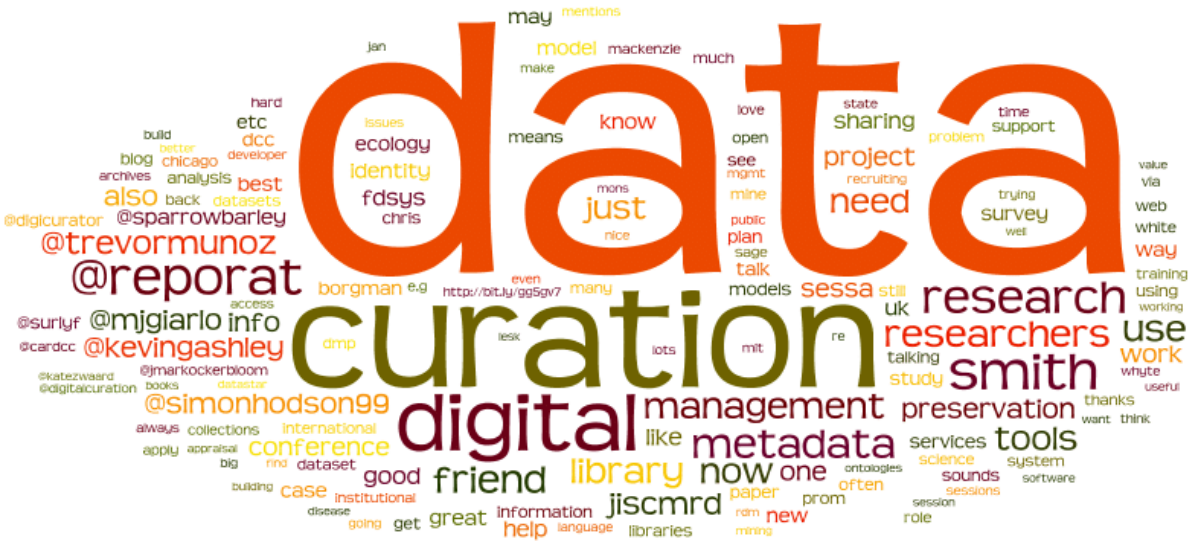
TABLE OF CONTENTS

ABSTRACT

Today in this ever-growing world having such a high population rate has led the development enormous number of data sources each having different qualities. Various organisations which are required to produce and maintain this data are needed for support by processes and technologies that allow them maintaining this enormous amount of created data. These organisations also have to make sure that the data is stored properly such at any time it can be accessed and analysed. Data curation is hence introduced in order to process, store and increase efficiency of the data. By using data curation techniques, data could be used to the maximum extent, and the speed of its elimination can be effectively slowed down. Data is collected from huge amount of data sets from conventional and electronic sources to differentiate the trends and patterns of a people's interest. That information is used by companies to improve what they know about customers' requirements and interests. The main motive of data curation is to set a certain se of questions to abstract information from the source such that this information can be useful in various motives other than its source motive for which the information was extracted.

INTRODUCTION

Data curation is a process by which the enormous data created from various data sources is to be managed and stored in such a that its usefulness and understandability is unaffected by any one in the organisation in the future. To provide these powers to the data certain steps are required to be taken such as data infrastructure, searchability, availability and preservation. One of the major motive of data curation is to enable us to discover and use the data. Requirements of data is changing over time so we have to enable data to be capable of being used in current technological requirements and future technological requirements.



One of the challenges of developing and maintaining a curation process is in making choices about where to invest limited resources and predicting likely needs for the data. This will then inform the continuum of complexity around the data, imposing factors on the data curation process based on the available resources and anticipated needs.

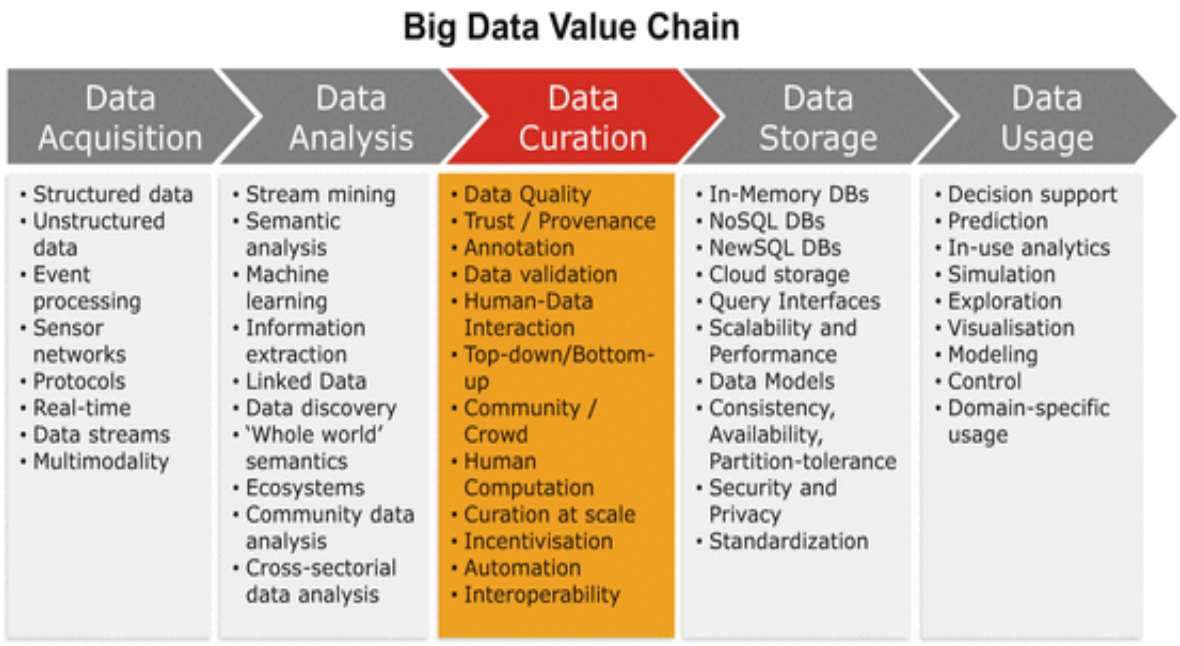
Today in this digital world has created a giant data source available for maintaining one such source is the social networking sites such as Facebook, Twitter, WhatsApp etc. These platforms gather a lot of information from the user which is required to be stored and analysed to further development of the platforms as per present scenario or interest of majority of people.

The emergence of new platforms for decentralized data creation such as sensor and mobile platforms, the increasing availability of open data on the web, added to the increase in the number of data sources inside organizations, brings an unprecedented volume of data to be managed. In addition to the data volume, data consumers in the big data era need to cope with data variety, as a consequence of the decentralized data generation, where data is created under different contexts and requirements.

Review

(3.1) What is Big Data Curation

Data curation is a process, a method of maintaining data throughout its life cycle right from starting i.e. from data creation to data storage then data analysis and when data becomes obsolete its deletion is all part of data curation.



Data curation is the process of turning independent structured and semi structured data sources into data sets ready for analytics. This process involves various steps such as: -

(1) Identifying data sources of interest as different organisations both internal and external have different requirements of data.

(2) Verification of information contained in the data is very important as there is no usefulness of false data.

(3) Cleaning of incoming data is also an important step, for example a data entry of pin code cannot have an entry of 9999999.

(4) Transforming the data is also done by various different organisations located in different parts of the world for example the organisation located in France would store the data in French language for its easy use whereas an organisation in USA would store it in English.

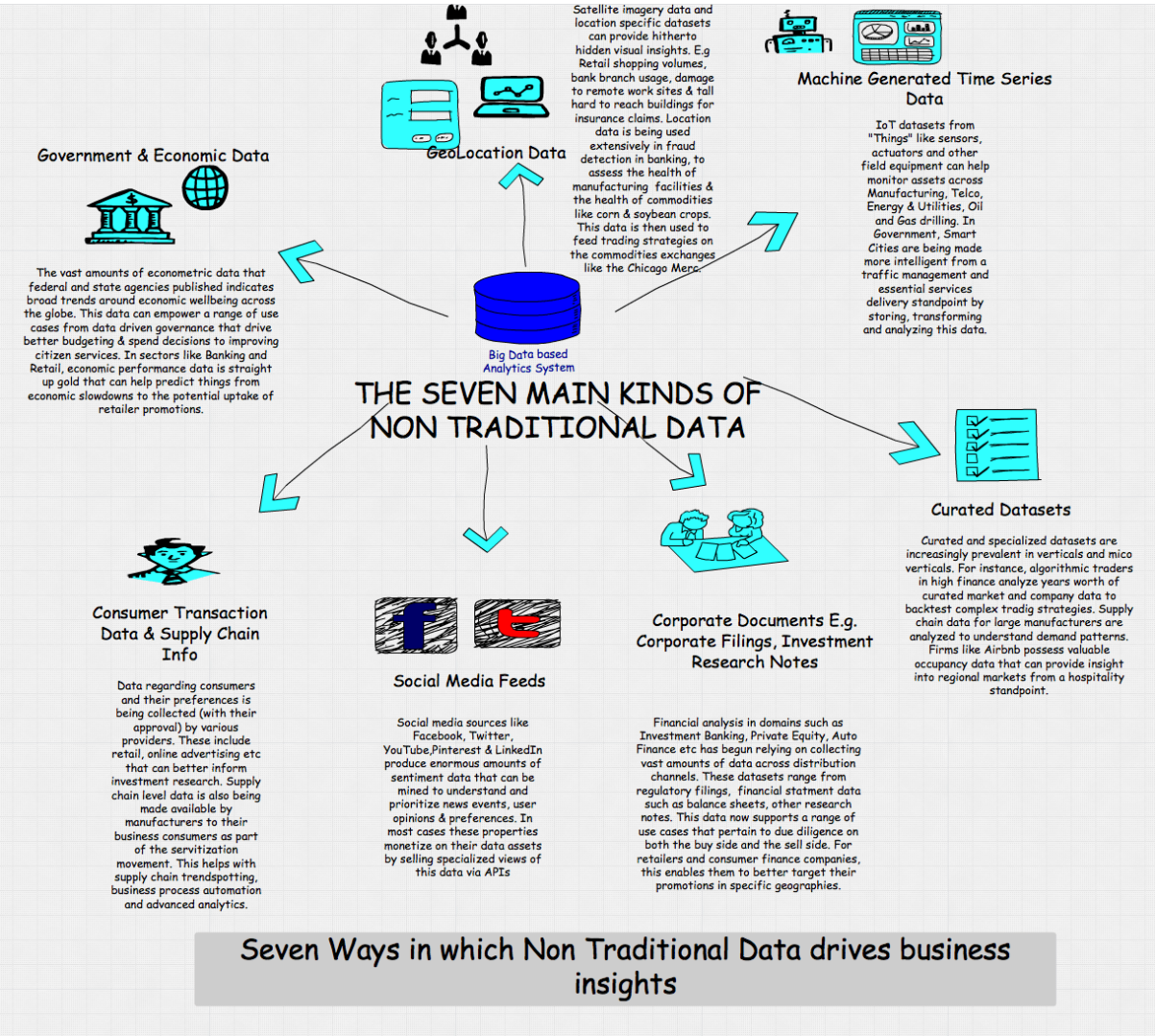
(5) Integrating the data from various data sources is also done to reduce the time consumption in its analysis

(6) Deduplicating of data is also done to remove any duplicate information from the data so as to reduce the size of storage of data.

(3.2) Key insights of Big Data Curation

Initially the idea behind the creation and requirement of data curation was from the backgrounds of government and science. The reason behind it was that the government was required to keep up the record of all of its citizens in the country in the form of Data , The maintenance and generation of data for country like India having such a high population was a big and difficult task and physical storage of such data was entirely not possible. The government is required to keep up the information of all individuals' passports, addhar card etc. This task was one of major cause that led to the development of data curation. This technological development proved to be a major helping hand in the maintaining of the data not only this much it also helped in data creation as it helped in setting up of protocols/set of questions which were required to get the minimal extraction of information but maximum possible use of it. One of the examples of this useful tool is the data of passport holders in India can be used as a helpful source for census calculation, number of people turning 18 i.e. number of voter ID card generation. This information can act as a one of data source as per the requirement. These sectors have acted as the innovators and visionaries of data curation and made it acceptable throughout the world.

Some other places where the data curation got easily accepted and evolved ever since are pharmaceutical industry which required the data curation to reduce the cost of a drug by reducing its time to market and also its discovery pipeline is made cheaper. Another industry which was influenced by this development is media organisation which required large scale unstructured data collection, to get an idea about the current demand, current popularity of a product and hence reduce the time of new product generation.

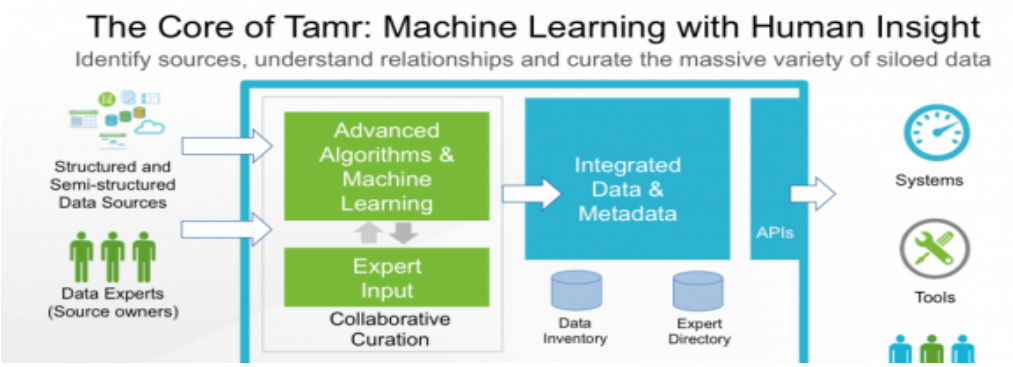


Today the target set to be achieved by data curation is to cope up with such huge data sources and set methods, protocols to extract information effectively and concisely and do not overload the information and the data to be maintained in such a manner such that its use in future also could be done effectively.

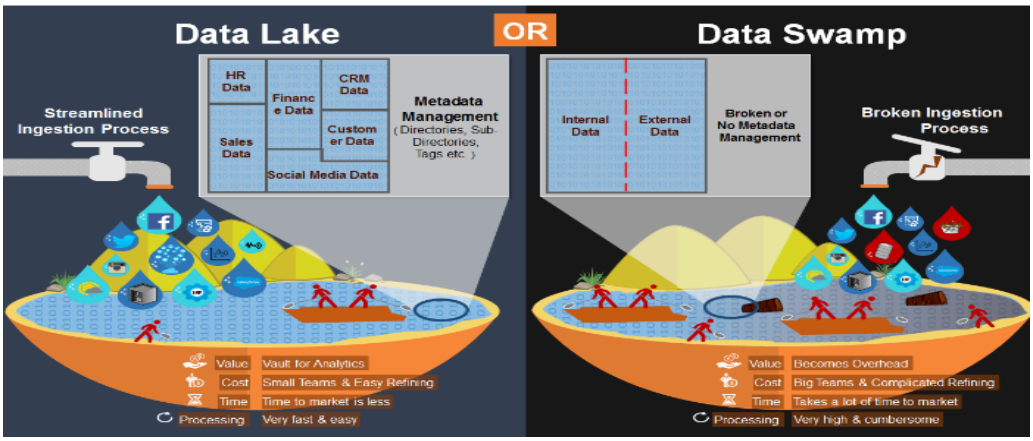
(3.3) Data Curation In Data Industry

Data curation enhances the life of the stored data such that it is available for high quality research for a long time. This technological advancement has given boost to certain number of industries with its compatibility for its use by them some of such industries are: -

(1) It has made Machine Learning more effective as we know that machine learning or artificial intelligence is the ability of a computer/device to respond to a query without the source of the query understanding weather the query was answered by a human or a device. The data curation technology has enabled the machine to access the stored data containing information about the source(user) interest, recent activities, recent searches, recent web surfing history etc. This has enabled the machine to learn more about humans as it has access over a lot of human activities.



(2) Data curation is the only quality that has been abled to differentiate or help convert data lakes into data swamps. Data Swamp is a form of data that has no management it is stored in the form of raw data with no plan of it being used in the future. Whereas Data Lake is the data stored in curated form i.e. it has been managed, arranged, analysed and is maintained for its use in the future.



(3) Data curation has also helped in educating a certain audience as during the curation of data the data is processed and is made concise it has been made quickly responsive and hence emphasise the information which is necessary and make it crisp and easy to understand for the audience.

(3.4) Data Curation Software's

Data Curation Tools: -

(1) Alation: - It is a type of tool in which any person of the company/organisation can work upon and extract the data that they want.

Key Features:

(1.1) It automatically indexes the data by its source

(1.2) It combines insights about the data

(1.3) It gathers metadata from data set which are sent up for analysis

(1.4) It has more fundamental way towards data governance

(2) Talend: - It is a type of open source data integration platform which help integrate, mask, cleanse and profile data.

Key Features:

(2.1) It is enabled to handle large number of source systems by using standard connectors.

(2.2) It offers master data management functionality.

(2.3) It provides an accurate perception of key enterprise data.

(3) Stitch Data: - It is a cloud first and developer centred platform.

Key Features:

(3.1) It is compatible with moving data into Amazon redshift, S3, Big query, Panoply, PostgreSQL and more.

(3.2) It offers effective error handling and automated error correction when possible.

(3.3) It has easy scheduling for data replication.

(4) Informatica: - It provides variety of products such as data masking, data replica, data quality and data virtualisation.

Key Features:

(4.1) It has cloud-based performance.

(4.2) It is designed to be use by non-technical users in master data management.

(4.3) It merges and cross references data from new types and sources.

(5) Ataccama ONE: - It is an AI driven data curation tool. It combines data processing engine, Machine learning, Multiple deployment options and enterprise proven capabilities.

Key Features:

(5.1) It has whole data configuration process automated.

(5.2) It has machine learning based data curation.

(6) Alteryx: - It is great at deploying and sharing analytics in a scalable manner at a very fast speed.

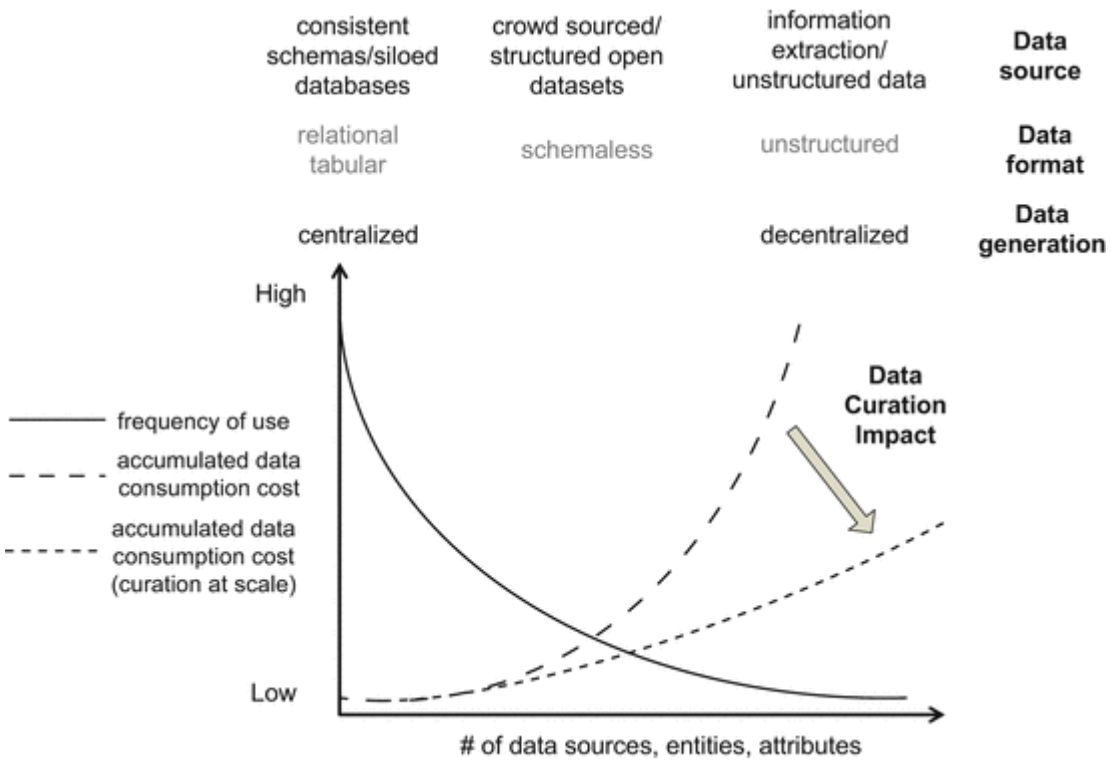
Key Features:

(6.1) It enables to draw data from multiple sources such as cloud, spreadsheet etc

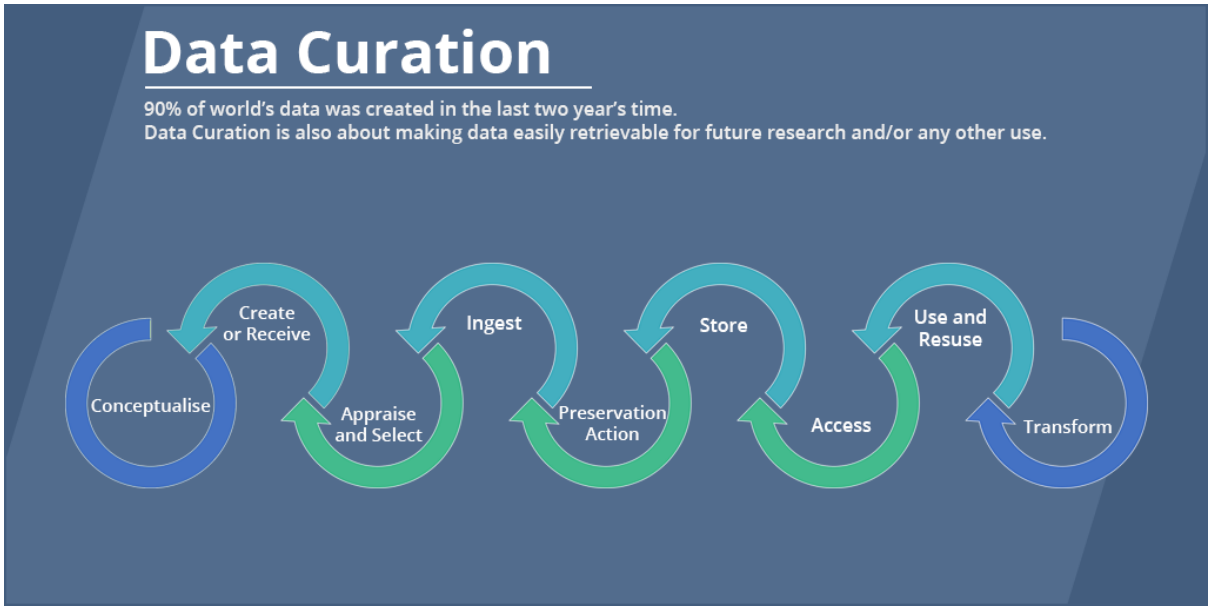
(6.2) It is user friendly as it has drag and drop interface

(6.3) It provides ability to create custom work flow

CONCLUSION



In this ever-enlarging world the demand and acceptance of data curation has increased rapidly in the past few years as it can be seen from the graph. The creation of data curation techniques is an important step for ensuring good data quality in this big data era.



It is concluded from the Term paper that "Data curation" is a powerful tool on today's emerging world it helps us to understand the Data variety and quantity pattern and analyse them to develop the techniques for and any Organisation. It also gives the analytics records which is very supportive for Marketing Agencies which will help them to target the correct audience with the majority requirement from the data created. Data curation also have important role in understanding the behaviour of the people as it takes data from all over places like Hospitals, Fitness Centre and determine the particular behaviour of the pattern of the people this helps the firms (like Pharmaceuticals)to develop the medicines/drug more effective way and help to tackle the cure of disease. In todays current scenario of COVID 19 with which the world has been hit, Data curation has helped a lot in maintain the records of all the patients and their case study which is surely going to help in its cure development. Also, the Arogya setu app which takes efficient entries from all the users to determine various quality of information for data generation which is used by the government in various aspects.

Reference

() http://www.w3.org/DesignIssues/LinkedData

() http://big-project.eu/text-interviews

() https://sanwen.net/a/asexvqo

()https://github.com/chrisquince/DESMAN

()https://www.researchgate.net/publication/280625426_Big_Data_Curation

()https://www.datasciencecentral.com/profiles/blogs/the-role-of-data-curation-in-big-data

()https://simplicable.com/new/data-curation

https://conferences.oreilly.com/strata/stratany2014/public/schedule/detail/36021