ROBHOOT
Open Discovery Network
Whitepaper v.2.0

February 18, 2020

1 Summary

Global sustainability is a major goal of humanity. Many studies have shown global sustainability could be achieved by strengthening transparency and feedbacks between social, ecological, technological and governance systems. Sustainability goals, however, strongly depend on global access to evidence- and research-based knowledge gaps. Yet, the science ecosystem lacks open-source technologies narrow down knowledge gaps. We introduce an open discovery network targeting knowledge gaps throughout open-access generation of fully reproducible science reports. Open discovery network encompasses a hybrid-automated-technology to lay out the foundation of an open-science ecosytem strengthening the robustness, decentralization, reproducibility and social accessibility of science. The project summarized here is not set out to deliver a finished open discovery network, but to provide the architecture of a science-enabled technology as a proof-of-principle to connect decentralized and neutral-knowledge generation with knowledge-inspired societies.

2 The Science Ecosystem

Science and technology contain multiple steps of information transfer among trusted and untrusted peers. As a consequence, science generates knowledge with specific features. Which are the desirable features of human-generated knowledge? Should such features be aligned with taking informed decisions in complex social, governance, environmental and technological problems? Can global transparency in knowledge generation be achieved to facilitate sustainability goals of humanity? Currently public funded science is highly centralized [1, 2], prone to errors [3], difficult to reproduce [4], and contains many biases [5]. This makes science an ecosystem building up problems to decrease the evidence- and research-based knowledge gaps in humanity [6]. Here we propose an open research network fully accounting for the research cycle. The goal of such a technology is to reduce global knowledge gaps while accounting for centralization, bias, error-prone, non-reproducibility and lack of incentives in the existing science and technology ecosystem (Figure 1 and Table 1).

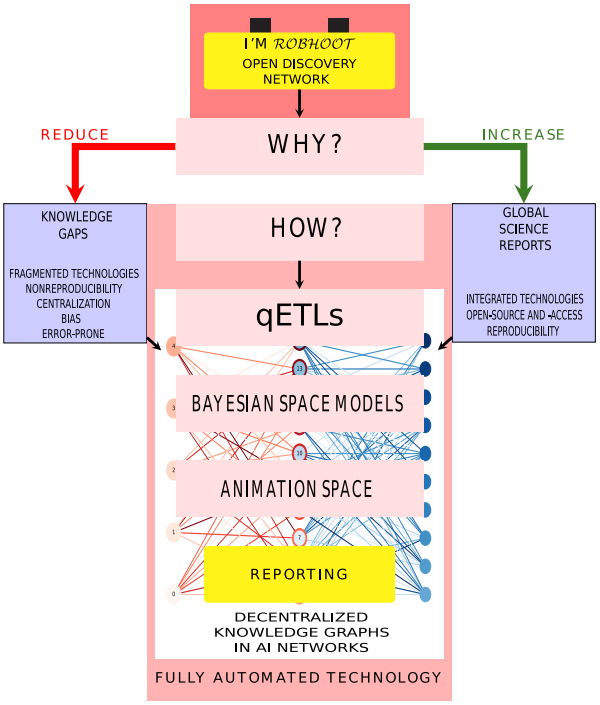

Figure 1: Open discovery network technology. *ROBHOOT* targets global knowledge gaps (red path) and open-access fully reproducible discovery reports (green path). *ROBHOOT* is open-source science-based automated technology decentralizing reproducible knowledge graphs. *ROBHOOT* targets 1) **qETLs** connecting question-based knowledge gaps with data extraction, transformation and loading algorithms for database integration and complexity reduction, 2) **Bayesian Space Models** accounting for open-ended model optimization techniques, 3) **Animation Space** visualizing fitting procedures and pattern generation connecting empirical data and open-ended models, and 4) open and globally accessible **Reporting** formalized in natural processing language.

The core feature of an open research network is to embed knowledge-generation into reproducible automation and decentralization. Currently, many studies focusing on decentralized ecosystems are producing an immense gain of knowledge in a variety of sectors about scalability, security and decentralization trade-offs [7, 8, 9, 10, 11]. In the open science ecosystem, only a few implementations of decentralized technologies exist [2]. Automation and AI technologies represent the other angle from which many advances are rapidly occurring in digital ecosystems

Features	Science Ecosystem	*ROBHOOT*
Decentralization	No	Yes
Full automation	No	Yes
Open-access	Mostly No	Yes
Immutability	No	Yes
Robustness	Mostly No	Yes
Reproducibility	Mostly No	Yes
Owner-Controlled assets	No	Yes

Table 1: *ROBHOOT is designed to resolve desirable properties of science: Open-access, immutability, robustness, reproducibility, and owner-controlled assets. These features will be added during the different stages of development of the project (section "Design Goals").*

[12, 13, 14]. While the existing technological paradigm in many sectors is rapidly shifting towards science-based decentralization and automation technologies, the science ecosystem currently lack decentralized, neutral and open-source knowledge-inspired technologies strongly impacting knowledge-inspired societies (Figure 1 and Tables 1 to 3). Rapid advances of research platforms automating parts of the research cycle are currently under development Automating [15, 16].[1] Most of the existing research projects aiming to automate part of the research cycle are being built around close-source software. Therefore, open-source, reproducible, decentralized and automated technologies accounting for the research cycle are at a very incipient stage of development. To move forward open-source technologies accounting for the research cycle we need to compactly integrate knowledge-generation (Figure 2a) to automated tools connecting knowledge graphs (Figure 2b) and deep learning networks (Figure 2c) in fully decentralized ecosystems (Figure 2d).

3 Design Goals

The open research network will be developed in four stages. The most advanced version is to provide real-time open-access reporting in a decentralized network. Figures 1 to 4 show goals and architecture, milestones, the digital ecosystem and the timeline for each of the stages, respectively. The overall objectives including tools, methods, and territory to explore in each stage for each of the four major versions are the following:

3.1 *ROBHOOT* v.1.0: Automated Research Cycle

- **Question generator** finds the weaknesses of question knowledge graphs to discover relevant topics across discipines.
- **Universal ETLs** connects open-source generalized fault-tolerance algorithms connecting relevant questions to the extraction, transformation and load of data with unique features (i.e., formats, historical-real time, storage, dimensions, size, sampling bias and spatiotemporal resolution, etc) (Figures 1 and 2a, two top layers).
- **Bayesian space models** explore open-ended language of models combining Bayesian networks and optimization methods. The Bayesian space models module will search, evaluate models, trading-off complexity, fit to data and quantify resource usage (Figures 1 and 2a, inference and validation layers).
- **Animation Space** will connect open-source visualization software to the exploration of open-ended models to make the whole search transparent, highly visual and reproducible (Figures 1 and 2a, visualization layer).
- **Reporting** will develop a procedure to automatically explain the structure of the Bayesian space modeling module. It will also communicate the module using visualizations of the procedures followed by the Universal ETLs and Bayesian space models modules (Figures 1 and 2a, reporting layer).
- Robhoot 1.0 testnet is an automated reporting generation on "Biodiversity, Global Change and Sustainability Research" to explore the robustness of the automated research cycle accounting for **Universal ETLs**, **Bayesian space models**, **Animation space** algorithms and **Reporting** in natural processing languages.
- **Tools and Methods**: Multilayer networks metrics, Bayesian Networks, Julia, Python, Open-source software protocols, Gitchain, ETLs open-source software, Kafka, Clickhouse, Fluentd, Hadoop.
- **Novel territory**: Develop universal open-source ETLs algorithms and Bayesian space models and connect them to reporting automation in a "Biodiversity, Global Change and Sustainability Research" case study.

3.2 *ROBHOOT* v.2.0: Reproducible Knowledge Graphs

- Implementation of algorithms tracking paths of the research cycle with Reproducible Knowledge Graphs (KGs) (Figure 2b).
- Robustness and stability of searching and fitting procedures following a suite of open-source lineage client-tracker algorithms.
- **Tools and Methods**: Reproducible knowlegde graph algorithms and open-source packages (i.e., Renku and others).
- **Novel Territory**: Contrasting a set of Reproducible Knowledge Graphs algorithms to quantify the reproducibility, reusability, and recovery properties of the full research cycle.

[1] This is by no means an exhaustive list but it gives an indication of the many projects currently in place: NakamotoT,BigQuery,Automated statistician,Modulos,Google AI,Iris,easeml,datarobot,aito,eureqa

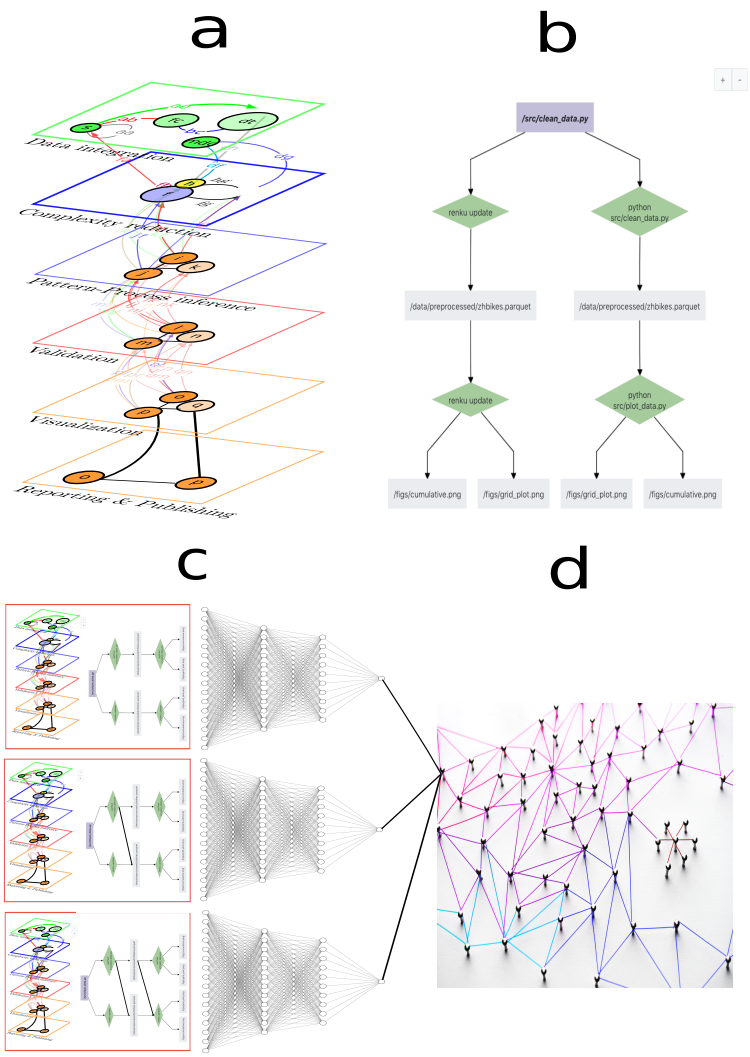

Figure 2: Milestones. **a**) $\mathcal{ROBHOOT}$ **v.1.0** features an automated research cycle connecting question generation to data integration and reporting. **b**) $\mathcal{ROBHOOT}$ **v.2.0** traces research paths as shown in **a** using reproducible knowledge graphs (KGs). **c**) $\mathcal{ROBHOOT}$ **v.3.0** contrasts deep leaning networks to explore populations of KGs for gaining undestanding of the process-based patterns contained in the data, and **d**) $\mathcal{ROBHOOT}$ **v.4.0** deploys KGs in a decentralized network of trusting/untrusting peers with every peer maintaining the population of the KGs.

3.3 $\mathcal{ROBHOOT}$ v.3.0: Deep learning networks

- Deployment of deep learning algorithms to sample paths of the research cycle to produce populations of Knowledge Graphs (KGs) (Figures 2a-c).
- Exploration of the robustness of the automated research cycle combining optimization algorithms and the population of Knowledge Graphs (Figure 2c).

- **Tools and Methods**: Neural Biological Networks, Spiking networks, Bayesian Networks, Deep learning networks. Optimization algorithms.

- **Novel Territory**: Join Bayesian networks models to biology inspired deep-learning networks to efficiently explore constrained model space and the robustness properties of the populations of KGs along ensembles of the research cycle.

3.4 $\mathcal{ROBHOOT}$ v.4.0: Distributed ledger network

- Deployment of a permissioned-permissionless distributed ledger technology to guarantee decentralization, open-access, neutral-knowledge-based network generation and prior confidenciality/posterior reproducibility of the KGs populations (Figures 2c and 2d).
- Exploration of a suite of consensus algorithms and smart contracts among trusted-untrusted peer-to-peer interactions to infer macroscopic metrics of the open research network (Figure 2d).
- Quantification of metrics to study the scalability-security-decentralization trade-offs when storing KGs in the research network (Figure 2d).
- Testnet case study to explore the interaction between consensus protocols and the scalability-security-decentralization trade-offs when committing the KGs to the distributed ledger.
- Mainnet to cryptographically link each population of

Word	Meaning
Reproducible knowledge graph	Algorithms accounting for the research cycle to make it fully transparent
Evidence-based knowledge gap	Factors limiting scientific transfer to benefit society
Research-based knowledge gap	Factors limiting access to reproducibility in science
Question-based knowledge graph	Algorithms detecting questions poorly explored but important for multi-disciplinarity in science
Automation	Interdependent algorithms executing orders targeting minimal human-driven interference
Knowledge-inspired society	Open access reproducible reports to all society
Neutral-knowledge generation	Open and fully reproducible reports making explicit methods accounting for the many biases of the scientiic process

Table 2: Glossary of terms.

KGs to previous KGs-ledger to create an historical KGs-ledger chain that goes back to the genesis ledger of the open research network. Launching of the mainnet to connect multiple database integration with real-time open-access citizen data science and knowledge-inspired societies.

- **Tools and Methods**: Distributed computing algorithms, Blockchain and consensus algorithms, BighainDB, Gitchain. Telegram open network, Golem.

- **Novel Territory**: Deployment of contrasting functional consensus algorithms to explore decentralization and robustness properties of the KGs populations along ensembles of the research cycle space.

$\mathcal{ROBHOOT}$ aims to be a hybrid-technology accounting for many features (Tables 1 to 3). Producing such a multi-feature technology requires multidisciplinarity teams making contributions for each of the Robhoot features while integrating all these features in a rapidly evolving digital ecosystem. In this regard, the project aims to put together data and computer scientists (i.e., distributed computing, open-source software development), scientists from the physics of complex systems (i.e., multilayer networks), artificial intelligence (i.e., deep learning and automation) and the biology, ecology and evolution of social, natural and technological ecosystems.

4 Robhoot in Digital Ecosystems

The science ecosystem currently lack technologies fully automating the research cycle into the open-source digital ecosystem. Despite public institutions are demanding more reproducibility and openness of the data and the scientific process, and overall a shifting towards open and reproducible scientific and engineering landscapes, there are not currently open and integrated technologies aiming to compactly facilitate and distribute the scientific and engineering knowledge in open, reproducible and immutable knowledge networks (Tables 1 and 2).

Automating knowledge-generation requires the integration of many distinct features. Usually, knowledge-generation comes from interactions within- and between-layers of the scientific process (Figure 2a). The feedbacks occurring within and among layers in the science and technology ecosystem also provide unexpected behaviors that are difficult to anticipate. Therefore many feedbacks and interactions within- and between-layers are not easy to reproduce if not properly accounted for. We will take advantage of the open-source software community to explore knowledge graphs, optimization, automation, and decentralization algorithms together to study the robustness and reproducibility properties of the scientific process (Figures 1 and 2).

One way of visualizing the dimensionality of $\mathcal{ROBHOOT}$ in the digital ecosystem is to connect each layer of the scientific process (Figure 2a) to open-source software to gain functionality of the open research network (Figure 3). For example, Node 0 (left column, Figure 3) can be the Data Integration layer in Figure 2a. This node is connected to seven nodes representing open-source ETLs open-source software (i.e., central column, Figure 3). Connections between Node 0 and nodes 5, 6, 8, 9, 10, 12 and 13 can be rapidly evolving (i.e., indicated by the different red tones of the connections). Indeed, open-source ETLs are rapidly evolving towards accounting for many heterogeneous aspects of data integration (i.e., formats, historical-real time, storage, dimensions, size, bias and spatiotemporal resolution). ETLs can also be connected to a gradient of reporting generation (i.e., right column, Figure 3) noting reports containing only a subset of the interactions of the digital ecosystem network. The network of the fully automated research cycle can be one where Nodes 0, 1, 2, 3, and 4 represent the different layers of the research cycle (left column, Figure 3 and Figure 2a) connected to the open-source software of the digital ecosystem (central column, Figure 3) to generate full populations of reports (right column, Figure 3).

Feature	*ROBHOOT*
Long-term vision	Global open-access to a fully reproducible knowledge-generation inspired technology
Breakthrough scientific and technological target	Collapsing evidence- and research-based knowledge gaps for a sustainable knowledge-inspired society
Novelty	Science-based technology emerging from targeted algorithmic discovery at the interface of multilayer networks, knowledge graphs, deep-learning, and consensus mechanisms
Foundational	Neutral-knowledge inspired technology for an emerging open science of science and science-society research disciplines
High-risk	Adapted to explore new terrirories into the open-science-technology-society interface ecosystem
Interdisciplinarity	Hybridizing expertise from distributed computing and deep learning to multilayer networks and the ecology and evolution of natural and digital ecosystems (Table 1)

Table 3: *ROBHOOT features along its developmental stages.*

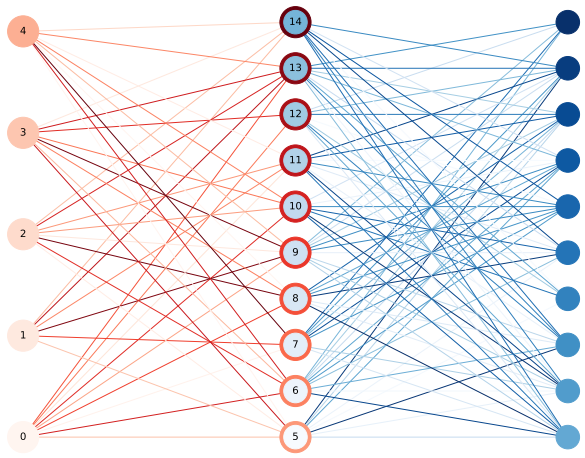

Figure 3: Robhoot in Digital Ecosystems: **Left column**: *ROBHOOT* **v.1.0** representing the research cycle as nodes from number 0 to 4: Data integration (0), Complexity Reduction (1), Inference (2), Validation (3), and Visualization(4)). **Central column**: Nodes representing the research cycle in the left column are connected to open-source software in the digital ecosystem. Connections with node number 0 in the left column can, for example, represent the ETLs open-source software interactions required to generate the **Universal ETLs** module. The same meaning applies to the different nodes of the left column. **Right column**: Each node represents a report meaning there is a reporting gradient generated by the connections to the open-source software from where each report is generated only using a subset of the research layers and open-source software.

5 Conclusion

Science and technology ecosystems are in need of accounting for the uncertainties, reproducibility and immutability related to the complexity of the research process (Table 1). Such needs are not just for a specific stage of the research cycle, but from data acquisition and integration to automated reporting generation because knowledge-inspired societies and decentralized governance will demand full research cycle transparency to solve complex social, environmental and technological problems (Tables 2 and 3). Reducing knowledge-gaps at global scales in knowledge-inspired societies bring many challenges to our research proposal because obtaining robust knowledge from integrating many layers of the research cycle, each containing its own set of methods and uncertainties, can generate divergent, fragile and contradictory outcomes.

We will develop a flexible and adaptive research method focusing step by step in increasing levels of complexity (i.e., from *ROBHOOT* v.1.0 to v.4.0, Figure 4). Our motivation will be to provide a first open-access proof of concept of how the technology works: we will automate reproducible research paths (*ROBHOOT* v.1.0) to sample the KGs (*ROBHOOT* v.2.0) contrasting deep learning algorithms to estimate the uncertainty of the ruled-based inference obtained by fitting predictions to simulated data (*ROBHOOT* v.3.0). Accounting for the uncertainties of each of the research stages when sampling the KGs comes from the many distinct paths within and across the layers in the research cycle (Figure 2a). *ROBHOOT* v.4.0 will test a variety of consensus algorithms to explore the degree of security, decentralization and scalability of the ledger knowledge network using the generated population of KGs.

Despite our focus will be bias towards the algorithmic robustness during the four stages of development, we will implement a domain-specific case study, a "Biodiversity, Global Change and Sustainability Research", to test the robustness of the rule-based inference obtained by fitting the KGs to empirical patterns. The high risk associated to robustly automate the full research cycle for producing immutable open knowledge will be buffered to a great extend because the existing digital ecosystem of highly reliable open-source software tools (Figure 3).

References

[1] H. Inhaber. Changes in centralization of science. *Research Policy*, 6(2):178–193, apr 1977. ISSN 0048-7333. doi: 10.1016/0048-7333(77)90024-5. URL `https://www.sciencedirect.com/science/article/abs/pii/0048733377900245`.

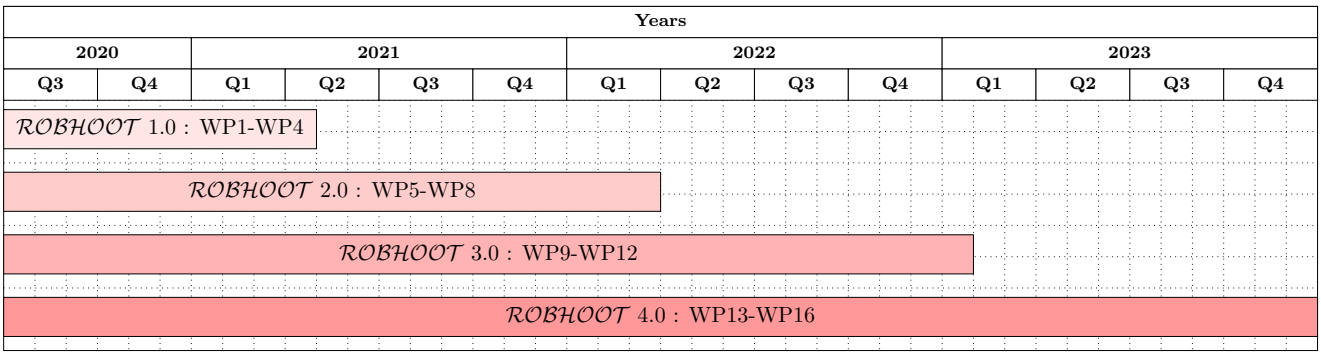

Figure 4: Roadmap: *ROBHOOT* **v.1.0** working packages **WP1** to **WP4** will take care of the integration of **Universal ETLs**, **Bayesian Space Models**, **Animation Space**, and **Reporting** to automate fully the research cycle (Figure 2a). *ROBHOOT* **v.2.0 WP5** to **WP8** will deploy knowledge graphs (KGs) into a fully traceable research cycle (Figure 2b). *ROBHOOT* **v.3.0 WP9** to **WP12** will explore deep learning networks to sample KGs populations to gain understanding of the robustness of the patterns in the data under distinct research paths (Figure 2c). *ROBHOOT* **v.4.0 WP13** to **WP16** will deploy KGs populations into a decentralized network of mutually trusting/untrusting peers with every peer maintaining the population of the KGs.

[2] Vlad Günther and Alexandru Chirita. " Scienceroot " Whitepaper. 2018. URL https://www.scienceroot.com/.

[3] Ferric C Fang and Arturo Casadevall. Retracted Science and the Retraction Index. *Infection and Immunity*, 79(10):3855 LP – 3859, oct 2011. doi: 10.1128/IAI.05661-11. URL http://iai.asm.org/content/79/10/3855.abstract.

[4] Tom E. Hardwicke, Maya B. Mathur, Kyle MacDonald, Gustav Nilsonne, George C. Banks, Mallory C. Kidwell, Alicia Hofelich Mohr, Elizabeth Clayton, Erica J. Yoon, Michael Henry Tessler, Richie L. Lenne, Sara Altman, Bria Long, and Michael C. Frank. Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, 5(8):180448, sep 2018. ISSN 20545703. doi: 10.1098/rsos.180448. URL https://doi.org/10.1098/rsos.180448.

[5] John P a Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, aug 2005. ISSN 1549-1676. doi: 10.1371/journal.pmed.0020124. URL http://www.ncbi.nlm.nih.gov/pubmed/16060722.

[6] Matías E Mastrángelo, Natalia Pérez-Harguindeguy, Lucas Enrico, Elena Bennett, Sandra Lavorel, Graeme S Cumming, Dilini Abeygunawardane, Leonardo D Amarilla, Benjamin Burkhard, Benis N Egoh, Luke Frishkoff, Leonardo Galetto, Sibyl Huber, Daniel S Karp, Alison Ke, Esteban Kowaljow, Angela Kronenburg-García, Bruno Locatelli, Berta Martín-López, Patrick Meyfroidt, Tuyeni H Mwampamba, Jeanne Nel, Kimberly A Nicholas, Charles Nicholson, Elisa Oteros-Rozas, Sebataolo J Rahlao, Ciara Raudsepp-Hearne, Taylor Ricketts, Uttam B Shrestha, Carolina Torres, Klara J Winkler, and Kim Zoeller. Key knowledge gaps to achieve global sustainability goals. *Nature Sustainability*, 2(12):1115–1121, 2019. ISSN 2398-9629. doi: 10.1038/s41893-019-0412-1. URL https://doi.org/10.1038/s41893-019-0412-1.

[7] Golem. The Golem Project Crowdfunding Whitepaper. *Golem.Network*, (November):1–28, 2016. URL https://golem.network/crowdfunding/Golemwhitepaper.pdf.

[8] Nikolai Durov. Telegram Open Network. pages 1–132, 2017.

[9] Elli Androulaki, Artem Barger, Vita Bortnikov, Srinivasan Muralidharan, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Chet Murthy, Christopher Ferris, Gennady Laventman, Yacov Manevich, Binh Nguyen, Manish Sethi, Gari Singh, Keith Smith, Alessandro Sorniotti, Chrysoula Stathakopoulou, Marko Vukolić, Sharon Weed Cocco, and Jason Yellick. Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains. *Proceedings of the 13th EuroSys Conference, EuroSys 2018*, 2018-Janua, 2018. doi: 10.1145/3190508.3190538.

[10] Ocean Protocol Foundation, BigchainDB GmbH, and DEX Pte. Ltd. Ocean Protocol: A Decentralized Substrate for AI Data & Services Technical Whitepaper. pages 1–51, 2018. URL https://oceanprotocol.com/.

[11] BigchainDB GmbH. BigchainDB: The blockchain database. *BigchainDB. The blockchain database.*, (May):1–14, 2018. doi: 10.1111/j.1365-2958.2006.05434.x. URL https://www.bigchaindb.com/whitepaper/bigchaindb-whitepaper.pdf.

[12] J Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[13] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and & Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*. ISSN 0028-0836. doi: 10.1038/s41586-019-0912-1. URL www.nature.com/nature.

[14] Yolanda Gil, Bart Selman, Marie Desjardins, Ken Forbus, Kathy Mckeown, Dan Weld, Tom Dietterich, Fei Fei Li, Liz Bradley, Daniel Lopresti, Nina Mishra, David Parkes, and Ann Schwartz Drobnis. A 20-Year Community Roadmap for Artificial Intelligence Research in the US Roadmap Co-chairs: Workshop Chairs: Steering Committee. Technical report, 2019. URL https://bit.ly/2ZNVBVb.

[15] Christian Steinruecken, Emma Smith, David Janz, James Lloyd, and Zoubin Ghahramani. The Automatic Statistician. pages 161–173.

[16] Roger Guimerà, Ignasi Reichardt, Antoni Aguilar-Mogas, Francesco A. Massucci, Manuel Miranda, Jordi Pallarès, and Marta Sales-Pardo. A bayesian machine scientist to aid in the solution of challenging scientific problems. *Science Advances*, 6(5), 2020. doi: 10.1126/sciadv.aav6971. URL https://advances.sciencemag.org/content/6/5/eaav6971.