IOTA untangled

IOTA ideas

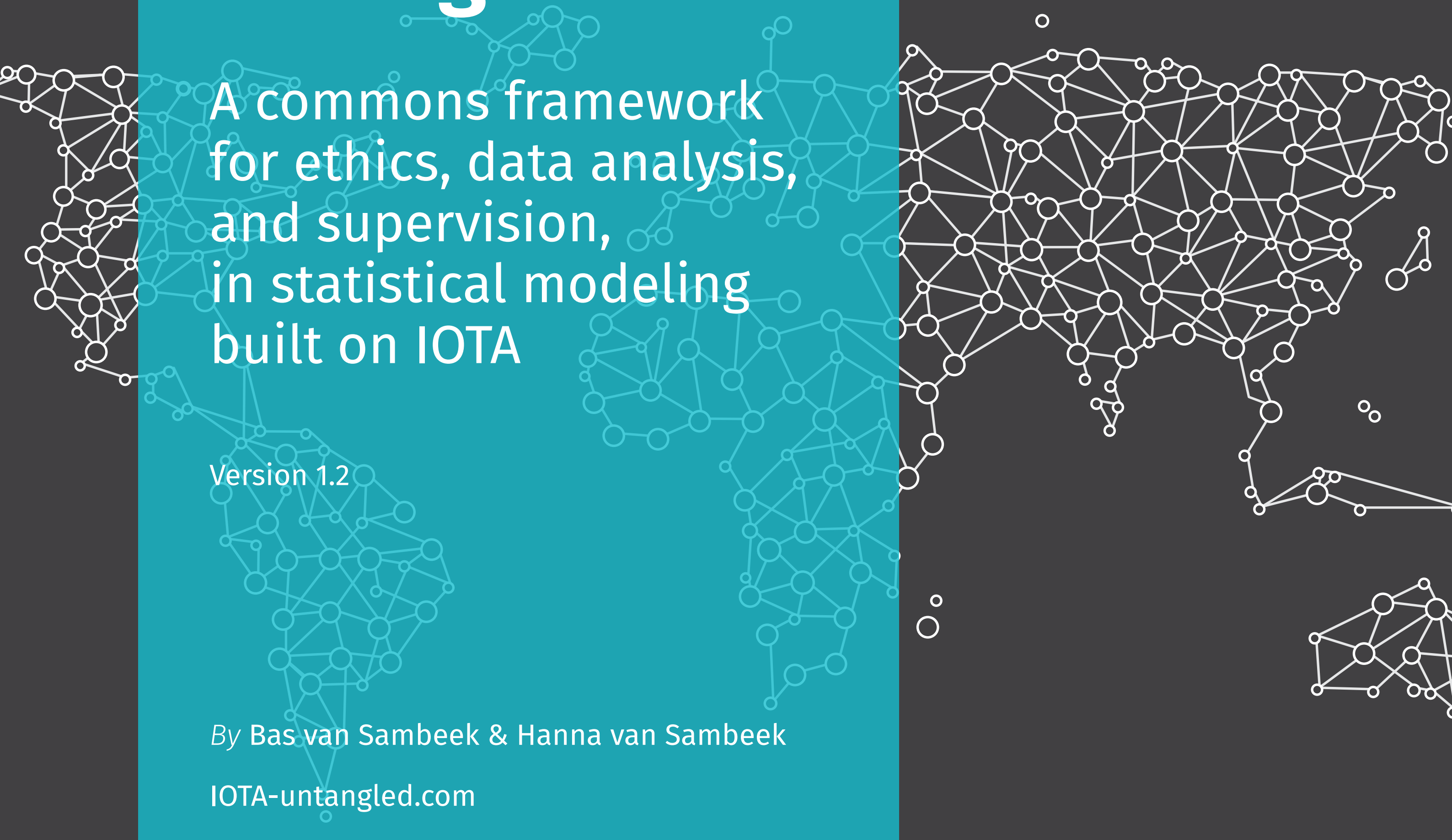
Trialogue

A commons framework for ethics, data analysis, and supervision, in statistical modeling built on IOTA

Version 1.2

By Bas van Sambeek & Hanna van Sambeek

IOTA-untangled.com



Trialogue summary

Make AI, Machine Learning & algorithms more ethical

Trialogue is a commons framework for *ethics, data analysis and supervision in statistical modeling*, built on IOTA. Ideally it will become a standard like ISO 9001 for quality management or ISO 27001 for infosec.

There is no moral standard for algorithms.
Trialogue lets data scientists find and set such norms in conjunction with clients. Afterwards it helps with auditing if models are built to specification.

A key feature is that the framework itself doesn't define moral behavior or good data hygiene itself, but instead lets users determine what is best – in an auditable way.

Trialogue is intended to aid data scientists in their communication with clients when discussing quality standards for algorithms, Machine Learning and AI's.

By using open standards for *ethics*, *data analysis* and *operational supervision* – which will have to be established – you don't need to know how a model exactly works to corroborate an opinion on the result.

Table of Contents

Problem statement

There is no moral standard for algorithms

Automation is a process that started with the use of tools, clothing and fire. It gave us a tremendous edge in survival, but now it is turning into a double edged sword.

With AI, Machine Learning and algorithms taking over day-to-day life, we are increasingly governed by automated decisions. These decisions can come from simple decision trees, to self-learning algorithms, and even rudimentary AI's. While automations can create huge efficiencies, they are not without faults. And because their inner workings are abstract to most people, chances are that errors are not corrected, or not corrected in time. Setting standards for ethics, data analysis and supervision could help balance the pros and cons.

While slightly controversial, George Box' saying "all models are wrong" is helpful because this notion is *not* completely obvious to those less versed in statistical modeling – which includes most of today's society. People have a tendency to believe whatever a model predicts, regardless of its statistical validity. **The important question now is: how wrong can a model be to lose its usefulness? And more importantly: are externalities for all parties affected taken into account when assessing its usefulness for a certain application.**

> "The most that can be expected from any model is that it can supply a useful approximation to reality: All models are wrong; some models are useful".
>
> George Box – statistician

Problem examples

Unintended consequences have many forms

Dystopian sci-fi novels have been warning of the dangers of automation. And while we're not seeing Terminators yet, there are many examples of unintended outcomes.

- **Microsoft chatbot Tay becomes Nazi on Twitter**
- **Uber's self driving car runs red light in unauthorized test**
- **Even good bots fight: The case of Wikipedia**
- **Amazon scraps secret AI recruiting tool that showed bias against women**
- **Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech**

And the question is for how long Terminators and other cyborgs killing humans will remain fictional:

- **Will #BlackLivesMatter to Robocop?**

These are some famous examples of automation gone wrong in well known organizations. These stories naturally get attention, but with the increased opportunity for smaller entities to utilize algorithms, Machine Learning and AI's, such errors will not get called out enough. An inherent lack of transparency might enable people deciding on the scope and quality of statistical models to cut corners, or even engage in unethical behavior.

Algorithms, Machine Learning and AI's are wonderful tools in potential, but we have to keep them in check.

> “There is no economic law that says that everyone, or even most people, automatically benefit from technological progress.”
>
> Nicholas Carr, The Glass Cage: Automation and Us

Solution

Set a standard for ethics, data hygiene & supervision

With the problems that are already occurring, preventive measures seem essential to make the use of algorithms, Machine Learning and AI's safe and morally correct.

We propose a framework that encapsulates **decentralized ledger technology (DLT)** and **shared values within a particular community** to set a standard against which statistical models are checked. Using DLT and open standards it creates the transparency needed to keep an eye on unwanted externalities from increased automation.

A key feature is that the framework doesn't define moral behavior or good data hygiene itself, but instead lets users set norms for their preferred standards beforehand and allow others to check if the result matches with these self selected norms. The immutability of the DLT prevents altering after-the-fact of the scope, rules and process.

This framework is designed to support trained data scientists. It helps communicate clearly with laymen on the scope and limits of a model, and proving the quality of their result. It is also meant to assist developers and programmers venturing into data science and machine learning, who might lack a solid statistical background.

Using open standards laymen can make sense of the general preconditions of a model. (Semi-)public audits further validate the quality of the model in regards to its intended purpose. An additional element of confidence is both the maker's and the auditor's trustworthiness score from their total body of work through a web-of-trust.

> “We believe that by far most people in this world are good and try to be honest.
> Most faults and errors can be traced back to miscommunication and misinterpretation.”
>
> Bas & Hanna van Sambeek

1: Ethical charters

What is “good”, and who decides that?

Establishing ethical standards will be costly, but what does it cost us if we don’t? Instead of trying to define ethics, the framework democratizes this dilemma and lets users decide for themselves.

Applications and models are made to work by programmers and data scientists. We can check if they work rather easily, but are they good? And what is good? Who defines this? The builder? The person who pays? The people who use it? Or we as a society?

The framework doesn’t dictate morality or good ethics, but instead allows users to set their own preferred standards. Various ethical standards can be used as a boilerplate on how to engage with data, outcomes of modeling and their implications in real life. Defining this goes beyond the scope of this proposal, because it is something we as a society should do, not just data scientists.

Ethical charters of common values, morals and ethics don’t seem to exist yet in the context of a data framework. It is likely that different charters are needed for various cultures around the world, based on general ideologies. Also a form of granularity might be needed within a culture, as values can differ according to needs and wants. Not every application is equal.

> “Ethics is knowing the difference between what you have a right to do and what is right to do.”
>
> Potter Stewart, US Supreme Court Justice

2: Data analysis standards

How do you set standards for the quality of a model?

Tamperproof hardware and IOTA solve the problem of collection and immutability of data. But such data do not automatically transform into useful information.

This framework is not about proving something right or wrong. **Instead its goal is to assess if the model is built according to specification**, defined in standards laymen can familiarize themselves with. This step adds to the likelihood an outcome is true, and in what situations.

Every model is wrong. However what we can do is define what is an acceptable abstraction for a model with regards to its intended purpose. We should also define beforehand where and how a model/AI/ML-algorithm becomes too wrong.

A quality standard for this purpose will likely shape itself like creative commons licenses, where every increment adds requirements in a linear fashion. In interviews with data scientist we have found indications that simple metrics – for instance coding standards and the way of dealing with missing value imputation – will likely result in a more useful model.

The data analysis standards still need to be established and certified, because they do not exist yet in the required format.

> “We are entering an age where fraudulent data cost lives and billions of dollars in losses at any moment. Our entire society rests on automatic decisions based on data, data which is currently easy to corrupt.”
>
> David Sønstebø on IOTA Discord

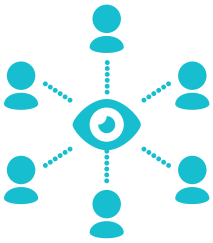
3: Operational supervision

Who guards the guardians?

With algorithms controlling many aspects of our society, often in an autonomous fashion, the question of supervision is becoming increasingly important.

After a statistical model is completed, the real work start. Even with a model of the highest quality, a flawless result in a live setting is not guaranteed. Practice shows that over time algorithms can show quirks that deviate their results from the intended outcome. With the increased reliance on these models to govern parts of our society, a way to check like Trialogue should be welcomed.

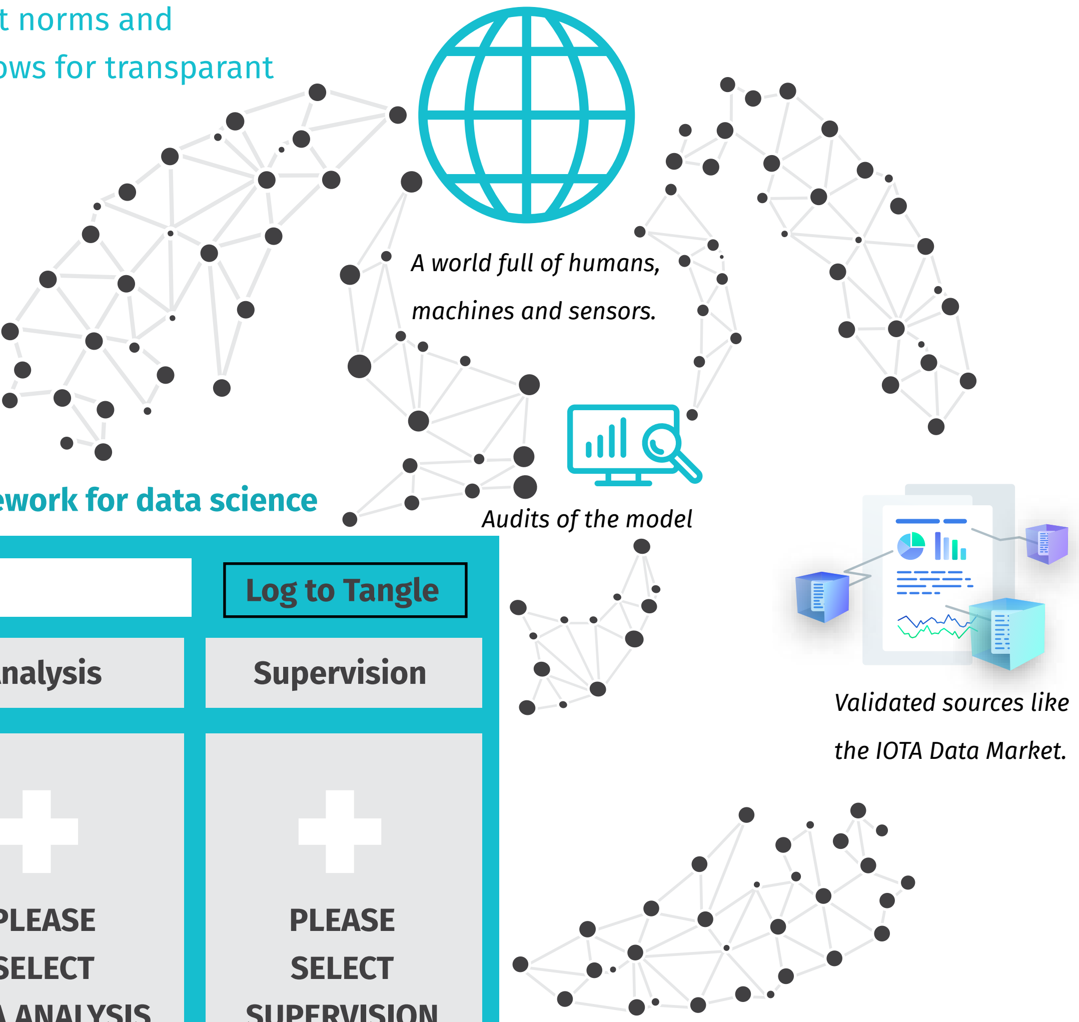
Monitoring and fine-tuning for AI's and (self-learning) algorithms is not defined yet in this proposal, as well as who should do this monitoring. It seems logical to have the original creator of the model do this, but this might be inefficient and not utilizing the creators qualities. The task of Supervision is likely to be cut up in different parts, such as observing, evaluating, and adjusting the model.



Trialogue visualization

The framework within a secure ecosystem

Trialogue is a framework to accommodate clear communication in data science. The combination of self set norms and immutable logging allows for transparant and auditable models.



A world full of humans, machines and sensors.

Audits of the model

Validated sources like the IOTA Data Market.

Trialogue framework for data science

Define purpose: | **Log to Tangle**

Ethics	Analysis	Supervision
+ PLEASE SELECT ETHICS CHARTER	+ PLEASE SELECT DATA ANALYSIS STANDARD	+ PLEASE SELECT SUPERVISION PROTOCOL

The framework allows users to set the purpose and conditions of the model and log choices immutably. A plug-in in the statistics program also logs the model's code including comments, and any alterations to the model for future auditing. Logs are kept in a data storage, with hashes of the data stored immutably on the Tangle.

(Semi-)public audits

Even if it's legal that doesn't mean you should do it

The law describes legal limits of what is allowd. Trialogue fills the gap between what is lawful and what is deemed moral. Audits – peer reviews, public or semi-public – will certify that self-set norms are met.

An essential component of this framework will be the ability to audit. The standards to which a model is set are always publicly published on the Tangle, as are the outcomes of audits. This way the general public can check if the model is built to reasonable standards, and if those standards were met and still being kept.

Auditors can be professionals similar to accountants, or fellow scientists in the case of peer reviews. In the case of supervision over the algorithm, auditors can be less of an expert, for instance in the case of monitoring performance. They still certify publicly on the Tangle the correct working of the algorithm.

There is even a possibility for the general public to audit the model in a "peer audit", although there is the problem of a lack of understanding of statistics. We'll leave this as a possibility, but don't investigate this for now.

The auditor compares the model to the predefined standards and "ticks off" the boxes. This proposal does not specify the process in detail, but a desired outcome would be a scale with the amount of compliance, instead of a binary good or bad. The binary option, in our opinion, would oversimplify the audit and trigger a minimal-effort approach in some cases. The scale in contrast would reward an increased effort, inspiring a better model.

Even proprietary models can be checked by various auditors, and get a public "stamp of approval" without publishing the underlying code. Here a diversity of auditors over time can create a pattern of reliability in the form of a web of trust, because in case of calamities the whole audit trail can be opened up, and malicious behavior can be uncovered, tainting reputations of everyone involved

The added value of IOTA

Immutable, feeless data transport and payments

IOTA's Tangle is an efficient way to register the norm set in the Trialogue framework, to log changes and comments to any model, and to acquire the raw data in a secure way.

Trialogue will mostly use the data transfer and settlement function of IOTA, while token payments will be an option in certain usecases. This proposal does not cover the technical implementation, but the use of MAM flash channels is almost guaranteed.

IOTA's Tangle is not intended to be used for raw data storage. The bulk of the data will likely be stored locally during the build, and when an algorithm goes live a cloud storage or decentralized solution involving IPFS could be used here. The Tangle will be used for data transport and validation of that data.

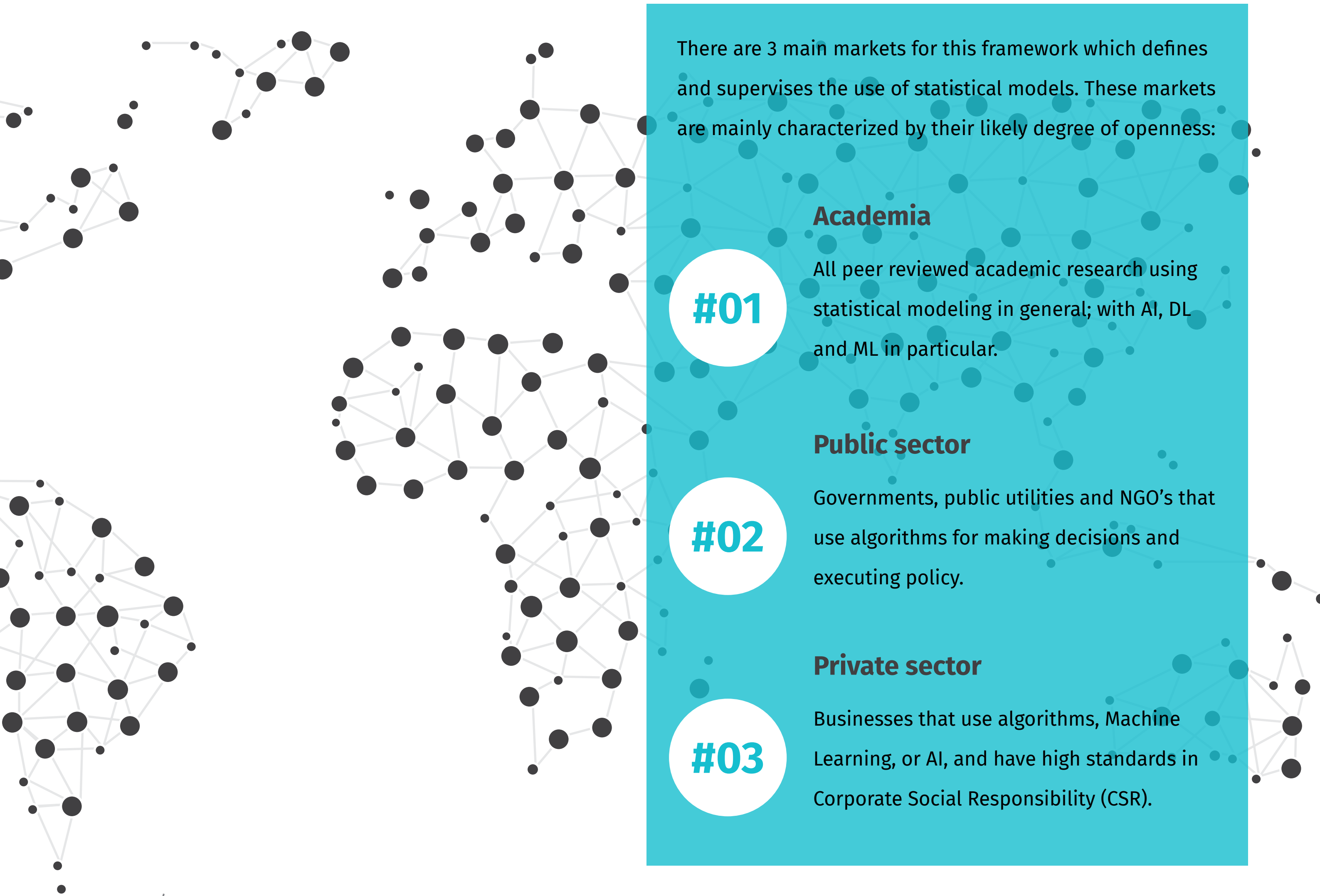
IOTA's huge benefit of feeless transactions has a drawback: transactions are not stored indefinitely. Because of the large amounts of transactions, full nodes prune the Tangle regularly to keep the database size manageable. A permanode service might be needed to audit a model. This permanode will almost certainly charge the auditor for the requested transaction data. This is reasonable, but it is an aspect that should be taken into account.

A huge benefit for data scientist and auditors is the IOTA ecosystem. The ecosystem supports the development of extra functionality on top of the lightweight base protocol. Extra functionality can be for example the proposed Qubic (Quorum Based Computing) to outsource computing in a decentralized fashion, the Data Market to acquire a wide variety of data in a secure and traceable way, and the advantages that come with interconnectivity as the IOTA protocol is set to be the standard for IoT.

Market description

Algorithms are taking over the world

The framework should be considered in all usecases of automation worldwide. Because of its open design it can accommodate for every culture, level of quality and purpose.



There are 3 main markets for this framework which defines and supervises the use of statistical models. These markets are mainly characterized by their likely degree of openness:

#01

Academia

All peer reviewed academic research using statistical modeling in general; with AI, DL and ML in particular.

#02

Public sector

Governments, public utilities and NGO's that use algorithms for making decisions and executing policy.

#03

Private sector

Businesses that use algorithms, Machine Learning, or AI, and have high standards in Corporate Social Responsibility (CSR).

Road to adoption

Possible implementation of the commons framework

The preferred way to implement is in close cooperation with data scientists. Their feedback, experience, needs and wants will be invaluable in establishing a useful and rigorous standard.

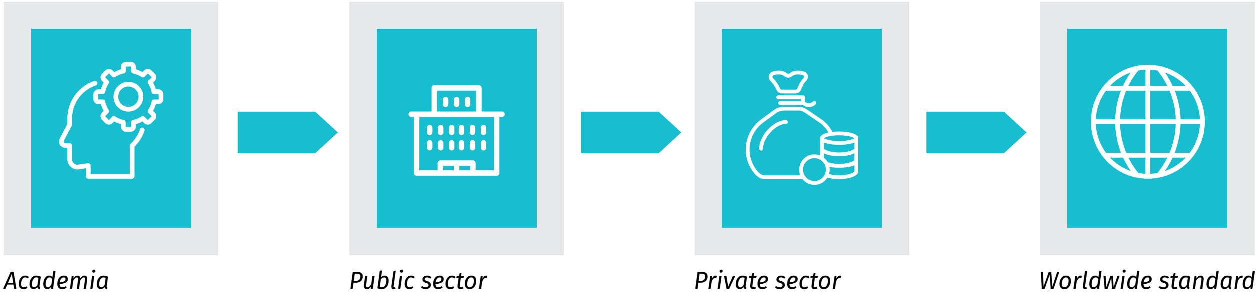
From explorative interviews there seems to be a demand for a tool that helps professionals in the data science field to guide the less well versed in statistical modeling.

Looking at the three main markets – academia, the public sector, and the private sector – a reasonable estimation is that academia is most likely to adopt first, since the framework is mimicking academic practices used for peer review and is beneficial without much overhead.

The public sector is likely to follow suit, because entities here are usually required to engage in some form of transparency, and have to account for their policies and outcomes.

The private sector might be reluctant to accept the scrutiny and overhead of such a framework at first. However two approaches are likely to help adoption here: public tenders and Corporate Social Responsibility (CSR) programs. Tenders from organizations following the framework might require companies in the private sector to comply. And a CSR program that does not account for auditing AI, ML or similar technology should trigger increased scrutiny from the broader community.

A future possibility is that using such a framework becomes mandated by law. In that case it might be that certain aspects of the framework that are now open to the preferences of client and maker will be more predefined.



Academia

Public sector

Private sector

Worldwide standard