Queriosity: Automated Data Exploration [Vision]

Abdul Wasay
Harvard University
awasay@seas.harvard.edu

Manos Athanassoulis
Harvard University
manos@seas.harvard.edu

Stratos Idreos
Harvard University
stratos@seas.harvard.edu

***Abstract*—Curiosity, a fundamental drive amongst higher living organisms, is what enables exploration, learning and creativity. In our increasingly data-driven world, data exploration, i.e., making sense of mounting haystacks of data, is akin to intelligence for science, business and individuals. However, modern data systems – designed for data retrieval rather than exploration – only let us retrieve data and ask *if it is interesting*. This makes knowledge discovery a game of hit-and-trial which can only be orchestrated by expert data scientists.**

We present the vision toward Queriosity*, an automated and personalized data exploration system. Designed on the principles of autonomy, learning and usability, Queriosity envisions a paradigm shift in data exploration and aims to become a a personalized "data robot" that provides a direct answer to *what is interesting* in a user's data set, instead of just retrieving data. Queriosity autonomously and continuously navigates toward interesting findings based on trends, statistical properties and interactive user feedback.

I. Introduction

The Growing Data Stacks. The past decade has seen an explosive trend in collecting, curating and consuming data. Data originates from a growing number of diverse sources which range from smart watches to the Large Hadron Collider, and gets piled up in data banks. In these mounting haystacks of data, needles are few and far between and with the prevalence of paradigms such as data-intensive science [3], Internet of Things [9] and information governance [7], useful knowledge is getting even more dispersed in data; this creates the need to *sieve* these data haystacks for useful knowledge [4].

New Needs, Yet Old Systems. Data exploration, the process of sifting through data to extract meaningful information, is inherently hard as we do not always have a clear idea of what we are looking for [4], [11]. Data exploration is analogous to the search for an *interesting haystack*, rather than a needle, in bigger multiple haystacks. In this context, data systems play a fundamental role; they allow us to store and retrieve data. Existing data systems, however, are ineffective for haystack-in-a-haystack queries. From a user's point of view, exploratory data analysis on existing data systems is an excruciating game of hit-and-trial, which involves several iterations of data retrieval and analysis to decide what is interesting. In addition to this, exploring huge domain-specific data sets requires a rare combination of data systems expertise and domain-specific knowledge. To bridge the gap between data systems and data exploration requirements, larger institutions dedicate a lot of resources to data exploration whereas smaller institutions, with less resources, end up wasting a lot of time to extract knowledge.

*A portmanteau of query and curiosity.
Website: http://daslab.seas.harvard.edu/queriosity

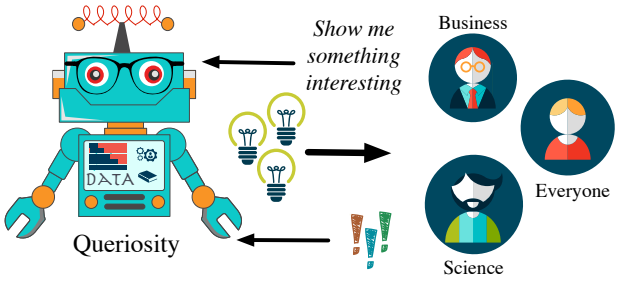

Fig. 1: Queriosity lets you ask what is interesting in your data set.

Queriosity – A Paradigm Shift. We present the vision of a data exploration system designed, from ground up, to act on the principles of autonomy, learning and usability. We call this data system Queriosity, a portmanteau of query and curiosity. With Queriosity a paradigm shift in data exploration is proposed, as depicted in Figure 1: users can ask "what is interesting" in their data set, instead of first retrieving a part of it and then asking "is it interesting", usually in the form of a complex data analysis task. Queriosity sets to work as soon as data is available. By automatically ranking statistical attributes of the underlying data set, Queriosity discovers interesting features and presents them to users in a consumable manner. As users observe results and provide feedback, Queriosity learns and adapts its strategy; it is a highly personalized data system which despite being guided by user's interest, requires no active supervision.

We envision Queriosity finding application in virtually all domains as a personal *data scientist* that assists businesses, scientists, and people in their every day lives, who try to make sense of the data around them. For example, below we describe how we envision the impact of Queriosity in two drastically different data-driven scenarios.

Large Data Colliders. Each of the 600 million collisions per second in the Large Hadron Collider aggregates to 30 petabytes of data annually. This is sifted through by a large team of scientists to determine if the collisions resulted in any interesting physics. Imagine a system which does so

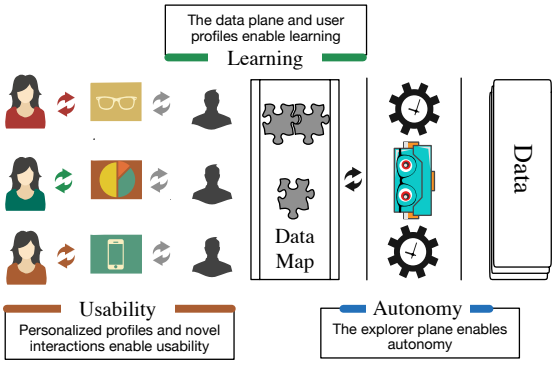

Fig. 2: Each of the principles translate into concrete components of Queriosity, revealing opportunities for innovative research to design an automated data exploration system.

automatically and interactively, freeing numerous man-hours. This enables scientists to design future experiments instead of deciphering those of the past.

Data-Intensive Living. Data generated by smartphones, sensors as well as personal medical and exercise history gets collected and curated. Another example is personal DNA sequencing. In the near future, as the sequencing process becomes affordable, every individual will be able to have access to a database of their genetic make-up. Imagine a system that weaves these scattered threads of information into a complete tapestry of knowledge about one's physical health and enables us to make decisions about our health and lifestyle based on data and not just intuition.

Curiosity to Queriosity. Exploration and curiosity are widely studied notions in artificial intelligence and machine learning within computer science, as well as in other fields such as neuroscience and behavioral psychology. The major focus has been to study and design systems that *make sense* of an *alien* and *dynamic* environment. The growing data landscape presents analogous problems in a the context of big data processing. Petabytes of data collected everyday represent an *alien* landscape for both users and conventional data systems. We find analogy between these distinct yet related paradigms. Queriousity heralds a paradigm shift in data exploration where the data system is designed to be *curious* i.e, to inherently value and conduct data exploration.

Contributions and Roadmap. The way to Queriosity is paved with a number of design and research challenges. In Section II we present the necessary research classified by the three design principles and we give concrete examples for each research theme. In Section III we present related research that serves both as inspiration and an initial toolbox towards the vision of curious data systems, and in Section IV we conclude.

II. Toward Queriosity

In this section, we discuss the roadmap toward curious data systems. The design of Queriosity takes a system-centric approach to data exploration, being drastically different from existing data systems. Queriosity's design rests on the principles of autonomy, learning and usability. Stemming from these principles are numerous research questions. For instance, quantifying interest and relevance, incorporating user interaction to inform unsupervised exploration and combining all these features in a data system that explores independently and communicates with the user intuitively are some of the basic challenges – Figure 2 depicts a high level view of how the basic concepts drive the design and research path towards Queriosity. For example, to achieve autonomy, a system should be designed, from the ground up, to work without human intervention. To achieve learning, a system should be designed, from the ground up, to autonomously search a big search space while continuously logging and taking into account incoming user preferences. To achieve usability a system should be designed, from the ground up to provide results in the form of interactive insights. The remainder of this section provides more details for the research path toward Queriosity along each of the design principles.

A. Autonomy

Haystack-in-a-haystack data exploration results in search space explosion and it can be better served by systems performing offline, automatic exploration. Based on this observation, Queriosity, at its core, is envisioned to require very little human supervision as it *provides* and not *demands* guidance during the data exploration process.

Queriosity explores data continuously and requires no active supervision.

Conventional data systems are designed to work under human supervision – a major bottleneck in conducting large scale data exploration. While long running and batched jobs are common in modern data systems, users need to know what they seek and define it prior to runtime. This dependance on human supervision makes data exploration using conventional data systems ineffective, as by definition in a data exploration scenario we do not know what we are looking for until we find it.

Never Stop Exploring. Queriosity removes this bottleneck in the data exploration process by introducing an *exploration plane*. The exploration plane takes charge of going over the data once the system is idle and users are not active. For instance, the exploration plane uses past user decisions and analyses to anticipate future exploration patterns. It can then interact with the rest of the system and orchestrate a set of exploratory operations such as discovering interesting statistical properties or exploring other areas of the data set. As a result, next time when users are ready to receive information, Queriosity will be in a better shape to provide it to them.

The research goal is to design and efficiently implement exploratory *consciousness* in the system that leverages every available system resource and opportunity for data exploration. First, we design storage and access methods such as lattices of statistical information on the existing data set. Second, in a multi-user environment, Queriosity aggregates information as a user operates on it. This information is used to sieve out common trajectories; Queriosity continuously exploits offline time (or idle CPU cycles) to optimize access and computation along common paths. Third, given resource constraints (such as energy or time), Queriosity automatically determines the optimal set of statistics which maximizes a user-defined or inherent utility function.

Quantifying Relevance. An unsupervised exploration paradigm, such as Queriosity, requires a strategy to rank observed data in its order of relevance, which serves as a way of validating its exploration strategy. For instance, Queriosity exploring seismic activities of an area should rate a series of spikes differently from a series of constant values. Relevance tends to be domain-specific, as a result, a major research challenge is to devise a scheme that quantifies relevance with no or small prior knowledge of the data set. Further, this scheme should be able to incorporate varying levels of domain-specific information, in case it is provided by users. In Queriosity, this scheme could be in the form of a navigable lattice of statistics for various data regions. These statistics can be used to infer the rank of a particular data region. A domain-expert specifies what statistics to calculate and how to compose them for ranking purposes. This has repercussions from a systems perspective in terms of efficiently storing and accessing such ranking metadata. For instance, what representation of the meta-data is most effective to directly drive the explorer plane and how can it be represented in a concise manner to effeciently utilize the memory resource.

B. Learning

The effectiveness of exploratory systems lies in their ability to balance exploration with exploitation of acquired knowledge. Statistical information and user-input will serve as a feedback loop to improve Queriosity's ability to explore.

The more Queriosity explores, the better it becomes at exploring.

Never Stop Learning. Queriosity designed as a system that learns both from user interactions and explored data, requires a formulation of *learning* in the context of data exploration. A typical unsupervised learner proceeds by observing its environment, trying out *actions* and accumulating *rewards*. An unsupervised learner with no prior knowledge about the surrounding environment tries out a set of, possibly, randomized actions and converges to a near-optimal exploration strategy. As a first step, these actions as well as associated rewards need to be defined in the context of data exploration. Based on this formulation, strategies that require the least amount of resources to converge can be formulated and implemented.

Building a Data Map. Unsupervised exploration coupled with user input, results in a progressively improved understanding of the underlying data set. Queriosity retains this understanding in the form of a global *data map*, composed of metadata and associated access structures. The metadata is a concise representation of metrics such as statistical properties and user interest gathered from both unsupervised exploration and user interaction. In a multi-user environment, the data map serves as a central repository of explored information and data insights, which can be shared across users.

Personalization. Queriosity is a data system that interacts with multiple users having, potentially, different interests and focus areas within the same data set. Queorisity attaches a profile to every user, which is a set of metadata comprising of user-specific information such as area of focus (within the data set), statistical properties of relevance, and a compressed provenance of user interaction revealing past exploration patterns. The local user profile serves as an entryway into the global data map which sits in between the data and the profiles. When a user interacts with Queriosity, the interaction is handled by their profile which delegates it to the global data map. This two-tiered approach yields a data system that is highly personalized yet efficiently shares information and computed metrics across various profiles. In this case, the global data map serves as a central repository of user-independent statistics whereas a user's profile maintains state of their interest.

Learning in Queriosity happens along two dimensions. Learning from user interactions ensures a progressively personalized data system and is important because the user's past interactions with the systems are used as an information to facilitate their future interactions. Along the other dimension, learning about the data results in a richer understanding of the data set and is important because it makes the system more efficient.

C. Usability

To enable users to draw meaningful inferences from explored data, Queriosity aims to provide extracted information in a consumable manner. The interface between user and the data system lies at the core of our data exploration paradigm. It is what enables Queriosity to inform the user about the underlying data set.

Queriosity engages the user to answer current questions and exposes future exploration directions.

Conventional data systems provide a linear retrieval-based interface where the burden to ascertain extremely specific queries rests solely on the user. Recent research advocates for an abstraction on top of conventional systems that guides the user or provides an intuitive interface for displaying results. On the other hand, Queriosity, lays emphasis on user interaction as a design principle rather than an afterthought. Not only does this asks for novel paradigms of interaction, but also requires systems where user interaction is not just another layer of abstraction but, rather, is built in the system to provide fast, efficient and intuitive user-experience.

Insights Not Results. As we move to a paradigm where insights, rather than query results, are output from a data system, we need to rethink how to store, access and update data. Relevance of data can inform data organization, and data structures can be optimized for extracting statistical information about the data and not the actual values. For instance, the internal nodes of a tree-based index can store metrics about the underlying data that can be used to decide which direction to explore in the tree-based structure. As an example, if a collection of statistics exists on top of the data set, we can then device interaction paradigms where the user queries for statistical properties and can directly ask for functional dependencies amongst the stored data set.

Seeing is Believing. Recent advancements in human-computer interaction such as touch screens, virtual reality, and gesture control have revolutionized the way we interact with computer systems. Data exploration naturally presents cases where these techniques are useful. Queriosity can both benefit from and expand on research in this direction. For instance, Queriosity

can combine data exploration with virtual or augmented reality devices to provide real-time insights to a scientist going over a large collection of scientific data. For example, insights from astronomical data can be visualized and related with coordinates in space. As an astronomer looks into the open skies, insights, such as time series of energy, gathered from years of telescope observations can be superimposed to enable real-time exploration. A major reason why such applications have not yet surfaced is the lack of data systems that can keep up with their information needs in terms of speed and variety. *How should one combine virtual and augmented reality with data exploration* is a key research challenge for Queriosity. The interface between Queriosity and the applications on top should be designed in a way that exposes varying granularity of information about the underlying data.

III. Related Work

Queriosity draws inspiration from a large range of related work in data systems design, artificial intelligence and machine learning. Here we briefly discuss related work[†], as well as how Queriosity brings new opportunities and research challenges to past work.

Data Exploration Efforts. In the data systems community, there has been significant work toward approximate processing [1] as well as work toward systems that are tailored for exploration [5]. Such work allows for fast approximate searches on big data as well as proposes novel systems that are designed for visual analytics, gestural interfaces and algorithms tailored for exploration. Queriosity is inspired by this work and moves it one step further by introducing the notion of systems that are tailored for automatic exploration as opposed to systems that are guided by the users queries.

Data Summarization. Recent work in artificial intelligence has led to automatic data summarization [10] where input data is classified using non-parametric regression models to generate a concise natural language report [2]. Automatic data summarization enables a synoptic understanding of the underlying data, however, being a static approach it can neither take into account user interests nor sieve through the data set for interesting patterns automatically. Queriosity automates the entire process of data exploration and not just one specific operation.

Reinforcement Learning. Efficient exploration is widely studied in reinforcement learning, a type of unsupervised learning where an agent learns from experience [6]. Various reinforcement learning techniques have seen successful applications in games, economics as well as robotics [8]. While the underlying ideas explored in these works are of direct relevance to Queriosity, the context is very different. Applying reinforcement learning to data exploration systems present new challenges which can be tackled by a combination of data systems understanding and reinforcement learning theory.

[†]Due to limited space, we mostly cite surveys that represent significant collections of work in areas relevant to Queriosity, instead of citing individual technical papers.

IV. Summary

We introduce Queriosity, a data exploration system to be designed on the principles of autonomy, learning, and usability. Initial data exploration efforts and modern data systems offer technical inspiration for Queriosity, however, today the data exploration workflow begins by retrieving a subset of the data, and then asking "is it interesting". Contrary to this approach, Queriosity turns the table around and allows users to ask "what is interesting" from the underlying data set, and presents this information to users in a consumable manner. Queriosity aims to achieve this by introducing the notion of automatic data exploration that requires minimal user supervision and explores the data set in an independent, increasingly, smarter way.

From a research point of view, there is room for advancement spanning multiple Computer Science areas such as systems, databases, machine learning, algorithms as well as other disciplines such as statistics. From an application point of view, auto-exploration systems allow data scientists to focus more on observing data patterns that are automatically created as opposed to spending most of their energy trying to tune systems and putting together workflows to explore the data. Similarly, we can think of exciting applications where auto-exploration data systems learn from multiple data scientists so they can leverage collective experience. Overall, Queriosity opens an exciting research path to rethink several problems and solutions in the big data era as well as to redesign the toolkit of modern data scientists.

References

[1] G. Cormode, M. N. Garofalakis, P. J. Haas, and C. Jermaine. Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. *Foundations and Trends in Databases*, 4(1-3):1–294, 2012.

[2] D. K. Duvenaud, J. R. Lloyd, R. B. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 28, pages 1166–1174, 2013.

[3] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[4] S. Idreos. *Big Data Exploration*. Taylor and Francis, 2013.

[5] S. Idreos, O. Papaemmanouil, and S. Chaudhuri. Overview of Data Exploration Techniques. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Tutorial*, 2015.

[6] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4(1):237–285, 1996.

[7] G.-H. Kim, S. Trimi, and J.-H. Chung. Big-data Applications in the Government Sector. *Communications of the ACM*, 57(3):78–85, 2014.

[8] N. Kohl and P. Stone. Policy Gradient Reinforcement Learning for Fast Quadrupedal Locomotion. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2619–2624, 2004.

[9] H. Kopetz. Internet of Things. In *Real-Time Systems*, pages 307–323. Springer, 2011.

[10] J. R. Lloyd, D. K. Duvenaud, R. B. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Automatic Construction and Natural-Language Description of Nonparametric Regression Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1242–1250, 2014.

[11] C. A. Lynch. Jim Gray's fourth paradigm and the construction of the scientific record. In T. Hey, S. Tansley, and K. M. Tolle, editors, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, pages 177–183. Microsoft Research, 2009.