MARi for Researchers

Tools for Acquiring and Using Properly Anonymized Datasets with Strong Provenance and Minimized Bias to Support Peer-Reviewed Research

Summary

Academic and commercial research often requires large amounts of data about individuals, and that data must be properly acquired, anonymized, aggregated, and attributed in a manner that provides both credibility and repeatability. The MARi platform eases the burden of such tasks by providing a unique opportunity to the research community that allows access to large collections of personal data spanning a wide variety of data domains. Researchers can use MARi to construct data snapshots for collaborative or competitive team-based research projects, or they can use live MARi data to access up-to-date streams of data being captured by numerous applications throughout the MARi ecosystem in order to create new predictive models that add value to the descriptions of MARi users. MARi has been built from the ground up to support a variety of actors, each of whom has a unique place in the MARi ecosystem. The researcher is a fundamental actor, and as such MARi takes the role of the researcher very seriously. Support for the researcher has been a primary concern, and will continue to be as the platform develops. This introduction to MARi from the researcher's point of view provides an insight into how the MARi platform can benefit a research team. The MARi team looks forward to the interesting innovations and collaborations the platform will bring to the academic world.

Introduction

Among the many challenges posed to researchers, a primary hurdle includes the acquisition of a statistically significant dataset to test hypotheses on. After positing a theory, the researcher must go through the often rigorous task of acquiring data that has been properly anonymized, has strong provenance, and minimized bias. Only then can the researcher's findings be corroborated and peer-reviewed. The MARI Platform provides a unique opportunity to the researcher by allowing access to a vast array of personal data submitted by an authorized user through both academic and commercial channels.

The MARi Platform

MARi behaves as an ecosystem, a repository that allows the exchange of data between many actors: end-users, developers, organizations, and researchers can all operate in the same space. At the core of MARI is the personal attribute (PA) each attribute behaves as a source attributed time series of data that describes some aspect of a user. The breadth of PAs is vast, covering measurable data points such as height and weight, testable data points such as mathematical skills, and even lifestyle data points such as books read and restaurants visited.

All PAs are associated with a user using this common link, correlations between PAs can be identified. New values are written to PAs frequently, and defining entirely new PAs is encouraged yet moderated to minimize noise. As such, the MARI system is an ideal repository for robust and well-structured data the perfect starting point for deep and impactful research.

The Researcher's Role

Each actor has a different role in the MARI system developers create applications that read and write PAs as authorized by a user. Organizations may capture large amounts of PA data through private programs and associate them with users. The researcher's role is unique PAs are sifted, sorted, and categorized in an experimental environment. The researcher creates predictive models that can assess patterns in the data and automatically produce predictive PA values.

The predictive models created by the researcher can then be made available to other actors at the researcher's discretion. If a researcher chooses to publish their model, the PA values it generates will be available for use by other actors in the system, providing an opportunity to legitimize the researcher's work through practical application and peer review.

Benefits to the Researcher: Summary

MARI provides an array of benefits to the researcher, including access to large amounts of data with invariant bias and strong provenance, an experimental sandbox, an easy-to-use API, visualization and classification tools, and a platform for storing and publicizing results. MARI aims to be more than just a repository for data as adoption of the MARI platform expands amongst all actors, the academic acceptability of the platform and the data it provides will grow. Ultimately, a strong sign of success would be a decreased emphasis in the defense of input data to an experiment, and increased emphasis on the findings the author’s justification can be handled by citing a MARI data set.

Benefits to the Researcher: Data

Statistically Significant

One of the most challenging aspects of observational science is the collection of data it can be extremely hard to find or create a statistically significant data set that captures the variables of an experiment while eliminating bias. MARI provides a significant boost to the researcher looking for comprehensive and sizable datasets. The researcher can quickly ascertain the quantity and quality of values associated with a PA, and determine if the sample size is significant. Cross products of multiple PAs can also be rapidly examined, and experimental hypotheses can be tested and validated with expediency.

Up-to-Date

MARI does not provide just a one-time view into a data set the data in MARI is constantly in flux. New PAs are added, PA values are being written to, and new users are joining regularly. This means that the PA data is up-to-date. Experimentation can be done in near real-time if desired, or on a recent view of the data. Of course, as research dictates, a historical view of the data can also be used.

Periphery Data

In addition to data series over time, MARI opens the researcher to a variety of feature-rich periphery data sets. Research may begin with a hypothesis regarding certain PAs due to the nature of PA values and their association to users, the researcher gains access to a large collection of PAs, many of which may seem unrelated during the initial hypothesis but are discovered to have some correlation with the experimental data. Had the researcher constructed a data set from scratch, or obtained a limited data set from elsewhere, this periphery data would not be available to investigate.

Anonymity

Of primary importance to all actors in the system are privacy and anonymity. MARI abstracts the data presented to the researcher from an actual user, while still allowing the researcher to explore other related PAs for the user. In essence, the researcher can see all data related to a user, and know with certainty that the data belongs to exactly one user, without knowing who the user is. Further, the researcher can add PA values back to that user, again entirely anonymously. At the discretion of the researcher, those PA values can even become available directly to the user, all while still maintaining a divide between the researcher and any personally identifiable information.

Invariant Bias

In addition to anonymity, a feature of data collection that is of paramount importance is the identification and removal of bias. While MARI does not, per se, remove all bias inherent in any data set, it does provide data with invariant bias. In other words, data collected by MARI has the same bias for one researcher as it does for another. Regardless of the origin of the data, it can be used by multiple researchers who can demonstrate both corroborative results or (in the case of competition) improved predictions.

Provenance

To assuage concerns of any given bias, each value of a PA has complete provenance associated with it. Any actor that writes to a PA will sign the value written the researcher can investigate these signatures, and identify ones that come from more trusted sources (perhaps other researchers, rather than an arbitrary mobile app). PA values can be requested only from certain sources if desired, providing stronger experimental footing and increasing the trustworthiness of the evidence in question.

Benefits to the Researcher: Tools

API

While the data set provided by MARI is of paramount interest to the researcher, MARI also provides a series of tools and systems designed to further lighten the load required to begin testing a hypothesis. Of most importance, MARI offers a robust API designed to help the researcher quickly connect to MARI's data and begin reasoning over it. Rather than providing a flat data file with cryptic or absent parsing instructions, MARI's API allows the researcher to very rapidly and easily access the data they are most interested in.

Existing PAs

Likely, the researcher is interested in a particular subset of PAs. The key to the success of MARI's data model is clean and accessible data. As such, the addition of new PAs is a moderated activity (to avoid duplication). This provides a strong benefit to the researcher: once a PA of interest is located, it is guaranteed to contain the comprehensive scope of all user data that fits its semantic description. Consequently, locating a PA of interest is extremely important. MARI provides a variety of tools, including natural language searching, to identify existing PAs or to help the researcher point out where a new PA is needed.

Visualization

As the researcher begins to explore the data encapsulated in MARI's PAs, they can use some of the built-in visualization tools to get a better feel for possible intersections or points of interest. The MARI platform contains core visualization systems out of the box in addition, using the API, developers and researchers can extend MARI's visualization capabilities (and can make these extensions available as desired).

Classification

Hand-in-hand with visualization is the ability to rapidly test proven machinelearning techniques (such as clustering and classification). MARI intends to make a large collection of classifiers available to the researcher in the near future in a very user-friendly interface. These classifiers can be used during the hypothesis construction phase, as well as during experimentation (and possibly included in the final functional model if desired).

Data Storage

As the researcher produces results using their functional model, this data should be stored somewhere. MARI provides builtin data storage for all users: from the researcher's point of view, PA values were placed there by someone utilizing MARI's data storage (be it a user directly, an app developer, or another researcher). The newly produced values can, in turn, be stored under a user's PA, making those values available to anyone with rights to access the data.

Data Snapshots

As the data points in PAs are in constant flux from new and incoming values, it can be difficult to nail down a repeatable dataset for experimentation. MARI has solved this problem by allowing the researcher to create (and share) data snapshots. Snapshots serve as a window or view into the data that is frozen in time, and may have been created with a variety of other filters (including user filters, PA provenance filters, or value filters). Once the snapshot has been created, it can be considered a static collection of MARI data that operates identically to the complete MARI data model, but without change. The researcher can then use this data to experiment with the full expectation of repeatable results.

Experimental Sandbox

All of these tools, and others, combine to create an experimental sandbox. MARI has been designed with researchers in mind, so the policies and tools in place have been created to maximally benefit experimentation. Starting with a solid and accessible data set, the researcher can easily, visually explore the data and identify possible hypotheses. The data can be frozen and experimented with using an intuitive API, and the resulting model can be published for use by the community.

MARI Researcher Use Cases

As a researcher, MARI is an instrumental platform for hypothesis creation and detection. While the creative uses for MARI are boundless, we highlight three likely scenarios for how the researcher can interact with the MARI environment.

Experimentation

Experimentation using MARI most likely occurs using existing MARI data. The researcher begins by exploring the PA system they dig around looking for PAs of interest to their research. This exploratory phase uses MARI's builtin PA tools, such as a natural language search interface, to locate the desired PAs.

Experimentation (Continued)

Once the PAs have been found, the researcher can then load those PAs into the existing visualization and classification tools. These tools, and others, are designed to help the researcher very rapidly identify areas of interest. The researcher should be able to quickly and easily get a feel for possible correlations amongst PAs, and can build a hypothesis, or corroborate an existing hypothesis rapidly.

Next, the researcher can take this hypothesis and begin a more rigorous experiment. Likely, the researcher will want to create a snapshot of anonymized PA data over an appropriate user population this will involve identifying sufficient provenance (perhaps the researcher only wants data from other researchers, or a particular institution), as well as selecting a large enough sample size and appropriate timeframe constraints.

With the data snapshot constructed, the researcher can now begin to build the experimental tools using MARi's API. The resulting functional model can now be tested against available data from MARi. At this point the researcher can opt to iterate on the functional model, tweaking the formulae and retesting the hypothesis. When the researcher is satisfied with the model, it can be made available and/or the results can be calculated for publication or report.

Alternatively, the researcher may come to MARi with a dataset already in hand. In which case, the first task involves aligning the existing data with MARi PAs. Again, using the toolsets available, the researcher can conveniently locate existing PAs that match their data points. Should a PA not yet exist in the system, the researcher can request the PA be added this is a moderated task to guarantee clean and consistent data.

Once the PAs have been located or created, the researcher must now align the individuals in their existing data set with MARi users. If this data is available as an effort of the study, then the task is trivial (each MARi user has a unique ID that can be used) if this data is unavailable then the researcher can create users unique to their dataset. It should be cautioned that this will result in sparse connections to other PAs, as the users created will only have the PAs provided by the researcher.

After the user alignment has occurred, the researcher can import their data this can either be done in a batch process, or may be streamed in as made available. As an example of streaming, the researcher may opt to set up a testing system for developing math skills. The testing system can, using MARi's API, require the user to log in prior to testing. As each test is conducted, the results can be recorded immediately to the appropriate PA. After all tests are complete, the dataset is ready to be reasoned over.

Experimentation (Continued)

Once the researcher has finished importing their external data, the experimentation can continue as before the visualization and classification tools can help the researcher explore their data, and a functional model can be created using the MARI API.

Collaboration and Competition

MARI enables researchers to collaborate as well as to construct friendly competitions. When collaborating, researchers can create a snapshot of the data to form the basis of their experimental discussions. Functional models created by any collaborating team member can be hosted by MARI their results can be tested and explored by all team members by utilizing the advanced permissions functionality. The models need never be made publicly available, but they can be shared with selected entities.

Researchers can also collaborate with commercial entities and developers with a popular app that allows users to generate PA data that can integrate predictive models. The researcher can partner with the developer to create a better user experience. The user will continue to fuel directed data for the researcher, who in turn provides feedback through prediction, delivered via the app.

Competition can be organized in a variety of ways. Likely, the organizer will create a specific data snapshot for use in the competition. This data snapshot will serve to level the playing field all researchers will be using the same inputs, allowing the dominant theory to stand out. The organizer will also likely withhold a certain additional data snapshot as a scoring dataset.

Researchers can use the available snapshot to create and test theories, and eventually produce functional models the organizer can then test each of those functional models on the scoring dataset. Similar to collaboration, MARI can host the functional models, and they can be privately shared with the competition organizer, or with the entire competitor set if desired.

Peer Review

The MARI platform provides strong support for peer review of work done by the researcher. By making both the data and functional model available to the peer reviewer within the scope of the MARI platform, the researcher has provided strong evidence that work done is accurately reported and repeatable.

Peer Review (Continued)

MARI allows for version control of both the data (in the form of snapshots) and the functional model (in the form of separate model instances). The tools provided by MARI allow the reviewer to investigate the input data (verifying its consistency and integrity) and then process that data through the functional model (thus verifying the reported results). The reviewer can be reasonably assured that the process is producing the results automatically (and is not a hack) by selecting another, similar, snapshot of data (of their choosing) and verifying similarity in the results. As a result, MARI provides an unprecedented level of support for the reviewer, allowing for a level of peer review that has previously been inaccessible.

The MARi Experience

MARI has been built from the ground up to support a variety of actors, each of whom has a unique place in the MARI ecosystem. The researcher is a fundamental actor, and as such MARI takes the role of the researcher very seriously. Support for the researcher has been a primary concern, and will continue to be as the platform develops. Hopefully, this introduction to MARI from the researcher's point of view has provided an insight into how the MARI platform can benefit a research team. The MARI team looks forward to the interesting innovations and collaborations the platform will bring to the academic world.

Frequently Asked Questions

Cost

What is the cost to the researcher to access MARI data?

Currently, there is no outofpocket expense for researchers to use the MARI platform. MARI encourages research, and recognizes that the success of the platform hinges on all actors being engaged. As a result, researchers can access MARI's data, and host functional models on MARI at no cost

What will MARI do to ensure researchers are using the data seriously?

Researchers must request access to the data, which is manually approved. These requests must be accompanied by a description of the research task and an explanation of how and why the data is being used.

Data Ownership

Who owns the PA data?

An individual data point on a PA may have multiple levels of ownership. Any data that the researcher requests access to is not owned by the researcher. Conversely, any data added by the researcher is owned by the researcher. That data is being added to the profile of a specific user. The researcher may opt to share ownership of the data with the user in this case both researcher and user can access and use the data. However, as ownership is tiered, the researcher (the originator of the data) has greater control over the data. The user may not lock the researcher out of the data, while the researcher may revoke the data rights of the user.

Will there be a way to download any of the data?

Snapshots of the data may be downloaded, rather than retrieved through the MARI API only for published data sets. A data set may be published if the owner opts to make it available for download.

Model Ownership

Who owns the functional models?

The researcher maintains complete ownership of their models. MARI will provide hosting for the model and access to the data. At the researcher's discretion, the model may be made publicly available or privately shared. In all cases, the researcher still owns the model.

How will others (e.g., app developers) use a researcher's model?

In the MARI platform, the output of a model is a value (or set of values) for a PA. The prediction a model produces is written to a PA and associated with a user (whose input values fueled the predictive output). A model is "used" by an external party by having the party request the value of the PA. PAs have associated provenance, so the requester can filter the array of values by the model that produced them. Functionally, the requester has asked for a model to be run over a user, and is given the result in practice, the technical details vary, but the result is the same.

Model Ownership (Continued)

What is the benefit to the researcher to share their model?

The benefits to the researcher will vary depending on the researcher's goals the research will be established by publishing to the MARI platform; it may be used by external parties and validated by peers, users, and developers. MARI encourages publications of models as the entire ecosystem will thrive as a result more models means more adoption, and that means more data for existing and new models.

Model Privacy

Can researchers make their models private or invite-only?

MARI provides a complete permissions set for data and models. The researcher can make their model completely open, completely closed, or available to a whitelisted set of users at their discretion.