(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 April 2013 (18.04.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/055704 A1

(51) International Patent Classification:
G06F 19/12 (2011.01) *G01N 33/53* (2006.01)
G01N 33/48 (2006.01)

(21) International Application Number:
PCT/US2012/059398

(22) International Filing Date:
9 October 2012 (09.10.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/545,539 10 October 2011 (10.10.2011) US

(71) Applicant: **AYASDI, INC.** [US/US]; 330 Waverley Street, Suite 1, Palo Alto, CA 94301 (US).

(72) Inventors; and
(71) Applicants *(for US only)*: **LUM, Pek, Yee**; 330 Waverly Street, Suite 1, Palo Alto, CA 94301 (US). **CARLSSON, Gunnar**; 998 Cottrell Way, Stanford, CA 94305 (US). **SEXTON, Harlan**; 3473 Cowper Street, Palo Alto, CA 94306 (US). **SINGH, Gurjeet**; 841 University Avenue, Palo Alto, CA 94301 (US).

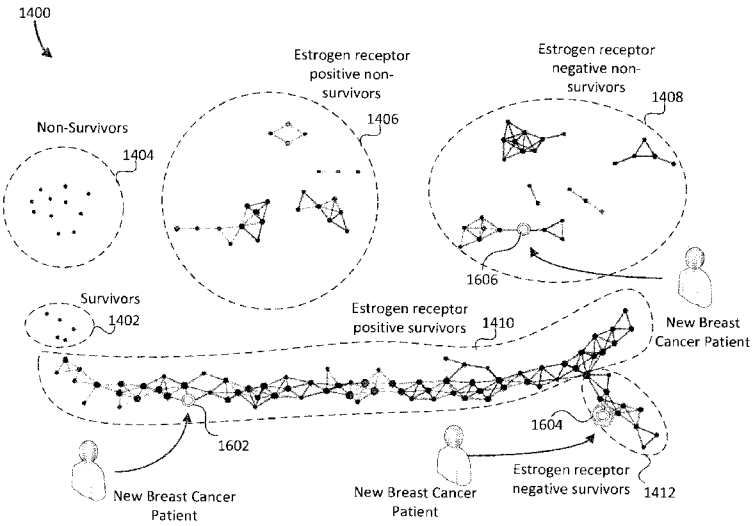
(74) Agents: **KLOKE, Daniel, C.** et al.; Sheppard, Mullin, Richter & Hampton LLP, 379 Lytton Avenue, Palo Alto, CA 94301 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) **Title**: SYSTEMS AND METHODS FOR MAPPING NEW PATIENT INFORMATION TO HISTORIC OUTCOMES FOR TREATMENT ASSISTANCE



(57) **Abstract**: Predictive visualization of patients is provided by a system including a map and a location engine. The map includes a plurality of groupings and interconnections of the groupings, each grouping having one or more patient members that share biological similarities, each interconnection linking groupings that share at least one common patient member, the map identifying a set of groupings and a set of interconnections having a set of medical characteristics. The location engine may be configured to determine whether a new patient shares the biological similarities with the one or more patient members of each grouping thereby enabling association of the new patient with one or more of the set of medical characteristics.

WO 2013/055704 A1

Systems and Methods for Mapping New Patient Information to Historic Outcomes for Treatment Assistance

BACKGROUND

1. Field of the Invention

[001] Embodiments of the present invention are directed to visualization of data analysis and more particularly to patient predictive visualization.

2. Related Art

[002] As the collection and storage data has increased, there is an increased need to analyze and make sense of large amounts of data. Examples of large datasets may be found in financial services companies, oil expiration, biotech, and academia. Unfortunately, previous methods of analysis of large multidimensional datasets tend to be insufficient (if possible at all) to identify important relationships and may be computationally inefficient.

[003] In one example, previous methods of analysis often use clustering. Clustering is often too blunt an instrument to identify important relationships in the data. Similarly, previous methods of linear regression, projection pursuit, principal component analysis, and multidimensional scaling often do not reveal important relationships. Existing linear algebraic and analytic methods are too sensitive to large scale distances and, as a result, lose detail.

[004] Further, even if the data is analyzed, sophisticated experts are often necessary to interpret and understand the output of previous methods. Although some previous methods allow graphs depicting some relationships in the data, the graphs are not interactive and require considerable time for a team of such experts to understand the relationships. Further, the output of previous methods does not allow for exploratory data analysis where the analysis can be quickly modified to discover new relationships. Rather, previous methods require the formulation of a hypothesis before testing.

[005] Etiologies of many diseases have a genetic basis. For example, many types of cancer arise when genes regulating cell growth and differentiation mutate, and the mutations are propagated during subsequent cell divisions, thereby causing uncontrolled cell growth and proliferation. Thus, techniques that measure the relative "activity" of genes (e.g., levels of

gene transcripts), called gene expression profiling techniques, can be used to assess which genes are involved in the etiology of a given type of cancer, or more generally, a disease that is caused by a genetic mutation or aberration.

[006] Gene expression profiling techniques estimate the activity of thousands of different genes simultaneously. Gene expression techniques typically measure the level of messenger RNA (mRNA) – molecules that are intermediaries between the genes encoded by DNA and proteins, the primary structural and functional components of cells – as a proxy for the activity of genes in cells under various conditions. Some gene expression profiling techniques, such as DNA microarray technologies, measure the relative activity of known target genes. Other gene expression techniques based on high-throughput sequencing technologies can measure the relative activity of any gene, including previously unidentified genes.

[007] Gene expression profiling techniques are currently used in the identification of specific types of cancer. Various cancer subtypes have been defined by the gene expression patterns, or molecular signatures, observed in various tumors. The cancer subtypes include the tissue or cell type giving rise to the tumor, and specific subtypes of cancer that arise from the same tissue or cell types. A patient's cancer subtype can thus be diagnosed when a doctor takes a biopsy of the patient's tumor and submits it for analysis using a gene expression profiling technique.

[008] Such diagnoses currently have limited therapeutic utility. It is not uncommon that the results of the diagnosis consists of a single value that may indicate a likelihood of a specific cancer. Merely identifying a cancer or tumor subtype, however, does not necessarily provide guidance to the physician on the expected outcome of a patient with a certain cancer subtype, nor the appropriate treatment for a patient with a particular cancer subtype. Currently, a patient's prognosis and therapeutic options are typically determined by a doctor, using his or her experience alone, based on the diagnosis.

SUMMARY OF THE INVENTION(S)

[009] Exemplary systems and methods for visualization of data analysis are provided. In various embodiments, a method comprises accessing a database, analyzing the database to identify clusters of data, generating an interactive visualization comprising a plurality of nodes and a plurality of edges wherein a first node of the plurality of nodes represents a cluster and an edge of the plurality of edges represents an intersection of nodes of the plurality of nodes, selecting and dragging the first node in response to a user action, and reorienting the interactive visualization in response to the user action of selecting and dragging the first node.

[0010] In various embodiments, the method further comprises saving the data in the database associated with the selected first node. The method may comprise selecting a second node and displaying information regarding the first and second node. In some embodiments, the method may comprise receiving a selection of data identifiers of the database and highlighting some of the plurality of nodes associated with the selection.

[0011] The first node and a second node of the interactive visualization may be colored differently based on a selected first function. In one example, the first function is a filter. The method may further comprise receiving a second function selection and changing the color of the first and second nodes based on the second function selection.

[0012] In various embodiments, the method further comprises receiving an interval value and an overlap percentage, re-analyzing the database based on the interval value and the overlap percentage, and regenerating the interactive visualization based on the re-analysis. Further, the method may comprise displaying movement of the interactive visualization after generation, the movement being based on visual optimization of the plurality of nodes.

[0013] In some embodiments, the method may further comprise displaying statistical information about the first node and a selected second node. The analysis of the database may be a topological analysis. In some embodiments, the analysis of the database is a nonlinear data analysis.

[0014] An exemplary system comprises a processor, an input module, an analysis module, and a visualization module. The input module may be configured to access a database. The analysis module may be configured to analyze the database to identify clusters of data. The visualization module may be configured to generate an interactive visualization comprising a

plurality of nodes and a plurality of edges, wherein a first node of the plurality of nodes represents a cluster and an edge of the plurality of edges represents an intersection between nodes of the plurality of nodes, to select and drag the first node in response to a user action, and to reorient the interactive visualization in response to the user action of selecting and dragging the first node.

[0015] An exemplary computer readable medium may comprise instructions. The instructions may be executable by a processor to perform a method. The method may comprise accessing a database, analyzing the database to identify clusters of data, generating an interactive visualization comprising a plurality of nodes and a plurality of edges wherein a first node of the plurality of nodes represents a cluster and an edge of the plurality of edges represents an intersection of nodes of the plurality of nodes, selecting and dragging the first node in response to a user action, and reorienting the interactive visualization in response to a user action of selecting and dragging the first node.

[0016] Exemplary systems and methods for predictive visualization of patients are provided. In various embodiments, a system comprises a map and a location engine. The map includes a plurality of groupings and interconnections of the groupings, each grouping having one or more patient members that share biological similarities, each interconnection interconnecting groupings that share at least one common patient member, the map identifying a set of groupings and a set of interconnections having a medical characteristic of a set of medical characteristics. The location engine may be configured to determine whether a new patient shares the biological similarities with the one or more patient members of each grouping thereby enabling association of the new patient with one or more of the set of medical characteristics.

[0017] The biological similarities may represent similarities of measurements of gene expressions or similarities of sequencing.

[0018] In some embodiments, the map is generated by an analysis server configured to receive biological data associated with the one or more patient members, apply a filtering function to generate a reference space, generate a cover of the reference space based on a resolution, the cover including cover data associated with the filtered biological data, cluster the cover data based on a metric, and display the groupings and the interconnections based on

the clusters. The filtering function may be a density estimation function. The metric may be a Pearson correlation.

[0019] The location engine configured to determine whether the new patient shares the biological similarities with the one or more patient members of each grouping may comprise the patient location engine configured to determine a distance between biological data of each patient member and new biological data of the new patient, compare distances between the patient members of each grouping and the distances determined for the new patient, and determine a location of the new patient relative to at least one of the member patients.

[0020] In some embodiments, the location engine may be further configured to compare distances to one or more of the patient members closest to the new patient's filtered biological data with a diameter of at least one grouping and to indicate that the new patient is associated with the grouping based on the comparison. In various embodiments, the location engine is further configured to determine if the distance to one or more of the patient members closest to the new patient's filtered biological data is greater than a diameter of each grouping and to indicate that the new patient is not associated with each grouping based on the comparison.

[0021] The medical characteristic may comprise a clinical outcome.

[0022] An exemplary method comprises receiving biological data of a new patient, determining distances between biological data of patient members of map and new biological data from the new patient, the map including a plurality of groupings and interconnections of the groupings, each grouping having one or more of the patient members that share biological similarities, each interconnection interconnecting groupings that share at least one common patient member, the map identifying a set of groupings and a set of interconnections having a medical characteristic of a set of medical characteristics, comparing distances between the one or more patient members and the distances determined for the new patient, and determining a location of the new patient relative to the member patients of the map based on the comparison, thereby enabling association of the new patient with one or more of the set of medical characteristics.

[0023] An exemplary computer readable medium may comprise instructions. The instructions may be executable by a processor to perform a method. The method may comprise receiving biological data of a new patient, determining distances between biological

data of patient members of map and new biological data from the new patient, the map including a plurality of groupings and interconnections of the groupings, each grouping having one or more of the patient members that share biological similarities, each interconnection interconnecting groupings that share at least one common patient member, the map identifying a set of groupings and a set of interconnections having a medical characteristic of a set of medical characteristics, comparing distances between the one or more patient members and the distances determined for the new patient, and determining a location of the new patient relative to the member patients of the map based on the comparison, thereby enabling association of the new patient with one or more of the set of medical characteristics.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024] FIG. 1 is an exemplary environment in which embodiments may be practiced.

[0025] FIG. 2 is a block diagram of an exemplary analysis server.

[0026] FIG. 3 is a flow chart depicting an exemplary method of dataset analysis and visualization in some embodiments.

[0027] FIG. 4 is an exemplary ID field selection interface window in some embodiments.

[0029] FIG. 5 is an exemplary data field selection interface window in some embodiments.

[0030] FIG. 6 is an exemplary metric and filter selection interface window in some embodiments.

[0031] FIG. 7 is an exemplary filter parameter interface window in some embodiments.

[0032] FIG. 8 is a flowchart for data analysis and generating a visualization in some embodiments.

[0033] FIG. 9 is an exemplary interactive visualization in some embodiments.

[0034] FIG. 10 is an exemplary interactive visualization displaying an explain information window in some embodiments.

[0035] FIG. 11 is a flowchart of functionality of the interactive visualization in some embodiments.

[0036] FIG. 12 is a flowchart of for generating a cancer map visualization utilizing biological data of a plurality of patients in some embodiments.

[0037] FIG. 13 is an exemplary data structure including biological data for a number of patients that may be used to generate the cancer map visualization in some embodiments.

[0038] FIG. 14 is an exemplary visualization displaying the cancer map in some embodiments.

[0039] FIG. 15 is a flowchart of for positioning new patient data relative to the cancer map visualization in some embodiments.

[0040] FIG. 16 is an exemplary visualization displaying the cancer map including positions for three new cancer patients in some embodiments.

[0041] FIG. 17 is a flowchart of utilization the visualization and positioning of new patient data in some embodiments

[0042] FIG. 18 is an exemplary digital device in some embodiments.

DETAILED DESCRIPTION OF EXEMPLARY EMBODIMENTS

[0043] In various embodiments, a system for handling, analyzing, and visualizing data using drag and drop methods as opposed to text based methods is described herein. Philosophically, data analytic tools are not necessarily regarded as "solvers," but rather as tools for *interacting* with data. For example, data analysis may consist of several iterations of a process in which computational tools point to regions of interest in a data set. The data set may then be examined by people with domain expertise concerning the data, and the data set may then be subjected to further computational analysis. In some embodiments, methods described herein provide for going back and forth between mathematical constructs, including interactive visualizations (e.g., graphs), on the one hand and data on the other.

[0044] In one example of data analysis in some embodiments described herein, an exemplary clustering tool is discussed which may be more powerful than existing technology, in that one can find structure within clusters and study how clusters change over a period of time or over a change of scale or resolution.

[0045] An exemplary interactive visualization tool (e.g., a visualization module which is further described herein) may produce combinatorial output in the form of a graph which can be readily visualized. In some embodiments, the exemplary interactive visualization tool may be less sensitive to changes in notions of distance than current methods, such as multidimensional scaling.

[0046] Some embodiments described herein permit manipulation of the data from a visualization. For example, portions of the data which are deemed to be interesting from the visualization can be selected and converted into database objects, which can then be further analyzed. Some embodiments described herein permit the location of data points of interest within the visualization, so that the connection between a given visualization and the information the visualization represents may be readily understood.

[0047] FIG. 1 is an exemplary environment 100 in which embodiments may be practiced. In various embodiments, data analysis and interactive visualization may be performed locally (e.g., with software and/or hardware on a local digital device), across a network (e.g., via cloud computing), or a combination of both. In many of these embodiments, a data structure is accessed to obtain the data for the analysis, the analysis is performed based on properties and

parameters selected by a user, and an interactive visualization is generated and displayed. There are many advantages between performing all or some activities locally and many advantages of performing all or some activities over a network.

[0048] Environment 100 comprises user devices 102a-102n, a communication network 104, data storage server 106, and analysis server 108. Environment 100 depicts an embodiment wherein functions are performed across a network. In this example, the user(s) may take advantage of cloud computing by storing data in a data storage server 106 over a communication network 104. The analysis server 108 may perform analysis and generation of an interactive visualization.

[0049] User devices 102a-102n may be any digital devices. A digital device is any device that comprises memory and a processor. Digital devices are further described in FIG. 2. The user devices 102a-102n may be any kind of digital device that may be used to access, analyze and/or view data including, but not limited to a desktop computer, laptop, notebook, or other computing device.

[0050] In various embodiments, a user, such as a data analyst, may generate a database or other data structure with the user device 102a to be saved to the data storage server 106. The user device 102a may communicate with the analysis server 108 via the communication network 104 to perform analysis, examination, and visualization of data within the database.

[0051] The user device 102a may comprise a client program for interacting with one or more applications on the analysis server 108. In other embodiments, the user device 102a may communicate with the analysis server 108 using a browser or other standard program. In various embodiments, the user device 102a communicates with the analysis server 108 via a virtual private network. Those skilled in the art will appreciate that that communication between the user device 102a, the data storage server 106, and/or the analysis server 108 may be encrypted or otherwise secured.

[0052] The communication network 104 may be any network that allows digital devices to communicate. The communication network 104 may be the Internet and/or include LAN and WANs. The communication network 104 may support wireless and/or wired communication.

[0053] The data storage server 110 is a digital device that is configured to store data. In various embodiments, the data storage server 110 stores databases and/or other data structures.

The data storage server 110 may be a single server or a combination of servers. In one example the data storage server 110 may be a secure server wherein a user may store data over a secured connection (e.g., via https). The data may be encrypted and backed-up. In some embodiments, the data storage server 106 is operated by a third-party such as Amazon's S3 service.

[0054] The database or other data structure may comprise large high-dimensional datasets. These datasets are traditionally very difficult to analyze and, as a result, relationships within the data may not be identifiable using previous methods. Further, previous methods may be computationally inefficient.

[0055] The analysis server 108 is a digital device that may be configured to analyze data. In various embodiments, the analysis server may perform many functions to interpret, examine, analyze, and display data and/or relationships within data. In some embodiments, the analysis server 108 performs, at least in part, topological analysis of large datasets applying metrics, filters, and resolution parameters chosen by the user. The analysis is further discussed in FIG. 8 herein.

[0056] The analysis server 108 may generate an interactive visualization of the output of the analysis. The interactive visualization allows the user to observe and explore relationships in the data. In various embodiments, the interactive visualization allows the user to select nodes comprising data that has been clustered. The user may then access the underlying data, perform further analysis (e.g., statistical analysis) on the underlying data, and manually reorient the graph(s) (e.g., structures of nodes and edges described herein) within the interactive visualization. The analysis server 108 may also allow for the user to interact with the data, see the graphic result. The interactive visualization is further discussed in FIGs 9-11.

[0057] In some embodiments, the analysis server 108 interacts with the user device(s) 102a-102n over a private and/or secure communication network. The user device 102a may comprise a client program that allows the user to interact with the data storage server 106, the analysis server 108, another user device (e.g., user device 102n), a database, and/or an analysis application executed on the analysis server 108.

[0058] Those skilled in the art will appreciate that all or part of the data analysis may occur at the user device 102a. Further, all or part of the interaction with the visualization (e.g., graphic) may be performed on the user device 102a.

[0059] Although two user devices 102a and 102n are depicted, those skilled in the art will appreciate that there may be any number of user devices in any location (e.g., remote from each other). Similarly, there may be any number of communication networks, data storage servers, and analysis servers.

[0060] Cloud computing may allow for greater access to large datasets (e.g., via a commercial storage service) over a faster connection. Further, those skilled in the art will appreciate that services and computing resources offered to the user(s) may be scalable.

[0061] FIG. 2 is a block diagram of an exemplary analysis server 108. In exemplary embodiments, the analysis server 108 comprises a processor 202, input/output (I/O) interface 204, a communication network interface 206, a memory system 208, and a storage system 210. The processor 202 may comprise any processor or combination of processors with one or more cores.

[0062] The input / output (I/O) device 204 may comprise interfaces for various I/O devices such as, for example, a keyboard, mouse, and display device. The exemplary communication network interface 206 is configured to allow the analysis server 108 to communication with the communication network 104 (see FIG. 1). The communication network interface 206 may support communication over an Ethernet connection, a serial connection, a parallel connection, and/or an ATA connection. The communication network interface 206 may also support wireless communication (e.g., 802.11 a/b/g/n, WiMax, LTE, WiFi). It will be apparent to those skilled in the art that the communication network interface 206 can support many wired and wireless standards.

[0063] The memory system 208 may be any kind of memory including RAM, ROM, or flash, cache, virtual memory, etc. In various embodiments, working data is stored within the memory system 208. The data within the memory system 208 may be cleared or ultimately transferred to the storage system 210.

[0064] The storage system 210 includes any storage configured to retrieve and store data. Some examples of the storage system 210 include flash drives, hard drives, optical drives,

and/or magnetic tape. Each of the memory system 208 and the storage system 210 comprises a computer-readable medium, which stores instructions (e.g., software programs) executable by processor 202.

[0065] The storage system 210 comprises a plurality of modules utilized by embodiments of the present invention. A module may be hardware, software (e.g., including instructions executable by a processor), or a combination of both. In one embodiment, the storage system 210 comprises a processing module 212 which comprises an input module 214, a filter module 216, a resolution module 218, an analysis module 220, a visualization engine 222, and database storage 224. Alternative embodiments of the analysis server 108 and/or the storage system 210 may comprise more, less, or functionally equivalent components and modules.

[0066] The input module 214 may be configured to receive commands and preferences from the user device 102a. In various examples, the input module 214 receives selections from the user which will be used to perform the analysis. The output of the analysis may be an interactive visualization.

[0067] The input module 214 may provide the user a variety of interface windows allowing the user to select and access a database, choose fields associated with the database, choose a metric, choose one or more filters, and identify resolution parameters for the analysis. In one example, the input module 214 receives a database identifier and accesses a large multi-dimensional database. The input module 214 may scan the database and provide the user with an interface window allowing the user to identify an ID field. An ID field is an identifier for each data point. In one example, the identifier is unique. The same column name may be present in the table from which filters are selected. After the ID field is selected, the input module 214 may then provide the user with another interface window to allow the user to choose one or more data fields from a table of the database.

[0068] Although interactive windows may be described herein, those skilled in the art will appreciate that any window, graphical user interface, and/or command line may be used to receive or prompt a user or user device 102a for information.

[0069] The filter module 216 may subsequently provide the user with an interface window to allow the user to select a metric to be used in analysis of the data within the chosen data fields. The filter module 216 may also allow the user to select and/or define one or more filters.

[0070] The resolution module 218 may allow the user to select a resolution, including filter parameters. In one example, the user enters a number of intervals and a percentage overlap for a filter.

[0071] The analysis module 220 may perform data analysis based on the database and the information provided by the user. In various embodiments, the analysis module 220 performs an algebraic topological analysis to identify structures and relationships within data and clusters of data. Those skilled in the art will appreciate that the analysis module 220 may use parallel algorithms or use generalizations of various statistical techniques (e.g., generalizing the bootstrap to zig-zag methods) to increase the size of data sets that can be processed. The analysis is further discussed in FIG. 8. Those skilled in the art will appreciate that the analysis module 220 is not limited to algebraic topological analysis but may perform any analysis.

[0072] The visualization engine 222 generates an interactive visualization including the output from the analysis module 220. The interactive visualization allows the user to see all or part of the analysis graphically. The interactive visualization also allows the user to interact with the visualization. For example, the user may select portions of a graph from within the visualization to see and/or interact with the underlying data and/or underlying analysis. The user may then change the parameters of the analysis (e.g., change the metric, filter(s), or resolution(s)) which allows the user to visually identify relationships in the data that may be otherwise undetectable using prior means. The interactive visualization is further described in FIGs 9-11.

[0073] The database storage 224 is configured to store all or part of the database that is being accessed. In some embodiments, the database storage 224 may store saved portions of the database. Further, the database storage 224 may be used to store user preferences, parameters, and analysis output thereby allowing the user to perform many different functions on the database without losing previous work.

[0074] Those skilled in the art will appreciate that that all or part of the processing module 212 may be at the user device 102a or the database storage server 106. In some embodiments, all or some of the functionality of the processing module 212 may be performed by the user device 102a.

[0075] In various embodiments, systems and methods discussed herein may be implemented with one or more digital devices. In some examples, some embodiments discussed herein may be implemented by a computer program (instructions) executed by a processor. The computer program may provide a graphical user interface. Although such a computer program is discussed, those skilled in the art will appreciate that embodiments may be performed using any of the following, either alone or in combination, including, but not limited to, a computer program, multiple computer programs, firmware, and/or hardware.

[0076] FIG. 3 is a flow chart 300 depicting an exemplary method of dataset analysis and visualization in some embodiments. In step 302, the input module 214 accesses a database. The database may be any data structure containing data (e.g., a very large dataset of multidimensional data). In some embodiments, the database may be a relational database. In some examples, the relational database may be used with MySQL, Oracle, Micosoft SQL Server, Aster nCluster, Teradata, and/or Vertica. Those skilled in the art will appreciate that the database may not be a relational database.

[0077] In some embodiments, the input module 214 receives a database identifier and a location of the database (e.g., the data storage server 106) from the user device 102a (see FIG. 1). The input module 214 may then access the identified database. In various embodiments, the input module 214 may read data from many different sources, including, but not limited to MS Excel files, text files (e.g., delimited or CSV), Matlab .mat format, or any other file.

[0078] In some embodiments, the input module 214 receives an IP address or hostname of a server hosting the database, a username, password, and the database identifier. This information (herein referred to as "connection information") may be cached for later use. Those skilled in the art will appreciate that the database may be locally accessed and that all, some, or none of the connection information may be required. In one example, the user device 102a may have full access to the database stored locally on the user device 102a so the IP address is unnecessary. In another example, the user device 102a may already have loaded the database and the input module 214 merely begins by accessing the loaded database.

[0079] In various embodiments, the identified database stores data within tables. A table may have a "column specification" which stores the names of the columns and their data types. A "row" in a table, may be a tuple with one entry for each column of the correct type. In one example, a table to store employee records might have a column specification such as:

- employee_id primary key int (this may store the employee's ID as an integer, and uniquely identifies a row)
- age int
- gender char(1) (gender of the employee may be a single character either M or F)
- salary double (salary of an employee may be a floating point number)
- name varchar (name of the employee may be a variable-length string)

In this example, each employee corresponds to a row in this table. Further, the tables in this exemplary relational database are organized into logical units called databases. An analogy to file systems is that databases can be thought of as folders and files as tables. Access to databases may be controlled by the database administrator by assigning a username/password pair to authenticate users.

[0080] Once the database is accessed, the input module 214 may allow the user to access a previously stored analysis or to begin a new analysis. If the user begins a new analysis, the input module 214 may provide the user device 102a with an interface window allowing the user to identify a table from within the database. In one example, the input module 214 provides a list of available tables from the identified database.

[0081] In step 304, the input module 214 receives a table identifier identifying a table from within the database. The input module 214 may then provide the user with a list of available ID fields from the table identifier. In step 306, the input module 214 receives the ID field identifier from the user and/or user device 102a. The ID field is, in some embodiments, the primary key.

[0082] Having selected the primary key, the input module 214 may generate a new interface window to allow the user to select data fields for analysis. In step 308, the input module 214 receives data field identifiers from the user device 102a. The data within the data fields may be later analyzed by the analysis module 220.

[0083] In step 310, the filter module 216 identifies a metric. In some embodiments, the filter module 216 and/or the input module 214 generates an interface window allowing the user of the user device 102a options for a variety of different metrics and filter preferences. The interface window may be a drop down menu identifying a variety of distance metrics to be used in the analysis. Metric options may include, but are not limited to, Euclidean, DB Metric,

variance normalized Euclidean, and total normalized Euclidean. The metric and the analysis are further described herein.

[0084] In step 312, the filter module 216 selects one or more filters. In some embodiments, the user selects and provides filter identifier(s) to the filter module 216. The role of the filters in the analysis is also further described herein. The filters, for example, may be user defined, geometric, or based on data which has been pre-processed. In some embodiments, the data based filters are numerical arrays which can assign a set of real numbers to each row in the table or each point in the data generally.

[0085] A variety of geometric filters may be available for the user to choose. Geometric filters may include, but are not limited to:

- Density
- L1 Eccentricity
- L-infinity Eccentricity
- Witness based Density
- Witness based Eccentricity
- Eccentricity as distance from a fixed point
- Approximate Kurtosis of the Eccentricity

[0086] In step 314, the resolution module 218 defines the resolution to be used with a filter in the analysis. The resolution may comprise a number of intervals and an overlap parameter. In various embodiments, the resolution module 218 allows the user to adjust the number of intervals and overlap parameter (e.g., percentage overlap) for one or more filters.

[0087] In step 316, the analysis module 220 processes data of selected fields based on the metric, filter(s), and resolution(s) to generate the visualization. This process is discussed in FIG. 8.

[0088] In step 318, the visualization module 222 displays the interactive visualization. In various embodiments, the visualization may be rendered in two or three dimensional space. The visualization module 222 may use an optimization algorithm for an objective function which is correlated with good visualization (e.g., the energy of the embedding). The visualization may show a collection of nodes corresponding to each of the partial clusters in

the analysis output and edges connecting them as specified by the output. The interactive visualization is further discussed in FIGs. 9-11.

[0089] Although many examples discuss the input module 214 as providing interface windows, those skilled in the art will appreciate that all or some of the interface may be provided by a client on the user device 102a. Further, in some embodiments, the user device 102a may be running all or some of the processing module 212.

[0090] FIGs 4-7 depict various interface windows to allow the user to make selections, enter information (e.g., fields, metrics, and filters), provide parameters (e.g., resolution), and provide data (e.g., identify the database) to be used with analysis. Those skilled in the art will appreciate that any graphical user interface or command line may be used to make selections, enter information, provide parameters, and provide data.

[0091] FIG. 4 is an exemplary ID field selection interface window 400 in some embodiments. The ID field selection interface window 400 allows the user to identify an ID field. The ID field selection interface window 400 comprises a table search field 402, a table list 404, and a fields selection window 406.

[0092] In various embodiments, the input module 214 identifies and accesses a database from the database storage 224, user device 102a, or the data storage server 106. The input module 214 may then generate the ID field selection interface window 400 and provide a list of available tables of the selected database in the table list 404. The user may click on a table or search for a table by entering a search query (e.g., a keyword) in the table search field 402. Once a table is identified (e.g., clicked on by the user), the fields selection window 406 may provide a list of available fields in the selected table. The user may then choose a field from the fields selection window 406 to be the ID field. In some embodiments, any number of fields may be chosen to be the ID field(s).

[0093] FIG. 5 is an exemplary data field selection interface window 500 in some embodiments. The data field selection interface window 500 allows the user to identify data fields. The data field selection interface window 500 comprises a table search field 502, a table list 504, a fields selection window 506, and a selected window 508.

[0094] In various embodiments, after selection of the ID field, the input module 214 provides a list of available tables of the selected database in the table list 504. Similar to the table

search field 402 in FIG. 4, the user may click on a table or search for a table by entering a search query (e.g., a keyword) in the table search field 502. Once a table is identified (e.g., clicked on by the user), the fields selection window 506 may provide a list of available fields in the selected table. The user may then choose any number of fields from the fields selection window 506 to be data fields. The selected data fields may appear in the selected window 508. The user may also deselect fields that appear in the selected window 508.

[0095] Those skilled in the art will appreciate that the table selected by the user in the table list 504 may be the same table selected with regard to FIG. 4. In some embodiments, however, the user may select a different table. Further, the user may, in various embodiments, select fields from a variety of different tables.

[0096] FIG. 6 is an exemplary metric and filter selection interface window 600 in some embodiments. The metric and filter selection interface window 600 allows the user to identify a metric, add filter(s), and adjust filter parameters. The metric and filter selection interface window 600 comprises a metric pull down menu 602, an add filter from database button 604, and an add geometric filter button 606.

[0097] In various embodiments, the user may click on the metric pull down menu 602 to view a variety of metric options. Various metric options are described herein. In some embodiments, the user may define a metric. The user defined metric may then be used with the analysis.

[0098] In one example, finite metric space data may be constructed from a data repository (i.e., database, spreadsheet, or Matlab file). This may mean selecting a collection of fields whose entries will specify the metric using the standard Euclidean metric for these fields, when they are floating point or integer variables. Other notions of distance, such as graph distance between collections of points, may be supported.

[0099] The analysis module 220 may perform analysis using the metric as a part of a distance function. The distance function can be expressed by a formula, a distance matrix, or other routine which computes it. The user may add a filter from a database by clicking on the add filter from database button 604. The metric space may arise from a relational database, a Matlab file, an Excel spreadsheet, or other methods for storing and manipulating data. The

metric and filter selection interface window 600 may allow the user to browse for other filters to use in the analysis. The analysis and metric function are further described in FIG. 8.

[00100] The user may also add a geometric filter 606 by clicking on the add geometric filter button 606. In various embodiments, the metric and filter selection interface window 600 may provide a list of geometric filters from which the user may choose.

[00101] FIG. 7 is an exemplary filter parameter interface window 700 in some embodiments. The filter parameter interface window 700 allows the user to determine a resolution for one or more selected filters (e.g., filters selected in the metric and filter selection interface window 600). The filter parameter interface window 700 comprises a filter name menu 702, an interval field 704, an overlap bar 706, and a done button 708.

[00102] The filter parameter interface window 700 allows the user to select a filter from the filter name menu 702. In some embodiments, the filter name menu 702 is a drop down box indicating all filters selected by the user in the metric and filter selection interface window 600. Once a filter is chosen, the name of the filter may appear in the filter name menu 702. The user may then change the intervals and overlap for one, some, or all selected filters.

[00103] The interval field 704 allows the user to define a number of intervals for the filter identified in the filter name menu 702. The user may enter a number of intervals or scroll up or down to get to a desired number of intervals. Any number of intervals may be selected by the user. The function of the intervals is further discussed in FIG. 8.

[00104] The overlap bar 706 allows the user to define the degree of overlap of the intervals for the filter identified in the filter name menu 702. In one example, the overlap bar 706 includes a slider that allows the user to define the percentage overlap for the interval to be used with the identified filter. Any percentage overlap may be set by the user.

[00105] Once the intervals and overlap are defined for the desired filters, the user may click the done button. The user may then go back to the metric and filter selection interface window 600 and see a new option to run the analysis. In some embodiments, the option to run the analysis may be available in the filter parameter interface window 700. Once the analysis is complete, the result may appear in an interactive visualization which is further described in FIGs 9-11.

[00106] Those skilled in the art will appreciate that that interface windows in FIGs 4-7 are exemplary. The exemplary interface windows are not limited to the functional objects (e.g., buttons, pull down menus, scroll fields, and search fields) shown. Any number of different functional objects may be used. Further, as described herein, any other interface, command line, or graphical user interface may be used.

[00107] FIG. 8 is a flowchart 800 for data analysis and generating an interactive visualization in some embodiments. In various embodiments, the processing on data and user-specified options is motivated by techniques from topology and, in some embodiments, algebraic topology. These techniques may be robust and general. In one example, these techniques apply to almost any kind of data for which some qualitative idea of "closeness" or "similarity" exists. The techniques discussed herein may be robust because the results may be relatively insensitive to noise in the data, user options, and even to errors in the specific details of the qualitative measure of similarity, which, in some embodiments, may be generally refer to as "the distance function" or "metric." Those skilled in the art will appreciate that while the description of the algorithms below may seem general, the implementation of techniques described herein may apply to any level of generality.

[00108] In step 802, the input module 214 receives data **S**. In one example, a user identifies a data structure and then identifies ID and data fields. Data **S** may be based on the information within the ID and data fields. In various embodiments, data **S** is treated as being processed as a finite "similarity space," where data **S** has a real-valued function **d** defined on pairs of points s and t in S, such that:

$$\mathbf{d}(s, s) = 0$$
$$\mathbf{d}(s, t) = \mathbf{d}(t, s)$$
$$\mathbf{d}(s, t) >= 0$$

These conditions may be similar to requirements for a finite metric space, but the conditions may be weaker. In various examples, the function is a metric.

[00109] Those skilled in the art will appreciate that data **S** may be a finite metric space, or a generalization thereof, such as a graph or weighted graph. In some embodiments, data **S** be specified by a formula, an algorithm, or by a distance matrix which specifies explicitly every pairwise distance.

[00110] In step 804, the input module 214 generates reference space **R**. In one example, reference space **R** may be a well-known metric space (e.g., such as the real line). The reference space **R** may be defined by the user. In step 806, the analysis module 220 generates a map **ref()** from **S** into **R**. The map **ref()** from **S** into **R** may be called the "reference map."

[00111] In one example, a reference of map from **S** is to a reference metric space **R**. **R** may be Euclidean space of some dimension, but it may also be the circle, torus, a tree, or other metric space. The map can be described by one or more filters (i.e., real valued functions on **S**). These filters can be defined by geometric invariants, such as the output of a density estimator, a notion of data depth, or functions specified by the origin of **S** as arising from a data set.

[00112] In step 808, the resolution module 218 generates a cover of **R** based on the resolution received from the user (e.g., filter(s), intervals, and overlap – see FIG. 7). The cover of **R** may be a finite collection of open sets (in the metric of **R**) such that every point in **R** lies in at least one of these sets. In various examples, **R** is k-dimensional Euclidean space, where k is the number of filter functions. More precisely in this example, **R** is a box in k-dimensional Euclidean space given by the product of the intervals [min_k, max_k], where min_k is the minimum value of the k-th filter function on **S**, and max_k is the maximum value.

[00113] For example, suppose there are 2 filter functions, F1 and F2, and that F1's values range from -1 to +1, and F2's values range from 0 to 5. Then the reference space is the rectangle in the x/y plane with corners (-1,0), (1,0), (-1, 5), (1, 5), as every point s of **S** will give rise to a pair (F1(s), F2(s)) that lies within that rectangle.

[00114] In various embodiments, the cover of **R** is given by taking products of intervals of the covers of [min_k,max_k] for each of the k filters. In one example, if the user requests 2 intervals and a 50% overlap for F1, the cover of the interval [-1,+1] will be the two intervals (-1.5, .5), (-.5, 1.5). If the user requests 5 intervals and a 30% overlap for F2, then that cover of [0, 5] will be (-.3, 1.3), (.7, 2.3), (1.7, 3.3), (2.7, 4.3), (3.7, 5.3). These intervals may give rise to a cover of the 2-dimensional box by taking all possible pairs of intervals where the first of the pair is chosen from the cover for F1 and the second from the cover for F2. This may give rise to 2 * 5, or 10, open boxes that covered the 2-dimensional reference space. However, those skilled in the art will appreciate that the intervals may not be uniform, or that the covers of a k-dimensional box may not be constructed by products of intervals. In some

embodiments, there are many other choices of intervals. Further, in various embodiments, a wide range of covers and/or more general reference spaces may be used.

[00115] In one example, given a cover, C_1, ..., C_m, of **R**, the reference map is used to assign a set of indices to each point in **S**, which are the indices of the C_j such that **ref**(s) belongs to C_j. This function may be called **ref_tags**(s). In a language such as Java, **ref_tags** would be a method that returned an **int[]**. Since the C's cover **R** in this example, **ref**(s) must lie in at least one of them, but the elements of the cover usually overlap one another, which means that points that "land near the edges" may well reside in multiple cover sets. In considering the two filter example, if F1(s) is -.99, and F2(s) is .001, then **ref**(s) is (-.99, .001), and this lies in the cover element (-1.5, .5)x(-.3,1.3). Supposing that was labeled C_1, the reference map may assign s to the set {1}. On the other hand, if t is mapped by F1, F2 to (.1, 2.1), then **ref**(t) will be in (-1.5,.5)x(.7, 2.3), (- .5, 1.5)x(.7,2.3), (-1.5,.5)x(1.7,3.3), and (-.5, 1.5)x(1.7,3.3), so the set of indices would have four elements for t.

[00116] Having computed, for each point, which "cover tags" it is assigned to, for each cover element, C_d, the points may be constructed, whose tags include d, as set **S**(d). This may mean that every point s is in **S**(d) for some d, but some points may belong to more than one such set. In some embodiments, there is, however, no requirement that each **S**(d) is non-empty, and it is frequently the case that some of these sets are empty. In the non-parallelized version of some embodiments, each point x is processed in turn, and x is inserted into a hash-bucket for each j in **ref_tags**(t) (that is, this may be how **S**(d) sets are computed).

[00117] Those skilled in the art will appreciate that the cover of the reference space **R** may be controlled by the number of intervals and the overlap identified in the resolution (e.g., see FIG. 7). For example, the more intervals, the finer the resolution in **S** - that is, the fewer points in each **S**(d), but the more similar (with respect to the filters) these points may be. The greater the overlap, the more times that clusters in **S**(d) may intersect clusters in **S**(e) - this means that more "relationships" between points may appear, but, in some embodiments, the greater the overlap, the more likely that accidental relationships may appear.

[00118] In step 810, the analysis module 220 clusters each **S**(d) based on the metric, filter, and the space **S**. In some embodiments, a dynamic single-linkage clustering algorithm may be used to partition **S**(d). Those skilled in the art will appreciate that any number of clustering algorithms may be used with embodiments discussed herein. For example, the clustering

scheme may be k-means clustering for some k, single linkage clustering, average linkage clustering, or any method specified by the user.

[00119] The significance of the user-specified inputs may now be seen. In some embodiments, a filter may amount to a "forced stretching" in a certain direction. In some embodiments, the analysis module 220 may not cluster two points unless ALL of the filter values are sufficiently "related" (recall that while normally related may mean "close," the cover may impose a much more general relationship on the filter values, such as relating two points s and t if **ref**(s) and **ref**(t) are sufficiently close to the same circle in the plane). In various embodiments, the ability of a user to impose one or more "critical measures" makes this technique more powerful than regular clustering, and the fact that these filters can be anything, is what makes it so general.

[00120] The output may be a simplicial complex, from which one can extract its 1-skeleton. The nodes of the complex may be partial clusters, (i.e., clusters constructed from subsets of **S** specified as the preimages of sets in the given covering of the reference space **R**).

[00121] In step 812, the visualization engine 222 identifies nodes which are associated with a subset of the partition elements of all of the **S**(d) for generating an interactive visualization. For example, suppose that **S** = {1, 2, 3, 4}, and the cover is C_1, C_2, C_3. Then if **ref_tags**(1) = {1, 2, 3} and **ref_tags**(2) = {2, 3}, and **ref_tags**(3) = {3}, and finally **ref_tags**(4) = {1 , 3}, then **S**(1) in this example is {1, 4}, **S**(2) = {1,2}, and **S**(3)={1,2,3,4}. If 1 and 2 are close enough to be clustered, and 3 and 4 are, but nothing else, then the clustering for **S**(1) may be {1} {3}, and for **S**(2) it may be {1,2}, and for **S**(3) it may be {1,2}, {3,4}. So the generated graph has, in this example, at most four nodes, given by the sets {1}, {4}, {1,2}, and {3,4} (note that {1,2} appears in two different clusterings). Of the sets of points that are used, two nodes intersect provided that the associated node sets have a non-empty intersection (although this could easily be modified to allow users to require that the intersection is "large enough" either in absolute or relative terms).

[00122] Nodes may be eliminated for any number of reasons. For example, a node may be eliminated as having too few points and/or not being connected to anything else. In some embodiments, the criteria for the elimination of nodes (if any) may be under user control or have application-specific requirements imposed on it. For example, if the points are consumers, for instance, clusters with too few people in area codes served by a company could

be eliminated. If a cluster was found with "enough" customers, however, this might indicate that expansion into area codes of the other consumers in the cluster could be warranted.

[00123] In step 814, the visualization engine 222 joins clusters to identify edges (e.g., connecting lines between nodes). Once the nodes are constructed, the intersections (e.g., edges) may be computed "all at once," by computing, for each point, the set of node sets (not **ref_tags**, this time). That is, for each s in **S**, **node_id_set**(s) may be computed, which is an **int[]**. In some embodiments, if the cover is well behaved, then this operation is linear in the size of the set **S**, and we then iterate over each pair in **node_id_set**(s). There may be an edge between two node_id's if they both belong to the same **node_id_set**() value, and the number of points in the intersection is precisely the number of different node_id sets in which that pair is seen. This means that, except for the clustering step (which is often quadratic in the size of the sets **S**(d), but whose size may be controlled by the choice of cover), all of the other steps in the graph construction algorithm may be linear in the size of **S**, and may be computed quite efficiently.

[00124] In step 816, the visualization engine 222 generates the interactive visualization of interconnected nodes (e.g., nodes and edges displayed in FIGs 10 and 11).

[00125] Those skilled in the art will appreciate that it is possible, in some embodiments, to make sense in a fairly deep way of connections between various **ref()** maps and/or choices of clustering. Further, in addition to computing edges (pairs of nodes), the embodiments described herein may be extended to compute triples of nodes, etc. For example, the analysis module 220 may compute simplicial complexes of any dimension (by a variety of rules) on nodes, and apply techniques from homology theory to the graphs to help users understand a structure in an automatic (or semi-automatic) way.

[00126] Further, those skilled in the art will appreciate that uniform intervals in the covering may not always be a good choice. For example, if the points are exponentially distributed with respect to a given filter, uniform intervals can fail - in such case adaptive interval sizing may yield uniformly-sized **S**(d) sets, for instance.

[00127] Further, in various embodiments, an interface may be used to encode techniques for incorporating third-party extensions to data access and display techniques. Further, an

interface may be used to for third-party extensions to underlying infrastructure to allow for new methods for generating coverings, and defining new reference spaces.

[00128] FIG. 9 is an exemplary interactive visualization 900 in some embodiments. The display of the interactive visualization may be considered a "graph" in the mathematical sense. The interactive visualization comprises of two types of objects: nodes (e.g., nodes 902 and 906) (the colored balls) and the edges (e.g., edge 904) (the black lines). The edges connect pairs of nodes (e.g., edge 904 connects node 902 with node 906). As discussed herein, each node may represent a collection of data points (rows in the database identified by the user). In one example, connected nodes tend to include data points which are "similar to" (e.g., clustered with) each other. The collection of data points may be referred to as being "in the node." The interactive visualization may be two-dimensional, three-dimensional, or a combination of both.

[00129] In various embodiments, connected nodes and edges may form a graph or structure. There may be multiple graphs in the interactive visualization. In one example, the interactive visualization may display two or more unconnected structures of nodes and edges.

[00130] The visual properties of the nodes and edges (such as, but not limited to, color, stroke color, text, texture, shape, coordinates of the nodes on the screen) can encode any data based property of the data points within each node. For example, coloring of the nodes and/or the edges may indicate (but is not limited to) the following:

- Values of fields or filters
- Any general functions of the data in the nodes (e.g., if the data were unemployment rates by state, then GDP of the states may be identifiable by color the nodes)
- Number of data points in the node

[00131] The interactive visualization 900 may contain a "color bar" 910 which may comprise a legend indicating the coloring of the nodes (e.g., balls) and may also identify what the colors indicate. For example, in FIG. 9, color bar 910 indicates that color is based on the density filter with blue (on the far left of the color bar 910) indicating "4.99e+03" and red (on the far right of the color bar 910) indicating "1.43e+04." In general this might be expanded to show any other legend by which nodes and/or edges are colored. Those skilled in the art will appreciate that the, In some embodiments, the user may control the color as well as what the

color (and/or stroke color, text, texture, shape, coordinates of the nodes on the screen) indicates.

[00132] The user may also drag and drop objects of the interactive visualization 900. In various embodiments, the user may reorient structures of nodes and edges by dragging one or more nodes to another portion of the interactive visualization (e.g., a window). In one example, the user may select node 902, hold node 902, and drag the node across the window. The node 902 will follow the user's cursor, dragging the structure of edges and/or nodes either directly or indirectly connected to the node 902. In some embodiments, the interactive visualization 900 may depict multiple unconnected structures. Each structure may include nodes, however, none of the nodes of either structure are connected to each other. If the user selects and drags a node of the first structure, only the first structure will be reoriented with respect to the user action. The other structure will remain unchanged. The user may wish to reorient the structure in order to view nodes, select nodes, and/or better understand the relationships of the underlying data.

[00133] In one example, a user may drag a node to reorient the interactive visualization (e.g., reorient the structure of nodes and edges). While the user selects and/or drags the node, the nodes of the structure associated with the selected node may move apart from each other in order to provide greater visibility. Once the user lets go (e.g., deselects or drops the node that was dragged), the nodes of the structure may continue to move apart from each other.

[00134] In various embodiments, once the visualization module 222 generates the interactive display, the depicted structures may move by spreading out the nodes from each other. In one example, the nodes spread from each other slowly allowing the user to view nodes distinguish from each other as well as the edges. In some embodiments, the visualization module 222 optimizes the spread of the nodes for the user's view. In one example, the structure(s) stop moving once an optimal view has been reached.

[00135] Those skilled in the art will appreciate that the interactive visualization 900 may respond to gestures (e.g., multitouch), stylus, or other interactions allowing the user to reorient nodes and edges and/or interacting with the underlying data.

[00136] The interactive visualization 900 may also respond to user actions such as when the user drags, clicks, or hovers a mouse cursor over a node. In some embodiments, when the user

selects a node or edge, node information or edge information may be displayed. In one example, when a node is selected (e.g., clicked on by a user with a mouse or a mouse cursor hovers over the node), a node information box 908 may appear that indicates information regarding the selected node. In this example, the node information box 908 indicates an ID, box ID, number of elements (e.g., data points associated with the node), and density of the data associated with the node.

[00137] The user may also select multiple nodes and/or edges by clicking separate on each object, or drawing a shape (such as a box) around the desired objects. Once the objects are selected, a selection information box 912 may display some information regarding the selection. For example, selection information box 912 indicates the number of nodes selected and the total points (e.g., data points or elements) of the selected nodes.

[00138] The interactive visualization 900 may also allow a user to further interact with the display. Color option 914 allows the user to display different information based on color of the objects. Color option 914 in FIG. 9 is set to filter_Density, however, other filters may be chosen and the objects re-colored based on the selection. Those skilled in the art will appreciate that the objects may be colored based on any filter, property of data, or characterization. When a new option is chosen in the color option 914, the information and/or colors depicted in the color bar 910 may be updated to reflect the change.

[00139] Layout checkbox 914 may allow the user to anchor the interactive visualization 900. In one example, the layout checkbox 914 is checked indicating that the interactive visualization 900 is anchored. As a result, the user will not be able to select and drag the node and/or related structure. Although other functions may still be available, the layout checkbox 914 may help the user keep from accidentally moving and/or reorienting nodes, edges, and/or related structures. Those skilled in the art will appreciate that the layout checkbox 914 may indicate that the interactive visualization 900 is anchored when the layout checkbox 914 is unchecked and that when the layout checkbox 914 is checked the interactive visualization 900 is no longer anchored.

[00140] The change parameters button 918 may allow a user to change the parameters (e.g., add/remove filters and/or change the resolution of one or more filters). In one example, when the change parameters button 918 is activated, the user may be directed back to the metric and filter selection interface window 600 (see FIG. 6) which allows the user to add or remove

filters (or change the metric). The user may then view the filter parameter interface 700 (see FIG. 7) and change parameters (e.g., intervals and overlap) for one or more filters. The analysis node 220 may then re-analyze the data based on the changes and display a new interactive visualization 900 without again having to specify the data sets, filters, etc.

[00141] The find ID's button 920 may allow a user to search for data within the interactive visualization 900. In one example, the user may click the find ID's button 920 and receive a window allowing the user to identify data or identify a range of data. Data may be identified by ID or searching for the data based on properties of data and/or metadata. If data is found and selected, the interactive visualization 900 may highlight the nodes associated with the selected data. For example, selecting a single row or collection of rows of a database or spreadsheet may produce a highlighting of nodes whose corresponding partial cluster contains any element of that selection.

[00142] In various embodiments, the user may select one or more objects and click on the explain button 922 to receive in-depth information regarding the selection. In some embodiments, when the user selects the explain button 922, the information about the data from which the selection is based may be displayed. The function of the explain button 922 is further discussed with regard to FIG. 10.

[00143] In various embodiments, the interactive visualization 900 may allow the user to specify and identify subsets of interest, such as output filtering, to remove clusters or connections which are too small or otherwise uninteresting. Further, the interactive visualization 900 may provide more general coloring and display techniques, including, for example, allowing a user to highlight nodes based on a user-specified predicate, and coloring the nodes based on the intensity of user-specified weighting functions.

[00144] The interactive visualization 900 may comprise any number of menu items. The "Selection" menu may allow the following functions:

- Select singletons (select nodes which are not connected to other nodes)
- Select all (selects all the nodes and edges)
- Select all nodes (selects all nodes)
- Select all edges
- Clear selection (no selection)

- Invert Selection (selects the complementary set of nodes or edges)
- Select "small" nodes (allows the user to threshold nodes based on how many points they have)
- Select leaves (selects all nodes which are connected to long "chains" in the graph)
- Remove selected nodes
- Show in a table (shows the selected nodes and their associated data in a table)
- Save selected nodes (saves the selected data to whatever format the user chooses. This may allow the user to subset the data and create new datasources which may be used for further analysis.)

[00145] In one example of the "show in a table" option, information from a selection of nodes may be displayed. The information may be specific to the origin of the data. In various embodiments, elements of a database table may be listed, however, other methods specified by the user may also be included. For example, in the case of microarray data from gene expression data, heat maps may be used to view the results of the selections.

[00146] The interactive visualization 900 may comprise any number of menu items. The "Save" menu may allow may allow the user to save the whole output in a variety of different formats such as (but not limited to):

- Image files (PNG/JPG/PDF/SVG etc.)
- Binary output (The interactive output is saved in the binary format. The user may reopen this file at any time to get this interactive window again)

In some embodiments, graphs may be saved in a format such that the graphs may be used for presentations. This may include simply saving the image as a pdf or png file, but it may also mean saving an executable .xml file, which may permit other users to use the search and save capability to the database on the file without having to recreate the analysis.

[00147] In various embodiments, a relationship between a first and a second analysis output / interactive visualization for differing values of the interval length and overlap percentage may be displayed. The formal relationship between the first and second analysis output / interactive visualization may be that when one cover refines the next, there is a map of simplicial complexes from the output of the first to the output of the second. This can be displayed by applying a restricted form of a three-dimensional graph embedding algorithm, in which a graph

is the union of the graphs for the various parameter values and in which the connections are the connections in the individual graphs as well as connections from one node to its image in the following graph. The constituent graphs may be placed in its own plane in 3D space. In some embodiments, there is a restriction that each constituent graph remain within its associated plane. Each constituent graph may be displayed individually, but a small change of parameter value may result in the visualization of the adjacent constituent graph. In some embodiments, nodes in the initial graph will move to nodes in the next graph, in a readily visualizable way.

[00148] FIG. 10 is an exemplary interactive visualization 1000 displaying an explain information window 1002 in some embodiments. In various embodiments, the user may select a plurality of nodes and click on the explain button. When the explain button is clicked, the explain information window 1002 may be generated. The explain information window 1002 may identify the data associated with the selected object(s) as well as information (e.g., statistical information) associated with the data.

[00149] In some embodiments, the explain button allows the user to get a sense for which fields within the selected data fields are responsible for "similarity" of data in the selected nodes and the differentiating characteristics. There can be many ways of scoring the data fields. The explain information window 1002 (i.e., the scoring window in FIG. 10) is shown along with the selected nodes. The highest scoring fields may distinguish variables with respect to the rest of the data.

[00150] In one example, the explain information window 1002 indicates that data from fields day0-day6 has been selected. The minimum value of the data in all of the fields is 0. The explain information window 1002 also indicates the maximum values. For example, the maximum value of all of the data associated with the day0 field across all of the points of the selected nodes is 0.353. The average (i.e., mean) of all of the data associated with the day0 field across all of the points of the selected nodes is 0.031. The score may be a relative (e.g., normalized) value indicating the relative function of the filter; here, the score may indicate the relative density of the data associated with the day0 field across all of the points of the selected nodes. Those skilled in the art will appreciate that any information regarding the data and/or selected nodes may appear in the explain information window 1002.

[00151] Those skilled in the art will appreciate that the data and the interactive visualization 1000 may be interacted with in any number of ways. The user may interact with the data

directly to see where the graph corresponds to the data, make changes to the analysis and view the changes in the graph, modify the graph and view changes to the data, or perform any kind of interaction.

[00152] FIG. 11 is a flowchart 1100 of functionality of the interactive visualization in some embodiments. In step 1102, the visualization engine 222 receives the analysis from the analysis module 220 and graphs nodes as balls and edges as connectors between balls 1102 to create interactive visualization 900 (see FIG. 9).

[00153] In step 1104, the visualization engine 222 determines if the user is hovering a mouse cursor (or has selected) a ball (i.e., a node). If the user is hovering a mouse cursor over a ball or selecting a ball, then information is displayed regarding the data associated with the ball. In one example, the visualization engine 222 displays a node information window 908.

[00154] If the visualization engine 222 does not determine that the user is hovering a mouse cursor (or has selected) a ball, then the visualization engine 222 determines if the user has selected balls on the graph (e.g., by clicking on a plurality of balls or drawing a box around a plurality of balls). If the user has selected balls on the graph, the visualization engine 222 may highlight the selected balls on the graph in step 1110. The visualization engine 222 may also display information regarding the selection (e.g., by displaying a selection information window 912). The user may also click on the explain button 922 to receive more information associated with the selection (e.g., the visualization engine 222 may display the explain information window 1002).

[00155] In step 1112, the user may save the selection. For example, the visualization engine 222 may save the underlying data, selected metric, filters, and/or resolution. The user may then access the saved information and create a new structure in another interactive visualization 900 thereby allowing the user to focus attention on a subset of the data.

[00156] If the visualization engine 222 does not determine that the user has selected balls on the graph, the visualization engine 222 may determine if the user selects and drags a ball on the graph in step 1114. If the user selects and drags a ball on the graph, the visualization engine 222 may reorient the selected balls and any connected edges and balls based on the user's action in step 1116. The user may reorient all or part of the structure at any level of granularity.

[00157] Those skilled in the art will appreciate that although FIG. 11 discussed the user hovering over, selecting, and/or dragging a ball, the user may interact with any object in the interactive visualization 900 (e.g., the user may hover over, select, and/or drag an edge). The user may also zoom in or zoom out using the interactive visualization 900 to focus on all or a part of the structure (e.g., one or more balls and/or edges).

[00158] Further, although balls are discussed and depicted in FIGs 9-11, those skilled in the art will appreciate that the nodes may be any shape and appear as any kind of object. Further, although some embodiments described herein discuss an interactive visualization being generated based on the output of algebraic topology, the interactive visualization may be generated based on any kind of analysis and is not limited.

[00159] For years, researchers have been collecting huge amounts of data on breast cancer, yet we are still battling the disease. Complexity, rather than quantity, is one of the fundamental issues in extracting knowledge from data. A topological data exploration and visualization platform may assist the analysis and assessment of complex data. In various embodiments, a predictive and visual cancer map generated by the topological data exploration and visualization platform may assist physicians to determine treatment options.

[00160] In one example, a breast cancer map visualization may be generated based on the large amount of available information already generated by many researchers. Physicians may send biopsy data directly to a cloud-based server which may localize a new patient's data within the breast cancer map visualization. The breast cancer map visualization may be annotated (e.g., labeled) such that the physician may view outcomes of patients with similar profiles as well as different kinds of statistical information such as survival probabilities. Each new data point from a patient may be incorporated into the breast cancer map visualization to improve accuracy of the breast cancer map visualization over time.

[00161] Although the following examples are largely focused on cancer map visualizations, those skilled in the art will appreciate that at least some of the embodiments described herein may apply to any biological condition and not be limited to cancer and/or disease. For example, some embodiments, may apply to different industries.

[00162] FIG. 12 is a flowchart for generating a cancer map visualization utilizing biological data of a plurality of patients in some embodiments. In various embodiments, the processing

of data and user-specified options is motivated by techniques from topology and, in some embodiments, algebraic topology. As discussed herein, these techniques may be robust and general. In one example, these techniques apply to almost any kind of data for which some qualitative idea of "closeness" or "similarity" exists. Those skilled in the art will appreciate that the implementation of techniques described herein may apply to any level of generality.

[00163] In various embodiments, a cancer map visualization is generated using genomic data linked to clinical outcomes (i.e., medical characteristics) which may be used by physicians during diagnosis and/or treatment. Initially, publicly available data sets may be integrated to construct the topological map visualizations of patients (e.g., breast cancer patients). Those skilled in the art will appreciate that any private, public, or combination of private and public data sets may be integrated to construct the topological map visualizations. A map visualization may be based on biological data such as, but not limited to, gene expression, sequencing, and copy number variation. As such, the map visualization may comprise many patients with many different types of collected data. Unlike traditional methods of analysis where distinct studies of breast cancer appear as separate entities, the map visualization may fuse disparate data sets while utilizing many datasets and data types.

[00164] In various embodiments, a new patient may be localized on the map visualization. With the map visualization for subtypes of a particular disease and a new patient diagnosed with the disease, point(s) may be located among the data points used in computing the map visualization (e.g., nearest neighbor) which is closest to the new patient point. The new patient may be labeled with nodes in the map visualization containing the closest neighbor. These nodes may be highlighted to give a physician the location of the new patient among the patients in the reference data set. The highlighted nodes may also give the physician the location of the new patient relative to annotated disease subtypes.

[00165] The visualization map may be interactive and/or searchable in real-time thereby potentially enabling extended analysis and providing speedy insight into treatment.

[00166] In step 1202, biological data and clinical outcomes of previous patients may be received. The clinical outcomes may be medical characteristics. Biological data is any data that may represent a condition (e.g., a medical condition) of a person. Biological data may include any health related, medical, physical, physiological, pharmaceutical data associated with one or more patients. In one example, biological data may include measurements of gene

expressions for any number of genes. In another example, biological data may include sequencing information (e.g., RNA sequencing).

[00167] In various embodiments, biological data for a plurality of patients may be publicly available. For example, various medical health facilities and/or public entities may provide gene expression data for a variety of patients. In addition to the biological data, information regarding any number of clinical outcomes, treatments, therapies, diagnoses and/or prognoses may also be provided. Those skilled in the art will appreciate that any kind of information may be provided in addition to the biological data.

[00168] The biological data, in one example, may be similar to data **S** as discussed with regard to step 802 of FIG. 8. The biological data may include ID fields that identify patients and data fields that are related to the biological information (e.g., gene expression measurements).

[00169] FIG. 13 is an exemplary data structure 1302 including biological data 1304a-1304y for a number of patients 1308a-1308n that may be used to generate the cancer map visualization in some embodiments. Column 1302 represents different patient identifiers for different patients. The patient identifiers may be any identifier.

[00170] At least some biological data may be contained within gene expression measurements 1304a-1304y. In FIG. 13, "y" represents any number. For example, there may be 50,000 or more separate columns for different gene expressions related to a single patient or related to one or more samples from a patient. Those skilled in the art will appreciate that column 1304a may represent a gene expression measurement for each patient (if any for some patients) associated with the patient identifiers in column 1302. The column 1304b may represent a gene expression measurement of one or more genes that are different than that of column 1304a. As discussed, there may be any number of columns representing different gene expression measurements.

[00171] Column 1306 may include any number of clinical outcomes, prognoses, diagnoses, reactions, treatments, and/or any other information associated with each patient. All or some of the information contained in column 1306 may be displayed (e.g., by a label or an annotation that is displayed on the visualization or available to the user of the visualization via clicking) on or for the visualization.

[00172] Rows 1308a-1308n each contains biological data associated with the patient identifier of the row. For example, gene expressions in row 1308a are associated with patient identifier P1. As similarly discussed with regard to "y" herein, "n" represents any number. For example, there may be 100,000 or more separate rows for different patients.

[00173] Those skilled in the art will appreciate that there may be any number of data structures that contain any amount of biological data for any number of patients. The data structure(s) may be utilized to generate any number of map visualizations.

[00174] In step 1204, the analysis server may receive a filter selection. In some embodiments, the filter selection is a density estimation function. Those skilled in the art will appreciate that the filter selection may include a selection of one or more functions to generate a reference space.

[00175] In step 1206, the analysis server performs the selected filter(s) on the biological data of the previous patients to map the biological data into a reference space. In one example, a density estimation function, which is well known in the art, may be performed on the biological data (e.g., data associated with gene expression measurement data 1304a-1304y) to relate each patient identifier to one or more locations in the reference space (e.g., on a real line).

[00176] In step 1208, the analysis server may receive a resolution selection. The resolution may be utilized to identify overlapping portions of the reference space (e.g., a cover of the reference space **R**) in step 1210.

[00177] As discussed herein, the cover of **R** may be a finite collection of open sets (in the metric of **R**) such that every point in **R** lies in at least one of these sets. In various examples, **R** is k-dimensional Euclidean space, where k is the number of filter functions. Those skilled in the art will appreciate that the cover of the reference space **R** may be controlled by the number of intervals and the overlap identified in the resolution (e.g., see FIG. 7). For example, the more intervals, the finer the resolution in **S** (e.g., the similarity space of the received biological data) - that is, the fewer points in each **S**(d), but the more similar (with respect to the filters) these points may be. The greater the overlap, the more times that clusters in **S**(d) may intersect clusters in **S**(e) - this means that more "relationships" between points may appear, but, in some embodiments, the greater the overlap, the more likely that accidental relationships may appear.

[00178] In step 1212, the analysis server receives a metric to cluster the information of the cover in the reference space to partition **S**(d). In one example, the metric may be a Pearson Correlation. The clusters may form the groupings (e.g., nodes or balls). Various cluster means may be used including, but not limited to, a single linkage, average linkage, complete linkage, or k-means method.

[00179] As discussed herein, in some embodiments, the analysis module 220 may not cluster two points unless filter values are sufficiently "related" (recall that while normally related may mean "close," the cover may impose a much more general relationship on the filter values, such as relating two points s and t if **ref**(s) and **ref**(t) are sufficiently close to the same circle in the plane where **ref**() represents one or more filter functions). The output may be a simplicial complex, from which one can extract its 1-skeleton. The nodes of the complex may be partial clusters, (i.e., clusters constructed from subsets of **S** specified as the preimages of sets in the given covering of the reference space **R**).

[00180] In step 1214, the analysis server may generate the visualization map with nodes representing clusters of patient members and edges between nodes representing common patient members. In one example, the analysis server identifies nodes which are associated with a subset of the partition elements of all of the **S**(d) for generating an interactive visualization.

[00181] As discussed herein, for example, suppose that **S** = {1, 2, 3, 4}, and the cover is C_1, C_2, C_3. Suppose cover C_1 contains {1, 4}, C_2 contains {1,2}, and C_3 contains {1,2,3,4}. If 1 and 2 are close enough to be clustered, and 3 and 4 are, but nothing else, then the clustering for **S**(1) may be {1}, {4}, and for **S**(2) it may be {1,2}, and for **S**(3) it may be {1,2}, {3,4}. So the generated graph has, in this example, at most four nodes, given by the sets {1}, {4}, {1, 2}, and {3, 4} (note that {1, 2} appears in two different clusterings). Of the sets of points that are used, two nodes intersect provided that the associated node sets have a non-empty intersection (although this could easily be modified to allow users to require that the intersection is "large enough" either in absolute or relative terms).

[00182] As a result of clustering, member patients of a grouping may share biological similarities (e.g., similarities based on the biological data).

[00183] The analysis server may join clusters to identify edges (e.g., connecting lines between nodes). Clusters joined by edges (i.e., interconnections) share one or more member patients. In step 1216, a display may display a visualization map with attributes based on the clinical outcomes contained in the data structures (e.g., see FIG. 13 regarding clinical outcomes). Any labels or annotations may be utilized based on information contained in the data structures. For example, treatments, prognoses, therapies, diagnoses, and the like may be used to label the visualization. In some embodiments, the physician or other user of the map visualization accesses the annotations or labels by interacting with the map visualization.

[00184] The resulting cancer map visualization may reveal interactions and relationships that were obscured, untested, and/or previously not recognized.

[00185] FIG. 14 is an exemplary visualization displaying the cancer map visualization 1400 in some embodiments. The cancer map visualization 1400 represents a topological network of cancer patients. The cancer map visualization 1400 may be based on publicly and/or privately available data.

[00186] In various embodiments, the cancer map visualization 1400 is created using gene expression profiles of excised tumors. Each node (i.e., ball or grouping displayed in the map visualization 1400) contains a subset of patients with similar genetic profiles.

[00187] As discussed herein, one or more patients (i.e., patient members of each node or grouping) may occur in multiple nodes. A patient may share a similar genetic profile with multiple nodes or multiple groupings. In one example, of 50,000 different gene expressions of the biological data, multiple patients may share a different genetic profiles (e.g., based on different gene expression combinations) with different groupings. When a patient shares a similar genetic profile with different groupings or nodes, the patient may be included within the groupings or nodes.

[00188] The cancer map visualization 1400 comprises groupings and interconnections that are associated with different clinical outcomes. All or some of the clinical outcomes may be associated with the biological data that generated the cancer map visualization 1400. The cancer map visualization 1400 includes groupings associated with survivors 1402 and groupings associated with non-survivors 1404. The cancer map visualization 1400 also includes different groupings associated with estrogen receptor positive non-survivors 1406,

estrogen receptor negative non-survivors 1408, estrogen receptor positive survivors 1410, and estrogen receptor negative survivors 1412.

[00189] In various embodiments, when one or more patients are members of two or more different nodes, the nodes are interconnected by an edge (e.g., a line or interconnection). If there is not an edge between the two nodes, then there are no common member patients between the two nodes. For example, grouping 1414 shares at least one common member patient with grouping 1418. The intersection of the two groupings is represented by edge 1416. As discussed herein, the number of shared member patients of the two groupings may be represented in any number of ways including color of the interconnection, color of the groupings, size of the interconnection, size of the groupings, animations of the interconnection, animations of the groupings, brightness, or the like. In some embodiments, the number and/or identifiers of shared member patients of the two groupings may be available if the user interacts with the groupings 1414 and/or 1418 (e.g., draws a box around the two groupings and the interconnection utilizing an input device such as a mouse).

[00190] In various embodiments, a physician, on obtaining some data on a breast tumor, direct the data to an analysis server (e.g., analysis server 108 over a network such as the Internet) which may localize the patient relative to one or more groupings on the cancer map visualization 1400. The context of the cancer map visualization 1400 may enable the physician to assess various possible outcomes (e.g., proximity of representation of new patient to the different associations of clinical outcomes).

[00191] FIG. 15 is a flowchart of for positioning new patient data relative to a cancer map visualization in some embodiments. In step 1502, new biological data of a new patient is received. In various embodiments, an input module 214 of an analysis server (e.g., analysis server 108 of FIGs 1 and 2) may receive biological data of a new patient from a physician or medical facility that performed analysis of one or more samples to generate the biological data. The biological data may be any data that represents a biological data of the new patient including, for example, gene expressions, sequencing information, or the like.

[00192] In some embodiments, the analysis server 108 may comprise a new patient distance module and a location engine. In step 1504, the new patient distance module determines distances between the biological data of each patient of the cancer map visualization 1600 and the new biological data from the new patient. For example, the previous biological data that

was utilized in the generation of the cancer map visualization 1600 may be stored in mapped data structures. Distances may be determined between the new biological data of the new patient and each of the previous patient's biological data in the mapped data structure.

[00193] Those skilled in the art will appreciate that distances may be determined in any number of ways using any number of different metrics or functions. Distances may be determined between the biological data of the previous patients and the new patients. For example, a distance may be determined between a first gene expression measurement of the new patient and each (or a subset) of the first gene expression measurements of the previous patients (e.g., the distance between G1 of the new patient and G1 of each previous patient may be calculated). Distances may be determined between all (or a subset of) other gene expression measurements of the new patient to the gene expression measurements of the previous patients.

[00194] In various embodiments, a location of the new patient on the cancer map visualization 1600 may be determined relative to the other member patients utilizing the determined distances.

[00195] In step 1506, the new patient distance module may compare distances between the patient members of each grouping to the distances determined for the new patient. The new patient may be located in the grouping of patient members that are closest in distance to the new patient. In some embodiments, the new patient location may be determined to be within a grouping that contains the one or more patient members that are closest to the new patient (even if other members of the grouping have longer distances with the new patient). In some embodiments, this step is optional.

[00196] In various embodiments, a representative patient member may be determined for each grouping. For example, some or all of the patient members of a grouping may be averaged or otherwise combined to generate a representative patient member of the grouping (e.g., the distances and/or biological data of the patient members may be averaged or aggregated). Distances may be determined between the new patient biological data and the averaged or combined biological data of one or more representative patient members of one or more groupings. The location engine may determine the location of the new patient based on the distances. In some embodiments, once the closest distance between the new patient and the representative patient member is found, distances may be determined between the new patient

and the individual patient members of the grouping associated with the closest representative patient member.

[00197] In optional step 1508, a diameter of the grouping with the one or more of the patient members that are closest to the new patient (based on the determined distances) may be determined. In one example, the diameters of the groupings of patient members closest to the new patient are calculated. The diameter of the grouping may be a distance between two patient members who are the farthest from each other when compared to the distances between all patient members of the grouping. If the distance between the new patient and the closest patient member of the grouping is less than the diameter of the grouping, the new patient may be located within the grouping. If the distance between the new patient and the closest patient member of the grouping is greater than the diameter of the grouping, the new patient may be outside the grouping (e.g., a new grouping may be displayed on the cancer map visualization with the new patient as the single patient member of the grouping). If the distance between the new patient and the closest patient member of the grouping is equal to the diameter of the grouping, the new patient may be placed within or outside the grouping.

[00198] It will be appreciated that the determination of the diameter of the grouping is not required in determining whether the new patient location is within or outside of a grouping. In various embodiments, a distribution of distances between member patients and between member patients and the new patient is determined. The decision to locate the new patient within or outside of the grouping may be based on the distribution. For example, if there is a gap in the distribution of distances, the new patient may be separated from the grouping (e.g., as a new grouping). In some embodiments, if the gap is greater than a preexisting threshold (e.g., established by the physician, other user, or previously programmed), the new patient may be placed in a new grouping that is placed relative to the grouping of the closest member patients. The process of calculating the distribution of distances of candidate member patients to determine whether there may be two or more groupings may be utilized in generation of the cancer map visualization (e.g., in the process as described with regard to FIG. 12). Those skilled in the art will appreciate that there may be any number of ways to determine whether a new patient should be included within a grouping of other patient members.

[00199] In step 1510, the location engine determines the location of the new patient relative to the member patients and/or groupings of the cancer map visualization. The new location may

be relative to the determined distances between the new patient and the previous patients. The location of the new patient may be part of a previously existing grouping or may form a new grouping.

[00200] In some embodiments, the location of the new patient with regard to the cancer map visualization may be performed locally to the physician. For example, the cancer map visualization 1400 may be provided to the physician (e.g., via digital device). The physician may load the new patient's biological data locally and the distances may be determined locally or via a cloud-based server. The location(s) associated with the new patient may be overlaid on the previously existing cancer map visualization either locally or remotely.

[00201] Those skilled in the art will appreciate that, in some embodiments, the previous state of the cancer map visualization (e.g., cancer map visualization 1400) may be retained or otherwise stored and a new cancer map visualization generated utilizing the new patient biological data (e.g., in a method similar to that discussed with regard to FIG. 12). The newly generated map may be compared to the previous state and the differences may be highlighted thereby, in some embodiments, highlighting the location(s) associated with the new patient. In this way, distances may be not be calculated as described with regard to FIG. 15, but rather, the process may be similar to that as previously discussed.

[00202] FIG. 16 is an exemplary visualization displaying the cancer map including positions for three new cancer patients in some embodiments. The cancer map visualization 1400 comprises groupings and interconnections that are associated with different clinical outcomes as discussed with regard to FIG. 14. All or some of the clinical outcomes may be associated with the biological data that generated the cancer map visualization 1400. The cancer map visualization 1400 includes different groupings associated with survivors 1402, groupings associated with non-survivors 1404, estrogen receptor positive non-survivors 1406, estrogen receptor negative non-survivors 1408, estrogen receptor positive survivors 1410, and estrogen receptor negative survivors 1412.

[00203] The cancer map visualization 1400 includes three locations for three new breast cancer patients. The breast cancer patient location 1602 is associated with the clinical outcome of estrogen receptor positive survivors. The breast cancer patient location 1604 is associated with the clinical outcome of estrogen receptor negative survivors. Unfortunately, breast cancer patient location 1606 is associated with estrogen receptor negative non-survivors. Based on

the locations, a physician may consider different diagnoses, prognoses, treatments, and therapies to maintain or attempt to move the breast cancer patient to a different location utilizing the cancer map visualization 1400.

[00204] In some embodiments, the physician may assess the underlying biological data associated with any number of member patients of any number of groupings to better understand the genetic similarities and/or dissimilarities. The physician may utilize the information to make better informed decisions.

[00205] The patient location 1604 is highlighted on the cancer map visualization 1400 as active (e.g., selected by the physician). Those skilled in the art will appreciate that the different locations may be of any color, size, brightness, and/or animated to highlight the desired location(s) for the physician. Further, although only one location is identified for three different breast cancer patients, any of the breast cancer patients may have multiple locations indicating different genetic similarities.

[00206] Those skilled in the art will appreciate that the cancer map visualization 1400 may be updated with new information at any time. As such, as new patients are added to the cancer map visualization 1400, the new data updates the visualization such that as future patients are placed in the map, the map may already include the updated information. As new information and/or new patient data is added to the cancer map visualization 1400, the cancer map visualization 1400 may improve as a tool to better inform physicians or other medical professionals.

[00207] In various embodiments, the cancer map visualization 1400 may track changes in patients over time. For example, updates to a new patient may be visually tracked as changes in are measured in the new patient's biological data. In some embodiments, previous patient data is similarly tracked which may be used to determine similarities of changes based on condition, treatment, and/or therapies, for example. In various embodiments, velocity of change and/or acceleration of change of any number of patients may be tracked over time using or as depicted on the cancer map visualization 1400. Such depictions may assist the treating physician or other personnel related to the treating physician to better understand changes in the patient and provide improved, current, and/or updated diagnoses, prognoses, treatments, and/or therapies.

[00208] FIG. 17 is a flowchart of utilization the visualization and positioning of new patient data in some embodiments. In various embodiments, a physician may collect amounts of genomic information from tumors removed from a new patient, input the data (e.g., upload the data to an analysis server), and receive a map visualization with a location of the new patient. The new patient's location within the map may offer the physician new information about the similarities to other patients. In some embodiments, the map visualization may be annotated so that the physician may check the outcomes of previous patients in a given region of the map visualization are distributed and then use the information to assist in decision-making for diagnosis, treatment, prognosis, and/or therapy.

[00209] In step 1702, a medical professional or other personnel may remove a sample from a patient. The sample may be of a tumor, blood, or any other biological material. In one example, a medical professional performs a tumor excision. Any number of samples may be taken from a patient.

[00210] In step 1704, the sample(s) may be provided to a medical facility to determine new patient biological data. In one example, the medical facility measures genomic data such as gene expression of a number of genes or protein levels.

[00211] In step 1706, the medical professional or other entity associated with the medical professional may receive the new patient biological data based on the sample(s) from the new patient. In one example, a physician may receive the new patient biological data. The physician may provide all or some of the new patient biological data to an analysis server over the Internet (e.g., the analysis server may be a cloud-based server). In some embodiments, the analysis server is the analysis server 108 of FIG. 1. In some embodiments, the medical facility that determines the new patient biological data provides the biological data in an electronic format which may be uploaded to the analysis server. In some embodiments, the medical facility that determines the new patient biological data (e.g., the medical facility that measures the genomic data) provide the biological data to the analysis server at the request of the physician or others associated with the physician. Those skilled in the art will appreciate that the biological data may be provided to the analysis server in any number of ways.

[00212] The analysis server may be any digital device and may not be limited to a digital device on a network. In some embodiments, the physician may have access to the digital

device. For example, the analysis server may be a table, personal computer, local server, or any other digital device.

[00213] Once the analysis server receives the biological data of the new patient, the new patient may be localized in the map visualization and the information may be sent back to the physician in step 1708. The visualization may be a map with nodes representing clusters of previous patient members and edges between nodes representing common patient members. The visualization may further depict one or more locations related to the biological data of the new patient.

[00214] The map visualization may be provided to the physician or other associated with the physician in real-time. For example, once the biological data associated with the new patient is provided to the analysis server, the analysis server may provide the map visualization back to the physician or other associated with the physician within a reasonably short time (e.g., within seconds or minutes). In some embodiments, the physician may receive the map visualization over any time.

[00215] The map visualization may be provided to the physician in any number of ways. For example, the physician may receive the map visualization over any digital device such as, but not limited to, an office computer, Ipad, tablet device, media device, smartphone, e-reader, or laptop.

[00216] In step 1710, the physician may assess possible different clinical outcomes based on the map visualization. In one example, the map-aided physician may make decisions on therapy and treatments depending on where the patient lands on the visualization (e.g., survivor or non-survivor). The map visualization may include annotations or labels that identify one or more sets of groupings and interconnections as being associated with one or more clinical outcomes. The physician may assess possible clinical outcomes based on the position(s) on the map associated with the new patient.

[00217] FIG. 18 is a block diagram of an exemplary digital device 1800. The digital device 1800 comprises a processor 1802, a memory system 1804, a storage system 1806, a communication network interface 1808, an I/O interface 1810, and a display interface 1812 communicatively coupled to a bus 1814. The processor 1802 may be configured to execute

executable instructions (e.g., programs). In some embodiments, the processor 1802 comprises circuitry or any processor capable of processing the executable instructions.

[00218] The memory system 1804 is any memory configured to store data. Some examples of the memory system 1804 are storage devices, such as RAM or ROM. The memory system 1804 can comprise the ram cache. In various embodiments, data is stored within the memory system 1804. The data within the memory system 1804 may be cleared or ultimately transferred to the storage system 1806.

[00219] The storage system 1806 is any storage configured to retrieve and store data. Some examples of the storage system 1806 are flash drives, hard drives, optical drives, and/or magnetic tape. In some embodiments, the digital device 1800 includes a memory system 1804 in the form of RAM and a storage system 1806 in the form of flash data. Both the memory system 1804 and the storage system 1806 comprise computer readable media which may store instructions or programs that are executable by a computer processor including the processor 1802.

[00220] The communication network interface (com. network interface) 1808 can be coupled to a data network (e.g., data network 504 or 514) via the link 1816. The communication network interface 1808 may support communication over an Ethernet connection, a serial connection, a parallel connection, or an ATA connection, for example. The communication network interface 1808 may also support wireless communication (e.g., 1802.11 a/b/g/n, WiMax). It will be apparent to those skilled in the art that the communication network interface 1808 can support many wired and wireless standards.

[00221] The optional input/output (I/O) interface 1810 is any device that receives input from the user and output data. The optional display interface 1812 is any device that may be configured to output graphics and data to a display. In one example, the display interface 1812 is a graphics adapter.

[00222] It will be appreciated by those skilled in the art that the hardware elements of the digital device 1800 are not limited to those depicted in FIG. 18. A digital device 1800 may comprise more or less hardware elements than those depicted. Further, hardware elements may share functionality and still be within various embodiments described herein. In one

example, encoding and/or decoding may be performed by the processor 1802 and/or a co-processor located on a GPU.

[00223] The above-described functions and components can be comprised of instructions that are stored on a storage medium (e.g., a computer readable storage medium). The instructions can be retrieved and executed by a processor. Some examples of instructions are software, program code, and firmware. Some examples of storage medium are memory devices, tape, disks, integrated circuits, and servers. The instructions are operational when executed by the processor to direct the processor to operate in accord with embodiments of the present invention. Those skilled in the art are familiar with instructions, processor(s), and storage medium.

[00224] The present invention has been described above with reference to exemplary embodiments. It will be apparent to those skilled in the art that various modifications may be made and other embodiments can be used without departing from the broader scope of the invention. Therefore, these and other variations upon the exemplary embodiments are intended to be covered by the present invention.

CLAIMS

What is claimed is:

1. A system comprising:

a map including a plurality of groupings and interconnections of the groupings, each grouping having one or more patient members that share biological similarities, each interconnection interconnecting groupings that share at least one common patient member, the map identifying a set of groupings and a set of interconnections having a medical characteristic of a set of medical characteristics; and

a location engine configured to determine whether a new patient shares the biological similarities with the one or more patient members of each grouping,

thereby enabling association of the new patient with one or more of the set of medical characteristics.

2. The system of claim 1 wherein the biological similarities represent similarities of measurements of gene expressions.

3. The system of claim 1 wherein the biological similarities represent similarities of sequencing.

4. The system of claim 1 wherein the map is generated by an analysis server configured to receive biological data associated with the one or more patient members, apply a filtering function to generate a reference space, generate a cover of the reference space based on a resolution, the cover including cover data associated with the filtered biological data, cluster the cover data based on a metric, and display the groupings and the interconnections based on the clusters.

5. The system of claim 4 wherein the filtering function is a density estimation function.

6. The system of claim 4 wherein the metric is a Pearson correlation.

7. The system of claim 1 wherein the location engine configured to determine whether the new patient shares the biological similarities with the one or more patient members of each grouping comprises the patient location engine configured to determine a distance between biological data of each patient member and new biological data of the new patient, compare distances between the patient members of each grouping and the distances determined for the

new patient, and determine a location of the new patient relative to at least one of the member patients.

8. The system of claim 7 wherein the location engine is further configured to compare distances to one or more of the patient members closest to the new patient's filtered biological data with a diameter of at least one grouping and to indicate that the new patient is associated with the grouping based on the comparison.

9. The system of claim 7 wherein the location engine is further configured to determine if the distance to one or more of the patient members closest to the new patient's filtered biological data is greater than a diameter of each grouping and to indicate that the new patient is not associated with each grouping based on the comparison.

10. The system of claim 1 wherein the medical characteristic comprises a clinical outcome.

11. A method comprising:

receiving biological data of a new patient;

determining distances between biological data of patient members of a map and new biological data from the new patient, the map including a plurality of groupings and interconnections of the groupings, each grouping having one or more of the patient members that share biological similarities, each interconnection interconnecting groupings that share at least one common patient member, the map identifying a set of groupings and a set of interconnections having a medical characteristic of a set of medical characteristics;

comparing distances between the one or more patient members and the distances determined for the new patient; and

determining a location of the new patient relative to the member patients of the map based on the comparison, thereby enabling association of the new patient with one or more of the set of medical characteristics.

12. The method of claim 11 wherein the biological similarities represent similarities of measurements of gene expressions.

13. The method of claim 11 wherein the biological similarities represent similarities of sequencing.

14. The method of claim 11 further comprising:

receiving biological data associated with the one or more patient members;

applying a filtering function to generate a reference space, generate a cover of the reference space based on a resolution, the cover including cover data associated with the filtered biological data;

clustering the cover data based on a metric; and

displaying the groupings and the interconnections based on the clusters.

15. The method of claim 14 wherein the filtering function is a density estimation function.

16. The method of claim 14 wherein the metric is a Pearson correlation.

17. The method of claim 14 further comprising comparing distances to one or more of the patient members closest to the new patient's filtered biological data with a diameter of at least one grouping and indicating that the new patient is associated with the grouping based on the comparison.

18. The method of claim 14 further comprising determining if the distance to one or more of the patient members closest to the new patient's filtered biological data is greater than a diameter of each grouping and indicating that the new patient is not associated with each grouping based on the comparison.

19. The method of claim 11 wherein the medical characteristic comprises a clinical outcome.

20. A computer readable medium comprising instructions, the instructions being executable by a processor to perform a method, the method comprising:

receiving biological data of a new patient;

determining distances between biological data of patient members of a map and new biological data from the new patient, the map including a plurality of groupings and interconnections of the groupings, each grouping having one or more of the patient members that share biological similarities, each interconnection interconnecting groupings that share at least one common patient member, the map identifying a set of groupings and a set of interconnections having a medical characteristic of a set of medical characteristics;

comparing distances between the one or more patient members and the distances determined for the new patient; and

determining a location of the new patient relative to the member patients of the map based on the comparison, thereby enabling association of the new patient with one or more of the set of medical characteristics.

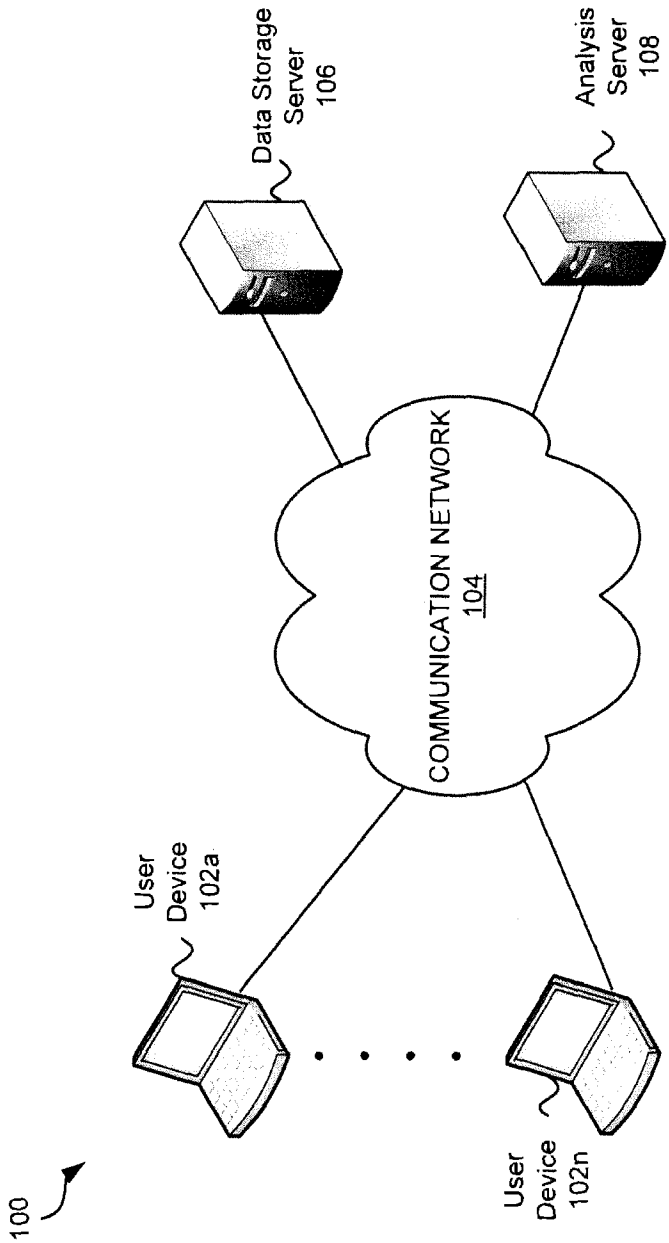

Data Storage Server 106
Analysis Server 108
COMMUNICATION NETWORK 104
User Device 102a
User Device 102n
100

FIG. 1

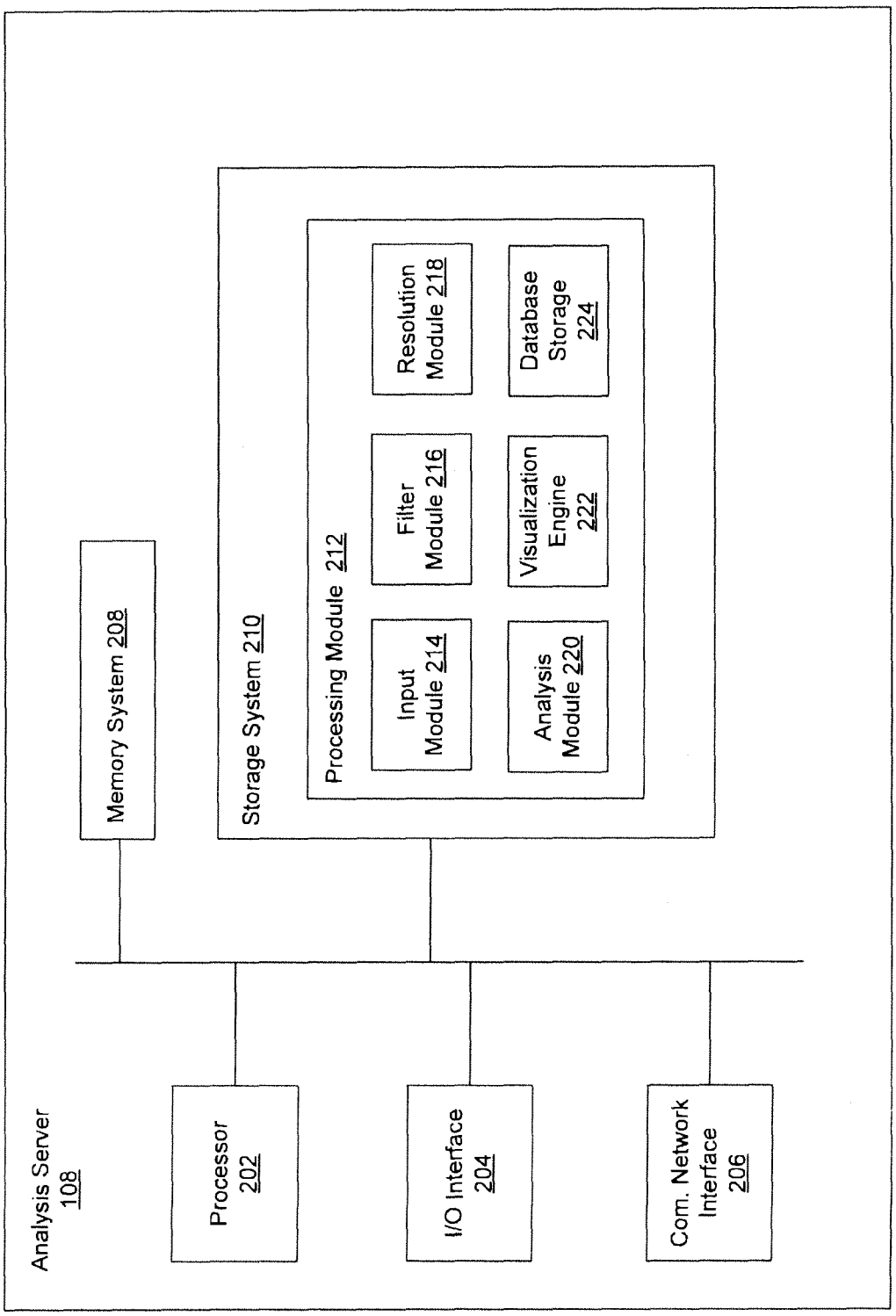

Analysis Server 108
Processor 202
I/O Interface 204
Com. Network Interface 206
Memory System 208
Storage System 210
Processing Module 212
Input Module 214
Filter Module 216
Resolution Module 218
Analysis Module 220
Visualization Engine 222
Database Storage 224

FIG. 2

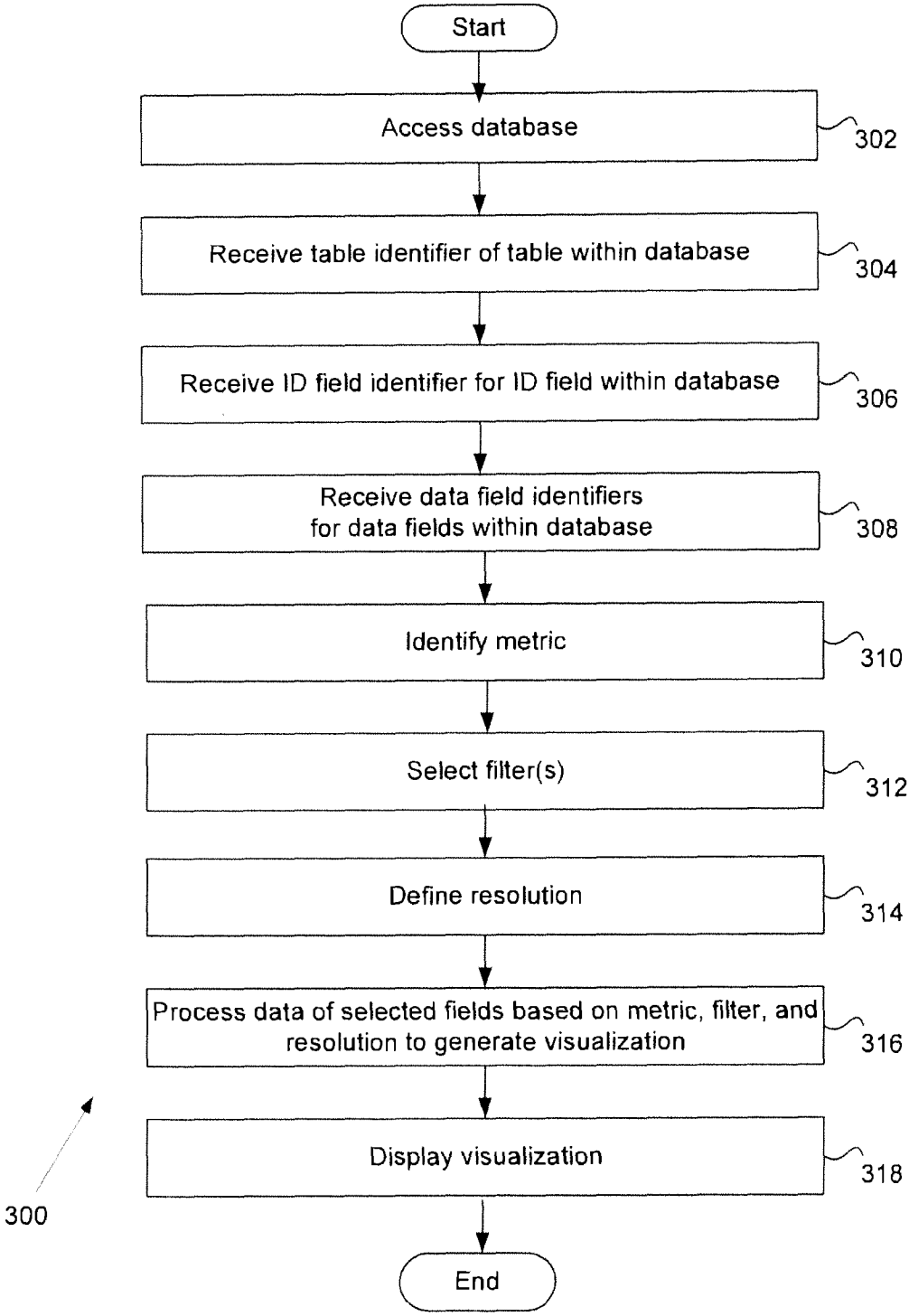
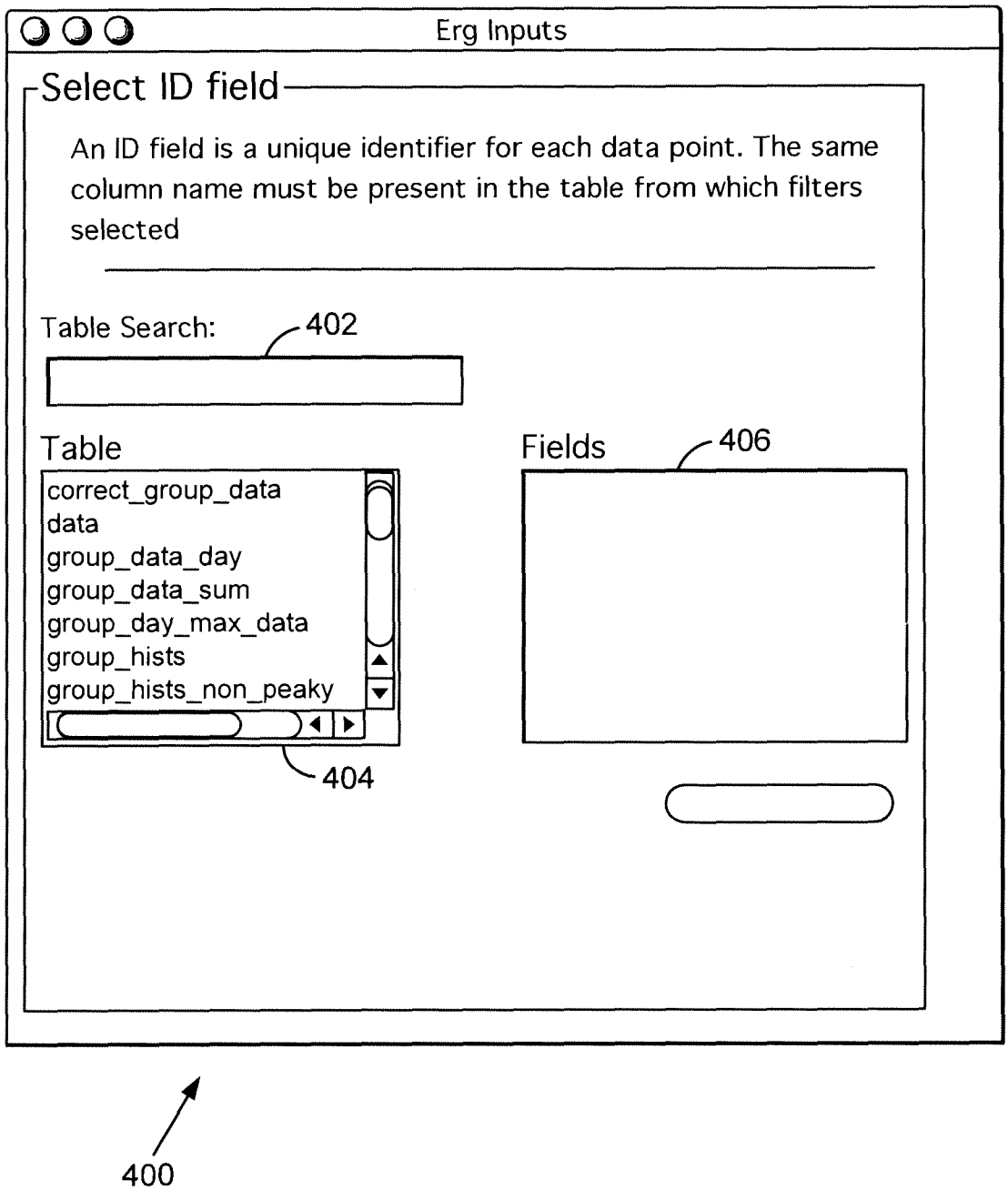

Start
Access database
302
Receive table identifier of table within database
304
Receive ID field identifier for ID field within database
306
Receive data field identifiers for data fields within database
308
Identify metric
310
Select filter(s)
312
Define resolution
314
Process data of selected fields based on metric, filter, and resolution to generate visualization
316
Display visualization
318
End
300

FIG. 3


Erg Inputs
Select ID field
An ID field is a unique identifier for each data point. The same column name must be present in the table from which filters selected
Table Search:
402
Table
correct_group_data
data
group_data_day
group_data_sum
group_day_max_data
group_hists
group_hists_non_peaky
404
Fields
406
400

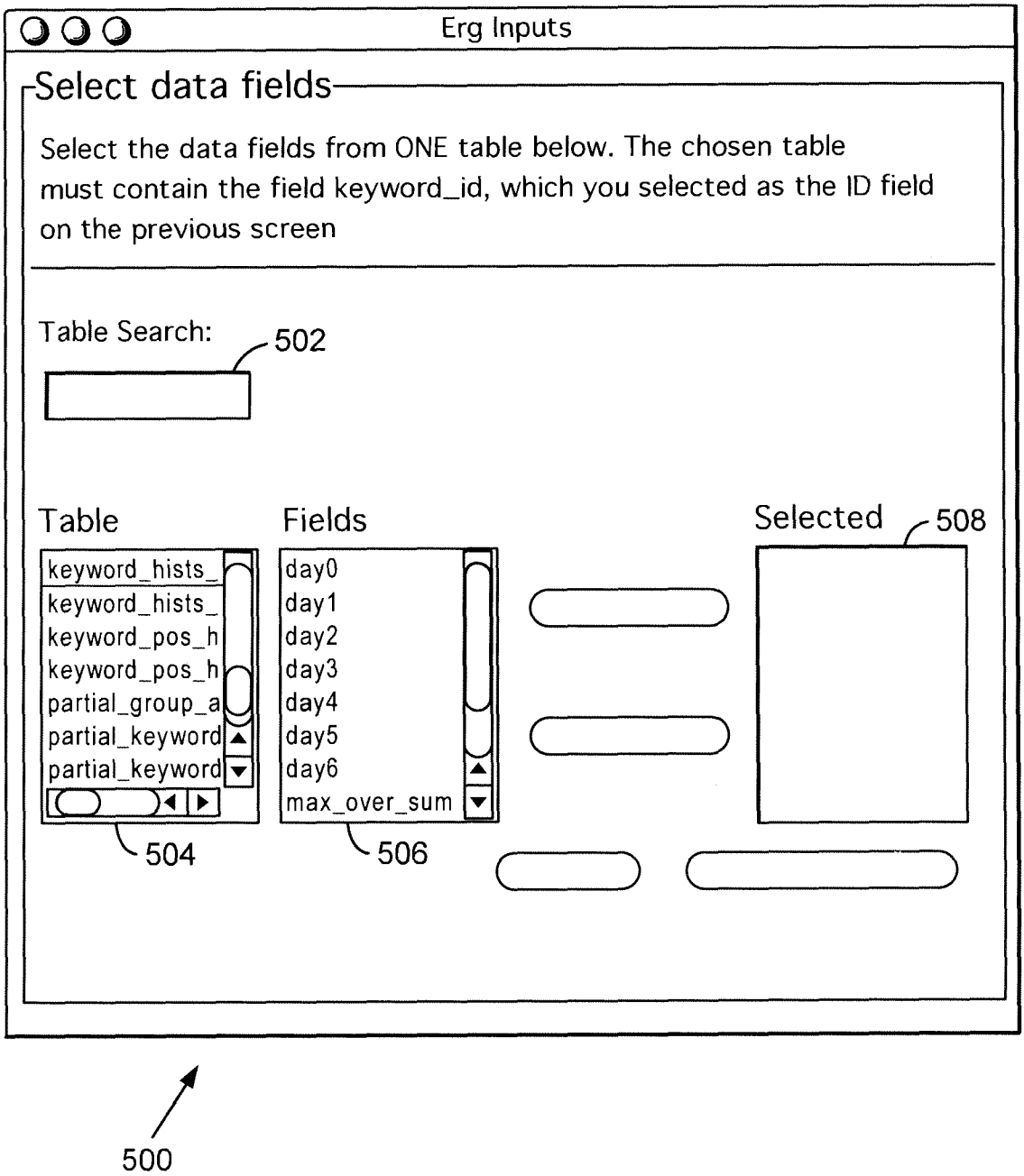

Erg Inputs
Select data fields
Select the data fields from ONE table below. The chosen table must contain the field keyword_id, which you selected as the ID field on the previous screen
Table Search:
502
Table
Fields
Selected
508
keyword_hists_
keyword_hists_
keyword_pos_h
keyword_pos_h
partial_group_a
partial_keyword
partial_keyword
day0
day1
day2
day3
day4
day5
day6
max_over_sum
504
506
500

FIG. 5

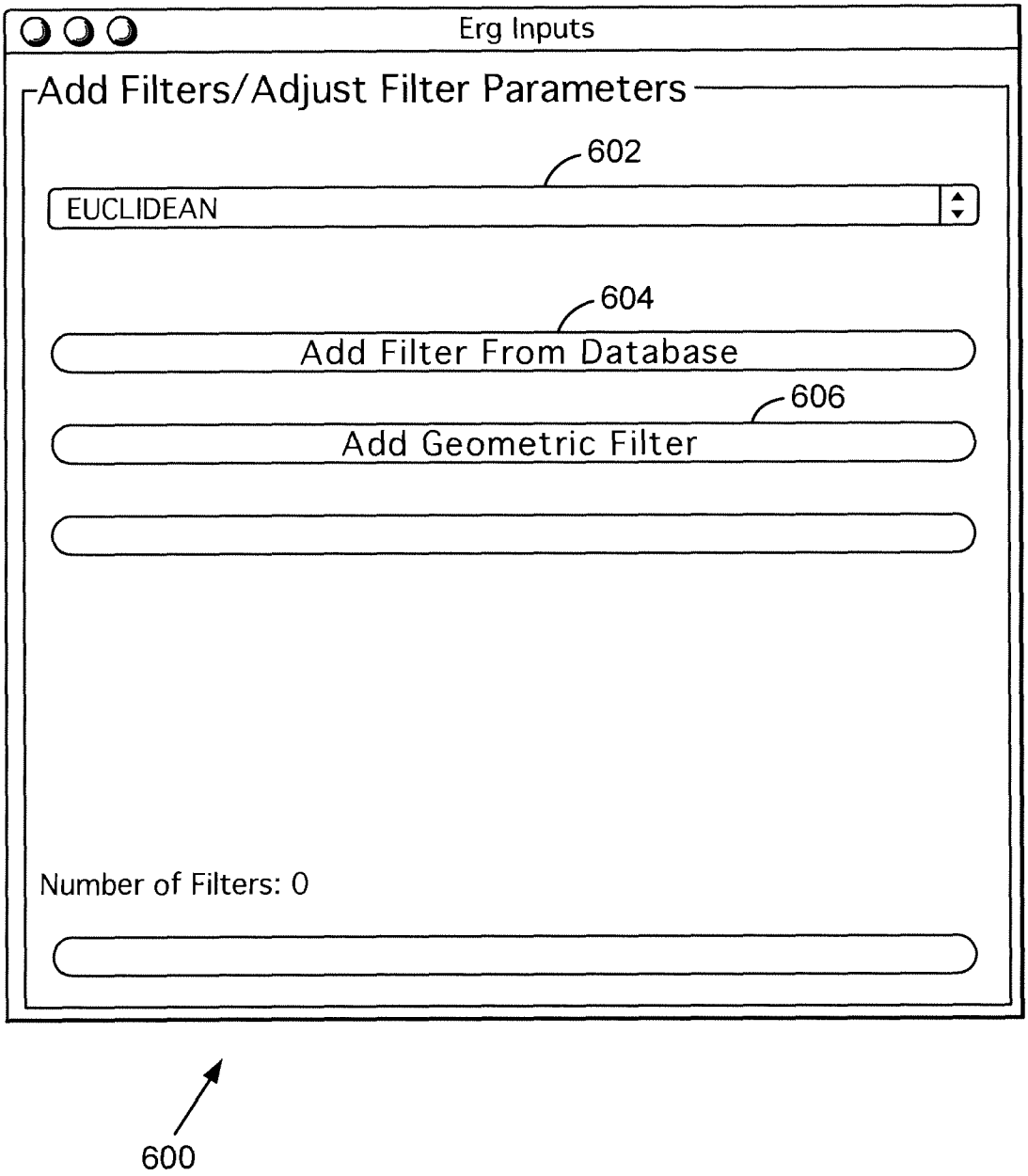

Erg Inputs
Add Filters/Adjust Filter Parameters
602
EUCLIDEAN
604
Add Filter From Database
606
Add Geometric Filter
Number of Filters: 0
600

FIG. 6

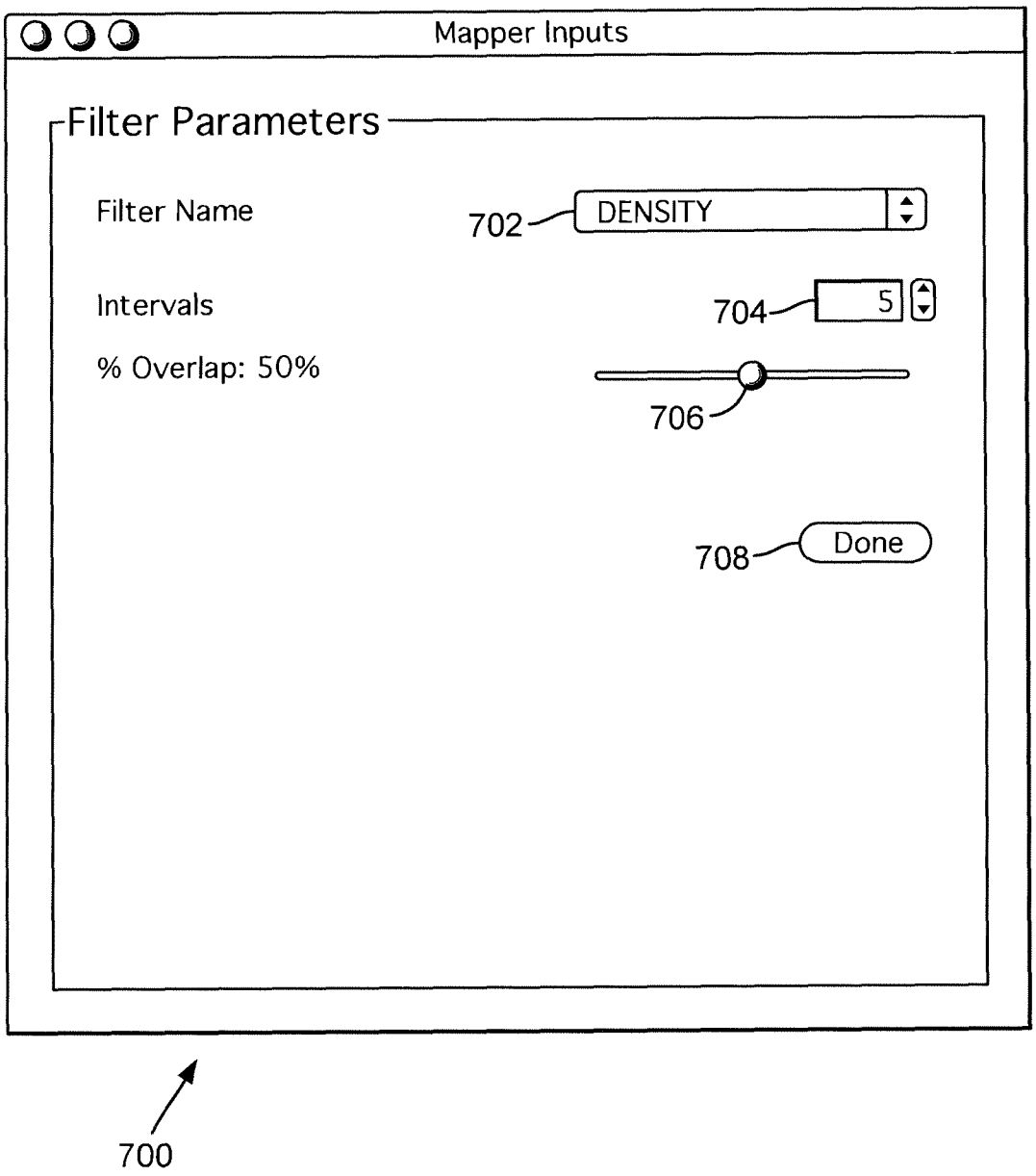

Mapper Inputs
Filter Parameters
Filter Name
702
DENSITY
Intervals
704
5
% Overlap: 50%
706
708
Done
700

FIG. 7

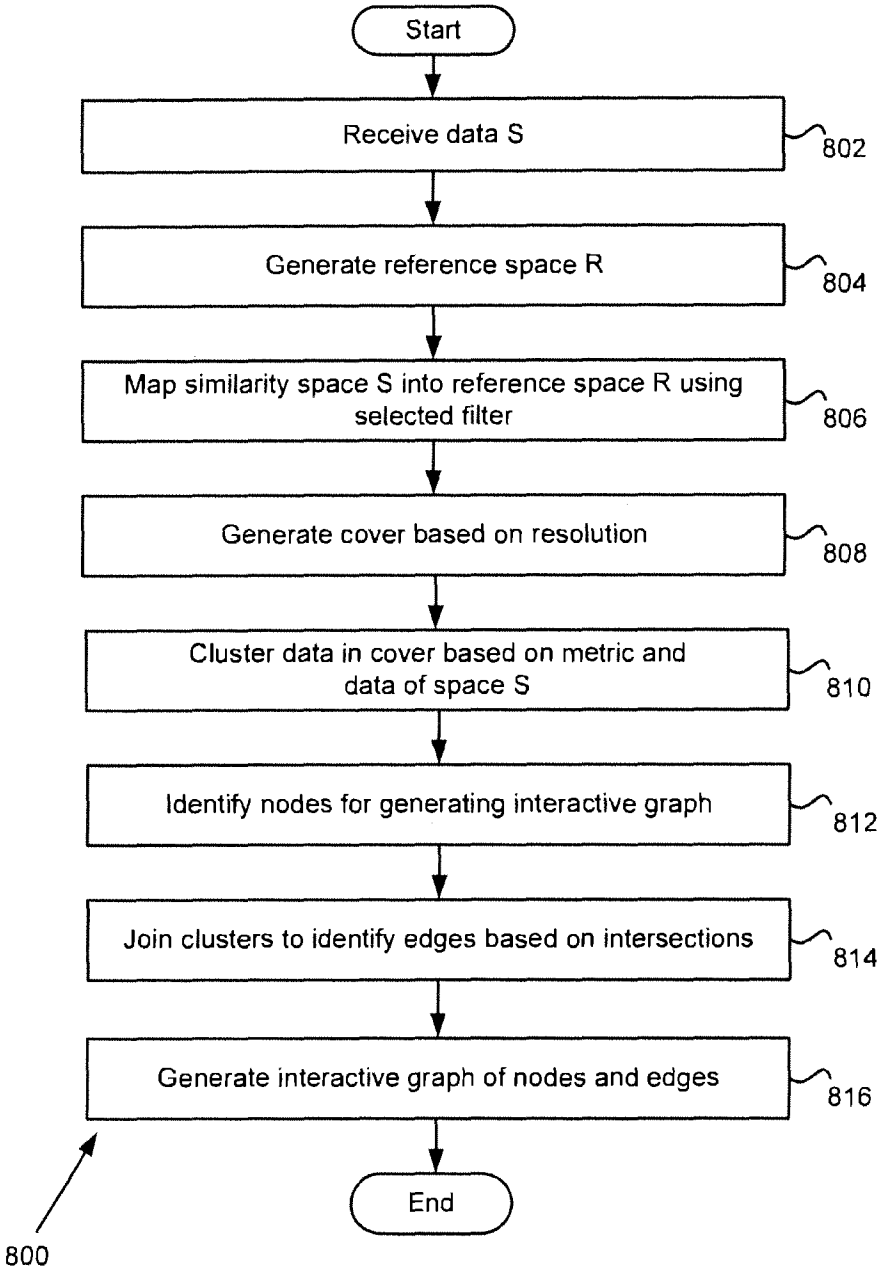

Start
Receive data S
802
Generate reference space R
804
Map similarity space S into reference space R using selected filter
806
Generate cover based on resolution
808
Cluster data in cover based on metric and data of space S
810
Identify nodes for generating interactive graph
812
Join clusters to identify edges based on intersections
814
Generate interactive graph of nodes and edges
816
End
800

FIG. 8

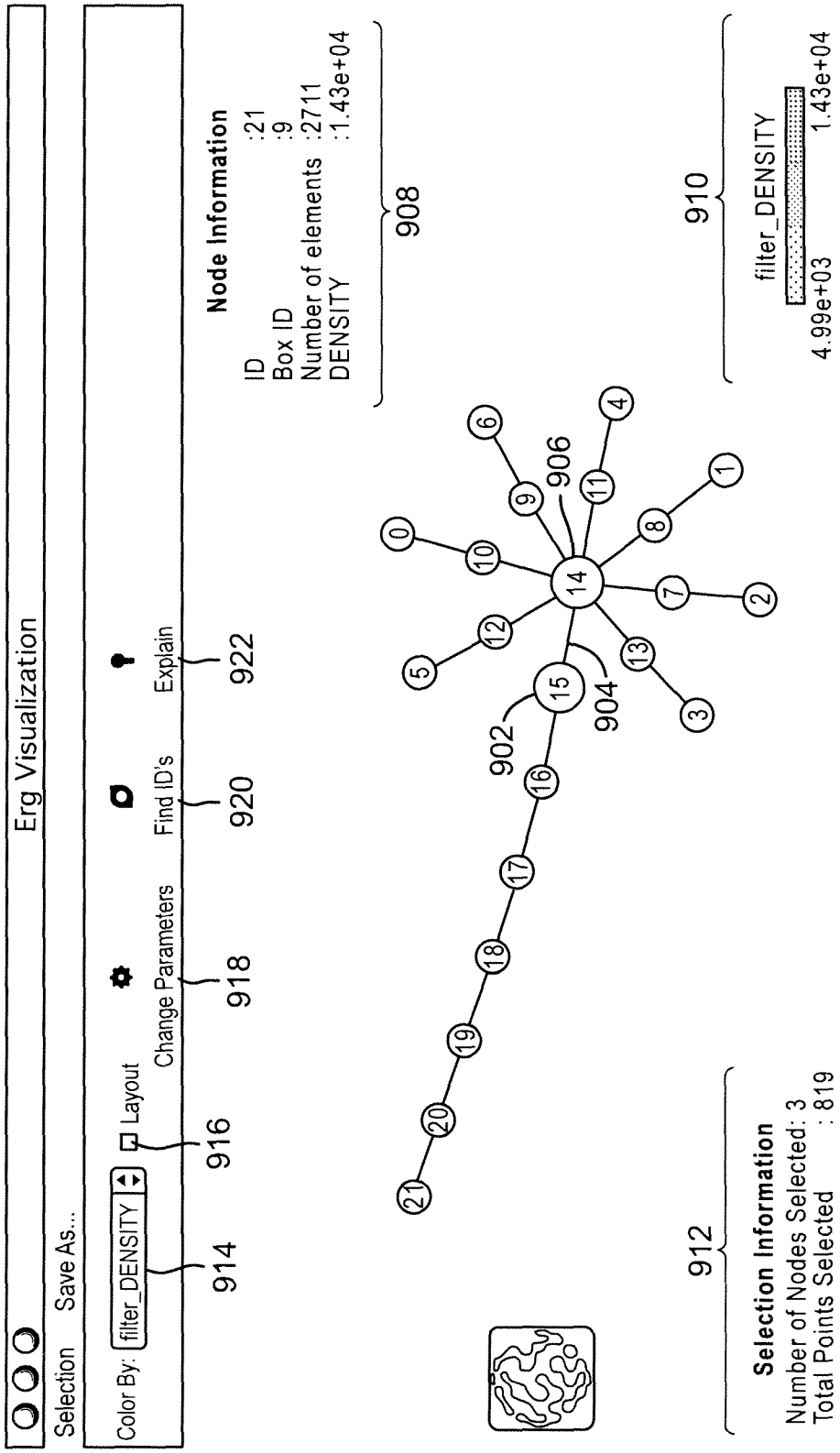

Erg Visualization
Selection
Save As...
Color By:
filter_DENSITY
Layout
Change Parameters
Find ID's
Explain
914
916
918
920
922
Node Information
ID :21
Box ID :9
Number of elements :2711
DENSITY :1.43e+04
908
910
filter_DENSITY
4.99e+03
1.43e+04
902
904
906
912
Selection Information
Number of Nodes Selected: 3
Total Points Selected : 819
900

FIG. 9

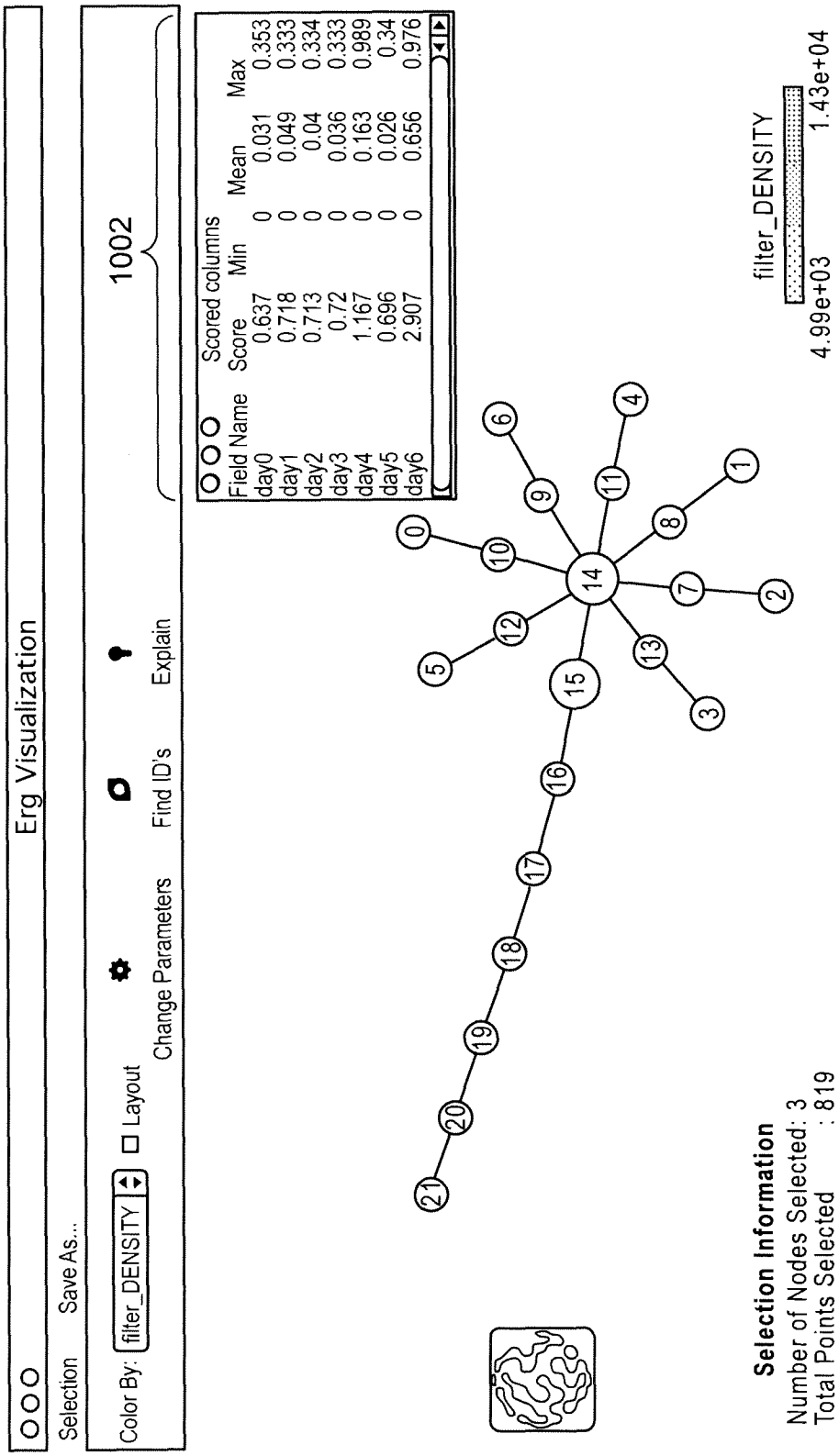

Erg Visualization
Selection
Save As...
Color By: filter_DENSITY
Layout
Change Parameters
Find ID's
Explain
1002
Scored columns
Field Name Score Min Mean Max
day0 0.637 0 0.031 0.353
day1 0.718 0 0.049 0.333
day2 0.713 0 0.04 0.334
day3 0.72 0 0.036 0.333
day4 1.167 0 0.163 0.989
day5 0.696 0 0.026 0.34
day6 2.907 0 0.656 0.976
filter_DENSITY
4.99e+03
1.43e+04
Selection Information
Number of Nodes Selected: 3
Total Points Selected : 819
1000

FIG. 10

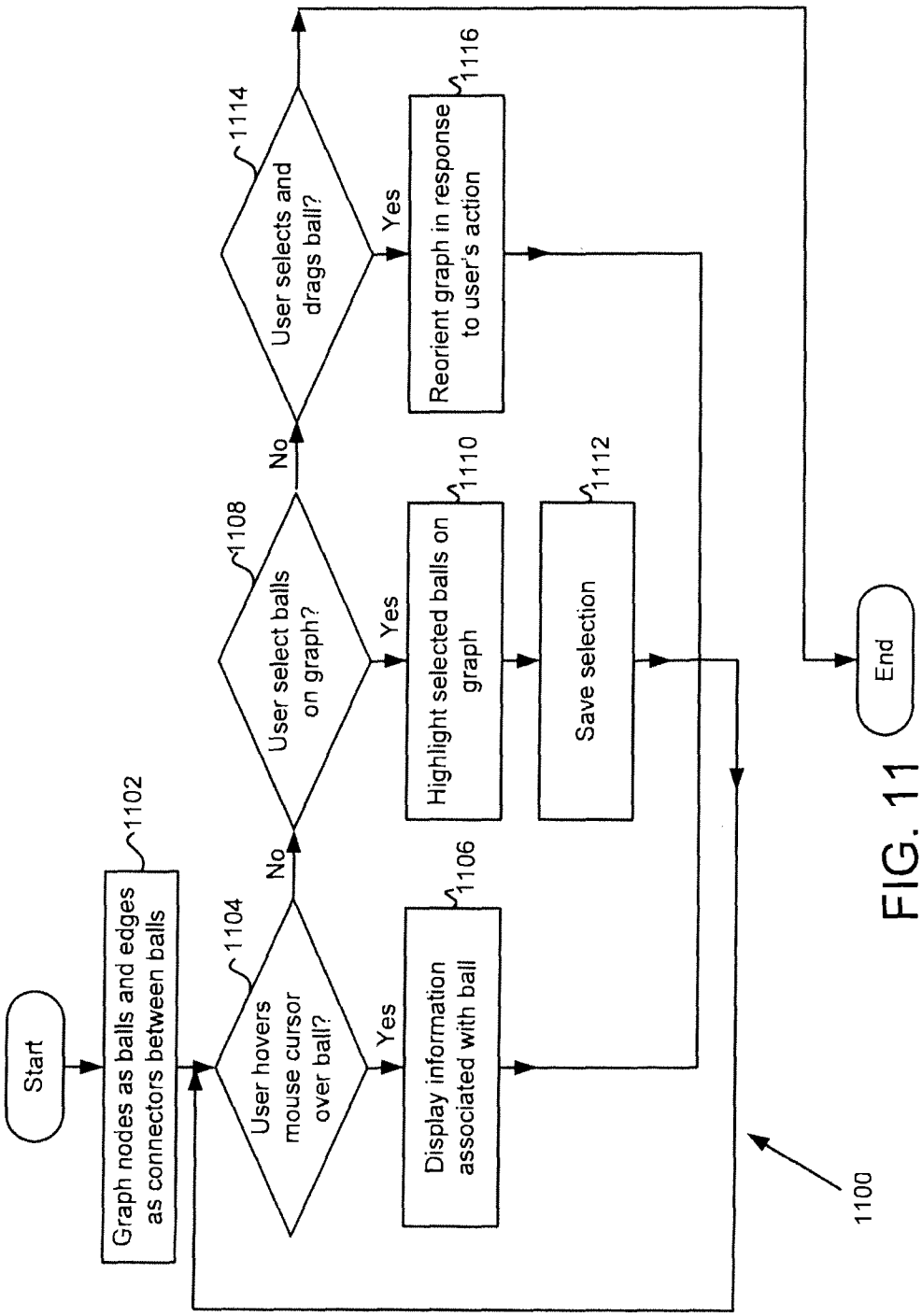

Start
Graph nodes as balls and edges as connectors between balls
1102
User hovers mouse cursor over ball?
1104
Yes
Display information associated with ball
1106
No
User select balls on graph?
1108
Yes
Highlight selected balls on graph
1110
Save selection
1112
No
User selects and drags ball?
1114
Yes
Reorient graph in response to user's action
1116
End
1100

FIG. 11

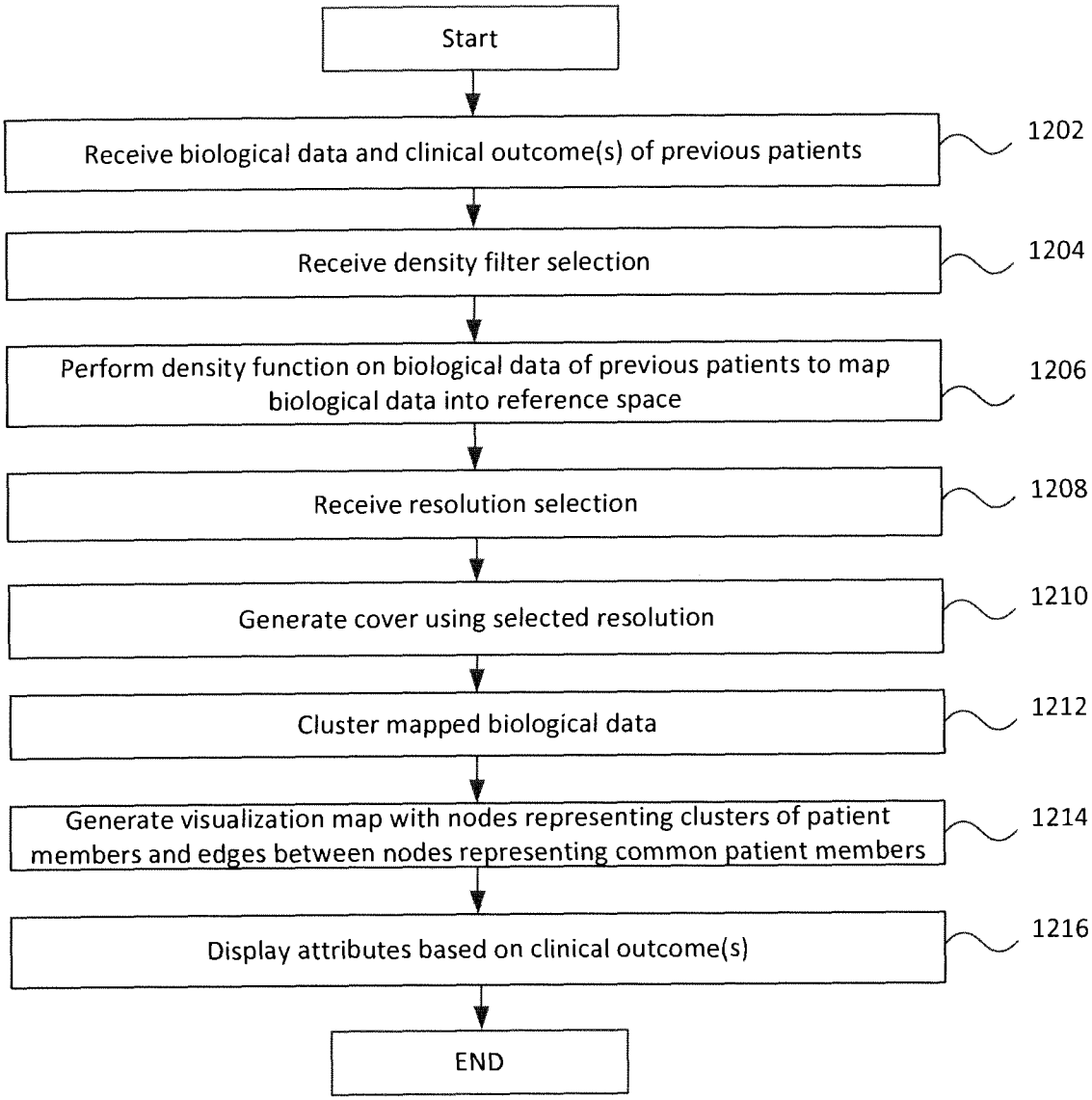

Start
Receive biological data and clinical outcome(s) of previous patients
1202
Receive density filter selection
1204
Perform density function on biological data of previous patients to map biological data into reference space
1206
Receive resolution selection
1208
Generate cover using selected resolution
1210
Cluster mapped biological data
1212
Generate visualization map with nodes representing clusters of patient members and edges between nodes representing common patient members
1214
Display attributes based on clinical outcome(s)
1216
END

FIG. 12

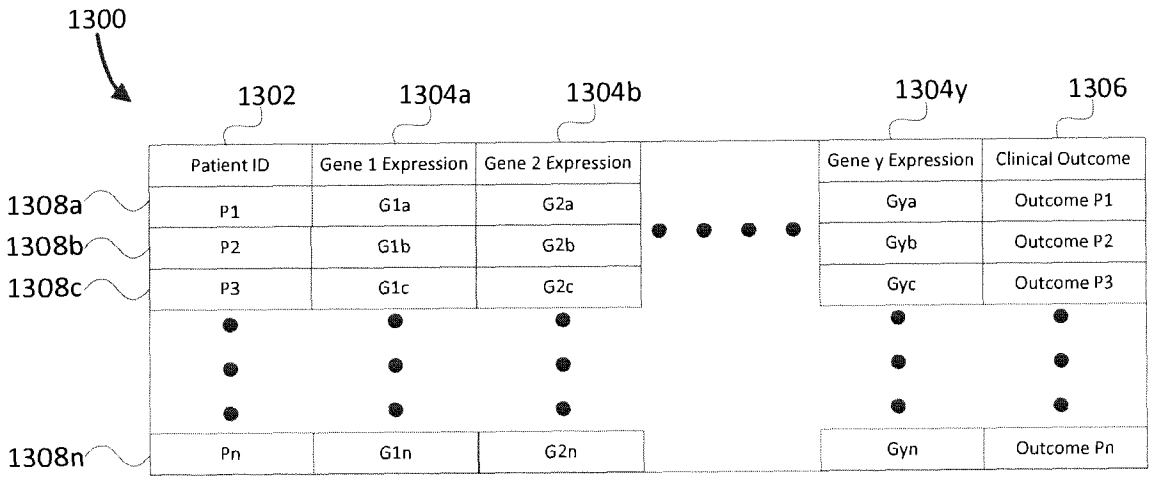

1300
1302
1304a
1304b
1304y
1306
Patient ID
Gene 1 Expression
Gene 2 Expression
Gene y Expression
Clinical Outcome
1308a
P1
G1a
G2a
Gya
Outcome P1
1308b
P2
G1b
G2b
Gyb
Outcome P2
1308c
P3
G1c
G2c
Gyc
Outcome P3
1308n
Pn
G1n
G2n
Gyn
Outcome Pn

FIG. 13

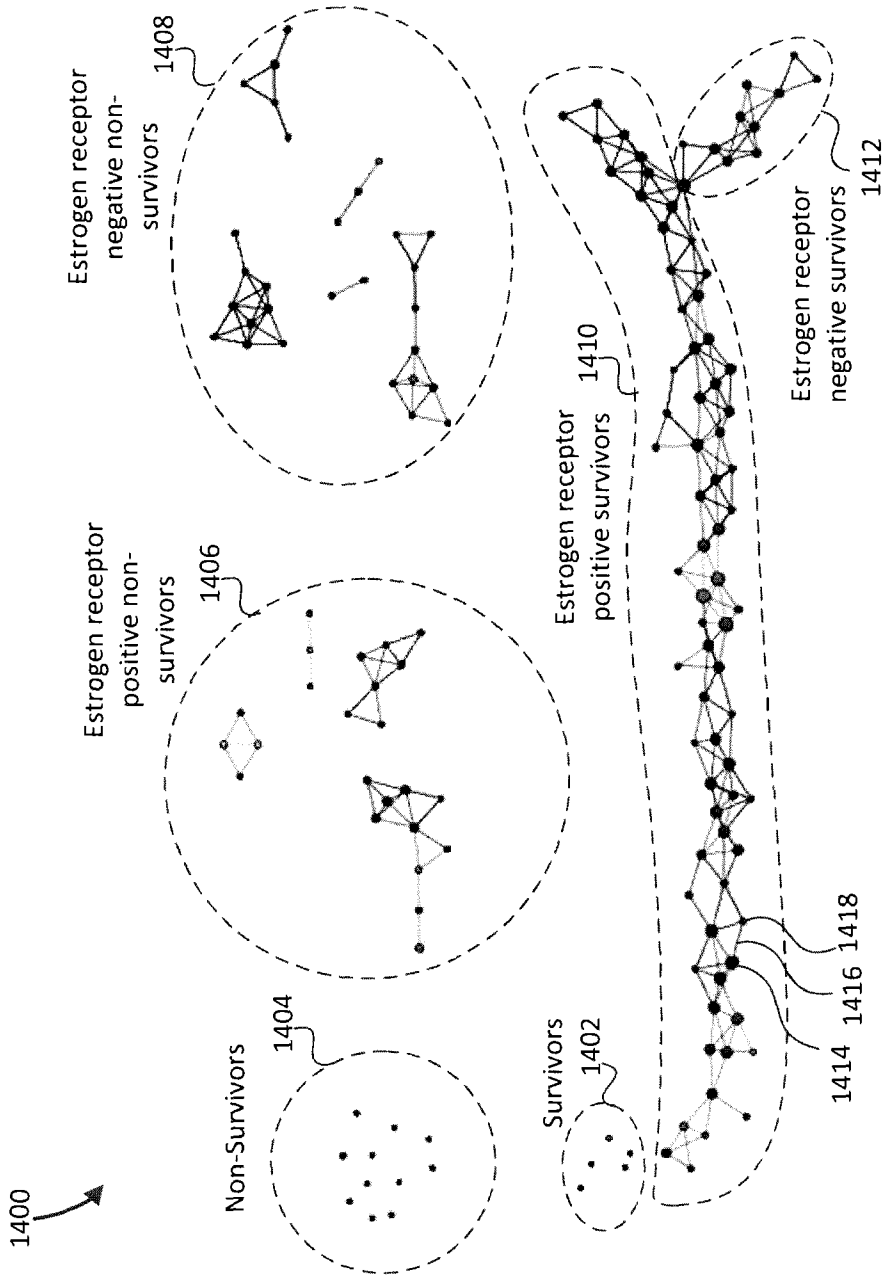

1400
Non-Survivors
1404
Survivors
1402
Estrogen receptor positive non-survivors
1406
Estrogen receptor positive survivors
1410
Estrogen receptor negative non-survivors
1408
Estrogen receptor negative survivors
1412
1414
1416
1418

FIG. 14

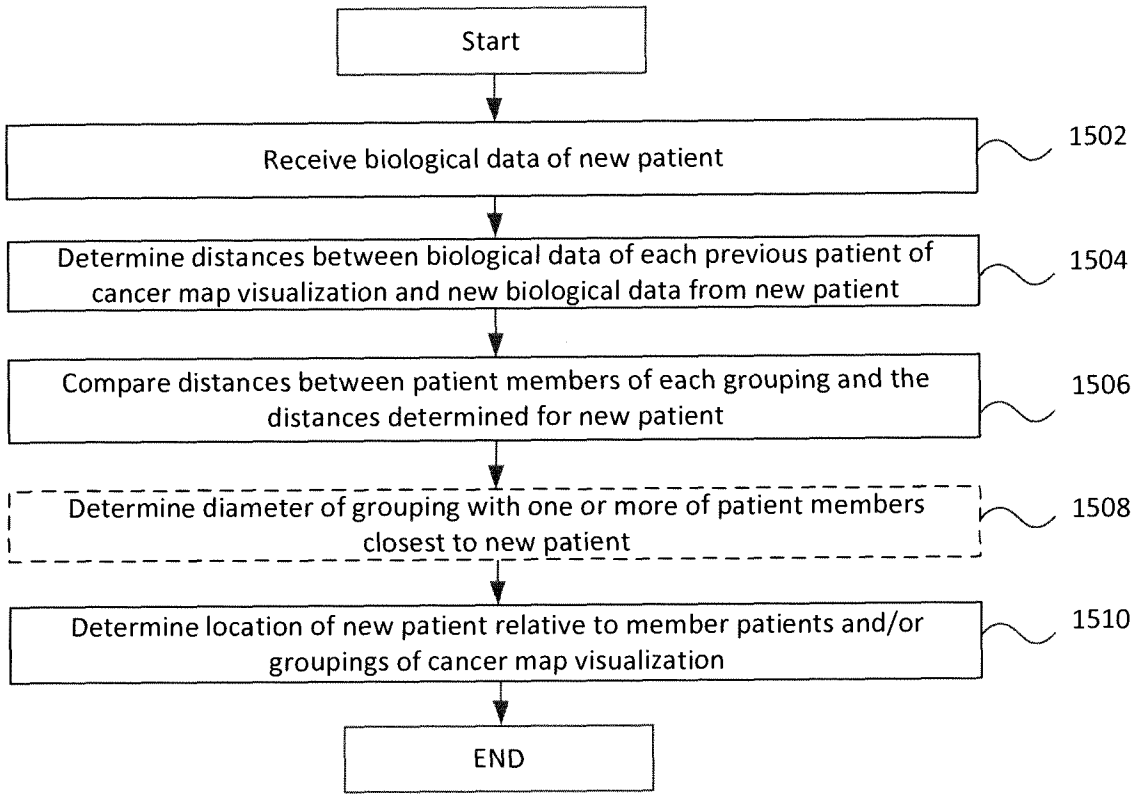

Start
Receive biological data of new patient
1502
Determine distances between biological data of each previous patient of cancer map visualization and new biological data from new patient
1504
Compare distances between patient members of each grouping and the distances determined for new patient
1506
Determine diameter of grouping with one or more of patient members closest to new patient
1508
Determine location of new patient relative to member patients and/or groupings of cancer map visualization
1510
END

FIG. 15

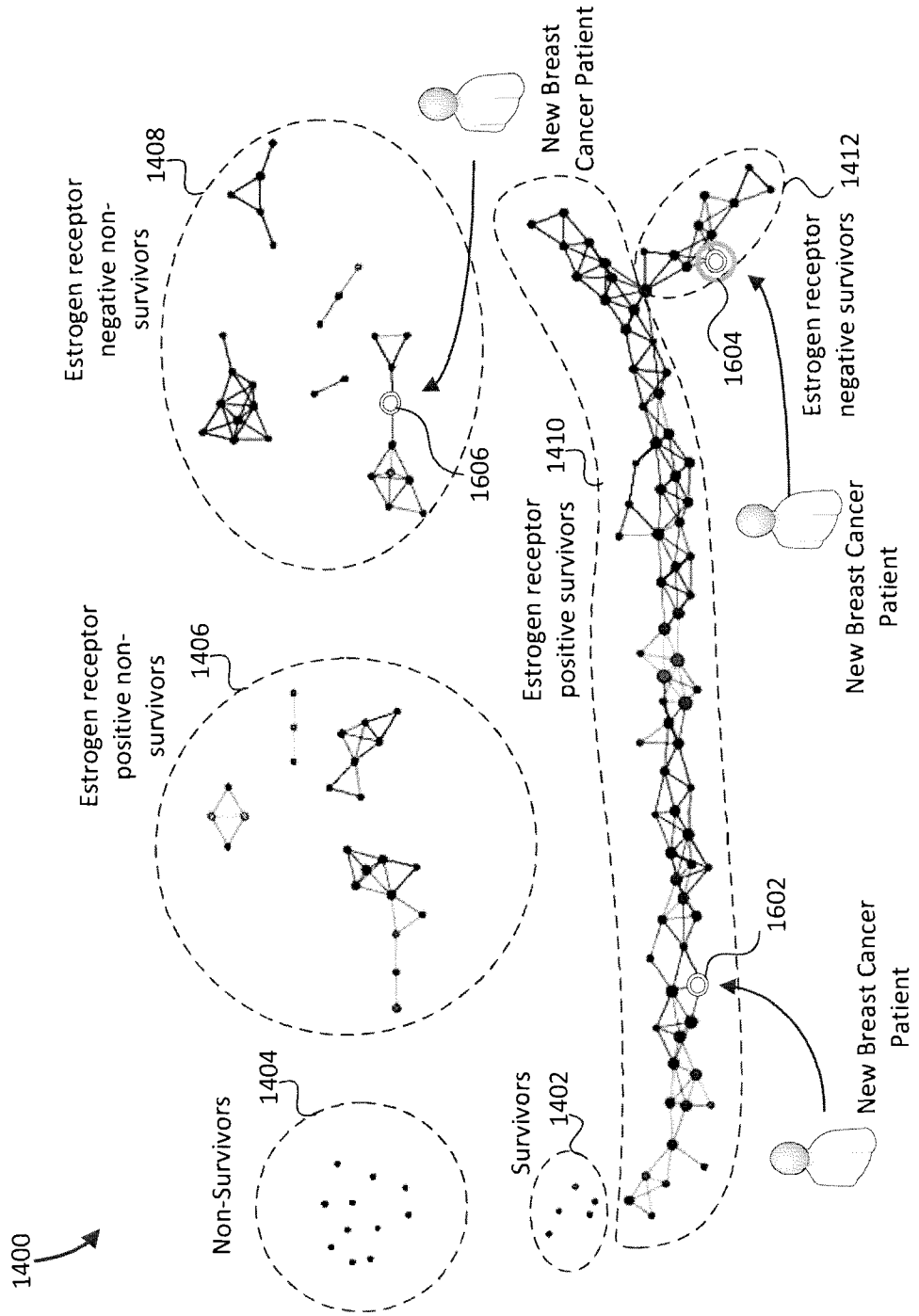

1400
Non-Survivors
1404
Survivors
1402
Estrogen receptor positive non-survivors
1406
Estrogen receptor negative non-survivors
1408
Estrogen receptor positive survivors
1410
Estrogen receptor negative survivors
1412
New Breast Cancer Patient
1602
1604
1606
New Breast Cancer Patient
New Breast Cancer Patient

FIG. 16

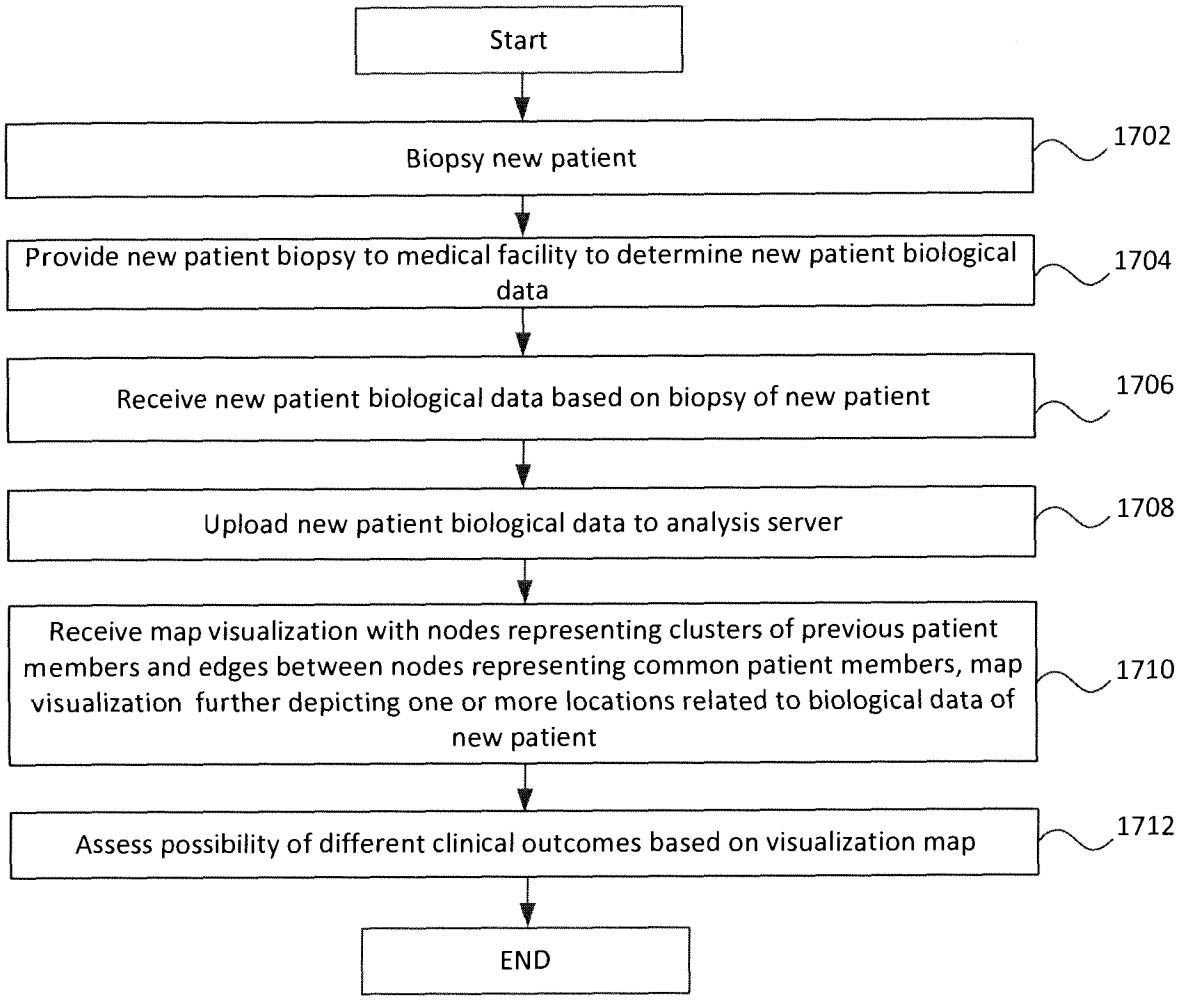

Start
Biopsy new patient
1702
Provide new patient biopsy to medical facility to determine new patient biological data
1704
Receive new patient biological data based on biopsy of new patient
1706
Upload new patient biological data to analysis server
1708
Receive map visualization with nodes representing clusters of previous patient members and edges between nodes representing common patient members, map visualization further depicting one or more locations related to biological data of new patient
1710
Assess possibility of different clinical outcomes based on visualization map
1712
END

FIG. 17

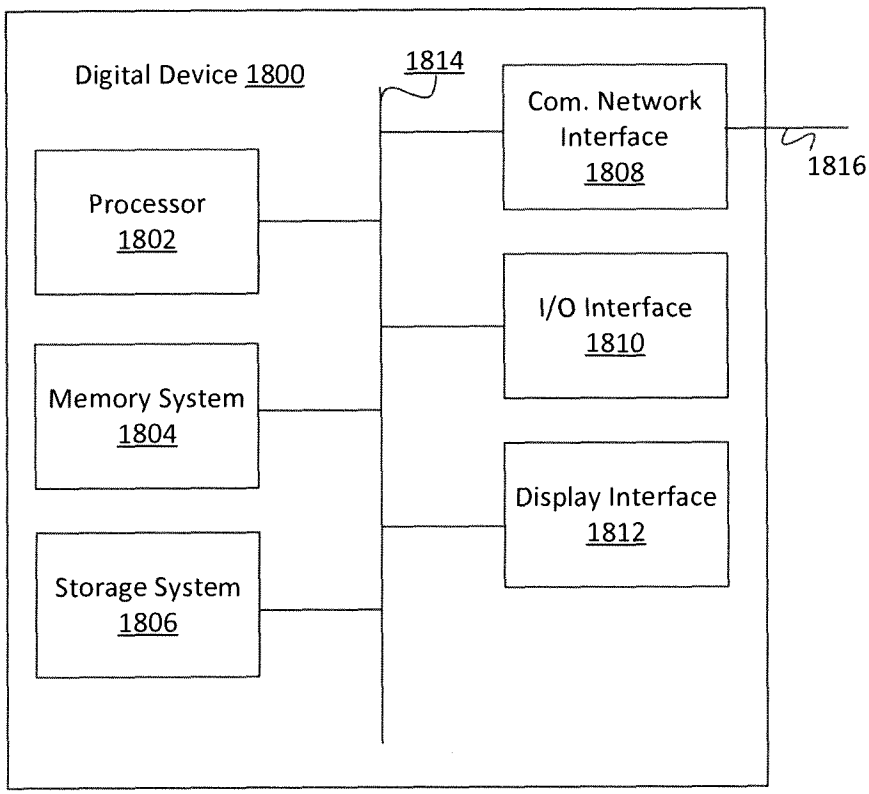

Digital Device 1800
1814
Com. Network Interface 1808
1816
Processor 1802
I/O Interface 1810
Memory System 1804
Display Interface 1812
Storage System 1806

FIG. 18

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 12/59398

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 19/12; G01N 33/48, 33/53 (2012.01)
USPC - 703/11; 702/19; 435/6.1
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8)- G06F 19/12; G01N 33/48, 33/53 (2012.01);
USPC- 703/11; 702/19; 435/6.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC- 702/20; 435/7.1, 436/501, 86, 94. keyword search, terms below.

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWest, PatBase (USPTO, EPO, JPO, WIPO), GoogleScholar (PL, NPL), FreePatentsOnline (USPTO, EPO, JPO, WIPO, NPL); search terms: map, graph, digraph, direct, undirect, node, link, tree, edge, network, associate, classify, category, group, subject, patient, clinic, sample, demographic, gene, express, sequence, density, pearson, engine

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2008/0114564 A1 (IHARA) 15 May 2008 (15.05.2008), Fig. 3; para [0021], [0026], [0082], [0105]-[0110], [0165], [0210]-[0215], [0229]-[0241], [0259]-[0268], [0276], [0290], [0293], [0310], [0313], [0322], [0341]	1-20
Y, P	US 2012/0158633 A1 (EDER) 21 June 2012 (21.06.2012), para [0007]-[0286]	1-20
Y	WO 2011/077353 A2 (OPFER et al.) 30 June 2011 (30.06.2011), pg 2-14	1-20
Y	US 2007/0166707 A1 (SCHADT et al.) 19 July 2007 (19.07.2007), para [0027]-[0986]	1-20
Y	US 2004/0005584 A1 (COHEN et al.) 08 January 2004 (08.01.2004), para [0022]-[0488]	1-20

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
20 November 2012 (20.11.2012)	08 JAN 2013
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)