial intelligence for social good. Providing information to applicants and our broader networks on our funding processes is a key part of ensuring an equitable proposal process that generates impactful datasets and outcomes in line with Lacuna Fund's principles.

In 2020, Lacuna Fund's agriculture and language funding processes received significant interest from across the African continent and beyond. Key observations from Lacuna Fund's first two calls for proposals include:

- Lacuna Fund received proposals from 121 teams and 67 teams respectively in response to the Agriculture and Language calls, indicating a need and interest in making machine learning datasets more representative and equitable.
- The strongest proposals emerged from engaged communities of researchers and interdisciplinary partnerships of both machine learning and domain experts. Projects selected for funding demonstrated clear objectives and use cases, the qualifications necessary to develop the proposed datasets, engagement of local actors in all aspects of the project, innovative and feasible approaches, and well justified budgets and timelines.
- In agriculture, proposals reflected key needs and capacity in the space: A vast majority addressed crop type, field boundary identification, and yield estimation use cases, with smaller but strong subsets of proposals addressing other use cases in crop and animal agriculture.
- In language, proposals addressed a range of use cases. A smaller subset of proposals addressed domain specific and multimodal applications.
- Requiring that lead applicants be headquartered in Africa or have a substantial partnership with organizations headquartered in Africa contributed to significant resources being directed to the target audience: data scientists and researchers in underserved communities.
- Proposals represented a diverse geographic distribution, coming from East, West, Central, and Southern Africa. The greatest number of agricultural proposals came from East Africa, whereas the greatest number of language proposals were submitted from West Africa.



- The majority of proposals submitted were led by men. This may in part be a reflection of underlying trends in the field. Female early career project leaders reflected a slight majority over male, which could signal a generational shift. Projects selected for funding feature greater gender balance, with 6 being led by female data scientists and 10 being led by male data scientists.
- Technical Advisory Panel composition was key to an effective proposal funding decision making process. Having TAP members with both domain and cross-cutting expertise enabled them to authoritatively evaluate proposals. Including TAP members from the focus geography provided knowledge of local context and needs.
- While the creation, expansion, and maintenance of datasets remains the explicit focus, the Secretariat and partners have observed the need to address issues such as interoperability, accessibility, and power dynamics in order to fully realize the benefits of the datasets created with support from Lacuna Fund.

Sixteen funded projects led by dozens of institutions across the African continent are now working to create, expand, and maintain impactful datasets for the training and evaluation of machine learning models. Lacuna Fund looks forward to continuing to work in the agriculture and language domains, as well as expanding to new domains and geographies in 2021 and beyond. To build greater gender parity and maintain geographic diversity in proposal submission and awards, Lacuna Fund will continue to work with local machine learning communities to engage a broad spectrum of applicants. In addition, the Fund plans to work with partners to support match-making and capacity building activities that facilitate interdisciplinary partnerships.

We are grateful for our Steering Committee, Technical Advisory Panels, and the many others who have driven Lacuna Fund's beginnings and growth. In addition, thank you to all who have taken the time to submit proposals to Lacuna Fund.

LACUNA FUND

BY THE NUMBERS

16

FUNDED PROJECTS

150+

PROPOSALS RECEIVED

4M+

POOLED FUND

LACUNA FUND:

WHAT, WHY, AND HOW

Lacuna Fund: Our Voice on Data is a collaborative effort dedicated to providing data scientists, researchers, and social entrepreneurs in low- and middle-income contexts globally with the resources they need to produce datasets for the training and evaluation of machine learning models that address urgent problems in their communities.

Lacuna Fund supports dataset creation, expansion, and maintenance through open funding processes. In 2020, Lacuna Fund issued calls for proposals for datasets in agriculture and language technologies in sub-Saharan Africa, and a call for expressions of interest for datasets to address inequities in health outcomes and health. In 2021, Lacuna Fund will issue funding calls in agriculture, health, and language.

Lacuna Fund's governance is structured to ensure that domain experts and communities the Fund serves guide our decisions and grantmaking.

- Lacuna Fund is governed by a representative Steering Committee comprised of 5-9 members, with a balance of perspectives that serves the principles of Lacuna Fund. The committee provides strategic direction and oversight for the Fund, working to ensure its impact and growth. The gender composition of Lacuna Fund's 2020 Steering Committee was 40% female, 60% male.
- Domain-specific Technical Advisory Panels (TAPs) provide technical guidance for the Fund. They shape the focus of requests for proposals within each domain area, select proposals for funding, and distill learnings from the funding process. See the following sections for additional information on the composition of 2020 TAPs.

Lacuna Fund began as a funder collaborative between The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre, with individual calls for proposals in 2020 also supported by the German development agency GIZ on behalf of the Federal Ministry for Economic Cooperation and Development (BMZ).

The Fund has since evolved into a multi-stakeholder engagement composed of technical experts, thought leaders, local beneficiaries, and end users. Collectively, we are committed to creating and mobilizing datasets that solve urgent local problems and lead to a step change in machine learning's potential worldwide.

Learn more about Lacuna Fund's contributors and governance.

2020

AGRICULTURE RFP

Lacuna Fund's first Request for Proposals (RFP) for labeled datasets in sub-Saharan Africa was issued in July 2020 along with the public launch of Lacuna Fund.

WHAT DID THE RFP ASK FOR?

Lacuna Fund's first RFP aimed to address a lack of ground truth labels, as well as a lack of datasets to address unique challenges in mapping smallholder farms in sub-Saharan Africa. This gap hinders further progress toward beneficial ML applications that can benefit underserved populations worldwide.

The Technical Advisory Panel (TAP) chose to keep this RFP broad. The RFP requested datasets related to a variety of use cases, including field boundary identification, crop type classification, yield estimation, pastoral migratory patterns and agricultural practices, and crop and animal pests and disease. See the full set of use cases in the RFP document.

The Agriculture TAP consisted of domain experts in machine learning across crop and animal agriculture, as well as business owners, government experts, and data users from Eastern, Western, and Southern Africa, Europe, and the U.S.

Thank you to our Technical Advisory Panel for their guidance in shaping the Fund's first RFP and their insight in selecting proposals.

WHO RESPONDED?

Use Cases:
Almost all eligible proposals included a use case related to either field boundary identification, crop type classification, or yield estimation ("crop" use cases). A significant but far fewer number of proposals included use cases related to soil. A smaller but robust set of proposals focused on animal agriculture. In some cases, proposals related to animal agriculture included a crop component.

Use Case	Quantity of Proposals That Selected ≥1 Use Case in Category
Crop	90
Crop Pest and Disease	27
Livestock	17
Marketing	17
Soil	34
Other**	17

**Many of the "Other" proposals included animal pest and disease datasets.

WHO RESPONDED? (CONT.)

Geographies:
Proposals originated from 26 countries across Africa and applicants often partnered to conduct work across the continent.

Country with Lead Institution	Number of Proposals
Burkina Faso	1
Congo, Democratic Republic	1
Cote d'Ivoire	1
Ethiopia	3
Germany	1
Ghana	6
India	1
Ireland	2
Italy	1
Kenya	20
Mali	1
Mauritius	1
Mexico	1
Morocco	1
Namibia	2
Netherlands	3
Niger	1
Nigeria	13
Philippines	1
Rwanda	2
Senegal	1
South Africa	2
Sudan	2
Swaziland	1
Switzerland	1
Tanzania	8
Uganda	4
United Kingdom	2
United States	6
Zambia	2

121
TEAMS SUBMITTED PROPOSALS

91
WERE ELEGIBLE

For the most part, proposals were considered ineligible if they were missing a strong machine learning component or a lead applicant or substantial partner headquartered in Africa. Many international or Western organizations submitted proposals with a wholly-owned or controlled subsidiary as the 'local' partner.

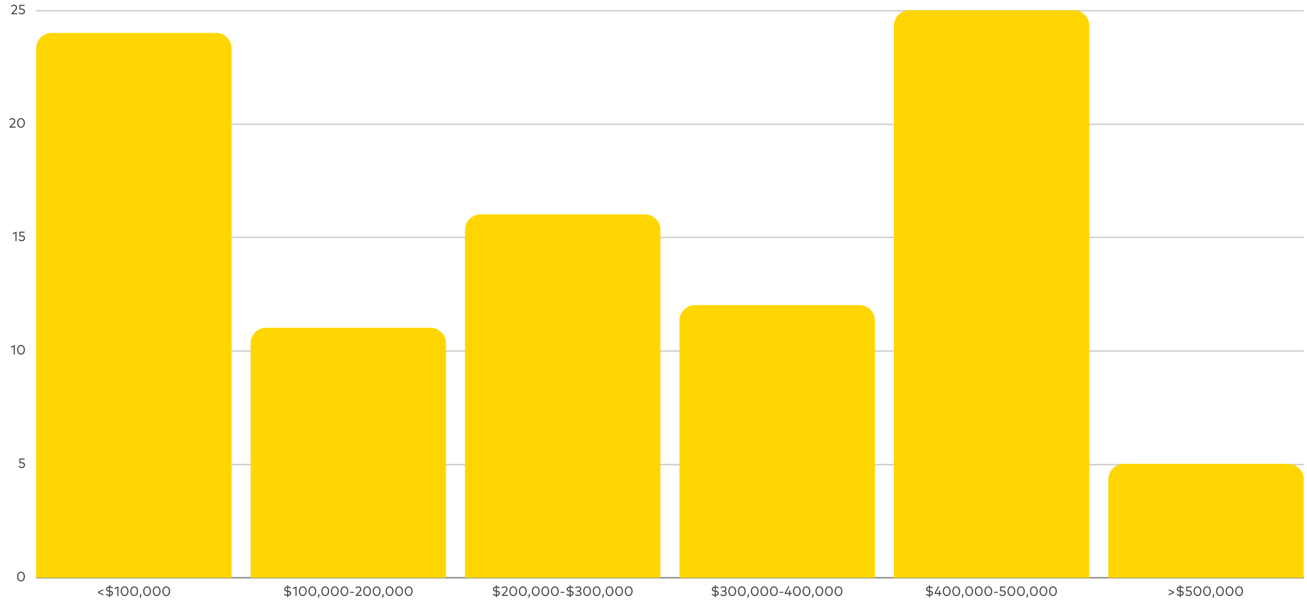
Sector:
Of the eligible pool:

- 50% of lead applicants were research institutions
- 30% were for-profit entities (including for-profit social enterprises, which were not specifically broken out)
- 15% were civil society organizations or other NGOs
- 5% were public sector entities

Note: Sector data was not collected for partner institutions for this call.

WHO RESPONDED? (CONT.)

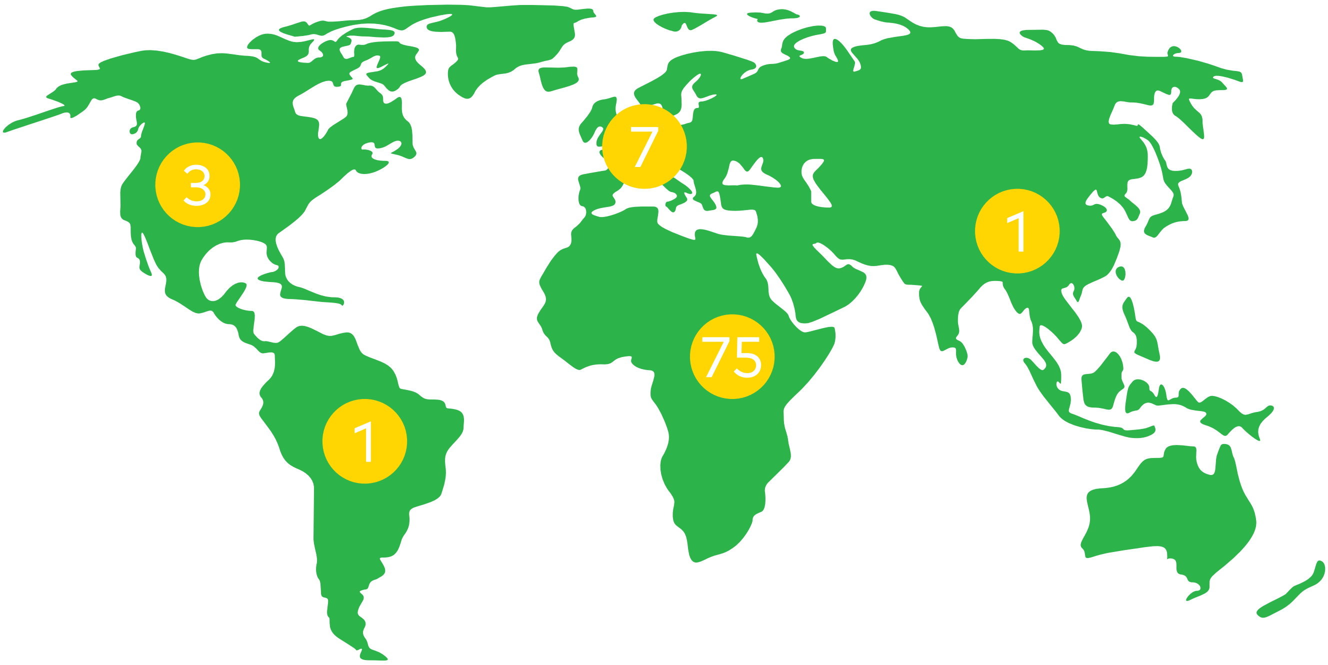
The RFP included a ceiling of 500,000 USD. The majority of eligible proposals fell on the ends of the cost range—either less than 100,000 or greater than 400,000 USD. The distribution of final costs for funded proposals is not disclosed for this call.



Demographics:
Note that the application portal only requested the demographics of the project leader and did not collect data on full project teams. Demographic information was not shared with the Technical Advisory Panel during selections.

- Location of PI

Of all elligible proposals:



**4 proposals did not disclose this information

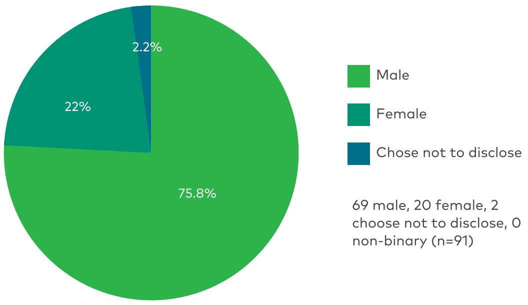
WHO RESPONDED? (CONT.)

- PI Gender

Male project leaders were disproportionately represented in the pool of eligible proposals.

Of projects selected for funding, 4 are led by females and 2 are led by males.

Eligible Projects - Gender Breakdown



69 male, 20 female, 2 choose not to disclose, 0 non-binary (n=91)

- Level of experience in the field

Almost all project leaders were experienced or mid-career, with a smaller contingent of early career researchers.

For eligible proposals:

Experience Level	Number of Proposals
10+ years in the field	51
Choose not to disclose	1
Early career researchers (<5 years in the field)	9
Mid-career researchers (5-10 years in the field)	29
(Blank)	1

Broken down by gender:

	Experience level	Number of Proposals
MALE	10+ years in the field	41
	Choose not to disclose	0
	Early career researchers (<5 years in the field)	4
	Mid-career researchers (5-10 years in the field)	24
FEMALE	10+ years in the field	9
	Choose not to disclose	1
	Early career researchers (<5 years in the field)	5
	Mid-career researchers (5-10 years in the field)	5

Notably, approximately one half of proposals submitted by early career researchers were considered ineligible, which represented a far higher proportion than proposals prepared by mid-career and established researchers.

HOW WERE PROPOSALS SELECTED?

Proposals undergo a three-stage review process:

An eligibility screen by the Secretariat

A review by the Technical Advisory Panel and partners

A deliberative review by the TAP to select a final portfolio of projects by consensus

The Secretariat then worked with projects to make revisions, if any, before the award was confirmed.

SELECTED PROJECTS

The recipients of this first round of funding will produce labeled training datasets in Eastern, Western, Central, and Southern Africa that will support a range of agricultural needs, including livestock and fisheries management, crop identification, yield estimation, and disease detection in crops that shore up food security efforts in the region—namely for cassava, maize, beans, bananas, pearl millet, and cocoa.

Read more about Lacuna Fund's supported projects in agriculture. Commonalities between the selected projects include:

- A clear objective and a team demonstrably qualified to achieve it
- Engagement of local actors in all parts of the proposal, not only in data collection or annotation
- Clear use cases and communities that will contribute to the sustainability of the dataset and downstream products
- Innovative and feasible approaches to dataset collection, augmentation, or maintenance
- Well justified and articulated budgets and timelines

Given the overwhelming interest in the call, the TAP was unable to support many strong proposals that shared some or all of these qualities. Learn more about our 2020 Agriculture TAP here. The gender composition of the TAP was 78% male, 22% female.

WHAT DID WE LEARN?

About the Field:

- We received many proposals from teams with strong experience in either agriculture or machine learning, but fewer from teams with expertise in both.
- In the agriculture call, a majority of proposals came from countries with more developed ML ecosystems, notably Ethiopia, Ghana, Kenya, Nigeria, South Africa, and Uganda. While the agriculture call was broad in both geography and scope, TAP members felt it was possible to distinguish among and construct a portfolio of proposals.
- Given the breadth of the proposal pool, having a variety of both domain and cross-cutting expertise on the TAP to authoritatively evaluate a proposal was critical.
- International organizations played a key role in providing transnational linkages between institutions and helped form competitive teams. In cases where proposals came from a single research institution, these partnerships were often less apparent and the potential impact of the proposed dataset weaker.
- Government and quasi-governmental organizations were partners or engaged in many proposals. However, some of these proposals that included partnerships between agricultural specialists at universities and government agricultural or statistical agencies would have been strengthened by engagement of a machine learning expert. This indicates a potential opportunity for match-making and/or capacity building.

About the Application Process:

- It was valuable to restrict the field of applicants to (teams of) organizations headquartered in Africa or who have a substantial partnership with organizations headquartered in Africa. The Secretariat and TAP found that this encouraged new and innovative collaborations and directed resources to the target audience: data scientists and researchers in underserved communities.
- In the agriculture RFP, anecdotally, the greatest opportunity to strengthen proposals was to provide additional specificity to persuade the Technical Advisory Panel of the technical soundness or feasibility of a project. Many proposals did not provide sufficient specificity on the proposed dataset to be collected, or project methods and timeline, for the Technical Advisory Panel to assess the feasibility of the project. To address this, the Secretariat emphasized the importance of providing specific information on the proposed dataset in the NLP call for proposals.
- In future RFPs, the Secretariat will work with TAPs to develop proposal prompts that allow the TAP to glean enough information about the feasibility of a project without privileging teams with already developed projects, funding, and other advantages. We continue to explore what key markers and indicators of success are and how Lacuna Fund can best collect information through the proposal process that allows for the assessment of these.

2020

LANGUAGE RFP

Lacuna Fund's second Request for Proposals (RFP) for language datasets in sub-Saharan Africa was issued in September 2020.

WHAT DID THE RFP ASK FOR?

Lacuna Fund's second RFP focused on datasets for natural language processing (NLP) in sub-Saharan Africa. The TAP aimed to address the gap in publicly available datasets for most African languages, supporting community-led efforts to create contextually and culturally relevant language tools across fundamental NLP tasks, text, speech, domain specific, and multi-modal datasets.

You can find the full RFP document, including the complete list of use cases here.

67

TEAMS SUBMITTED PROPOSALS

63

WERE ELEGIBLE

The lower percentage of ineligible proposals than in the agriculture call may be due to the more specialized field of natural language processing. Most proposals in this pool featured a strong ML component.

WHO RESPONDED?

Geographies:
Similar to the agriculture call, countries with particularly large or well-developed machine learning communities accounted for the bulk of applications.

Country with Lead Institution	Number of Proposals
Angola	1
Benin	1
Burkina Faso	1
Cameroon	3
Cote d'Ivoire	1
Djibouti	1
Ethiopia	4
France	1
Germany	3
Ghana	5
Kenya	3
Malawi	1
Mali	1
Netherlands	1
Nigeria	13
Rwanda	1
Senegal	1
South Africa	5
Tanzania	3
Uganda	4
United Kingdom	5
United States	3
Zambia	1

WHO RESPONDED? (CONT.)

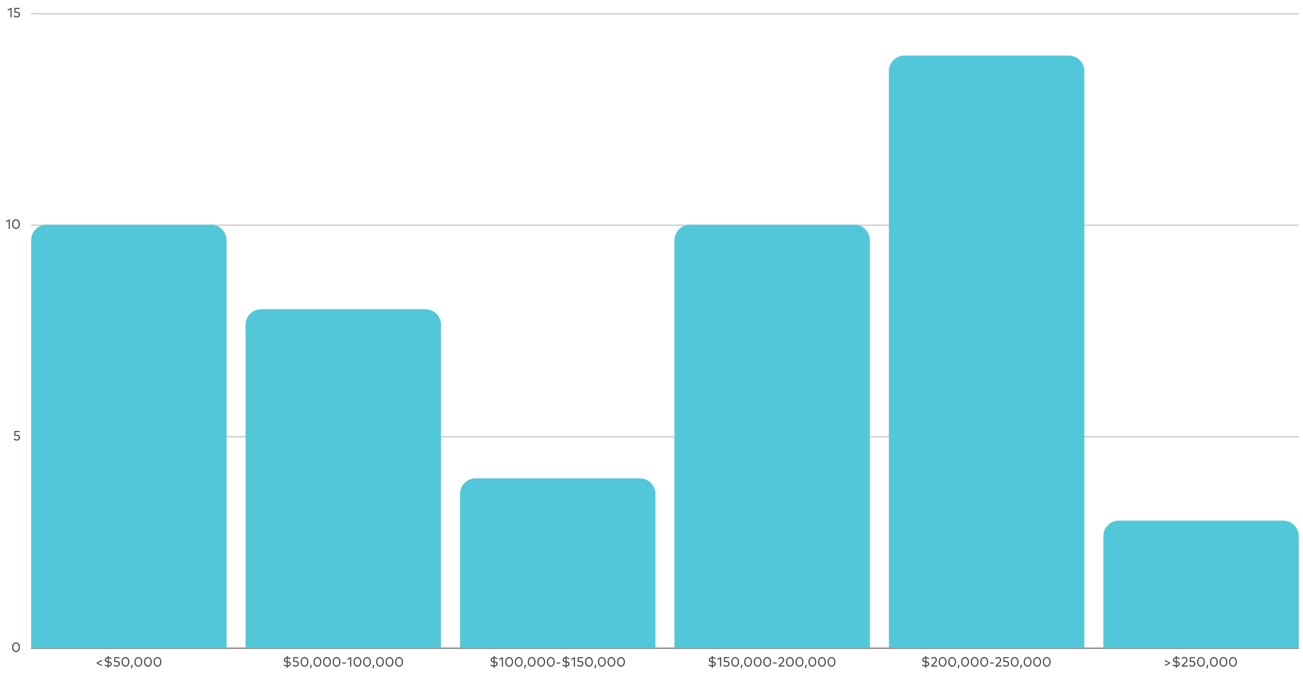
Use Cases:
Applicant-selected use cases revealed a broad spectrum of efforts, including speech and text data collection; work on named entity recognition, part-of-speech tagging, and other fundamental tasks; sentiment analysis and question answering; and various other downstream applications.

There were fewer proposals that aimed to create, expand, or maintain multimodal or domain-specific datasets. This may be due to the lack of available resources in most languages that efforts proposed, making it difficult to create more complex datasets.

Use Case	Quantity of Proposals That Selected ≥1 Use Case in Category
Speech	37
Text	36
MT	38
Fundamental tasks	25
Downstream tasks	25
Code-Switching	10
Improve Bias/Usability	14
Benchmark	30
Innovative applications/ AV	7
Domain Specific	13

Key areas of overlap included projects that aimed to create fundamental, speech, and text resources for a certain number of languages, as well as benchmark datasets for specific tasks.

Proposal Budgets:
The RFP specified that small to medium proposals focusing on one task or language should cost between 20,000 and 100,000 USD, and that larger proposals would be capped at 250,000 USD. The distribution of proposal budgets is below.



WHO RESPONDED? (CONT.)

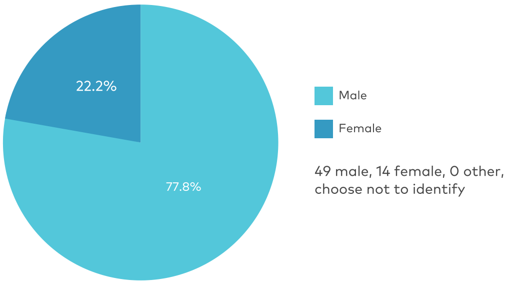
Demographics:
Note that the application portal only requested the demographics of the project leader and did not collect data on full project teams.

- Gender

Similar to the agriculture RFP, fewer project leaders were women.

Of projects selected for funding, 8 were led by males and 2 were led by females.

Eligible Projects - Gender Breakdown



49 male, 14 female, 0 other, choose not to identify

- Level of experience in the field

For eligible proposals, project leaders had:

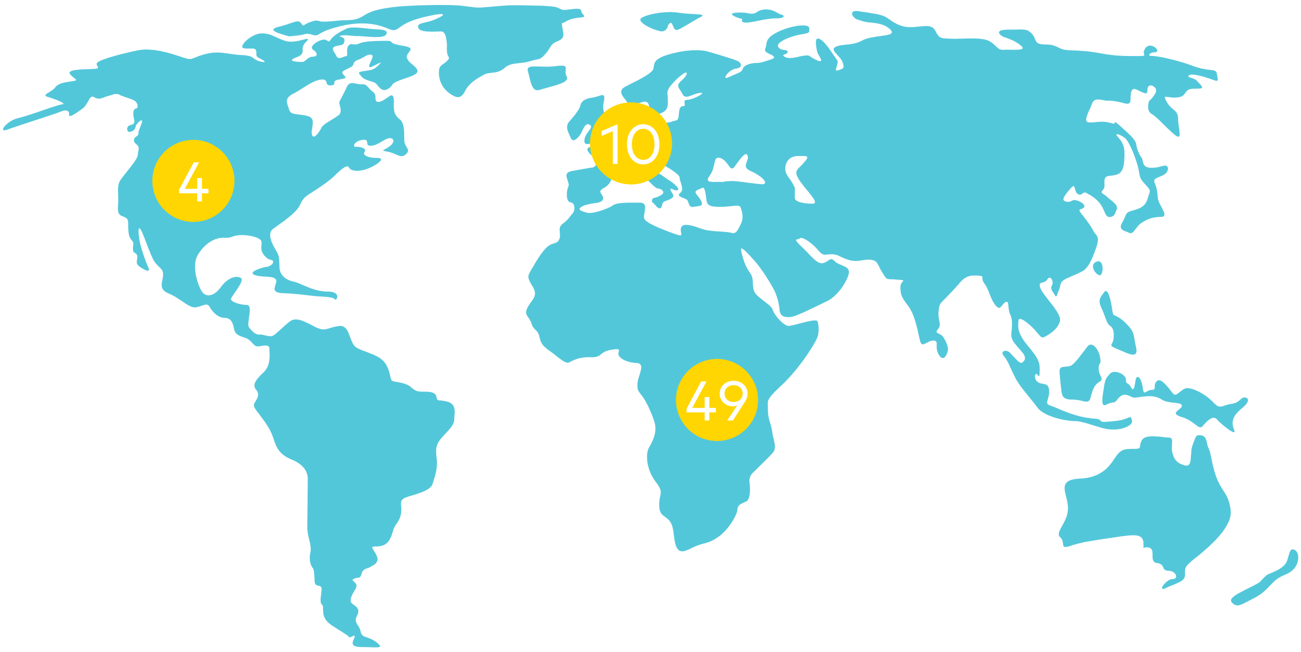
Experience level	Number of Proposals
<5 years of experience in the field	17
5-10 years	20
10+ years	26

Sector of primary institution:
Of the eligible pool, the number of proposals the came from the following types of institutions are:

- For-profit: 14
- Non-profit (research institution): 34
- Non-profit (civil society): 6
- Public sector entity: 4
- Other (IGOs, UN agencies, consortia): 5

- Location of project leaders

Of the eligible pool:



SELECTED PROJECTS

The Technical Advisory Panel awarded funding to 10 teams who are from, or are working in partnership with, organizations across Africa. These recipients will produce training datasets in Eastern, Western, and Southern Africa that will support a range of needs for low resource languages, including machine translation, speech recognition, named entity recognition and part-of-speech tagging, sentiment analysis, and multi-modal datasets.

See the full descriptions of funded projects. The selected projects will support language technologies that will serve hundreds of millions of first and second language speakers across the continent of Africa and beyond.

The projects include data efforts in the following languages:

Language	NER/POS	MT	ASR	Other
Amharic		X		
Bambara	X	X	X	
Bemba				Multimodal
Chichewa	X			
Ewe	X			
Fon	X			
Ga			X	
Ghomala	X			
Hausa	X	X		Sentiment
Igbo	X	X	X	Sentiment
Kinyarwanda	X			
Luganda	X	X		
Luhya		X		
Luo/Acholi	X	X		
Luusamba		X	X	
Moore	X			
Nija Pidgin	X			
Northern Sotho		X		
Runyankore-Rukiga		X	X	
Setswana	X			
Shona	X			
Swahili	X	X	X	
Twi	X		X	
Wolof	X			
Xhosa	X			
Yoruba	X	X		Sentiment
Zulu	X	X		

SELECTED PROJECTS (CONT.)

In addition to key strengths of proposals selected in the agriculture call, the Secretariat found that key elements of successful NLP proposals included:

- Efforts that catalyzed community efforts for sustainable data collection or further avenues for work. Benchmarks, strong community-based efforts, and datasets pioneering a new set of technologies in a particular language family are examples of this.
- Providing clarity on incentive structures for and feasibility of planned annotation or translation. (Many proposals that were not selected aimed to rely on volunteer annotators with little information on how those volunteers would be recruited and motivated to complete the task.)
- Partnerships with a variety of disciplinary expertise in machine learning, linguistics, and domain area expertise (if applicable).

Learn more about our 2020 Language TAP here. The gender composition of the TAP was 33% female, 66% male.

WHAT DID WE LEARN?

About the Field:

- In general, the TAP was impressed by the quality of NLP proposals. There were few easy opportunities to winnow the pool.
- Established communities of researchers facilitated teaming and delivery of compelling transnational proposals. In the field of African NLP, Masakhane is one of the most prominently established communities.
- For higher-resourced languages within Africa (such as Swahili and to a lesser extent Igbo, Yoruba, and Hausa and Amharic), the field of competitive proposals was strong. Most successful proposals in these languages had innovative aspects or clear pathways to scale or sustainability given the number of ongoing efforts in voice and text data collection and annotation. In languages that have historically received less attention from data science communities, proposals often focused on more fundamental data collection efforts. Across all languages, purely translation-oriented proposals that were successful or received high scores tended to have domain-specific or innovative aspects (e.g., scientific translation) or work to build a broad community, including education and training, to support resources in a language.
- Use of existing resources, such as open-source collection infrastructure, and strong relationships with media companies or others made for more cost-effective and feasible proposals. Lacuna Fund is considering how to allow potential future applicants more time to build these relationships.
- The Technical Advisory Panel found proposals that allowed for partnerships between private sector actors, researchers, and international organizations particularly compelling. Some proposals from private-sector actors principally aimed to generate data that would be helpful for their business activities but did not effectively connect this to the broader domain or community benefit.

About the Application Process:

- Engaging both expertise in particular groups of languages, as well as in multilingual NLP and transfer learning approaches on the Technical Advisory Panel was critical to selecting an impactful project portfolio. The TAP was able to propose revisions that strengthened both datasets' ability to support models in particular languages as well as be useful to the field as a whole.
- Requesting more detail from applicants (e.g. specifying the need to provide the number of proposed sentences or other metric) paid dividends in the NLP process. Even then, many reviewers in the initial stage noted that a lack of information (e.g., not specifying the number of hours of voice collection, sentence pairs) kept them from recommending a proposal given uncertainty over its feasibility and impact.
- However, this persistent gap in demonstrating feasibility of a proposal may indicate a capacity need in preparing the concise, complex proposals that a successful application to Lacuna Fund currently demands. From an anecdotal review of incomplete proposals in the application portal in the first two rounds, detailed budget and timeline information appears to be one of the largest barriers to applicants.

BROADER LANDSCAPE

OBSERVATIONS AND NEEDS

Through Lacuna Fund's first two funding processes, the Secretariat and partners have observed key capacity and interoperability needs to help the Fund achieve its mission. Lacuna Fund is also cognizant of broader power imbalances in machine learning that have an impact on downstream use of Lacuna-supported datasets.

Key Capacity Needs
Both domains Lacuna Fund has granted in have well established communities. Key messengers from those communities, as well as TAP members who could disseminate the RFP widely and scope it appropriately, drove many of the quality proposals that the Fund has received.

In both RFPs, the TAP took explicit steps to address power dynamics and help achieve the goal of Lacuna Fund —most importantly by limiting proposal eligibility to organizations headquartered in Africa or in substantial partnership with such an organization.

Interoperability Efforts and their Relevance to Lacuna Fund
Interoperability of Lacuna-supported datasets with existing work is of critical importance for achieving impact. Lacuna Fund's bottom-up approach to granting enables researchers to identify what is most needed in their communities; however, one drawback to this approach is potential lack of interoperability both across funded datasets and with existing resources in the field. The Secretariat sees a need for common infrastructure or policies across domains. This work is largely outside the scope of Lacuna Fund. However, Lacuna Fund can and will continue to encourage the use of existing standards and frameworks, and new ones as they become available.

Some areas have existing or emerging standards that are supported by dedicated organizations or communities and may present opportunities for partnership or coordination with Lacuna Fund. These communities, in many cases, are striving for more inclusive participation and representation in their standards but suffer from lack of support or complementary investments.

In the agriculture domain, earth observation data for ML has increasing coherency through largely inclusive processes. However, beyond earth observation training data, other areas, such as crop pest and disease identification, or ML for livestock management, have less established standards and are in need of further investment and harmonization. The Bill & Melinda Gates Foundation-funded Enabling Crop Analytics at Scale (ECAAS) project has a particular focus on developing and scaling innovative and cost-effective

methodologies for training data collection. They have a focus on crop pest and disease identification, crop type classification, and yield estimation, but not on ML for livestock.

Many types of NLP data have clear standards for interoperability, and proponents in the first NLP round were generally aware of them.

Please reach out to Lacuna Fund Secretariat if you are involved in an effort to foster greater interoperability in the domains where Lacuna Fund works.

Accessibility and Discoverability of Lacuna-Funded Datasets

Discoverability and accessibility of open data is critical to ensuring the impact of Lacuna Fund's supported projects. This relates not to the hosting of data, which in some cases is subject to data localization or other regulatory requirements, but to access to that data in a consistent and machine-readable form. Ideally, this will be facilitated through easily usable APIs, such as that already available from Radiant ML Hub in the earth observation for agriculture space. Clear metadata and guidelines for model inputs and use, such as model cards and datasheets, are also a critical component of discoverability.

In future funding rounds, Lacuna Fund intends to strengthen guidance related to sustainability and use planning for datasets. We welcome inputs and examples of success in ensuring the discoverability and responsible use of open data.

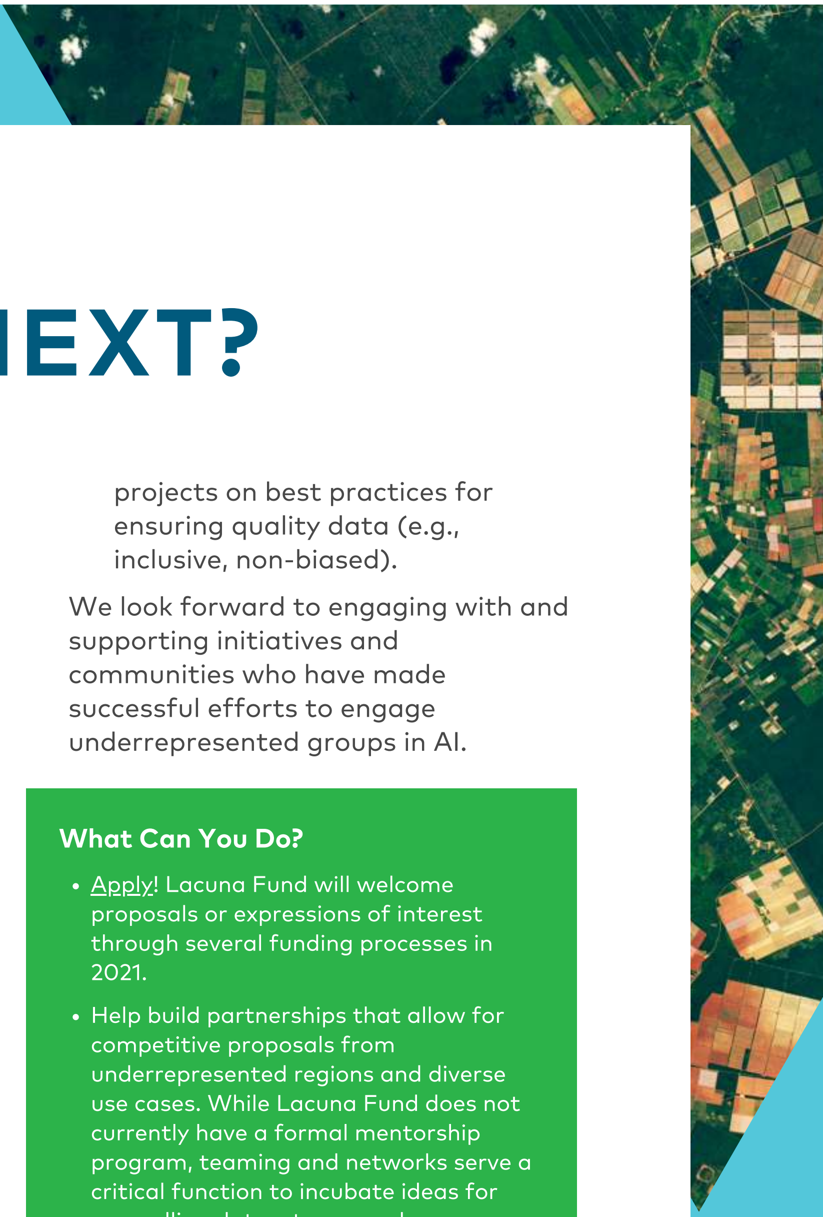
Beyond Data and Tech Solutionism

Lacuna Fund's remit is focused on dataset creation, expansion, and maintenance. Ensuring that representative, inclusive training and evaluation data are openly available fills a critical gap in achieving the promise of using machine learning for good worldwide. However, it is one step, and will not by itself address deep inequities and power imbalances in machine learning.

At present, the Secretariat requires that Lacuna-funded data be released under a CC-BY 4.0 license unless a more restrictive license is necessary to protect privacy, prevent harm, or otherwise maximize the potential for responsible release and downstream use. However, we recognize that data sharing may not always privilege the sovereignty and needs of communities who the data represents.[1] This is an active topic of research and discussion, and we look forward to engaging with new developments in the field.

As Lacuna Fund grows and matures, we look forward to working with partners and continuing to support the creation of machine learning applications by and for communities in underserved contexts worldwide.

[1] See Abebe et al, "Narratives and Counternarratives on Data Sharing in Africa." Proceedings of the 2021 ACM FAccT, March 3, 2021, 329–41, as well as a broader field of work on indigenous data sovereignty.



LOOKING AHEAD

WHAT'S NEXT?

We continue to believe that the three domain areas where Lacuna Fund has focused its initial investment—agriculture, health, and languages—have significant gaps in accessible and equitable datasets for the training and evaluation of machine learning models, and in 2021 will continue to fund in these domains.

Steps Lacuna Fund Will Take: The Lacuna Fund Secretariat recognizes that while gender disparities and varying levels of knowledge of machine learning in proposal submission may in part be a reflection of underlying trends in the field, working towards greater parity and interdisciplinary partnerships will improve the quality of proposals and the impact of the fund. To address these issues, Lacuna Fund plans to work with partners to:

- Disseminate opportunities to networks and points of contact focused on engaging and building the capacity of female and non-binary data scientists and researchers.
- Develop a matchmaking process to connect early career researchers and teams with similar project interests.
- Offer best practices for proposal development, with outreach through affinity groups and other targeted channels.
- Create more opportunities for capacity-building and outreach in French, and by and with machine learning communities in Francophone Africa.
- Increased focus on developing guidelines and providing support to projects on best practices for ensuring quality data (e.g., inclusive, non-biased).

We look forward to engaging with and supporting initiatives and communities who have made successful efforts to engage underrepresented groups in AI.

What Can You Do?

- Apply! Lacuna Fund will welcome proposals or expressions of interest through several funding processes in 2021.
- Help build partnerships that allow for competitive proposals from underrepresented regions and diverse use cases. While Lacuna Fund does not currently have a formal mentorship program, teaming and networks serve a critical function to incubate ideas for compelling dataset proposals.
- Develop (and make us aware) of standards that would be valuable to include in guidance that Lacuna Fund provides to applicants and grantees. This will help ensure that datasets that Lacuna Fund supports are as accessible and impactful as possible.
- Let us know if you are interested in convening and collaboration that would allow for dynamic and interdisciplinary teams to develop proposals for Lacuna Fund. We would like to partner with you!
- Connect Lacuna funded datasets to data users and build the capacity of local actors to create contextually relevant AI applications.
- Lacuna Fund welcomes interest from potential contributors to the Fund. Contact the Secretariat at secretariat@lacunafund.org.