Visual exploration of large geolocation-rich data sets using formal concept analysis

by

Tiaan du Toit

Thesis presented in fulfilment of the requirements for the degree of

Master of Arts in Socio-Informatics

in the Faculty of Arts and Social Sciences at Stellenbosch University

Supervisor:	Prof. Katarina Britz
Co-supervisor:	Prof. Bernd Fischer

April 2022

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: ....April 2022..............................

Abstract

Visual exploration of large geolocation-rich data sets using formal concept analysis

Thesis: MA Socio-Informatics
April 2022

The rate at which data is being generated is ever-increasing, resulting in an abundance of very large data sets of different structures, all requiring improved methods of data capture, processing and storage. This in turn requires improved representation and data exploration methods. In an age of web applications, one such structure that has gained popularity is semi-structured data. Unlike relational data which can be specifically queried, semi-structured data relies on different methods for data exploration and knowledge discovery. One such method is the visualisation of data beyond a purely granular textual format, in a dynamic and reactive manner. ConceptCloud, a flexible interactive web application for exploring, visualising, and analysing semi-structured data sets, uses a combination of an intuitive tag cloud visualisation with an underlying formal concept lattice to provide a formal structure for navigation through a data set. It is an effective, robust and scalable tool that allows for extension. The underlying formal concept lattice also allows for alternative possibilities of exploring data. This research describes the development and implementation of extensions made to the existing ConceptCloud tool, which are focused on improving the visualisation of, and the knowledge discovery through interaction with a data set, especially data with aspects that could be visualised in a more effective manner. These extensions include a map based viewer for visualising geolocation data, a graph based viewer for visualising the composition of the data as well as a REST API to allow for mobile application development and further unique visualisations. These extensions are demonstrated and evaluated by visualising and exploring two semi-structured data sets in the viticultural domain, namely atmospheric measurements of grape growing regions and wine reviews. It is shown how these extensions aid in and support data exploration and knowledge discovery of these multi- faceted semi-structured data sets. These visualisations can not only be refined and expanded upon to include other types of visualisations, but also improve data mining and knowledge discovery using ConceptCloud. This could result in further research and improvements in similar tools and processes.

Uittreksel

Visual exploration of large geolocation-rich data sets using formal concept analysis

Tesis: MA Socio-Informatics
April 2022

Die pas waarteen data gegenereer word, neem daagliks toe. Dit lei tot 'n oorvloed van baie groot data stelle van verskillende strukture, wat almal beter metodes van data vaslegging, verwerking en berging vereis. Op ‘n soortgelyke wyse, vereis dit op sy beurt weer beter voorstelling en data ondersoek metodes. In 'n era van webtoepassings , is semi-gestruktureerde data een so 'n struktuur wat gewild geword het. Anders as verwante data wat spesifiek ondersoek kan word, maak semi-gestruktureerde data staat op verskillende metodes vir data verkenning en kennis blootlegging. Een so 'n metode is die visualisering van data buite 'n suiwer tekstuele formaat, op 'n dinamiese en reaktiewe wyse. ConceptCloud, 'n buigsame interaktiewe webtoepassing vir die verkenning, visualisering en ontleding van semi-gestruktureerde data stelle, gebruik 'n kombinasie van 'n intuïtiewe merker wolk visualisering met 'n onderliggende formele konsep rooster om 'n formele struktuur vir navigasie deur 'n datastel te verskaf. Dit is 'n effektiewe, robuuste en skaalbare hulpmiddel , wat uitbreiding moontlik maak. Die onderliggende formele konsep rooster maak ook voorsiening vir alternatiewe moontlikhede om data te verken. Hierdie navorsing beskryf die ontwikkeling en implementering van uitbreidings , wat aan die bestaande ConceptCloud instrument gemaak is. Hierdie uitbreidings is gefokus op die verbetering , die visualisering van, en die ontdekking van kennis deur interaksie met 'n data stel, veral data met aspekte wat op 'n meer effektiewe wyse gevisualiseer kan word. Hierdie uitbreidings sluit 'n kaartgebaseerde observasie platform in, vir die visualisering van geoliggingdata , 'n grafiekgebaseerde observasie platform, vir die vertoon van die samestelling van die data , sowel as 'n REST API om voorsiening te maak vir mobiele toepassings ontwikkeling en verdere unieke visualiserings. Hierdie uitbreidings word gedemonstreer en geëvalueer deur twee semi-gestruktureerde data stelle in die wingerdbou domein te visualiseer en te verken, naamlik eerstens , atmosferiese metings van druiweverbouing streke , en tweedens, wynresensies. Dit word getoon hoe hierdie uitbreidings bydra tot , en ondersteuning van dataverkenning en kennis ontdekking van hierdie veelvlakkige semi-gestruktureerde data stelle bied. Hierdie visualiserings kan nie net verfyn en uitgebrei word om ander soorte visualiserings in te sluit nie , maar verbeter ook data ontginning en kennis blootlegging met behulp van ConceptCloud. Dit kan lei tot verdere navorsing en verbeterings in soortgelyke sagteware hulpmiddels en prosesse.

Publication

[1] Tiaan du Toit, Joshua Berndt, Katarina Britz and Bernd Fischer. 2019. ConceptCloud 2.0: Visualisation and Exploration of Geolocation-Rich Semi-Structured Data Sets. *Supplementary Proceedings of ICFCA 2019 Conference and Workshops*, Frankfurt, Germany, June 25-28, (2019). Eds: D. Cristea, F. Le Ber, R. Missaoui, L. Kwuida, B. Sertkaya, Volume 2379, CEUR-WS, http://ceur-ws.org/Vol-2379/

Acknowledgements

Thank you to both my supervisors for all their effort, patience and time spent in providing me with guidance in order to complete this thesis and contribute to the body of knowledge.

The financial assistance of the Centre for Artificial Intelligence Research (CAIR) is also gratefully acknowledged.

Contents

List of Figures

Chapter 1

Introduction

In the current Information Age, data is being generated at an unprecedented rate. In 2013, it was shown that ninety percent of data existing today had been created in the previous two years [18]. The widespread use of location-tracking and GPS-capable devices such as smartphones and tablets has resulted in the generation of vast amounts of geolocation-rich data. This abundance presents a new opportunity for knowledge discovery by exploring the geolocation aspect of data, but it also demands a specific visualisation approach [44]. Consequently, there is an ever-increasing demand to improve the methods for which data is captured, processed, and stored which will ultimately improve data exploration and representation. Furthermore, the increased generation of data has created the need for more efficient digital data storage structures as well as improved and specialised user-centred data visualisations.

1.1 Background and Context

In an age of web applications, one such structure that has gained popularity is semi-structured data. Traditionally, databases based on a relational model of data, were used for storing data. This relational model allows for managing data through a structure and language consistent with first-order logic where all data is represented in terms of tuples, grouped into relations. Most relational databases make use of the Structured Query Language (SQL), where the relational model, now more of an approximation than a pure relational mode, is used to provide a declarative method for specifying data relations and queries through the database management system software [32]. Semi-structured data differs from relational data as it does not have a rigid structure. Instead, it contains some self-describing meta-data in terms of object field titles [22].

Semi-structured data is typically organised within a tree-like structure in order to perform exploration on the data. Exploration (or browsing) allows the user to step through the data without prior knowledge of the structure or schema of the data. An example of this data structure is the JavaScript

Object Notation (JSON) format which is presented in Figure 1.1.

The data field description appears to the left of the colon and the value for that field on the right. JSON objects can be arbitrarily long and the values can be a string in double quotes, a number, a boolean value, another JSON object or an array of any of these data types. This means that JSON objects can be nested and cater for most data storage requirements making JSON a very flexible format. JSON is commonly used in modern web applications, and data sets in this format has grown exponentially in availability and size in recent times.

```
{
   "crimeid":2000,
   "stationid":"157190",
   "stationlocation":"CHUNGWA",
   "crimetype":"All theft not mentioned elsewhere",
   "crimecategory":"Theft",
   "crimelocation":"-32.83898, 27.15612"
},
```

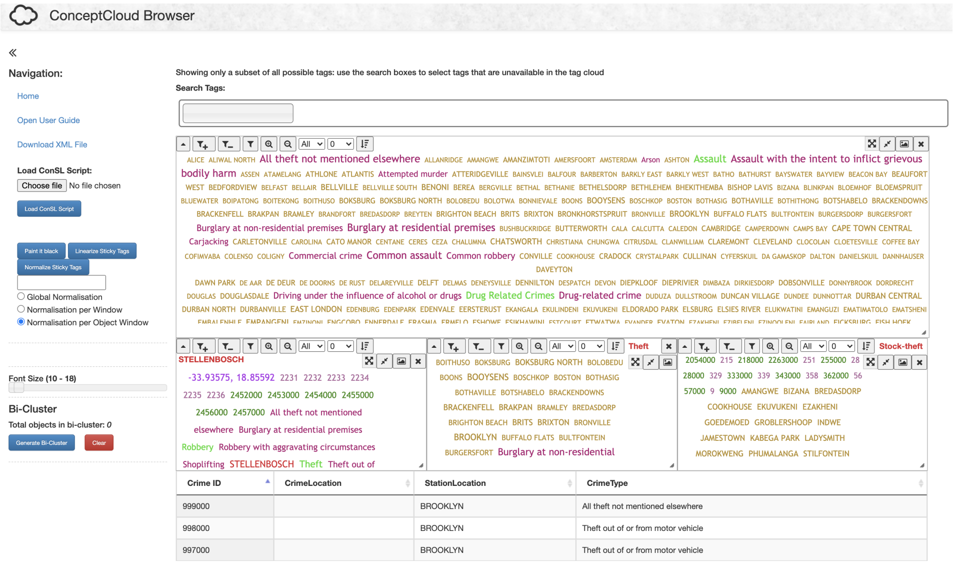
Figure 1.1: a Semi-structured (JSON) object of crime data.

The emergence of large data sets requires new innovative methods of exploring the data due to its increasing size and varying structures. Contrary to relational data, there are cases where unstructured and semi-structured data can not be queried effectively. Furthermore, unstructured and semi-structured data rely on the use of different methods for data exploration with the most effective method being visualisation of data beyond purely textual in table or JSON object format. There are various tools that can be used to visualise semi-structured data, some to a better extent than others. To account for the various challenges mentioned, these tools should be scalable, robust, and expandable to deal with the requirements of very large data sets.

A formal concept lattice, a visual representation of a formal context, is frequently used to explore and visualise semi-structured data. This is further discussed in Chapter 2. Although various tools implement this functionality, the majority of these tools scale poorly due to attempting to visualise the entire concept lattice or context table, which is not feasible for very large data sets. For example, when viewing the full lattice, it becomes difficult to interpret the detail of the lattice and it is overwhelming for the user from a visualisation perspective. Therefore, when dealing with large data sets, it is crucial to implement an alternate scalable visualisation to represent the data so that the data is never displayed in its' entirety. One such tool, ConceptCloud, has shown to possess these attributes and provide a good foundation for improving and extending the functionality of these tools. ConceptCloud is an effective, robust and scaleable tool that allows for extension [9].

1.2 Semi-structured data exploration with ConceptCloud

Formal concept analysis is used within ConceptCloud in order to hierarchically structure semi-structured data in a concept lattice [21]. ConceptCloud uses this concept lattice as an underlying navigation structure, and displays the data to the user in the form of an interactive tag cloud. Figure 1.2 shows ConceptCloud's user interface of this tag cloud. The tag cloud visualisation allows the user to refine or broaden a focus concept by selecting and deselecting tags which correspond to formal objects and attributes in the concept lattice respectively. This is process is called exploratory search, and allows for data exploration and navigation without constricting the user to pre-defined search paths, or requiring prior knowledge of the underlying structure of the data [21].


ConceptCloud Browser
«
Navigation:
Home
Open User Guide
Download XML File
Load ConSL Script:
Choose file
Load ConSL Script
Paint it black
Linearize Sticky Tags
Normalize Sticky Tags
Font Size (10 - 18)
Bi-Cluster
Generate Bi-Cluster
Clear
Search Tags:
All
0
ALICE ALIWAL NORTH All theft not mentioned elsewhere ALLANRIDGE AMANGWE AMANZIMTOTI AMERSFOORT AMSTERDAM Arson ASHTON Assault Assault with the intent to inflict grievous bodily harm ASSEN ATAMELANG ATHLONE ATLANTIS Attempted murder ATTERIDGEVILLE BAINSVLEI BALFOUR BARBERTON BARKLY EAST BARKLY WEST BATHO BATHURST BAYSWATER BAYVIEW BEACON BAY BEAUFORT WEST BEDFORDVIEW BELFAST BELLAIR BELLVILLE BELLVILLE SOUTH BENONI BEREA BERGVILLE BETHAL BETHANIE BETHELSDORP BETHLEHEM BHEKITHEMBA BISHOP LAVIS BIZANA BLINKPAN BLOEMHOF BLOEMSPRUIT BLUEWATER BOIPATONG BOITEKONG BOITHUSO BOKSBURG BOKSBURG NORTH BOLOBEDU BOLOTWA BONNIEVALE BOONS BOOYSENS BOSCHKOP BOSTON BOTHASIG BOTHAVILLE BOTHITHONG BOTSHABELO BRACKENDOWNS BRACKENFELL BRAKPAN BRAMLEY BRANDFORT BREDASDORP BREYTEN BRIGHTON BEACH BRITS BRIXTON BRONKHORSTSPRUIT BRONVILLE BROOKLYN BUFFALO FLATS BULTFONTEIN BURGERSDORP BURGERSFORT Burglary at non-residential premises Burglary at residential premises BUSHBUCKRIDGE BUTTERWORTH CALA CALCUTTA CALEDON CAMBRIDGE CAMPERDOWN CAMPS BAY CAPE TOWN CENTRAL Carjacking CARLETONVILLE CAROLINA CATO MANOR CENTANE CERES CEZA CHALUMNA CHATSWORTH CHRISTIANA CHUNGWA CITRUSDAL CLANWILLIAM CLAREMONT CLEVELAND CLOCOLAN CLOETESVILLE COFFEE BAY COFIMVABA COLENSO COLIGNY Commercial crime Common assault Common robbery CONVILLE COOKHOUSE CRADOCK CRYSTALPARK CULLINAN CYFERSKUIL DA GAMASKOP DALTON DANIELSKUIL DANNHAUSER DAVEYTON DAWN PARK DE AAR DE DEUR DE DOORNS DE RUST DELAREYVILLE DELFT DELMAS DENEYSVILLE DENNILTON DESPATCH DEVON DIEPKLOOF DIEPRIVIER DIMBAZA DIRKIESDORP DOBSONVILLE DONNYBROOK DORDRECHT DOUGLAS DOUGLASDALE Driving under the influence of alcohol or drugs Drug Related Crimes Drug-related crime DUDUZA DULLSTROOM DUNCAN VILLAGE DUNDEE DUNNOTTAR DURBAN CENTRAL DURBAN NORTH DURBANVILLE EAST LONDON EDENBURG EDENPARK EDENVALE EERSTERUST EKANGALA EKULINDENI EKUVUKENI ELDORADO PARK ELSBURG ELSIES RIVER ELUKWATINI EMANGUZI EMATIMATOLO EMATSHENI EMBALENHLE EMPANGENI EMZINONI ENGCOBO ENNERDALE ERASMIA ERMELO ESHOWE ESIKHAWINI ESTCOURT ETWATWA EVANDER EVATON EZAKHENI EZIBELENI EZINQOLENI FAIRLAND FICKSBURG FISH HOEK
STELLENBOSCH
-33.93575, 18.85592 2231 2232 2233 2234 2235 2236 2452000 2453000 2454000 2455000 2456000 2457000 All theft not mentioned elsewhere Burglary at residential premises Robbery Robbery with aggravating circumstances Shoplifting STELLENBOSCH Theft Theft out of
Theft
BOITHUSO BOKSBURG BOKSBURG NORTH BOLOBEDU BOONS BOOYSENS BOSCHKOP BOSTON BOTHASIG BOTHAVILLE BOTSHABELO BRACKENDOWNS BRACKENFELL BRAKPAN BRAMLEY BREDASDORP BRIGHTON BEACH BRITS BRIXTON BRONVILLE BROOKLYN BUFFALO FLATS BULTFONTEIN BURGERSFORT Burglary at non-residential
Stock-theft
2054000 215 218000 2263000 251 255000 28 28000 329 333000 339 343000 358 362000 56 57000 9 9000 AMANGWE BIZANA BREDASDORP COOKHOUSE EKUVUKENI EZAKHENI GOEDEMOED GROBLERSHOOP INDWE JAMESTOWN KABEGA PARK LADYSMITH MOROKWENG PHUMALANGA STILFONTEIN

Load ConSL Script: No file chosen

○ Global Normalisation
○ Normalisation per Window
● Normalisation per Object Window

Total objects in bi-cluster: *0*

Showing only a subset of all possible tags: use the search boxes to select tags that are unavailable in the tag cloud

Crime ID	CrimeLocation	StationLocation	CrimeType
999000		BROOKLYN	All theft not mentioned elsewhere
998000		BROOKLYN	Theft out of or from motor vehicle
997000		BROOKLYN	Theft out of or from motor vehicle

Figure 1.2: ConceptCloud User Interface

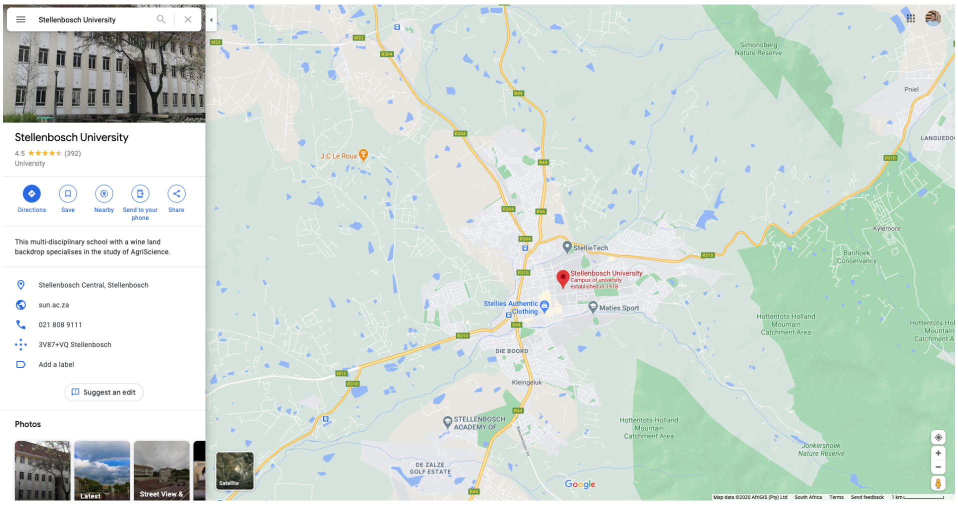
1.2.1 Extensibility

The existing ConceptCloud architecture provides a method of exploring semi-structured data through the use of an auto-generated and object-attribute tag cloud. The data is presented in an interactive *tag cloud*, providing the user with both an intuitive representation of the data set and allows for interactive navigation through the data [9]. Tag clouds are a representation of the number of occurrences and types of the attributes and objects in the data set, wherein the size of each tag displayed is based on the frequency of that tag within the data set and the colour is based on the tag type or category. Navigation is achieved by tag selection and deselection, removing the confines of predefined search paths which allows for an exploratory search.

A dynamic visualisation of the formal concept lattice during navigation also allows for alternative possibilities of exploring data, especially when it comes to specific data types such as geographical coordinates or numeric data. The modern trend of web application popularity places ConceptCloud in a strong position to be expanded upon and it will remain relevant and accessible for the foreseeable future. ConceptCloud's browser interface makes use of HTML, CSS and Javascript in the JAVA based Play! framework, which are some of the most widely used and popular programming languages among developers [15]. A visualisation of the formal concept lattice also allows for alternative possibilities of exploring data, especially when it comes to specific data types such as geographical coordinates or numeric data.

1.2.2 Map Visualisation

For data sets with geolocation aspects, one such possibility is a geographical centred exploration approach. Google Maps is a popular and widely used tool for geolocation data exploration and the user interface is easy to interpret. Since Google Maps is a well established application containing all required functionality and geolocation data needed, leveraging the data and integrating its functionality into ConceptCloud would provide a powerful method of exploring semi structured data with geolocation aspects. Figure 1.3 shows the Google Maps user interface. Google Maps is one of the most popular and widely used geographical data exploration tools on the web. Most users are familiar with the functionality the tool provides and how it operates. Beyond this, the company has allowed developers to implement and fully integrate the tool's functionality and visualisation capabilities into their own applications, making it a prime candidate when working with any geographical data or data with a geolocation aspect.



Stellenbosch University

Stellenbosch University

4.5 ★★★★★ (392)

University

This multi-disciplinary school with a wine land backdrop specialises in the study of AgriScience.

Stellenbosch Central, Stellenbosch

sun.ac.za

021 808 9111

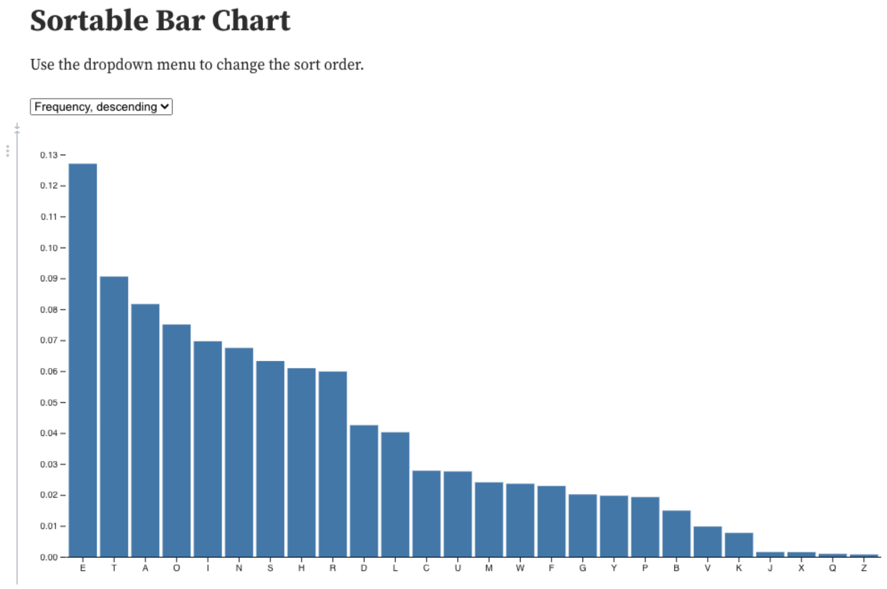
3V87+VQ Stellenbosch

Add a label

Figure 1.3: Google Maps User Interface

1.2.3 Graphs

As for numeric data, graphs and charts are the most effective method of visualising data sets with numeric aspects so that trends and patterns can be identified. ConceptCloud being a web application, allows for integration of advanced graph drawing libraries that can fully utilise the HTML canvas to display advanced and fully interactive graphs and charts. Figure 1.4 shows an example of a bar graph rendered by D3, a graph drawing library, which is widely used in modern web applications to render various types of graphs and charts. D3 is one of many libraries that offers this capability. D3 also allows for low level manipulation and customisation of graph objects, allowing for graphs to be drawn inside other graphs, or rendering three dimensional graphs which can be fully rotated and explored in the browser.



Sortable Bar Chart

Use the dropdown menu to change the sort order.

Frequency, descending

Figure 1.4: D3 Example Graphs

1.2.4 REST API and Mobile Applications

Another substantial benefit of modern web applications, is their ability to allow for the addition of an externally accessible, Representational State Transfer (REST) Application Programming Interface (API) to enable Cross Origin Resource Sharing (CORS). An externally accessible API allows developers to build independent applications that can query and present data from the server without having to do any resource intensive calculations or data processing itself. This not only makes it possible for web applications to make use of the functionality and capabilities of one another, but also allows mobile applications to leverage the capabilities of web applications without the mobile device needing to do any of the processing. ConceptCloud's existing server architecture allows for the development of a REST API making it possible to share its powerful lattice generation and processing capabilities with other applications. Figure 1.5 shows a high level architecture diagram of the web application server and API layer, and how it communicates with mobile devices.

The technology behind mobile devices is always advancing, more specifically smartphones and tablets and their touch screen user interfaces. This allows for new and unique visualisation opportunities as well as methods of interacting with the data. By implementing an externally accessible API, it becomes possible to develop a mobile application with much of the same functionality as ConceptCloud.

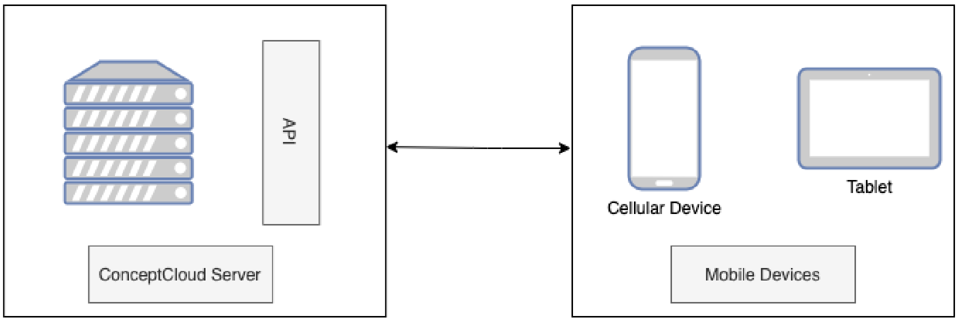


Figure 1.5: API and Mobile Device Architecture

1.3 Aims and Objectives

This thesis aims to improve ConceptCloud by expanding on the visualisation and exploration functions already present. This will be achieved by providing the user with additional data visualisations and exploration methods and by extending the tool's functionality to other applications and devices. These extensions will be implemented through a design science research methodology following a user-centric approach that maximises the benefits of data analysis and knowledge discovery experienced by the user making use of the tool. This involves considering which visualisations best suite certain types of data as well as how the user will interact with the visualisations in manner that improves analysis and sensemaking of the data while reducing or limiting cognitive load experienced by the user. Ultimately, these extensions will be evaluated by applying the tool to suitable data sets and conducting a case study in order to determine their effectiveness in analysing data and discovering knowledge.

Specifically, this thesis aims to design, develop and evaluate three key extensions to ConceptCloud, namely a map viewer, graph viewer and mobile interface. These extensions must adhere to certain requirements which can be stated below as:

- The extension should be clear and readable for a user with moderate levels of data literacy.
- The visualisations must remain an accurate representation of the data where opportunity for misinterpretation of the data by the user is minimised.
- The new extension must aid in data analysis by providing cognitive support and reducing the cognitive load experienced by the user as far as possible.
- The extension must aid in data exploration and knowledge discovery by providing different visualisations of the same data.

- The extension must allow the new visualisation as well as the *Tag Cloud* to drive exploration. Thus a pivotal function of this tool is *to ensure the data exploration works in a bidirectional manner*. In other words, to update the new visualisation when changes are made in one of the visualisations and vice versa.

- The extension must ensure that *the visualisation and exploration tool be highly scalable to process large data sets*, since the rate at which such data is being generated is ever increasing.

1.4 Methodology

This thesis follows a design science research methodology, which is used when a research problem, usually a business problem, requires the creation of an artefact to be solved. A 2007 paper by information systems researcher Hevner [34], discusses, implements and applies DSRM to research done in the information systems field.

Hevner [34] provides a good foundation for thinking about DSRM and states that there are three design science research cycles. These cycles are based on the twelve theses stated by author Juhani Iivari [39], and includes the relevance cycle, the rigor cycle and the design cycle [34]. As can be seen in Figure 1.6, these cycles either occur in or connect three different environments, which include the knowledge base, the design science research environment and the application domain

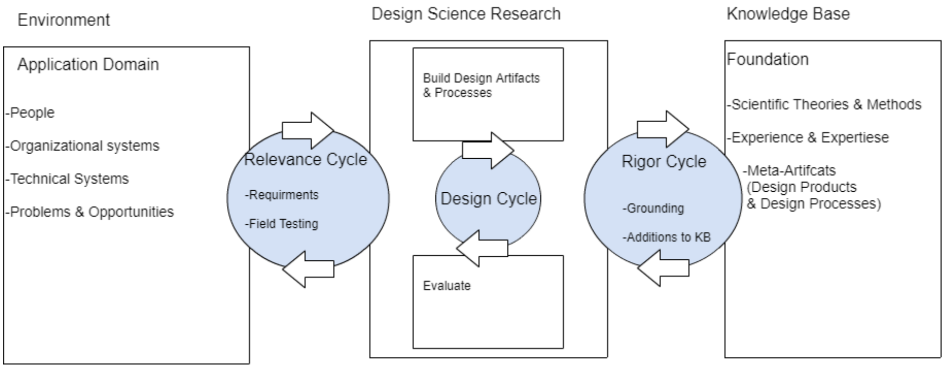


Figure 1.6: Design Science [34]

The relevance cycle ensures that the research is being done with an application context, solving real-world practical problems that provides both the requirements for the research as well as defines the evaluation criteria for results [34]. The rigor cycle is located in the knowledge base environment which serves as a source of past knowledge to the researcher to ensure innovation.

It is the researcher's responsibility to thoroughly evaluate the knowledge base and ensure that their research contributes to it, and that similar research has not been done before. The final cycle, the internal design cycle, involves the iteration of design, evaluation and feedback of the artefact, until a satisfactory design is achieved [34]. This cycle draws from both the relevance cycle and the rigor cycle.

This model provides a unique macro perspective on conducting this type of research and shows how the practical environment defined both the requirements and evaluation criteria for the required software artefact. It also emphasises the role of iteration, especially the internal design cycle, until the software artefact is fine tuned.

The methodology used in this thesis loosely follows the three cycle methodology recommended by Hevner [34]. Firstly, the relevance cycle was iterated through during the conceptualisation of the software extensions which solves the real-world practical problem of effective data visualisations. Secondly, a literature review of the knowledge base was conducted, thus iterating through the rigor cycle, before development started. During the development and evaluation of the extensions during the case study, the internal design cycle was iterated through.

When considering the relevance cycle and the requirements and evaluation criteria imposed by the practical environment when it comes to data visualisation, it is important to identify current trends and approaches in the field.

Modern web applications are becoming adept at processing, presenting and storing very large data sets via efficient digital data storage structures as well as improved user-centred data visualisations. These applications typically preprocesses and transforms the data before showing it to the user. They also makes use of third party APIs or libraries to provide the user with specialised user interfaces which makes visualising and navigating through the data as simple and intuitive as possible. This is important since not all users have the same technological skills or data literacy. This becomes a necessity as the amount of the data shown to the user increases.

The Cognitive Psychology component of Human-Computer Interaction is focused on how humans use software applications. Furthermore, it investigates the importance of creating impactful data visualisations to facilitate sensemaking and reduce the cognitive load when working with complex data. As the analysis of visualised data requires an individual to understand, analyse and learn, it is unsurprising that human cognition plays a vital role in visualising data [31]. Therefore, it is important to acknowledge the role of cognition within visualisation and when designing user interfaces and data visualisations, it becomes crucial to follow a user-centric approach. This approach describes principles in line with visual perception and cognitive processing which includes not only considering the end-user of the visualisation but also providing the user all of the information required to interpret the visualisations [33].

The guidelines in this approach might seem obvious but modern software applications have progressed far beyond static graphs and charts and now allow for complex visualisations for user-driven refinement [33]. Thus, developers of such applications must either decide to visualise as much data as possible or simplify the data visualisation to ensure they remain effective. Balancing the need for accuracy with the risk of confusing or overwhelming the user becomes a difficult task.

1.5 Structure of Thesis

Chapter 2 provides the reader with background theory on the field of formal concept analysis involving formal contexts as well as how concept lattices are generated. A brief background of data visualisation will also be given before exploring some formal concept analysis powered data visualisation and exploration tools. The initial version of ConceptCloud will then be formally introduced and described.

Chapter 3 of this thesis describes a map extension made to ConceptCloud which allows not only for more effective visualisation of geolocation data, but also improving the exploration of the data by allowing bidirectional navigation between the map and tag cloud, i.e., update the map when changes are made in the tag cloud and vice versa.

This chapter also involves the further extension to ConceptCloud's visualisation arsenal, an integrated graph viewer that allows the user to generate either a pie chart of bar graph of the data in order to get a quantitative view of tags in the data set.

Lastly, this chapter describes how the initial tool was extended to allow other developers to access the lattice navigating functionality of ConceptCloud in their own applications. To showcase this, a mobile application was built that re-creates the functionality of the initial tool with an updated touchscreen interface.

The fourth chapter of this thesis includes an evaluation of the extensions made to ConceptCloud through two case studies conducted on very large semi-structured data sets. The first is a data set of atmospheric indicators and viticulture indexes of weather conditions measured at various weather stations in the state of California in the United States of America, over a period of three years ranging from 2014 to 2016. The second is a wine review data sets containing textual descriptions of wine, the varietal and review scores of the wines among other aspects.

The final chapter summarises the changes made to ConceptCloud and their benefits. Thereafter, it discusses the findings from the evaluation. Finally, the thesis will conclude by proposing future work.

Chapter 2

Background

2.1 Overview

Chapter 2 covers the theoretical background of Formal Concept Analysis (FCA). A brief overview of FCA fundamentals as well as examples of the formal context and concept lattice is provided. Subsequently, the importance of using data visualisation is investigated by discussing their drawbacks and benefits. In addition, the various methods of data visualisation are investigated with a focus placed on a user-centric approach. Furthermore, data visualisation tools that use FCA to power data visualisations and exploration are introduced. A detailed discussion on ConceptCloud is provided as this data visualisation tool will be used throughout the thesis. To conclude, a discussion on ways to improve tag cloud sizing is provided.

2.2 Formal Concept Analysis

Traditional data structures are unable to make use of powerful data visualisation and exploration tools [26]. Therefore, the research conducted in this thesis makes use of the theory behind Formal Concept Analysis as a foundation on which to build these powerful data visualisation and exploration tools.

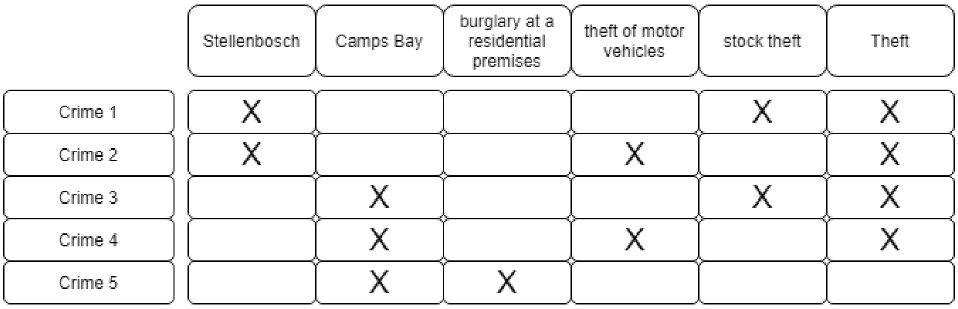
Formal concept analysis is a field of applied mathematics with specific focus placed on applied lattice theory. It is a method that is used to group, represent and analyse data sets. This is achieved by grouping data objects hierarchically with their common attributes. For example, when analysing a crime, the data object refers to the crime incidence and the details surrounding the crime incidence are its attributes.

Formal Concept Analysis was introduced by Rudolf Wille, a German researcher, in 1982 [52]. It has since been refined and is widely used in various fields, such as software development for data analysis, knowledge discovery, and information retrieval [51].

Mathematical set theory and order theory are both used to formulate the theory of formal concept analysis [26]. At its core, formal concept analysis consists of binary relations (partial order relations specifically) between sets of objects and attributes. There is a natural partial order relation of subset inclusion on a data set, in certain pairs of subsets of elements, neither set is included in the other. This is in contrast to a total orders relation, where all elements are ordered.

The data set is partitioned into two types of elements, namely formal objects and formal attributes. The adjective "formal" distinguishes the terms from their more classical definitions [52] and emphasises their connection to mathematical concepts.

Data representation as a formal context allows for a great deal of manipulation and practical use (such as exploratory searching through unknown data sets). An ideal way of representing a formal context is by making use of a cross table or context table. An example of such a cross table (or context table) is presented in Figure 2.1. The table summarises a number of crimes and their respective details. The formal objects are the crime incidences and the formal attributes are crime descriptors (location or type). The relationship between them is represented by a cross [52]. For example, *Crime 5* (the object) occurred in *Camps Bay* (attribute 1) and the crime committed was *burglary of a residential premises* (attribute 2).



	Stellenbosch	Camps Bay	burglary at a residential premises	theft of motor vehicles	stock theft	Theft
Crime 1	X				X	X
Crime 2	X			X		X
Crime 3		X			X	X
Crime 4		X		X		X
Crime 5		X	X			

Figure 2.1: a formal context of crimes committed in South Africa

In FCA, the binary relation on sets of objects and attributes is a partial order relation which has certain properties. This will be made more precise in Definition 4 below. Formally, a binary relation R on a set A is called a partial order relation (a partial order), if it satisfies the following conditions for all elements $a, b, c \in A$:

1. aRa (Reflexivity: every element is related to itself)

2. aRb and $a \neq b \rightarrow$ not bRa (Anti-symmetry: two distinct elements cannot be related in both directions)

3. aRb and $bRc \to aRc$ (Transitivity: if a first element is related to a second element, and, in turn, that element is related to a third element, then the first element is related to the third element)

Formally:

Definition 1. *A formal context is a triple* $(\mathcal{O}, \mathcal{A}, \mathcal{I})$ *where* $\mathcal{O}$ *and* $\mathcal{A}$ *are sets of objects and attributes, respectively, and* $\mathcal{I} \subseteq \mathcal{O} \times \mathcal{A}$ *is an arbitrary incidence relation.*

If we consider Figure 2.1, with sets:

$$\begin{aligned} \mathcal{O} :=& \{\textit{Crime 1, Crime 2, Crime 3, Crime 4, Crime 5}\ \} \\ \mathcal{A} :=& \{\textit{Stellenbosch, Camps Bay, burglary at a residential premises,} \\ & \textit{theft of a motor vehicle, stock theft, Theft}\}, \end{aligned} \tag{2.1}$$

the incidence relation, for *Crime 1* is defined by:

$$\{\textit{Crime 1}\} \times \{\textit{Stellenbosch, stock theft, Theft}\} \subseteq \mathcal{I} \tag{2.2}$$

Formal concept are certain subsets of the incidence relation $\mathcal{I}$. They can be viewed as maximal rectangles in the context table a term used in computational geometry involving finding a rectangle of maximal size to be placed among obstacles in a plane within the context table. A formal concept is therefore a maximal set of object-attribute pairs, having the property that adding another object to the formal concept will result in a loss of common attributes and adding an additional attribute will result in a loss of common objects [28].

The intent of a formal object is the set of formal attributes possessed by this object, and the extent of a formal attribute is the set of all formal objects that share this attribute. By considering the formal context table shown in Figure 2.1, the extent of the formal attribute *Theft* is the set T := {*Crime 1, Crime 2, Crime 3, Crime 4*} and the intent for the formal object *Crime 3* is the set G := {*Camps Bay, stock theft, Theft*}. This is formally generalised to sets of objects and attributes as follows:

Definition 2. *Let* $(\mathcal{O}, \mathcal{A}, \mathcal{I})$ *be a context,* $O \subseteq \mathcal{O}$*, and* $A \subseteq \mathcal{A}$*. The common attributes of* O *are defined by* $\alpha(O) = \{a \in \mathcal{A} | \forall o \in O : (o, a) \in \mathcal{I}\}$*, the common objects of* A *are denoted by* $\omega(A) = \{o \in \mathcal{O} | \forall a \in A : (o, a) \in \mathcal{I}\}$*.*

Applying this definition to the formal context in Figure 2.1 and sets $\mathcal{O}$ and $\mathcal{A}$, ({*stock theft, Theft*}, {*Crime 1, Crime 3*}) represents a formal concept, and adding an additional attribute, say *Stellenbosch* to the concept will result in a loss of the *Crime 3* object since it does not share this attribute.

This yields a new formal concept, as can be seen from the above example. It is possible to add or remove objects or attributes from the formal concept in order to make it more or less general, we can continue doing so until it results in either a universal or empty set of objects and attributes respectively. By removing formal attributes from the formal concept, the number of formal objects will increase resulting in a more general formal concept. Conversely, adding formal attributes will result in fewer and more specific or granular formal objects in the formal concept. The same holds true for adding or removing formal objects to a formal concept, thus formal objects and attributes can be used as a filter in order to make the formal concept more or less specific. This relation which exists between the attributes and objects of a formal concept, generates a partial order relation on the formal concepts, which is referred to as a subconcept, superconcept relation [52]. This is defined formally in Definition 3.

Definition 3. *Let $\mathcal{C}$ be a formal context. Then $c = (O, A)$ is called a concept of $\mathcal{C}$ if $\alpha(O) = A$ and $\omega(A) = O$.* $\pi_O(c) := O$ *and* $\pi_A(c) := A$ *are the called the extent and intent of c, respectively. The set of all concepts of $\mathcal{C}$ is denoted by $B(\mathcal{C})$.*

Definition 3 allows formal concepts to possess the property where a concept's extent includes the extent of all of its subconcepts, and the intent of a concept includes the intent of all of its super concepts. Returning to our example context, the formal context of crimes, this means that the formal attribute *Theft*, has a subconcept *Stellenbosch Theft* which has an extent of set Q := {*Crime 1, Crime 2*} which is contained in the extent *Theft* which is the Y:= {*Crime 1, Crime 2, Crime 3, Crime 4*}. The subconcept relation induces a partial order on the set of all formal concepts, and is formally defined as follows:

Definition 4. *Let $\mathcal{C}$ be a formal context, $c_1 = (O_1, A_1)$, $c_2 = (O_2, A_2)$ for $c_1, c_2 \in B(C)$. Then c_1 and c_2 are ordered by the subconcept relation, written $c_1 \leq c_2$ if $O_1 \subseteq O_2$ or equivalently, $A_2 \subseteq A_1$. The structure of $B(C)$ and $\leq$ is denoted by $\mathcal{B}(\mathcal{C})$.*

2.2.1 The Formal Concept Lattice

To represent a formal context, a concept lattice can be used instead of a context table to show the relationship between the concepts. This differs from a context table because the lattice shows the relation between the subconcepts and super-concepts visually [52]. An example of a formal concept lattice is provided, most specific to the concept shown, in Figure 2.2. The nodes of the lattice represent a formal concept, with formal objects below the node and formal attributes shown above. The extension (formal objects) of a formal concept is determined by tracing all paths that lead down from the node whereas the intention (formal attributes) of a formal concept is determined by tracing all paths that lead up from the node. Consequently, the top node of a lattice has all formal objects in its extension and the bottom node has all of the formal attributes in its intention [52].

Many algorithms exist that are used to compute formal concept lattices, with some being more efficient than others. An example of an algorithm that is used to draw a concept lattice is discussed below [38]:

Every formal concept of a context (G, M, I) has the form $(\omega(\alpha(X)), \alpha(X))$ for some subset X ⊆ G and the form $(\omega(Y), \alpha(\omega(Y))$ for some subset Y ⊆ M. Vice versa all such pairs of sets are formal concepts. Then, the proposed algorithm states:

1. β(G;M; I) := ∅
2. For every subset X of G, add $((\omega(\alpha(X)), \alpha(X)))$ to β (G,M, I)

Since FCA theory states that the structure induced by the partially-ordered formal concepts of a formal context is always a complete lattice, it can be automatically computed for any incidence relation between objects and attributes [38]. The greatest lower bound of a lattice is the lowest point in the lattice that still maintains a common concept. The meet and join can always be computed based on any position in the lattice. The formal concept lattice as well as the meet and join operation is described in the theorem below:

Theorem 1. *Let $\mathcal{C}$ be a context, then $\mathcal{B}(\mathcal{C})$ is a complete lattice, called the concept lattice of $\mathcal{C}$. Its meet and join operation for any set $\{\langle A_i, B_i\rangle | i \in I\} \subseteq B(\mathcal{C})$ of concepts are given by:*

$$\bigwedge_{i\in I}(O_i, A_i) = (\bigcap_{i\in I} O_i, \alpha\ (\omega(\bigcup_{i\in I} A_i)))$$

$$\bigvee_{i\in I}(O_i, A_i) = (\omega(\alpha(\bigcup_{i\in I} O_i)), \bigcap_{i\in I} A_i)$$

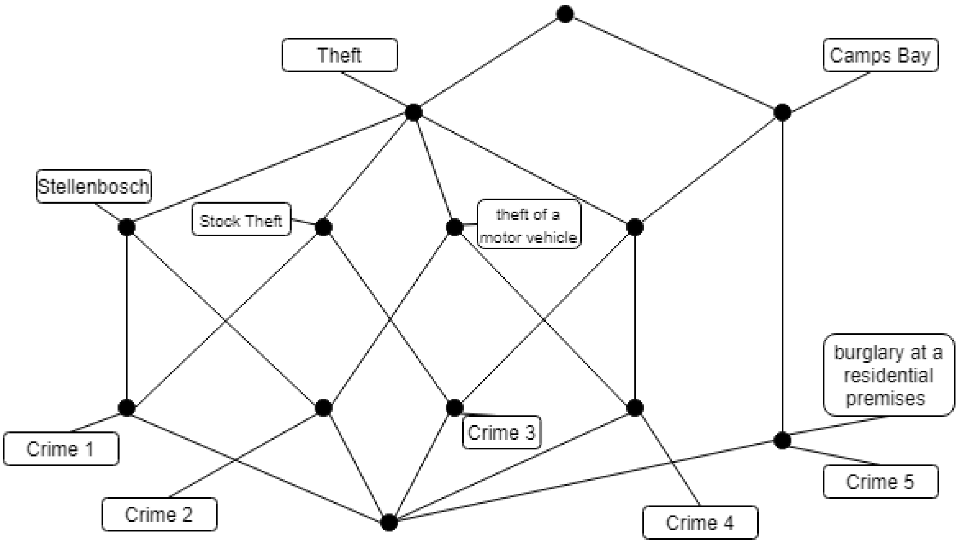


Figure 2.2: An example concept lattice of the Formal Context given in Figure 2.1

Since each formal object and formal attribute has a uniquely determined defining formal concept in the lattice, it is never necessary to draw or generate the entire formal concept lattice since these defining concepts can be calculated from the objects or attributes, instead of being searched for in the lattice [28]. This is formally defined as follows:

Definition 5. *Let $\mathcal{B}(\mathcal{O}, \mathcal{A}, \mathcal{I})$ be a concept lattice. The defining concept of an attribute $a \in \mathcal{A}$ is the greatest concept c such that $a \in \pi_A(c)$ holds, and is denoted by $\mu(a)$. The defining concept of an object $o \in O$ is the smallest concept c such that $o \in \pi_O(c)$ holds, and is denoted by $\sigma(o)$.*

Referring to Figure 2.2, the defining concept of the formal object is $\langle\{$*Crime 1*$\}, \{$*Stellenbosch, stock theft*$\}\rangle$ and the defining concept of the attribute *stock theft* is $\langle\{$*Crime 1, Crime 3*$\}, \{$*stock theft, Theft*$\}\rangle$. The ability of formal concept lattices to arrange formal concepts in a complex hierarchy has many practical uses and has become a popular method of facilitating knowledge discovery and ontology engineering in various fields. According to a survey published by Poelmans et al., multiple applications have been implemented which make use of FCA to facilitate data analysis, knowledge discovery and information retrieval. From the study, it was found that the most popular application field at the time was text mining.

2.3 Data Visualisation

In the age of ubiquitous technology, data is being generated at an unprecedented rate. In many cases, software applications generate this data from user input and allow users to explore it. These applications typically pre-process data and make use of specialised user interfaces so that users can more easily interpret data, since not all users have the same technological skills or data literacy. This becomes a necessity as the amount of the data shown to the user increases. It should also be noted that the visualisations being explored are dynamic and reactive to user input. These differ from post-processed still images and graphs visualisations. The visualisations change as the user interacts with the data, increasing the risk of confusing the user or generating. visualisation that is not human-readable.

2.3.1 Cognitive View of Data Analysis

When data is shown to a user, the data needs to be interpreted by the user through a process called data analysis. Data analysis is a process in which the data is first summarised and follows a methodology in which discrepancies in the data are exposed and analysed [31]. Furthermore, Huber (2012) breaks down data analysis into several different activities that one is required to fulfil [36]. These activities include: the collection and inspection of data, the process of error checking, modifying and transforming the data, identifying discrepancies by comparing data, modelling and simulation, interpretation of the results, and lastly, the presentation of the findings [36]. In a similar manner, sensemaking is a process of searching for a representation and encoding data in which it is used to answer a particular question [31].

It is stated that an individual can only store between two to six pieces of information at a single moment in time. Consequently, sensemaking relies heavily on the use of schemas [31]. A schema is defined as a mental model which stores a piece of information regarding a particular type of object or concept (different to a formal concept as discussed above) [31]. Furthermore, schemas exist with two forms of data analysis, namely confirmatory analysis and exploratory analysis [31].

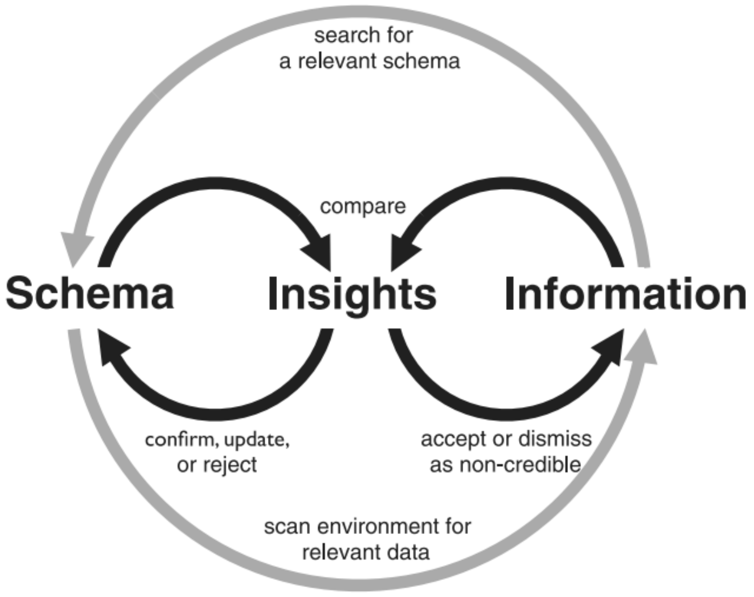


Figure 2.3: A simplified summary of the sensemaking process.

A sensemaking process that starts with schemas which lead to the collection of data is common in confirmatory analysis. In contrast, when sensemaking starts with data and proceeds to search for possible schema it is known as exploratory analysis [31].

From a cognitive perspective, data analysis is similar to sensemaking. This is reinforced by Grolemund and Wickham (2014) who state that tasks within data analysis can be compared to sensemaking, as they both use quantitative data to analyse information. Furthermore, both data analysis and sensemaking have shared goals which aim to produce reliable ideas of reality extracted from observed data [31]. To elaborate, they compare fact with theory, search for discrepancies between the two and lastly, modify the theory. Furthermore, they both stress the importance of "comparison, discrepancy and iteration" of new information and data [31]. In addition to shared goals, they also have shared methods as they both require individuals equipped with cognitive strategies which facilitate the working memory to interpret the data. The role of cognition is a pivotal component of data analysis, as assigning meaning to information is not a computational or statistical but rather a cognitive process [31]. To elaborate, Grolemund and Wickham (2014) go as far to state that the success of data analysis is dependent on the interaction of the above mentioned cognitive processes [31]. Figure 2.3 showcases the flow of the confirmatory and exploratory sensemaking processes.

2.3.2 Cognitive Load

The above theories of data analysis and sensemaking fall under the field of cognitive psychology which is focused on understanding mental processes in humans. In this field it is accepted that sensemaking requires cognitive effort, also referred to as cognitive load, and occurs in what is referred to as "working memory".

Human memory can be compared to an information processing system and consists of four core components, namely: short-term memory, working memory, sensory memory and long-term memory [35]. A proposed model of how data is analysed and understood, states that there are four major steps within the process. The first step includes external information being processed via the sensory memory. Once the information has been processed, it is transferred to short-term memory where it is processed and placed within working memory [35]. Figure 2.4 indicates this flow of incoming information into the various types of memory.

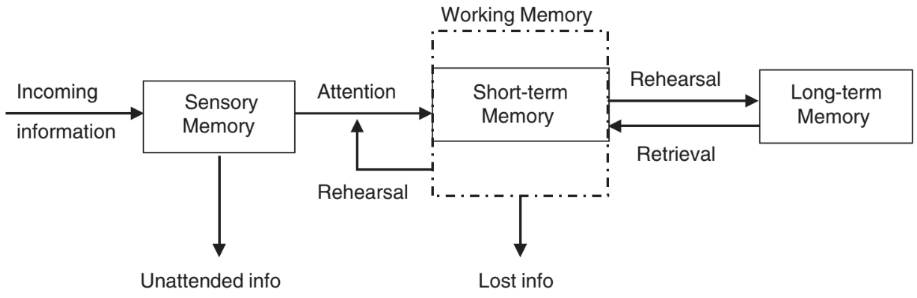


Figure 2.4: Information processing model of memory.

The theory of cognitive load places a great deal of emphasis on working memory [35]. The working memory is used when performing cognitive tasks, however, it is limited in both duration and capacity. The quantity of cognitive resources required to successfully perform a task is referred to as the cognitive load [35]. Cognitive load is associated with the learning of complex cognitive tasks [35]. Mental effort is suggested to mirror the cognitive load, as mental effort is the amount of cognitive capacity allocated to facilitate the cognitive demand initiated by a task.

However, there are more complexities to the process, in addition to several limitations regarding the working memory. To elaborate, working memory is limited in both duration and capacity [35]. The information that exists within the working memory is believed to be forgotten over time as it is most commonly used when performing tasks. These limitations are problematic for cognitive activity as often tasks require many elements to be held in working memory and require the processing of many elements simultaneously which causes memory overload and information loss [35].

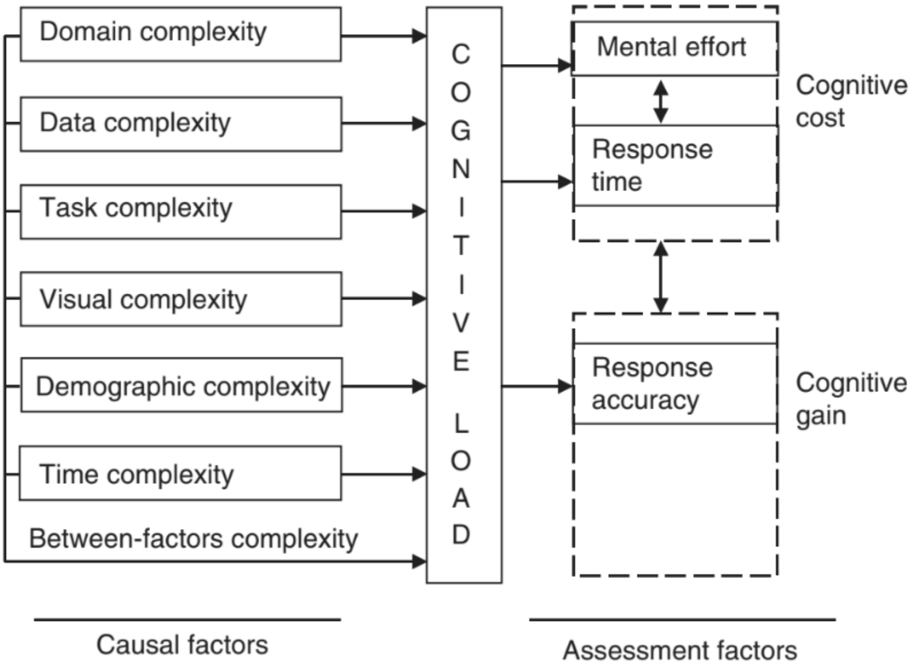


Figure 2.5: A visualisation of cognitive load for understanding visualisations.

When analysing cognitive load, it can be measured by using three different techniques. These include physiological measures (pupillary dilation), subjective measures (secondary tasks) and performance-based measures (rating scale) [35]. As mental effort can be measured, so too can mental effort improvement. Although having the techniques available to measure cognitive load, it is not the goal of this thesis to measure it, but rather to reduce it. Thus, it is important to acknowledge that it can be improved and to take lessons learnt and apply them accordingly.

Figure 2.5 illustrates all of the factors that influence cognitive load as well as the effects the load has on the individual, when processing some data visualisation. As can be seen from the figure, some causal factors such as "Time complexity" and "Visual complexity" can be controlled or reduced, while others unfortunately can not, namely "Domain complexity" and "Data complexity" among others. Further, some causal factors are set by the user such as "Task Complexity".

2.3.3 Visualising Data

The Cognitive Psychology component of Human-Computer Interaction is focused on how humans use software applications. Furthermore, it investigates the importance of creating impactful data visualisations to facilitate sensemaking and reduce the cognitive load when working with complex data. As the analysis of visualised data requires an individual to understand, analyse and learn, it is unsurprising that human cognition plays a vital role in visualising data [31]. Therefore, it is important to acknowledge the role of cognition within visualisation.

Data visualisations are useful since it supports the cognitive processes of sensemaking/ understanding, especially when a user has access to multiple visualisations of the same data. Visuals are often easier and faster to process than numeric or semantic information and it reduces the cognitive load experienced by said user. In order to do so however, the data first needs to undergo pre-processing and manipulation [31].

The final product of data visualisation is known as a "knowledge product". It requires that this product be interpreted easily by an individual [31]. A 'knowledge product' must contain meaning that the analyst can understand when visually examining data. To do so, the act of assigning meaning to the "knowledge product" is not classified as a computational or statistical step but rather a cognitive one [31]. To summarise, recognition of properties within visualisation is only one-half of the visualisation process. The other comprises of the cognitive processes involved in gaining meaning from visualisation [50].

The cognitive process of understanding data visualisation has two core components. Firstly, a process that scans through data for information, highlights relationships and visually displays the data in a way that can be easily understood by the individual [31]. Secondly, the nature and detail of the visualisation needs to be in line with the cognitive ability of the observer in a way that is appropriate to the intended target audience.

Individuals can gain access to attributes of the visualisation process, the properties of the results, or characteristics of the user's perceptual behaviours when analysing visualisations. The user uses such information to reduce the search space for optimal control parameters, making the interaction much more cost-effective. An inappropriate visualisation may impose a high cognitive load on the human cognitive system, thus overwhelming the viewer and undoing its benefits of cognitive support.

Therefore, the visualisation process includes representing data, information, and knowledge regarding the computational and cognitive space [16]. Visualisation compensates for the difficulties individuals face in acquiring a sufficient amount of information or knowledge directly from a data set [16]. Therefore, visualisation is a tool that facilitates an effective and efficient cognitive process that allows for the retrieval of information and meaning from a data set [16].

2.3.4 Benefits of Visualisation

A key benefit of visualisation is the ability to provide cognitive support to the user. This includes assistance in understanding data as well as the ability to communicate information about the data [35]. Furthermore, by visually representing data it allows for tasks that typically require a laborious cognitive process to be completed via a much simpler visual process such as simple perceptual operations [35]. This helps reduce the demand placed on the human memory. In addition, visual representation facilitates the use of recognition mechanisms, which are embedded within an individual's visual perception system, allowing for a quicker rate of recognition [50].

2.3.5 User-centric Approach

In order to benefit from data visualisation, a user-centric approach can be followed as it ensures clear and simple data visualisations. The user-centric approach provides some guidelines and principles in line with visual perception and cognitive processing by providing all the stages and actions to follow in order to uncover the information needed to interpret visualisations. For example, the key and legend on a graph must always clearly be visible and cryptic units or abbreviations should be avoided. One aspect of a user-centric approach is the consideration the user of the visualisation and whether it was successful in fulfilling their requirements. Data scientists may find vastly different visualisations more meaningful compared to standard software application users [33]. These stages and actions might seem obvious, but modern software applications have progressed far beyond static graphs and charts and now allow for complex visualisations for user-driven refinement [33]. Thus, developers of such applications must either decide to visualise as much data as possible or simplify the data visualisation. Balancing the need for accuracy with the risk of confusing or overwhelming the user becomes a difficult task.

According to Wassink et al. 2009, the visualisation design process can be divided into three specific parts. These are an *early envisioning phase*, in which various factors are analysed, such as the users, the environment in which they reside, and the tasks that they conduct within those environments. A *global specification phase*, in which the stakeholders are presented various solutions and propositions are made. Lastly, a *detailed specification phase*, which further expands on what is presented and proposed to the applicable stakeholders. It is important to note that this is and iterative process, and as such, a high level of refinement can be achieved over time. As the user-centric approach is implemented, each phase can consist of multiple iterations, each iteration consisting of three actions. Those being *analysis*, *design* and *evaluation* [55]. Analyses comprises of analysing the users and their environments in order to distinguish between their differences, individually and environmentally. Factors to consider when doing this analysis include that of, defining which spectrum of users will be used to the analysis and which characteristics define them [55]. Furthermore, which tasks they conduct or undertake, what needs to be visualised based on the objects or processes at hand, and lastly, which types of understandings this visualisation should provide. The Design stage emerges from the analysis stage, and comprises of using the knowledge gained from the outcomes of the analysis to develop solutions. These solutions can be expressed through the use of low and high-fidelity prototypes, and as such range from the use of sketches to examples of the final product which can be used for evaluation [55]. The last action that takes place in the iterative process is evaluation, in which interaction styles and the way in which the data visualised is evaluated. The aim of the evaluation is to uncover whether the correct user requirements have been met, through test the various designs and all of their complexity, as well as if there is an improved ability to discover perceptions within data sets, such as those which are heterogeneous in nature [55].

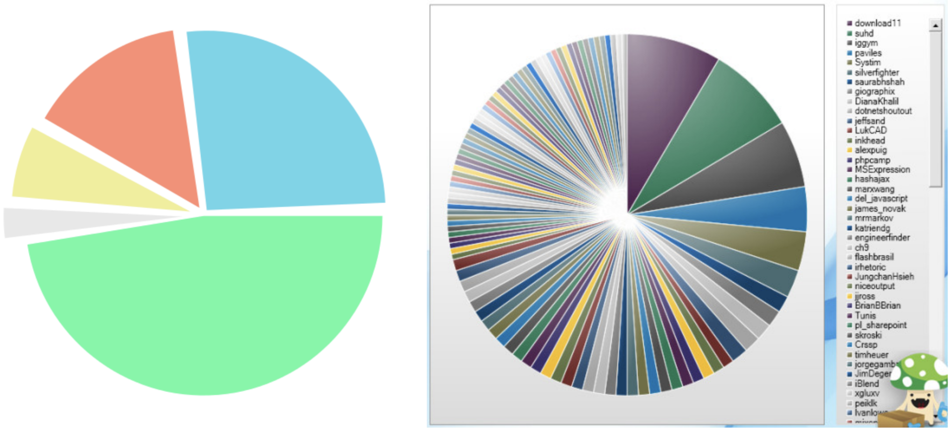


Figure 2.6: Poor Pie Chart Visualisations [52]

Figure 2.6 shows two pie charts, both being examples of poor data visualisation. The pie chart on the left has no title or legend describing what the chart represents, making it difficult for the user to interpret. Contrary to this, the pie chart on the right does have a legend but contains far too many unique segments to be legible to the user. At best these pie charts are not beneficial to the user and at worst they can distract and confuse the user.

2.3.6 Visualising Geolocation Data

Geolocation data can take a number of forms but is generally either street addresses or latitude and longitudinal coordinates, with coordinates being the more accurate of the two. Traditionally this type of data is visualised using maps and with the advancements made in web browser technology and web applications, these maps have evolved far beyond just static visualisations. There are a number of third-party libraries which dynamically render interactive 2D and 3D maps and allows for the plotting of geolocation data on them, either in the form of pins/markers, polygons, polylines or heatmaps[1]. These third-party libraries allow for full integration of these powerful map visualisations into web applications.

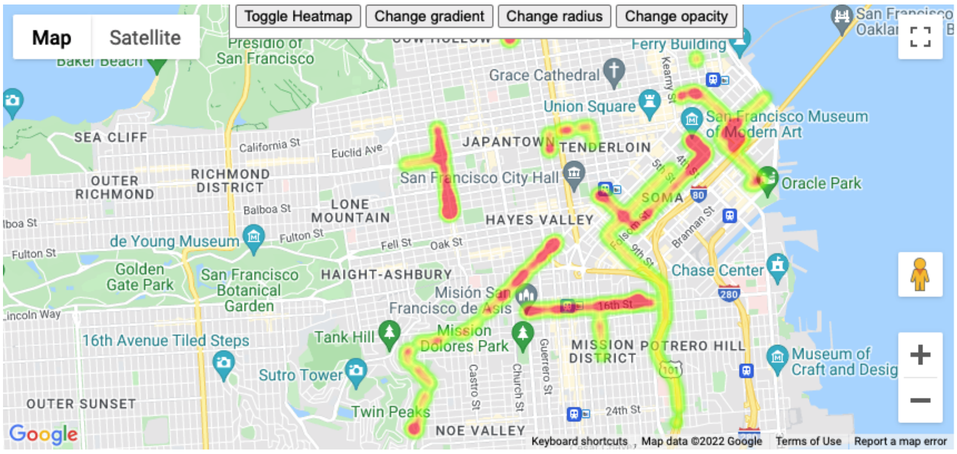


Figure 2.7: Geolocation data plotted as heat maps

Figure 2.7 shows a map visualisation with geolocation data plotted as heat maps. Heat maps are used to depict the intensity of data at geographical points. Areas of higher intensity will be coloured red, and areas of lower intensity will appear green.

2.3.7 Drawbacks of Data Visualisations

Despite the many benefits of visualisation, there are however, several important factors to consider when comparing the different visualisation approaches. For example, most users generally do not have much experience reading sophisticated graphs and charts such as the one shown in Figure 2.8, and overly complex visualisations can confuse the user and obscure meaning. The worst offender being visualisations that deliberately try to mislead the user [33] i.e., by making use of of inconsistent axis on a graph.

It is also important to consider if data should be transformed or visualised beyond just textually at all and it should not just be assumed that textual data is more cognitively demanding and less effective when compared to other visualisations [31]. Further, the visualisation of data, in either a graph or image format, does not ensure the improvement of human understanding of data. There are, in fact, many complementary factors that guide this understanding. Although visualisation may contribute to quicker task performance and a

[1] https://developers.google.com/maps/documentation/javascript/datalayer

decrease in errors, it is a seldom occurrence for visualisation to produce both high accuracy as well as a short response time [31].

Another risk with data visualisations relates to limited working memory. This capacity means that the user may experience a high cognitive load which causes the user to feel overwhelmed and ultimately negates the benefits provided by visualisation [35]. Consequently, the success of data visualisation is dependent on the ability of the user to perceive and process the embedded information efficiently and effectively [35].

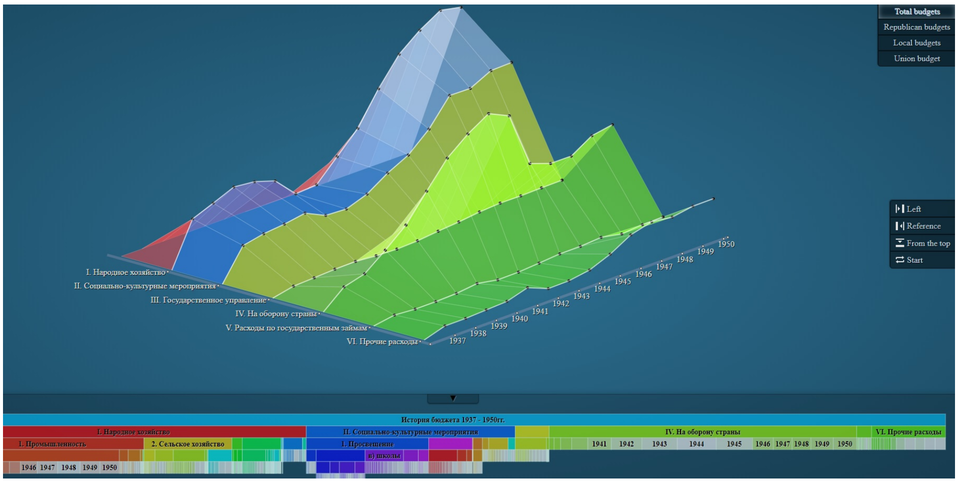


Figure 2.8: A complex modern graph visualisation [2]

2.3.8 Data Visualisations and Large Data Sets

Search queries can often take very long to be processed, especially when working with large data sets. As a result, emphasis is placed on improving the efficiency of visualisation tools to allow users to have an interactive search that is both fast and allows for the exploration of bigger parameter spaces [16]. Efficient visualisation tools will facilitate better knowledge discovery when working with large databases and ultimately aid complex problem-solving and interactive analysis [50].

2.3.9 Future of Visualisations

Chen et al. (2008), proposes the need to use knowledge-assisted visualisation in order to handle the vast amount of collected information. To elaborate, it is stated that knowledge-assisted visualisation provides an interface which merges information-visualisation and scientific-visualisation communities [16]. Furthermore, Chen et al. (2008) states that, due to the increase in both complexity and size of data, the use of visualisation for large data sets will be a necessity rather than an option [16].

2.4 FCA Data Visualisation Tools

The use of a formal concept lattice improves multiple areas of information retrieval and knowledge discovery [14]. The first area is query refinement, where a concept can be viewed as a query (the intent) with a set of retrieved results (the extent). Secondly, the combination of querying and navigation into a single process allows for functionality and data exploration not previously possible. For example, being able to determine both the results of a specific topic and the topics covered by certain results.

Finally, the ability to refine and add attributes to a query incrementally based on currently displayed results, allows for a more refined and valuable set of results. This is achieved by combining multiple views of semi-structured data in cases where data is classified according to various attributes. With concept lattices specifically, this is achieved by combining different lattices of partial views of the data and focusing on the points of intersection between them [14].

Within FCA, most software applications utilise the benefits of the formal context lattice for data visualisation and exploration. These applications use a formal context as input to automatically generate a lattice. The lattice is used to perform various calculations. Most of these applications visualise the data textually, with some also showing the user the entire lattice.

Generally, these tools function effectively, but a common shortcoming is their inability to visualise large data sets [48]. Attempting to generate and retain entire formal concept lattices in memory becomes very expensive, especially when attempting to perform complex calculations on the lattice. Furthermore, the larger the formal context, the more difficult it is to interpret the formal concept lattice. Consequently, the visualisation of it becomes less useful to the user.

A discussion of four tools is provided to evaluate their effectiveness in visualising data and empowering data exploration, especially in larger data sets. The tools are ToscanaJ, FCART, Fish-eye software, and ConceptCloud. The analysis of these tools will determine the current state of FCA-based data visualisation tools.

2.4.1 ToscanaJ

The ToscanaJ project is a collaborative project that aims to recreate the original Toscana lattice browser as a modern open-source desktop application for the Microsoft Windows Operating System [4]. The ToscanaJ suite contains the following tools:

- ToscanaJ: The main lattice viewer component.

- Elba: A tool which allows for the editing of conceptual schemas on relational databases and allows for the exporting of SQL scripts.
- Siena: Similar to Elba, Siena also allows for the editing of conceptual schemas, but for ones which store their data in memory.
- Lucca: Another editor that makes use of implication analysis of SQL clauses in order to create and explore database-connected systems.

Currently, ToscanaJ is the most popular tool in the suite. The software displays predefined diagrams of conceptual structures and allows the user to browse and navigate through complex data sets, using a simple graphical interface [8]. Figure 2.9 provides a view of the user interface which shows nested diagrams that have been defined by some knowledge engineer. The user is then able to analyse any predefined diagram or combination of diagrams to obtain more detailed information. Clicking on one node results in a filter process, consequently, only the objects belonging to the selected node will be used for the following analysis.

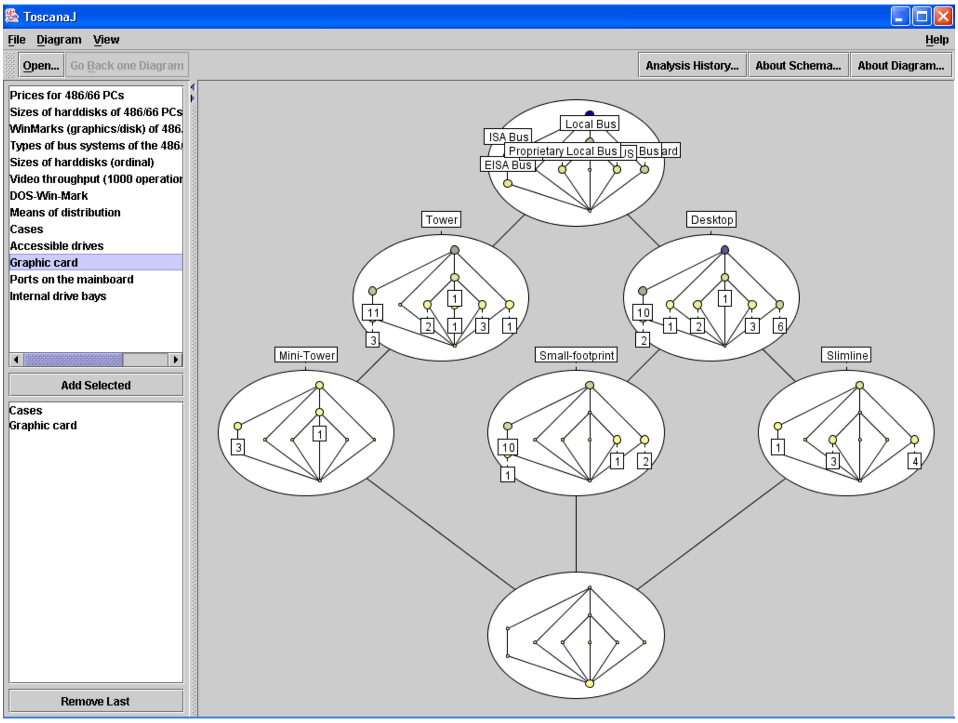
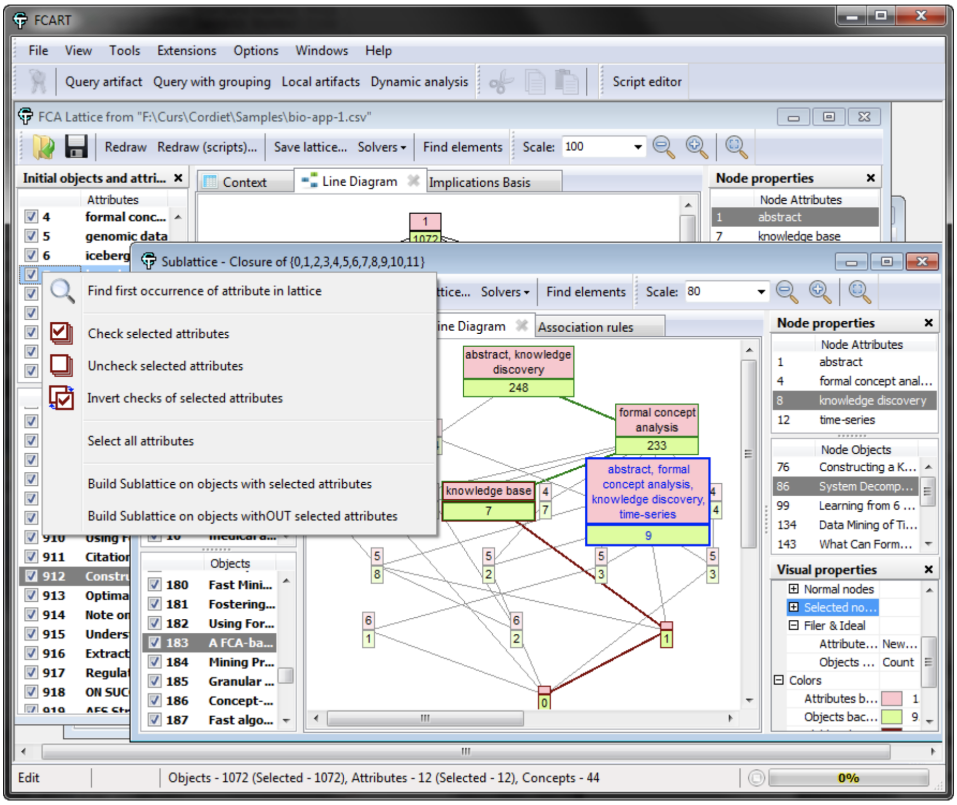


Figure 2.9: The ToscanaJ user interface with nested diagrams [8]

2.4.2 FCART

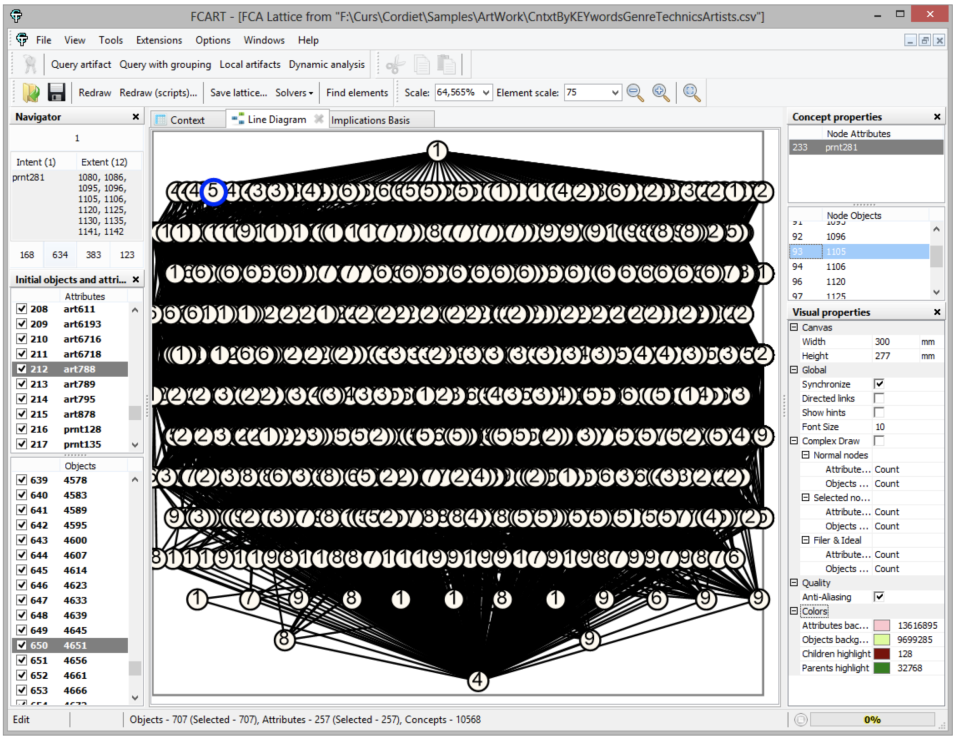
The Formal Concept Analysis Research Toolbox (FCART) was released in 2013 as a tool for information retrieval and knowledge discovery from various data sources, with a focus on unstructured (textual) data [48]. FCART is a local Microsoft Windows Application that provides the user with tools for iterative analysis of their data with adjustable queries and analytic artefacts. It accepts various data formats as input which the software uses to construct a binary context. This binary context then proceeds to generate a concept lattice. Figure 2.10 shows FCART's user interface of a generated concept lattice.



Objects - 1072 (Selected - 1072), Attributes - 12 (Selected - 12), Concepts - 44

Figure 2.10: FCART Concept Lattice Visualizer

Although FCART allows for complex visualisations, the tool has some drawbacks. Prior knowledge of the data is required to write effective queries. The user is also never given a complete overview of their data. Furthermore, as seen in Figure 2.11, the interface becomes challenging to read when working with larger data sets. A visualisation that becomes overpopulated makes it difficult for the user to distinguish between the various labels within the visualisation or the links between the items. Consequently, gaining any intuition into the data set becomes near impossible.



FCART - [FCA Lattice from "F:\Curs\Cordiet\Samples\ArtWork\CntxtByKEYwordsGenreTechnicsArtists.csv"]

Scale: 64,565% Element scale: 75

Navigator

1

Intent (1)	Extent (12)
prnt281	1080, 1086, 1095, 1096, 1105, 1106, 1120, 1125, 1130, 1135, 1141, 1142

168	634	383	123

	Attributes
☑ 208	art611
☑ 209	art6193
☑ 210	art6716
☑ 211	art6718
☑ 212	art788
☑ 213	art789
☑ 214	art795
☑ 215	art878
☑ 216	prnt128
☑ 217	prnt135

	Objects
☑ 639	4578
☑ 640	4583
☑ 641	4589
☑ 642	4595
☑ 643	4600
☑ 644	4607
☑ 645	4614
☑ 646	4623
☑ 647	4633
☑ 648	4639
☑ 649	4645
☑ 650	4651
☑ 651	4656
☑ 652	4661
☑ 653	4666

	Node Attributes
233	prnt281

	Node Objects
92	1096
93	1105
94	1106
96	1120
97	1125

Visual properties		
⊟ Canvas		
Width	300	mm
Height	277	mm
⊟ Global		
Synchronize	☑	
Directed links	□	
Show hints	□	
Font Size	10	
⊟ Complex Draw	□	
⊟ Normal nodes		
Attribute...	Count	
Objects ...	Count	
⊟ Selected no...		
Attribute...	Count	
Objects ...	Count	
⊟ Filer & Ideal		
Attribute...	Count	
Objects ...	Count	
⊟ Quality		
Anti-Aliasing	☑	
⊟ Colors		
Attributes bac...	🟥	13616895
Objects backg...	🟩	9699285
Children highlight	🟥	128
Parents highlight	🟩	32768

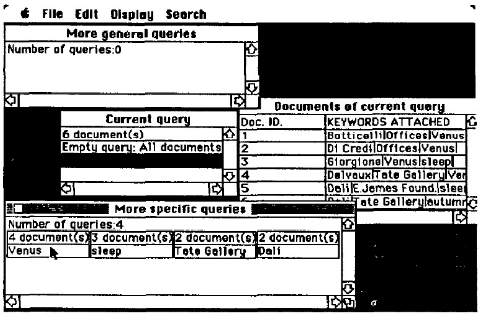
Objects - 707 (Selected - 707), Attributes - 257 (Selected - 257), Concepts - 10568

Figure 2.11: Large FCART Concept Lattice

2.4.3 Fish-eye Viewer Software

This tool was developed in 1986 for the Apple Macintosh operating system and is not able to run on current hardware. The reason for its inclusion is due to its unique approach to the visualisation of large data sets due to technological restrictions. The tool's user interface allows gradual enlargement or refinement of the user's query by browsing through a graph of term and document subsets [27]. The graph is obtained from a lattice which is automatically generated from a document-term relation data set. Each vertex in the graph represents a query formed of a conjunction of terms with the retrieved documents, similar to formal objects and attributes. A fish-eye view is employed to deal with the problem of managing a large amount of output data. By analogy with a wide-angle fish-eye lens, the idea is to display all details of the current focus while gradually reducing detail as you move away from it. This method allows for the decrease in detail to be displayed when moving further from the focal point [27]. In a similar manner to a word cloud, a vertex is displayed according to its degree of interest (DOI). This means that the importance of the vertex is relative to the distance from the focal area [27]. A fish-eye view of a controllable size can be obtained by showing the most interesting points within a given DOI threshold.

There are two different ways of interacting with the user interface which is dependent on the level of familiarity a user has with the underlying data set. Users with previous experience may directly navigate to a particular vertex in the graph by preemptively deciding on the subject which the user wants to query [27]. Furthermore, by interacting with the graph, a user can alter the initial query [27], as is shown in Figure 2.12, Figure 2.13 and Figure 2.14. Alternately, a user with no prior knowledge of the data set may interact with the lattice in order to navigate to the desired results without needing to input a specific query [27].



More general queries

Number of queries:0

Current query

6 document(s)

Empty query: All documents

Documents of current query

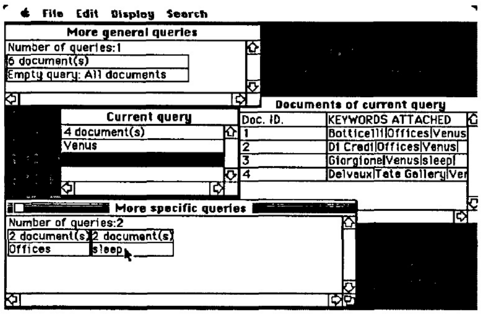
Doc. ID.	KEYWORDS ATTACHED
1	Botticelli|Offices|Venus
2	Di Credi|Offices|Venus|
3	Giorgione|Venus|sleep|
4	Delvaux|Tate Gallery|Ver
5	Dali|E.James Found.|slee
	Tate Gallery|autumn

More specific queries

Number of queries:4

4 document(s)	3 document(s)	2 document(s)	2 document(s)
Venus	sleep	Tate Gallery	Dali

Figure 2.12: User interface when the user clicks the more specific query "Venus" in the "More specific queries" window.



More general queries

Number of queries:1

6 document(s)

Empty query: All documents

Current query

4 document(s)

Venus

Documents of current query

Doc. ID.	KEYWORDS ATTACHED
1	Botticelli|Offices|Venus
2	Di Credi|Offices|Venus|
3	Giorgione|Venus|sleep|
4	Delvaux|Tate Gallery|Ver

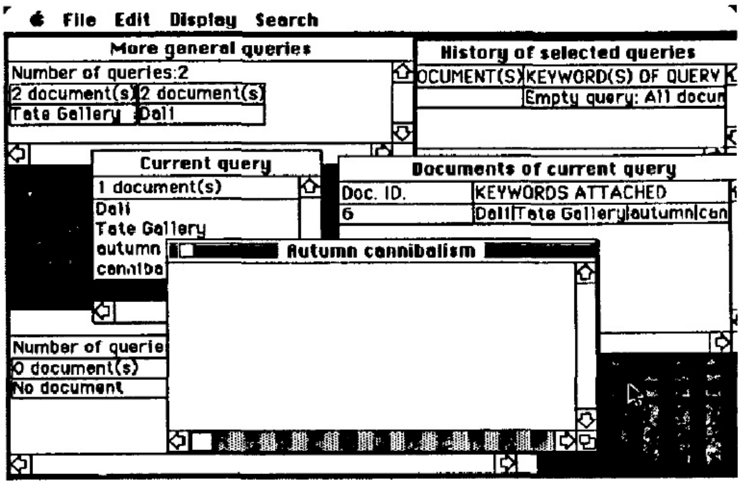
More specific queries

Number of queries:2

2 document(s)	2 document(s)
Offices	sleep

Figure 2.13: View resulting from the user selecting the more specific query "Venus and sleep". N.B. Only the additional terms with respect to the current query are displayed in the "More specific queries" window".

When evaluating the tool's ability to process large data sets, the tool understandably has difficulty generating large complex graphs. The researchers state that devices with limited space and display resolution often give rise to issues as the structure of the graph is typically too large. As a result, a window is used in order to display overflow queries. Godin et al. (1986) states that due to the graph's complexity, it may be difficult to manage and as a result a compromise or schematisation is often necessary.



Number of queries:2
2 document(s) 2 document(s)
Tate Gallery Dali

OCUMENT(S) KEYWORD(S) OF QUERY
Empty query: All docu

1 document(s)
Dali
Tate Gallery
autumn
cannibа

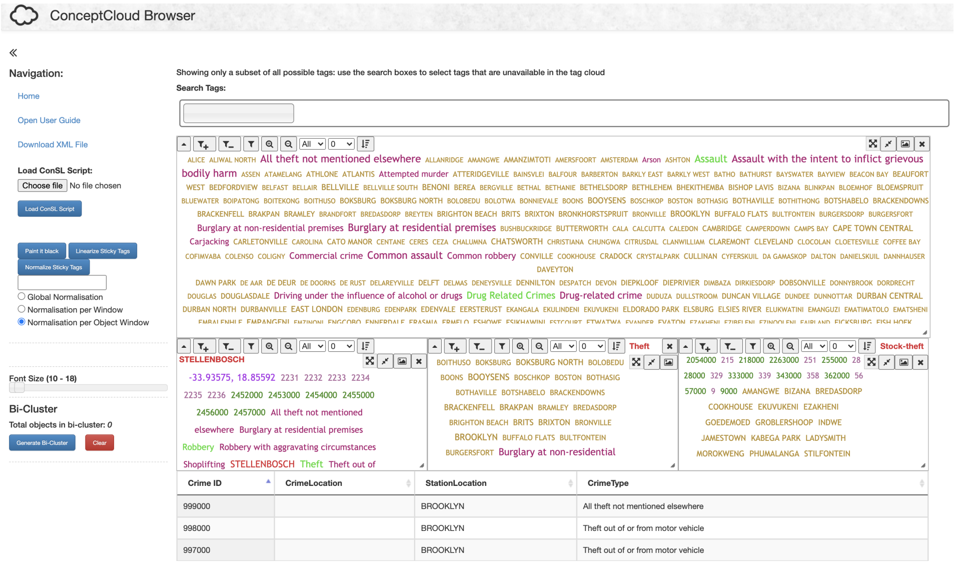
Doc. ID. KEYWORDS ATTACHED
6 Dali Tate Gallery autumn can

Number of querie
0 document(s)
No document

Figure 2.14: The contents of the selected document selected from the "Documents of current query" window

2.5 ConceptCloud

The final FCA data visualisation tool to be discussed is ConceptCloud, the focus of this thesis. ConceptCloud uses a formal concept lattice generated from the input data as the underlying navigation structure [10, 26, 29]. The data is presented in an interactive *tag cloud*, providing the user with both an intuitive representation of the data set and allowing for interactive navigation through the data. Figure 2.15 shows ConceptCloud's user interface. Tag clouds are representations of the number of occurrences and types of attributes and objects in a data set, where the size of each tag displayed is determined by its frequency within the data set, as is the colour depending on the type of tag displayed. Navigation is achieved by tag selection and deselection, removing the confines of predefined search paths and allows for an explorative search process to be followed.



Crime ID	CrimeLocation	StationLocation	CrimeType
999000		BROOKLYN	All theft not mentioned elsewhere
998000		BROOKLYN	Theft out of or from motor vehicle
997000		BROOKLYN	Theft out of or from motor vehicle

Figure 2.15: ConceptCloud User Interface

The tool allows the user to iteratively select an attribute or object tag in the tag cloud, which updates the tag cloud to display all other objects or attribute tags possessing the selected attribute tag(s). This is achieved by making use of a *focus concept* from which a tag cloud is created.

Formally, the *focus concept* $c := \langle O, A \rangle$ is the concept whose extent is the set of objects that share the set of currently selected attributes, $F \subset A$, within the tag cloud. The focus concept can be further refined by iteratively adding elements to F. When an additional attribute is added to F, the focus concept is updated by computing the meet of the current focus concept (c) and the concept introduced by the additional attribute. This strategy obviates the need to compute an entire concept lattice, which is not feasible for large data sets [10].

The explorative search process described above simulates the process of navigating through a concept lattice, wherein the selection of an attribute moves the user to the point in the lattice where all linked objects contain that attribute. As more attributes are selected and the query is refined, the user moves further down the lattice. The reverse of this process is also possible where the user may deselect previously selected tags and move back up the lattice. This process corresponds to the refinement of the focus concept by adding and removing elements from F.

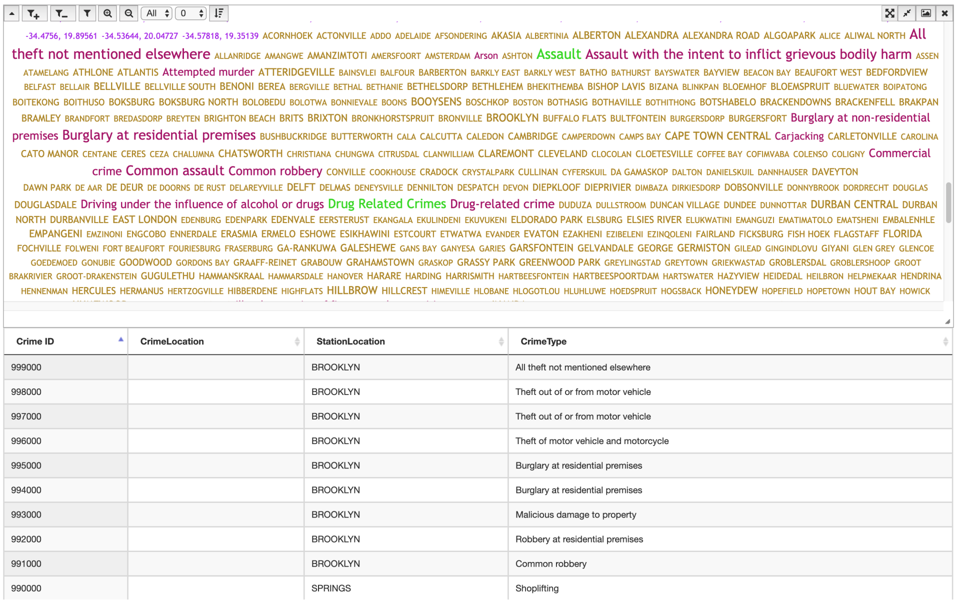
Definition 6. *The focus concept $c = \langle O, A \rangle$ is the concept whose extent is the set of objects that share the set of currently selected attributes, F, within the tag cloud, such that $\alpha(\omega(F)) = \pi_A(c) = A$.*

Tag clouds are constructed by taking the extent of the focus concept, $c = \langle O, A \rangle$. For each $o_i \in O$, its defining concept c_i is determined, and all the intents (attributes) and extents (objects) of these defining concepts are displayed.

The initial focus concept has no selected attributes, and thus the tag cloud created from it will contain tags representing all attributes and objects. Formally this can be defined as (in this case $\uplus$ denotes a multiset union):

Definition 7. *The tag cloud from a concept* $c = (O, A) \in B(\mathcal{C})$ *is defined as* $\tau(c) = O \uplus \biguplus_{o \in O} \pi_A(\sigma(o))$.

Through construction of tags in the tag cloud, subconcepts of the focus concept is induced with concepts having a non-bottom meet with that focus concept.

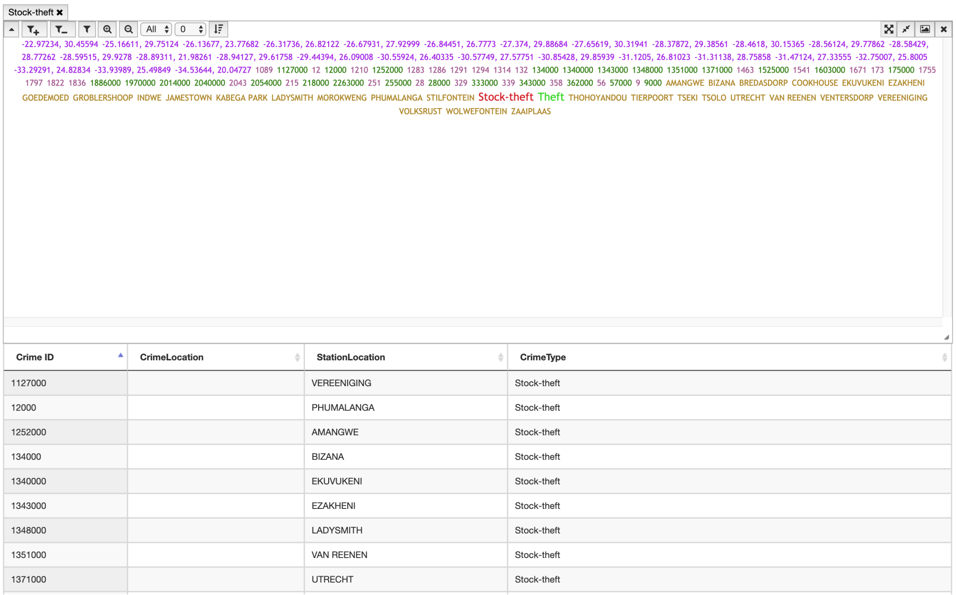


-34.4756, 19.89561 -34.53644, 20.04727 -34.57818, 19.35139 ACORNHOEK ACTONVILLE ADDO ADELAIDE AFSONDERING AKASIA ALBERTINIA ALBERTON ALEXANDRA ALEXANDRA ROAD ALGOAPARK ALICE ALIWAL NORTH All
theft not mentioned elsewhere ALLANRIDGE AMANGWE AMANZIMTOTI AMERSFOORT AMSTERDAM Arson ASHTON Assault Assault with the intent to inflict grievous bodily harm ASSEN
ATAMELANG ATHLONE ATLANTIS Attempted murder ATTERIDGEVILLE BAINSVLEI BALFOUR BARBERTON BARKLY EAST BARKLY WEST BATHO BATHURST BAYSWATER BAYVIEW BEACON BAY BEAUFORT WEST BEDFORDVIEW
BELFAST BELLAIR BELLVILLE BELLVILLE SOUTH BENONI BEREA BERGVILLE BETHAL BETHANIE BETHELSDORP BETHLEHEM BHEKITHEMBA BISHOP LAVIS BIZANA BLINKPAN BLOEMHOF BLOEMSPRUIT BLUEWATER BOIPATONG
BOITEKONG BOITHUSO BOKSBURG BOKSBURG NORTH BOLOBEDU BOLOTWA BONNIEVALE BOONS BOOYSENS BOSCHKOP BOSTON BOTHASIG BOTHAVILLE BOTHITHONG BOTSHABELO BRACKENDOWNS BRACKENFELL BRAKPAN
BRAMLEY BRANDFORT BREDASDORP BREYTEN BRIGHTON BEACH BRITS BRIXTON BRONKHORSTSPRUIT BRONVILLE BROOKLYN BUFFALO FLATS BULTFONTEIN BURGERSDORP BURGERSFORT Burglary at non-residential
premises Burglary at residential premises BUSHBUCKRIDGE BUTTERWORTH CALA CALCUTTA CALEDON CAMBRIDGE CAMPERDOWN CAMPS BAY CAPE TOWN CENTRAL Carjacking CARLETONVILLE CAROLINA
CATO MANOR CENTANE CERES CEZA CHALUMNA CHATSWORTH CHRISTIANA CHUNGWA CITRUSDAL CLANWILLIAM CLAREMONT CLEVELAND CLOCOLAN CLOETESVILLE COFFEE BAY COFIMVABA COLENSO COLIGNY Commercial
crime Common assault Common robbery CONVILLE COOKHOUSE CRADOCK CRYSTALPARK CULLINAN CYFERSKUIL DA GAMASKOP DALTON DANIELSKUIL DANNHAUSER DAVEYTON
DAWN PARK DE AAR DE DEUR DE DOORNS DE RUST DELAREYVILLE DELFT DELMAS DENEYSVILLE DENNILTON DESPATCH DEVON DIEPKLOOF DIEPRIVIER DIMBAZA DIRKIESDORP DOBSONVILLE DONNYBROOK DORDRECHT DOUGLAS
DOUGLASDALE Driving under the influence of alcohol or drugs Drug Related Crimes Drug-related crime DUDUZA DULLSTROOM DUNCAN VILLAGE DUNDEE DUNNOTTAR DURBAN CENTRAL DURBAN
NORTH DURBANVILLE EAST LONDON EDENBURG EDENPARK EDENVALE EERSTERUST EKANGALA EKULINDENI EKUVUKENI ELDORADO PARK ELSBURG ELSIES RIVER ELUKWATINI EMANGUZI EMATIMATOLO EMATSHENI EMBALENHLE
EMPANGENI EMZINONI ENGCOBO ENNERDALE ERASMIA ERMELO ESHOWE ESIKHAWINI ESTCOURT ETWATWA EVANDER EVATON EZAKHENI EZIBELENI EZINQOLENI FAIRLAND FICKSBURG FISH HOEK FLAGSTAFF FLORIDA
FOCHVILLE FOLWENI FORT BEAUFORT FOURIESBURG FRASERBURG GA-RANKUWA GALESHEWE GANS BAY GANYESA GARIES GARSFONTEIN GELVANDALE GEORGE GERMISTON GILEAD GINGINDLOVU GIYANI GLEN GREY GLENCOE
GOEDEMOED GONUBIE GOODWOOD GORDONS BAY GRAAFF-REINET GRABOUW GRAHAMSTOWN GRASKOP GRASSY PARK GREENWOOD PARK GREYLINGSTAD GREYTOWN GRIEKWASTAD GROBLERSDAL GROBLERSHOOP GROOT
BRAKRIVIER GROOT-DRAKENSTEIN GUGULETHU HAMMANSKRAAL HAMMARSDALE HANOVER HARARE HARDING HARRISMITH HARTBEESFONTEIN HARTBEESPOORTDAM HARTSWATER HAZYVIEW HEIDEDAL HEILBRON HELPMEKAAR HENDRINA
HENNENMAN HERCULES HERMANUS HERTZOGVILLE HIBBERDENE HIGHFLATS HILLBROW HILLCREST HIMEVILLE HLOBANE HLOGOTLOU HLUHLUWE HOEDSPRUIT HOGSBACK HONEYDEW HOPEFIELD HOPETOWN HOUT BAY HOWICK

Crime ID	CrimeLocation	StationLocation	CrimeType
999000		BROOKLYN	All theft not mentioned elsewhere
998000		BROOKLYN	Theft out of or from motor vehicle
997000		BROOKLYN	Theft out of or from motor vehicle
996000		BROOKLYN	Theft of motor vehicle and motorcycle
995000		BROOKLYN	Burglary at residential premises
994000		BROOKLYN	Burglary at residential premises
993000		BROOKLYN	Malicious damage to property
992000		BROOKLYN	Robbery at residential premises
991000		BROOKLYN	Common robbery
990000		SPRINGS	Shoplifting

Figure 2.16: Initial Tag Cloud with no Focus Concept

Figure 2.16 and Figure 2.17 show the tag cloud and object table, applied to a crime data set, before and after an attribute has been selected. In this case the user has selected the "Stock Theft" attribute and the tag cloud and table has been updated to represent the new lattice location. The table below the tag cloud shows the various values of the attributes of the objects in the data set with the attribute category appearing in the header of the table.



-22.97234, 30.45594 -25.16611, 29.75124 -26.13677, 23.77682 -26.31736, 26.82122 -26.67931, 27.92999 -26.84451, 26.7773 -27.374, 29.88684 -27.65619, 30.31941 -28.37872, 29.38561 -28.4618, 30.15365 -28.56124, 29.77862 -28.58429, 28.77262 -28.59515, 29.9278 -28.89311, 21.98261 -28.94127, 29.61758 -29.44394, 26.09008 -30.55924, 26.40335 -30.57749, 27.57751 -30.85428, 29.85939 -31.1205, 26.81023 -31.31138, 28.75858 -31.47124, 27.33555 -32.75007, 25.8005 -33.29291, 24.82834 -33.93989, 25.49849 -34.53644, 20.04727 1089 1127000 12 12000 1210 1252000 1283 1286 1291 1294 1314 132 134000 1340000 1343000 1348000 1351000 1371000 1463 1525000 1541 1603000 1671 173 175000 1755 1797 1822 1836 1886000 1970000 2014000 2040000 2043 2054000 215 218000 2263000 251 255000 28 28000 329 333000 339 343000 358 362000 56 57000 9 9000 AMANGWE BIZANA BREDASDORP COOKHOUSE EKUVUKENI EZAKHENI GOEDEMOED GROBLERSHOOP INDWE JAMESTOWN KABEGA PARK LADYSMITH MOROKWENG PHUMALANGA STILFONTEIN Stock-theft Theft THOHOYANDOU TIERPOORT TSEKI TSOLO UTRECHT VAN REENEN VENTERSDORP VEREENIGING VOLKSRUST WOLWEFONTEIN ZAAIPLAAS

Crime ID	CrimeLocation	StationLocation	CrimeType
1127000		VEREENIGING	Stock-theft
12000		PHUMALANGA	Stock-theft
1252000		AMANGWE	Stock-theft
134000		BIZANA	Stock-theft
1340000		EKUVUKENI	Stock-theft
1343000		EZAKHENI	Stock-theft
1348000		LADYSMITH	Stock-theft
1351000		VAN REENEN	Stock-theft
1371000		UTRECHT	Stock-theft

Figure 2.17: Tag Cloud with Attribute Selected

2.5.1 Architecture

ConceptCloud adopts a client-server architecture visualised in Figure 2.18. Its server is written in Java and uses the Play! framework[2] along with a PostgreSQL database to generate the concept lattice. Since the lattice is only used for explorative search, it is not necessary to generate the entire lattice at any one time. The *Colibri* java library is used to generate the required segments of lattice at each navigation step, and these segments are generated and discarded as the user navigates the lattice [9]. As a result, ConceptCloud is capable of manipulating large lattices and visualising very large data sets. The application can run on any Unix based system and the front-end can be accessed using any web browser.

It is also worth noting that the context table displayed by ConceptCloud is not a traditional binary context table, but rather a a multi-valued context where each attribute may have any particular value. Multi-valued contexts are much more human-readable and practical to display than a binary context, especially with larger data sets. Figure 2.16 shows a multi-valued context table. In this example, `Attribute:CrimeType Value:Shoplifting` is displayed instead of `Attribute:Shoplifting Value:X` [9].

In terms of loading data into ConceptCloud, it is not specialised to any specific data set. It accepts any well formed JSON data set or CSV data set, and processes it in order to generate the concept lattice, and subsequently the tag cloud. This process is described below:

[2] https://www.playframework.com/

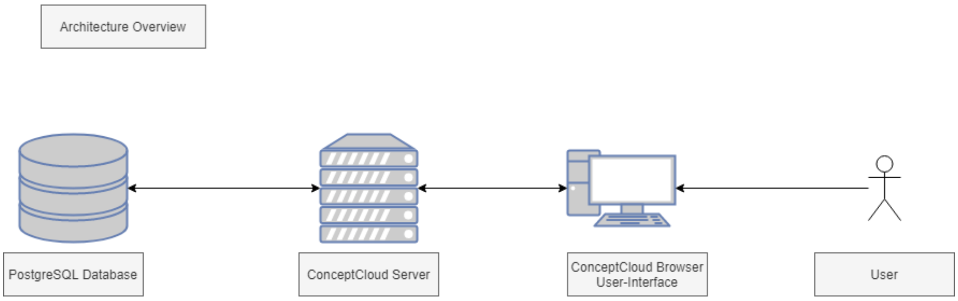


Figure 2.18: ConceptCloud Architecture Overview

2.5.1.1 Data Preparation and Manipulation

- **Data Source:** The first step in uploading data to ConceptCloud is preparing a JSON file of the desired data. The data must be broken into individual objects (For each row or entry in the set) and formatted according to the JSON standards [9]. The JSON objects may not have nested objects.

- **Data Configuration Script:** The configuration script is responsible for setting the initial run configuration for the project, which includes setting the directory path of the data source, the data fields in which key-phrase extraction should be performed and the attributes to be displayed in the context table [9].

- **Data Extraction:** During the data extraction process, the application obtains the data structure in order to generate the caching database and reads in the data from the input data source [9]. Additionally, this component performs the key phrase extraction and any required NLP operations on the data before passing the data through to the table management and lattice builder operations.

- **Database Table Generation and Population:** This step involves the generation of the context table in the PostgreSQL database and the insertion of the extracted data in these new tables [9].

- **Concept Lattice Generator:** During this step the concept lattice is created from the Context table. In order to maintain scalability and keep resource usage low, only the required sub-sections of the lattice are generated as the user navigates the lattice via the tag cloud [9].

- **User Interface:** The final step of the process is to send the data to the User Interface in the form of a tag cloud through which the user interacts with the underlying concept lattice and exploratively searches through the data set [9].

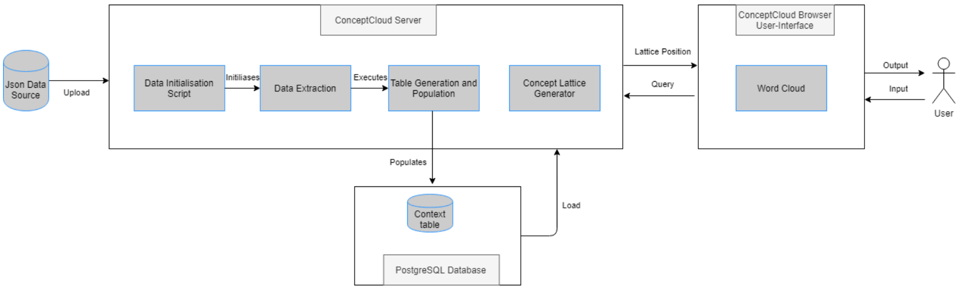
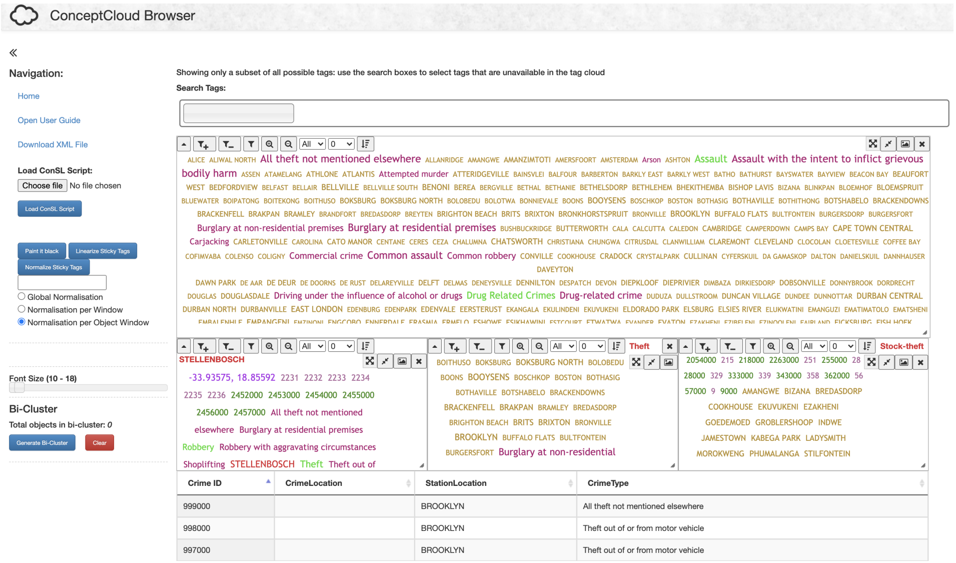


Figure 2.19: ConceptCloud Low Level Architecture

ConceptCloud's architecture, visualised in Figure 2.19, and the described process above allows the application to process very large data sets efficiently and in a timely manner. This also allows the user to upload any semi structured data into ConceptCloud as long as it is of a suitable format.

2.5.2 User Interface

ConceptCloud's user interface consists of two main elements, namely a *Tag Cloud* and *Object Table*. When accessing the user interface for the first time, the user will see a tag cloud of attributes and objects for the uploaded data set as well as a table that represents each formal object as a row.



No file chosen

○ Global Normalisation
○ Normalisation per Window
● Normalisation per Object Window

Bi-Cluster

Total objects in bi-cluster: *0*

Showing only a subset of all possible tags: use the search boxes to select tags that are unavailable in the tag cloud

Search Tags:

ALICE ALIWAL NORTH All theft not mentioned elsewhere ALLANRIDGE AMANGWE AMANZIMTOTI AMERSFOORT AMSTERDAM Arson ASHTON Assault Assault with the intent to inflict grievous bodily harm ASSEN ATAMELANG ATHLONE ATLANTIS Attempted murder ATTERIDGEVILLE BAINSVLEI BALFOUR BARBERTON BARKLY EAST BARKLY WEST BATHO BATHURST BAYSWATER BAYVIEW BEACON BAY BEAUFORT WEST BEDFORDVIEW BELFAST BELLAIR BELLVILLE BELLVILLE SOUTH BENONI BEREA BERGVILLE BETHAL BETHANIE BETHELSDORP BETHLEHEM BHEKITHEMBA BISHOP LAVIS BIZANA BLINKPAN BLOEMHOF BLOEMSPRUIT BLUEWATER BOIPATONG BOITEKONG BOITHUSO BOKSBURG BOKSBURG NORTH BOLOBEDU BOLOTWA BONNIEVALE BOONS BOOYSENS BOSCHKOP BOSTON BOTHASIG BOTHAVILLE BOTHITHONG BOTSHABELO BRACKENDOWNS BRACKENFELL BRAKPAN BRAMLEY BRANDFORT BREDASDORP BREYTEN BRIGHTON BEACH BRITS BRIXTON BRONKHORSTSPRUIT BRONVILLE BROOKLYN BUFFALO FLATS BULTFONTEIN BURGERSDORP BURGERSFORT Burglary at non-residential premises Burglary at residential premises BUSHBUCKRIDGE BUTTERWORTH CALA CALCUTTA CALEDON CAMBRIDGE CAMPERDOWN CAMPS BAY CAPE TOWN CENTRAL Carjacking CARLETONVILLE CAROLINA CATO MANOR CENTANE CERES CEZA CHALUMNA CHATSWORTH CHRISTIANA CHUNGWA CITRUSDAL CLANWILLIAM CLAREMONT CLEVELAND CLOCOLAN CLOETESVILLE COFFEE BAY COFIMVABA COLENSO COLIGNY Commercial crime Common assault Common robbery CONVILLE COOKHOUSE CRADOCK CRYSTALPARK CULLINAN CYFERSKUIL DA GAMASKOP DALTON DANIELSKUIL DANNHAUSER DAVEYTON DAWN PARK DE AAR DE DEUR DE DOORNS DE RUST DELAREYVILLE DELFT DELMAS DENEYSVILLE DENNILTON DESPATCH DEVON DIEPKLOOF DIEPRIVIER DIMBAZA DIRKIESDORP DOBSONVILLE DONNYBROOK DORDRECHT DOUGLAS DOUGLASDALE Driving under the influence of alcohol or drugs Drug Related Crimes Drug-related crime DUDUZA DULLSTROOM DUNCAN VILLAGE DUNDEE DUNNOTTAR DURBAN CENTRAL DURBAN NORTH DURBANVILLE EAST LONDON EDENBURG EDENPARK EDENVALE EERSTERUST EKANGALA EKULINDENI EKUVUKENI ELDORADO PARK ELSBURG ELSIES RIVER ELUKWATINI EMANGUZI EMATIMATOLO EMATSHENI EMBALENHLE EMPANGENI EMZINONI ENGCOBO ENNERDALE ERASMIA ERMELO ESHOWE ESIKHAWINI ESTCOURT ETWATWA EVANDER EVATON EZAKHENI EZIBELENI EZINQOLENI FAIRLAND FICKSBURG FISH HOEK

STELLENBOSCH

-33.93575, 18.85592 2231 2232 2233 2234 2235 2236 2452000 2453000 2454000 2455000 2456000 2457000 All theft not mentioned elsewhere Burglary at residential premises Robbery Robbery with aggravating circumstances Shoplifting STELLENBOSCH Theft Theft out of

BOITHUSO BOKSBURG BOKSBURG NORTH BOLOBEDU BOONS BOOYSENS BOSCHKOP BOSTON BOTHASIG BOTHAVILLE BOTSHABELO BRACKENDOWNS BRACKENFELL BRAKPAN BRAMLEY BREDASDORP BRIGHTON BEACH BRITS BRIXTON BRONVILLE BROOKLYN BUFFALO FLATS BULTFONTEIN BURGERSFORT Burglary at non-residential

2054000 215 218000 2263000 251 255000 28 28000 329 333000 339 343000 358 362000 56 57000 9 9000 AMANGWE BIZANA BREDASDORP COOKHOUSE EKUVUKENI EZAKHENI GOEDEMOED GROBLERSHOOP INDWE JAMESTOWN KABEGA PARK LADYSMITH MOROKWENG PHUMALANGA STILFONTEIN

Crime ID	CrimeLocation	StationLocation	CrimeType
999000		BROOKLYN	All theft not mentioned elsewhere
998000		BROOKLYN	Theft out of or from motor vehicle
997000		BROOKLYN	Theft out of or from motor vehicle

Figure 2.20: ConceptCloud User Interface

ConceptCloud's user interface consists of the following elements (as can be seen in Figure 2.20):

- **Tag Cloud Window:** The focal tag cloud is on display in this window. On initial execution, the initial tag cloud viewer with the default focus concept (the default being the top of the lattice) is displayed. The tag cloud within the main window displays the top 5000 most relevant tags. By selecting a relevant tag, the user can recalculate the focus concept [10]. As a result, the new point of focus in the lattice displayed in the window. Selected tags appear at the top of this window and a user can deselect these tags if they wish to move back up the lattice.

- **Navigation Window:** This window provides the user with navigation links to the home screen and user documentation. Furthermore, it gives the user the ability to upload a ConSL script which allows for automation of what is displayed in the tag windows [28]. The user is also able to change the scaling of the tags within the tag cloud by dragging the slider on the left of the screen to change the scaling constant [10].

- **Search Bar:** The search bar allows the user to search for a specific attribute or object that is not displayed in the initial tag window. By using a caching database, ConceptCloud provides users with an auto-complete list of terms and their categories as they enter their search terms. Selecting one of these terms has the same effect as selecting a tag in the main window [10]. For example, it updates the focus concept as well as any tag cloud viewer containing the selected term as its focus concept.

- **Sticky Tag Cloud Viewers:** In addition to the main tag cloud window, sub-windows can be generated from *Sticky Tags* and appear below the main window. A sticky tag is an object or attribute to which the viewer is fixed, and each Sticky Tag Cloud Viewer contains the displayed tags for the sticky concept of that window [10]. These sub-windows allow the user to have multiple views/tag clouds of their data. Selecting a new focus concept will adjust these windows to use the union of the sticky tag and select the focus concept as their unique focus concept.

- **Object Table:** The table at the bottom of the interface represents each formal object as a row, with the columns being the attribute/tag categories. The table is linked to the tag cloud viewer and is updated whenever a tag is selected or deselected to reflect the multi-valued context table corresponding to the current focus concept [10].

2.6 Chapter Summary

Chapter 2 introduced the theory behind Formal Concept Analysis and provided a discussion on the Formal Concept Lattice: a method used to represent a formal context by showing the visual relationship between the sub-concepts and super-concepts. Secondly, the importance of visualising data with the use of a user-centric approach was discussed as well as the benefits and drawbacks of data visualisation. Moreover, a brief discussion on the future of visualisations was provided in which Chen et al. 2008 proposed the need to use knowledge-assisted visualisation to handle the vast amount of collected information. Thirdly, a discussion of four data visualisation tools was provided namely: ToscanaJ, FCART, Fish-eye software, and ConceptCloud. Of these tools, none provide specialised visualisations for specific types of data such as numeric or geolocation data. ToscanaJ and FCART both attempt to visualise the entire lattice, resulting in poor performance and cluttered visualisations for larger and more diverse the data set. ConceptCloud, while also not having specialised visualisations for different data types, is the only tool that does not attempt to visualise the entire lattice, but instead provides the user with more intuitive tag cloud. ConceptCloud only generates the section of the lattice currently required to visualise the tag cloud, making it the top choice for visualising large and diverse data sets. Thus, focus was placed on ConceptCloud as this data visualisation tool will be extended and used to evaluate data sets in this thesis. To understand the theory behind ConceptCloud, background information on the architecture and user interface was provided.

Chapter 3

Visualisations

3.1 Overview

This chapter describes the various extensions made to the initial ConceptCloud architecture. The extensions provide the user with additional and improved visualisations of the data, thus enabling more efficient data exploration and knowledge discovery. The first extension made to ConceptCloud is a fully integrated map-based viewer window that allows for the visualisation and exploration of semi-structured data with some geolocation aspects. The second extension, an integrated graph viewer, is then implemented on the ConceptCloud front-end to provide the user with a view of the composition of their data set as well as serving any additional visualisations of their data. The third extension focuses on the changes made to ConceptCloud's server architecture to implement the map and graph viewers. To conclude, an introduction to the development of a mobile application that mimics the browser user interface and tag visualisation is provided.

The work discussed in the Map Visualisation section of this chapter is a re-working of the work covered in a previous research paper: *ConceptCloud 2.0: Visualisation and Exploration of Geolocation-Rich Semi-Structured Data Sets* du Toit et al., 2019 co-authored with Joshua Berndt who implemented the required server architecture modifications to support map visualisation.

3.2 Extending ConceptCloud

ConceptCloud is a modern web application with a browser-based user interface that allows for an extension of its functionality and integration with other web applications as well as third-party JavaScript libraries. ConceptCloud's user interface consists of two main components, namely the *Tag Cloud* and *Object Table*. As described in Chapter 2, both the *Tag Cloud* and *Object Table* textually visualises the formal objects and attributes, with the *Tag Cloud* driving the navigation through the lattice. When considering extending the functionality of ConceptCloud in terms of new data visualisations, both navigation and maintaining a consistent view of the user's position in the concept lattice are important factors in maintaining an effective user interface and providing cognitive support during data analysis.

Consequently, the extension must adhere to a few requirements to fully integrate new visualisations with ConceptCloud's user interface. These requirements are as follows:

- The extension should be clear and readable for a user with moderate levels of data literacy.
- The visualisations must remain an accurate representation of the data where opportunity for misinterpretation of the data by the user is minimised.
- The new extension must aid in data analysis by providing cognitive support and reducing the cognitive load experienced by the user as far as possible.
- The extension must aid in data exploration and knowledge discovery by providing different visualisations of the same data.
- The extension must allow the new visualisation as well as the *Tag Cloud* to drive exploration. Thus a pivotal function of this tool is *to ensure the data exploration works in a bidirectional manner*. In other words, to update the new visualisation when changes are made in one of the visualisations and vice versa.
- The extension must ensure that *the visualisation and exploration tool be highly scalable to process large data sets*, since the rate at which such data is being generated is ever increasing.

In order to design, develop and implement these extensions, an iterative design science research methodology was adopted. A user-centric approach was followed to ensure the extensions adhered to the requirements set out above.

Three key visualisation extensions were developed, namely a map viewer, graph viewer and mobile application. This required substantial changes be made to the existing ConceptCloud architecture, mentioned in Section 2.5.1, which resulted in the final ConceptCloud architecture shown in Figure 3.1, with the additions marked in green.

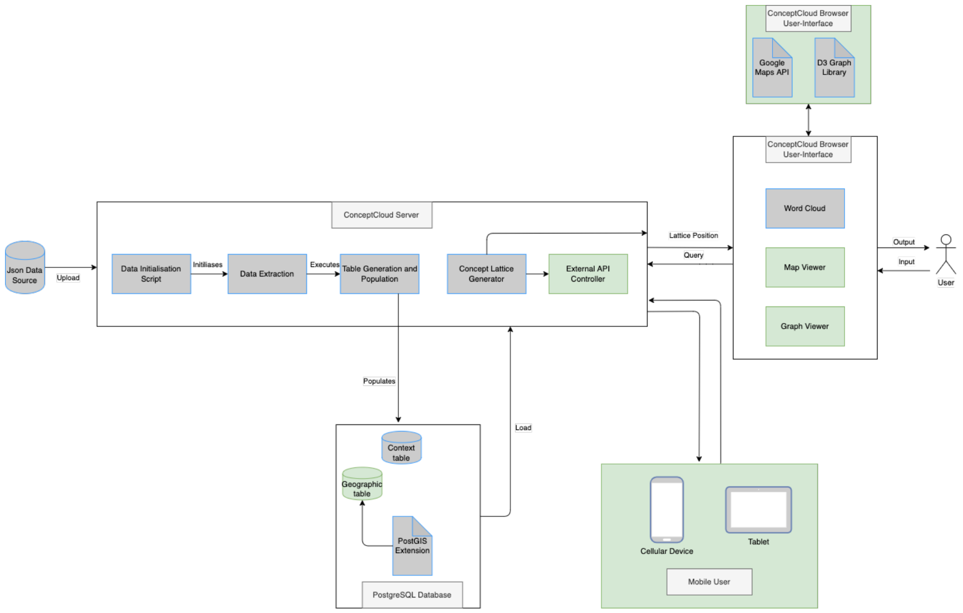


Figure 3.1: ConceptCloud Mobile Architecture Extension

The architecture now includes the following:

3.2.1 Final ConceptCloud Architecture

- **PostGIS Extension:** A spatial database extender for the PostgreSQL object-relational database which allows for optimised geographic queries.
- **Geographic database column:** An additional column in the table in the database for the transformed geolocation aspect of the data, stored as a PostGIS Geography Type in order to make use of more efficient look up functions.
- **Map Viewer:** An additional data visualisation window that makes use of the Google Maps JavaScript API to render a fully interactive map and pin objects.
- **Graph Viewer:** An additional data visualisation window that makes use of the D3 API to render customisable and refineable graphs.
- **External API Controller:** The API controller forms part of ConcepCloud's server and is accessible by external applications and clients who may call any of the methods described in Section 3.5.4.1 for data or lattice computation.
- **Mobile Applications:** ConceptCloud Mobile exists as a separate, Operating System agnostic mobile application that replicates the functionality found in the browser version of ConceptCloud.

3.3 Map Visualisation

The widespread use of location-tracking and GPS-capable devices such as smartphones and tablets has resulted in the generation of vast amounts of geolocation-rich data. This abundance presents a new opportunity for knowledge discovery by exploring the geolocation aspect of data, but it also demands a specific visualisation approach [44].

When considering an optimal way to visualise this type of data in ConceptCloud, maps are an immediately obvious choice since they have been used historically and are superior to textually displaying latitude and longitudinal coordinates in a tag cloud. Furthermore, tag clouds provide an effective method for facilitating knowledge discovery and data visualisation for semi-structured data. Thus, the overall goal of the extension is to merge these two successful methods into an integrated and scalable system.

3.3.1 Implementation

As a starting point, ConceptCloud's tag-cloud viewer makes up the main section of the user interface. The viewer window has several controls along the top border that allow the user to manipulate and adjust the contents of the window in various ways. For example, the user can resize the window, zoom in and out, and filter the tags. To maintain a consistent visual aesthetic and user experience, the map viewer is implemented in this same type of viewer window.

An effective map viewer was implemented by considering various third-party libraries based on functionality, customisability, and ease of use. The Google Maps JavaScript API[1] was chosen as it has extensive functionality and documentation available, a proven track record, and user familiarity. In the newly added viewer, the Google Maps API is used to render a fully interactive map. Furthermore, the geolocation attribute of the data objects is used to generate map pins, where available.

The initial location of the map and zoom level is customisable and the user is able to navigate to different areas of the map by clicking and dragging their cursor. The user is also able to change the zoom level of the map by either clicking on the icons on the right hand side of the map or by scrolling their mouse.

The user is also able to switch between a “Map” and “Satellite” view, visible in Figure 3.2, by clicking the labelled buttons in the top left corner of the map viewer. The “Map” option provides the user with an illustrated map view, which is the default view, indicating place names, roads, places of interest and major landmarks, with the ability to toggle the “Terrain” setting which will display the topography and terrain of the area being viewed. If the

[1] https://developers.google.com/maps/documentation

user selects the "Satellite" view, the map will be rendered as satellite images with the option to toggle labels on or off.

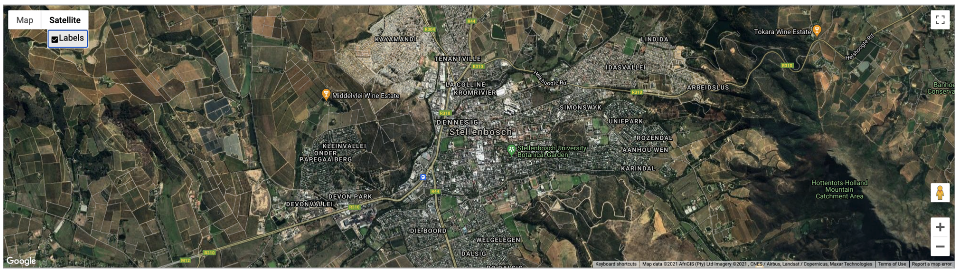
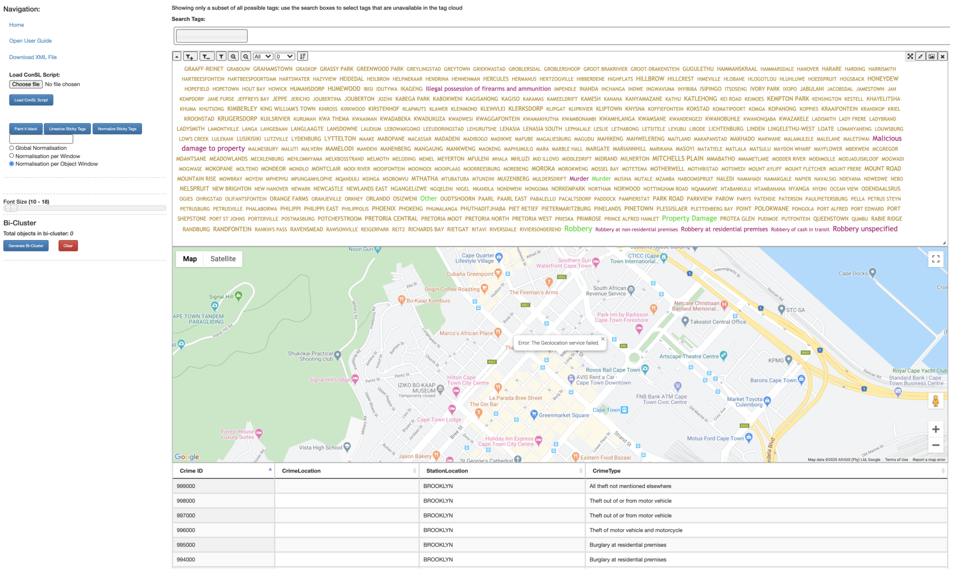


Figure 3.2: Satellite view of the map

Figure 3.3 shows the ConceptCloud user interface with the new map viewer applied to crime data. A specific crime data set contains, for each crime, its category (e.g., *Theft*) or sub-category (e.g., *Stock theft*) and the name and location of the police station where the crime was reported. Map pins are drawn on the map at a fixed size. At certain zoom levels, the pins begin to overlap one another which introduces ambiguity and compromises the visualisation of the data. The map viewer clusters the pins automatically, based on the current zoom level of the map, to prevent them from obstructing one another. It also displays the number of pins in each cluster. The further the user zooms in on the map, the fewer pins will be contained in each cluster until individual pins become visible.


Navigation:
Home
Open User Guide
Download XML File
Choose file
Load ConSL Script
Paint it black
Linearize Sticky Tags
Normalize Sticky Tags
Generate Bi-Cluster
Clear
All
0
GRAAFF-REINET GRABOUW GRAHAMSTOWN GRASKOP GRASSY PARK GREENWOOD PARK GREYLINGSTAD GREYTOWN GRIEKWASTAD GROBLERSDAL GROBLERSHOOP GROOT BRAKRIVIER GROOT-DRAKENSTEIN GUGULETHU HAMMANSKRAAL HAMMARSDALE HANOVER HARARE HARDING HARRISMITH
HARTBEESFONTEIN HARTBEESPOORTDAM HARTSWATER HAZYVIEW HEIDEDAL HEILBRON HELPMEKAAR HENDRINA HENNENMAN HERCULES HERMANUS HERTZOGVILLE HIBBERDENE HIGHFLATS HILLBROW HILLCREST HIMEVILLE HLOBANE HLOGOTLOU HLUHLUWE HOEDSPRUIT HOGSBACK HONEYDEW
HOPEFIELD HOPETOWN HOUT BAY HOWICK HUMANSDORP HUMEWOOD IBISI IDUTYWA IKAGENG Illegal possession of firearms and ammunition IMPENDLE INANDA INCHANGA INDWE INGWAVUMA INYIBIBA ISIPINGO ITSOSENG IVORY PARK IXOPO JABULANI JACOBSDAL JAMESTOWN JAN
KEMPDORP JANE FURSE JEFFREYS BAY JEPPE JERICHO JOUBERTINA JOUBERTON JOZINI KABEGA PARK KABOKWENI KAGISANONG KAGISO KAKAMAS KAMEELDRIFT KAMESH KANANA KANYAMAZANE KATHU KATLEHONG KEI ROAD KEIMOES KEMPTON PARK KENSINGTON KESTELL KHAYELITSHA
KHUMA KHUTSONG KIMBERLEY KING WILLIAM'S TOWN KINROSS KIRKWOOD KIRSTENHOF KLAPMUTS KLAWER KLEINMOND KLEINVLEI KLERKSDORP KLIPGAT KLIPRIVIER KLIPTOWN KNYSNA KOFFIEFONTEIN KOKSTAD KOMATIPOORT KOMGA KOPANONG KOPPIES KRAAIFONTEIN KRANSKOP KRIEL
KROONSTAD KRUGERSDORP KUILSRIVIER KURUMAN KWA THEMA KWAAIMAN KWADABEKA KWADUKUZA KWADWESI KWAGGAFONTEIN KWAMAKHUTHA KWAMBONAMBI KWAMHLANGA KWAMSANE KWANDENGEZI KWANOBUHLE KWANONQABA KWAZAKELE LADISMITH LADY FRERE LADYBRAND
LADYSMITH LAMONTVILLE LANGA LANGEBAAN LANGLAAGTE LANSDOWNE LAUDIUM LEBOWAKGOMO LEEUDORINGSTAD LEHURUTSHE LENASIA LENASIA SOUTH LEPHALALE LESLIE LETHABONG LETSITELE LEVUBU LIBODE LICHTENBURG LINDEN LINGELETHU-WEST LOATE LOMANYANENG LOUWSBURG
LOW'S CREEK LULEKANI LUSIKISIKI LUTZVILLE LYDENBURG LYTTELTON MAAKE MABOPANE MACASSAR MADADENI MADIBOGO MADIKWE MAFUBE MAGALIESBURG MAGUDU MAHIKENG MAHWELERENG MAITLAND MAKAPANSTAD MAKHADO MAKWANE MALAMULELE MALELANE MALETSWAI Malicious
damage to property MALMESBURY MALUTI MALVERN MAMELODI MANDENI MANENBERG MANGAUNG MANKWENG MAOKENG MAPHUMULO MARA MARBLE HALL MARGATE MARIANNHILL MARIKANA MASOYI MATATIELE MATLALA MATSULU MAYDON WHARF MAYFLOWER MBEKWENI MCGREGOR
MDANTSANE MEADOWLANDS MECKLENBURG MEHLOMNYAMA MELKBOSSTRAND MELMOTH MELODING MEMEL MEYERTON MFULENI MHALA MHLUZI MID ILLOVO MIDDLEDRIFT MIDRAND MILNERTON MITCHELLS PLAIN MMABATHO MMAMETLAKE MODDER RIVER MODIMOLLE MODJADJISKLOOF MOGWADI
MOGWASE MOKOPANE MOLTENO MONDEOR MONDLO MONTCLAIR MOOI RIVER MOOIFONTEIN MOOINOOI MOOIPLAAS MOORREESBURG MOREBENG MOROKA MOROKWENG MOSSEL BAY MOTETEMA MOTHERWELL MOTHIBISTAD MOTSWEDI MOUNT AYLIFF MOUNT FLETCHER MOUNT FRERE MOUNT ROAD
MOUNTAIN RISE MOWBRAY MOYENI MPHEPHU MPUNGAMHLOPHE MQANDULI MSINGA MSOBOMVU MTHATHA MTUBATUBA MTUNZINI MUIZENBERG MULDERSDRIFT Murder Murder MUSINA MUTALE MZAMBA NABOOMSPRUIT NALEDI NAMAHADI NAMAKGALE NAPIER NAVALSIG NDEVANA NDWEDWE NEBO
NELSPRUIT NEW BRIGHTON NEW HANOVER NEWARK NEWCASTLE NEWLANDS EAST NGANGELIZWE NGQELENI NIGEL NKANDLA NONDWENI NONGOMA NORKEMPARK NORTHAM NORWOOD NOTTINGHAM ROAD NQAMAKWE NTABANKULU NTAMBANANA NYANGA NYONI OCEAN VIEW ODENDAALSRUS
OGIES OHRIGSTAD OLIFANTSFONTEIN ORANGE FARMS ORANJEVILLE ORKNEY ORLANDO OSIZWENI Other OUDTSHOORN PAARL PAARL EAST PABALELLO PACALTSDORP PADDOCK PAMPIERSTAT PARK ROAD PARKVIEW PAROW PARYS PATENSIE PATERSON PAULPIETERSBURG PELLA PETRUS STEYN
PETRUSBURG PETRUSVILLE PHALABORWA PHILIPPI PHILIPPI EAST PHILIPPOLIS PHOENIX PHOKENG PHUMALANGA PHUTHADITJHABA PIET RETIEF PIETERMARITZBURG PINELANDS PINETOWN PLESSISLAER PLETTENBERG BAY POINT POLOKWANE PONGOLA PORT ALFRED PORT EDWARD PORT
SHEPSTONE PORT ST JOHNS PORTERVILLE POSTMASBURG POTCHEFSTROOM PRETORIA CENTRAL PRETORIA MOOT PRETORIA NORTH PRETORIA WEST PRIESKA PRIMROSE PRINCE ALFRED HAMLET Property Damage PROTEA GLEN PUDIMOE PUTFONTEIN QUEENSTOWN QUMBU RABIE RIDGE
RANDBURG RANDFONTEIN RANKIN'S PASS RAVENSMEAD RAWSONVILLE REIGERPARK REITZ RICHARDS BAY RIETGAT RITAVI RIVERSDALE RIVIERSONDEREND Robbery Robbery at non-residential premises Robbery at residential premises Robbery of cash in transit Robbery unspecified
Map
Satellite
Noon Gun
Cape Quarter Lifestyle Village
Southern Sun Waterfront Cape Town
CTICC (Cape Town International...
Cubaña Greenpoint
Origin Coffee Roasting
The Fireman's Arms
South African Revenue Service
Cape Docks
Signal Hill
CAPE TOWN TANDEM PARAGLIDING
Bo-Kaap Kombuis
Netcare Christiaan Barnard Memorial...
Park Inn by Radisson Cape Town Foreshore
Takealot Central Office
STC-SA
Marco's African Place
Shukokai Practical Shooting club
Artscape Theatre Centre
KPMG
Rovos Rail Cape Town
AVIS Rent a Car - Cape Town Downtown
Signal Hill Lodge
Hilton Cape Town City Centre
IZIKO BO-KAAP MUSEUM
Temporarily closed
La Parada Bree Street
The Gin Bar
Greenmarket Square
Cape Town
FNB Bank ATM Cape Town Civic Centre
Barons Cape Town
Market Toyota Culemborg
Royal Cape Yacht Club
Standard Bank | Cape Town Business Centre
Cape Town Lodge
Forest House Luxury Suites
Vista High School
Holiday Inn Express Cape Town City Centre
Motus Ford Cape Town
Jason Bakery
St George's Cathedral
Eastern Food Bazaar
Google
Map data ©2020 AfriGIS (Pty) Ltd, Google
Terms of Use
Report a map error

Load ConSL Script: No file chosen

○ Global Normalisation
○ Normalisation per Window
● Normalisation per Object Window

Font Size (10 - 18)

Bi-Cluster

Total objects in bi-cluster: *0*

Showing only a subset of all possible tags: use the search boxes to select tags that are unavailable in the tag cloud

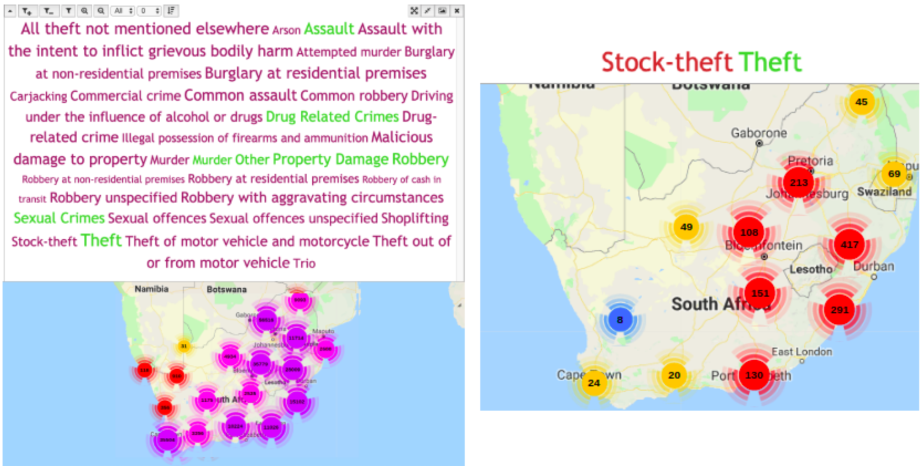
Search Tags:

Error: The Geolocation service failed.

Crime ID	CrimeLocation	StationLocation	CrimeType
999000		BROOKLYN	All theft not mentioned elsewhere
998000		BROOKLYN	Theft out of or from motor vehicle
997000		BROOKLYN	Theft out of or from motor vehicle
996000		BROOKLYN	Theft of motor vehicle and motorcycle
995000		BROOKLYN	Burglary at residential premises
994000		BROOKLYN	Burglary at residential premises

Figure 3.3: Interface with Map Viewer applied to Crime Data

Then, the pins function the same way as the tags in the tag cloud. The user can thus explore the data by selecting an attribute in the tag cloud, which will update the cloud as before [29]. Furthermore, the selection will also update the map view to display only the objects (resp. their pins) with that selected (focus) attribute. Figure 3.4 shows the interface before (left) and after (right) the user has selected the *Stock theft* sub-category from the tag cloud. The tag cloud now contains only the selected *Stock theft* tag with the map viewer containing only the (clustered) pins corresponding to the *Stock theft* objects. Furthermore, the *Theft* crime category tag is still visible since *Stock theft* belongs to this category. Alternatively, the user can select an individual pin that will "drill down" the map view, i.e., decrease the map's scale, redraw and re-cluster the pins, and update the tag cloud to reflect only the objects still represented on the map view.



All theft not mentioned elsewhere Arson Assault Assault with the intent to inflict grievous bodily harm Attempted murder Burglary at non-residential premises Burglary at residential premises Carjacking Commercial crime Common assault Common robbery Driving under the influence of alcohol or drugs Drug Related Crimes Drug-related crime Illegal possession of firearms and ammunition Malicious damage to property Murder Murder Other Property Damage Robbery Robbery at non-residential premises Robbery at residential premises Robbery of cash in transit Robbery unspecified Robbery with aggravating circumstances Sexual Crimes Sexual offences Sexual offences unspecified Shoplifting Stock-theft Theft Theft of motor vehicle and motorcycle Theft out of or from motor vehicle Trio

Stock-theft Theft

Figure 3.4: Left: Interface applied to Crime Data; Right: Data filtered for Stock-theft

The geolocation attribute values of the data objects are used to display the pin objects in the map viewer, although they also appear in the tag cloud as latitude and longitude pairs. When considering the underlying context table, each lat-long pair is an attribute, with each object having at most one geolocation attribute.

A single specialised server call populates the map and marker objects in the browser-based client. The size of the visible map in the Google Map viewer is dependent on two factors, namely the size of the viewer window and the zoom level. Based on the dimensions of this viewer, the visible radius is calculated and used along with the zoom level and centre coordinates of the map viewer to only retrieve the map pins that are visible to the user. Map viewer movement is tracked, and extra server calls are made when the map is moved to a new location.

ConceptCloud makes use of a PostgreSQL database which allows for the addition of a powerful spatial database extension named PostGIS [2]. It adds support for geographic objects allowing SQL to run location queries. Utilising the optimisation that PostGIS offers, ConceptCloud first indexes the data before clustering it by coordinate. This process speeds up access to the data by ensuring that records that are likely to be retrieved together in the same result set, such as objects located in the same geographical area, are located in similar physical locations on the hard disk. This allows for much faster, optimised lookup queries, which becomes a necessity for larger data sets. The "ST_D within" method, specifically, is utilised to query the geographic data required. This method takes three parameters, namely a latitude, a longitude, and a radius/distance measured in SRID (A spatial reference identifier) to query the data.

3.3.2 Biclustering

The pin clusters mentioned above also have a formal correspondence in the concept lattice in the form of *biclusters*. A bicluster is defined as a pair (A, B) of inclusion-maximal sets of objects and attributes, where each object, A, contains almost all of its attributes, B. This technique has been implemented and applied successfully to mine numeric data sets using triadic formal concept analysis [43].

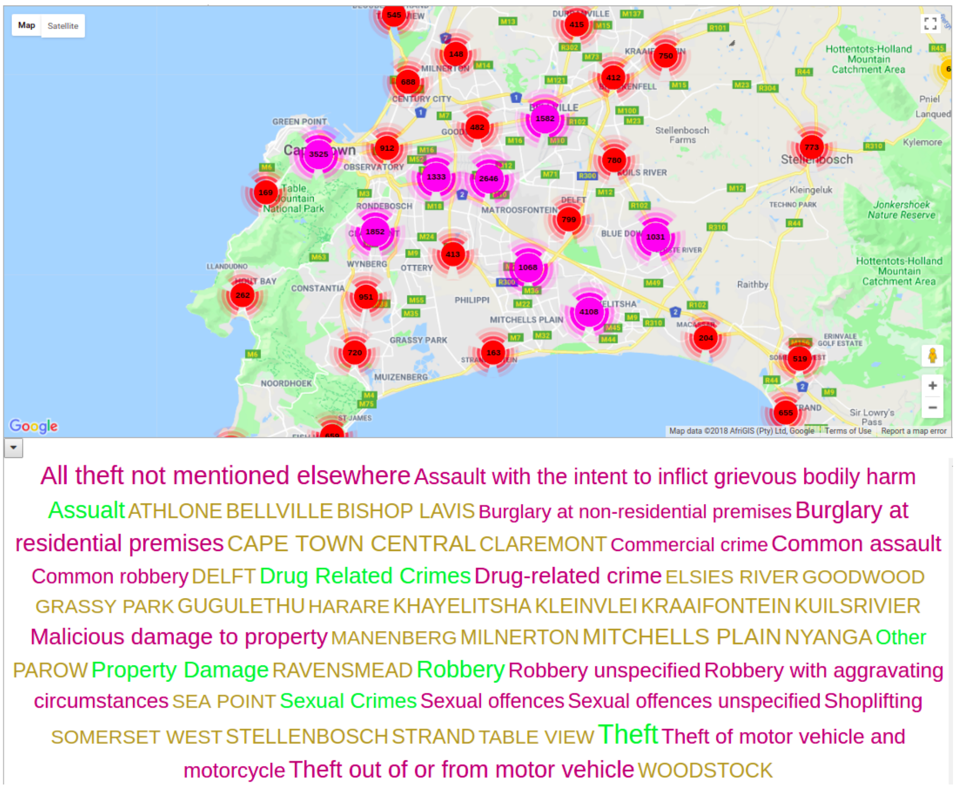
When pins representing the formal objects are grouped into clusters on the map view, a bicluster emerges as an element of the concept lattice by forcing the inclusion of all the geolocation attributes from the object in a pin cluster to all the attributes in that pin cluster.

When considering the map viewer in Figure 3.5, multiple pin clusters can be seen around the central Cape Town area. The objects represented by the pins in the cluster all share geolocation attributes which are similiar enough that they were grouped into a single cluster or *bicluster*. These naturally occurring biclusters can be "mined" to view common trends where possible.

Figure 3.5 shows the new tag cloud generated from biclusters from the central Cape Town area below the map viewer. The user can generate the corresponding bicluster tag cloud by right-clicking either single or multiple pin clusters as well as individual pins.

This bicluster tag cloud corresponds to a smaller lattice of selected objects, created from the original lattice, with its own focus concept, namely all of the objects that share the new "Selected" attribute assigned during the process mentioned above (The objects in the lattice technically also share the same relative geolocation aspect but this is no longer relevant once the objects have been selected and the new tag cloud is generated). This new lattice can then be navigated independently of the main lattice.

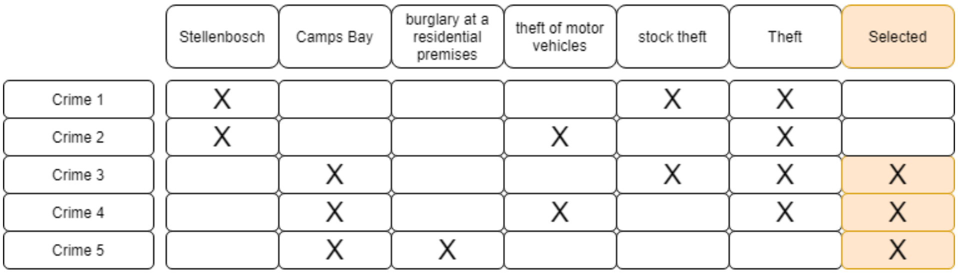
[2] https://postgis.net



All theft not mentioned elsewhere Assault with the intent to inflict grievous bodily harm Assault ATHLONE BELLVILLE BISHOP LAVIS Burglary at non-residential premises Burglary at residential premises CAPE TOWN CENTRAL CLAREMONT Commercial crime Common assault Common robbery DELFT Drug Related Crimes Drug-related crime ELSIES RIVER GOODWOOD GRASSY PARK GUGULETHU HARARE KHAYELITSHA KLEINVLEI KRAAIFONTEIN KUILSRIVIER Malicious damage to property MANENBERG MILNERTON MITCHELLS PLAIN NYANGA Other PAROW Property Damage RAVENSMEAD Robbery Robbery unspecified Robbery with aggravating circumstances SEA POINT Sexual Crimes Sexual offences Sexual offences unspecified Shoplifting SOMERSET WEST STELLENBOSCH STRAND TABLE VIEW Theft Theft of motor vehicle and motorcycle Theft out of or from motor vehicle WOODSTOCK

Figure 3.5: Central Cape Town Bicluster

When dealing with biclusters, a Boolean disjunctive selection is used to maintain a single focus information retrieval navigation algorithm [30]. This technique involves modifying the underlying context table of the lattice, as can be seen in Figure 3.6, and generating a new temporary concept lattice of only the selected objects from the data set and displaying them in a new tag cloud. The tags in the new cloud are either from objects in the grouped bicluster, or object pins selected by the user, or multiple pin biclusters selected by the user.



	Stellenbosch	Camps Bay	burglary at a residential premises	theft of motor vehicles	stock theft	Theft	Selected
Crime 1	X				X	X	
Crime 2	X			X		X	
Crime 3		X			X	X	X
Crime 4		X		X		X	X
Crime 5		X	X				X

Figure 3.6: Context Table with additional attribute

Computing the disjunction of two or more objects involves assigning a new meta-tag to the selected objects, and in doing so, instantly generating a new temporary lattice. This new lattice consisting of the merged objects becomes the subject of the new tag cloud window. As a result, the user is free to explore the desired objects without introducing concept broadening [30].

3.3.3 Geo-location Rich Data Exploration

A geolocation-rich data set is required to make full use of the newly added viewer and its functionality. ConceptCloud allows the user to pre-configure the data attributes they wish to appear in each map pin object, allowing for a far smaller, more responsive server call to create each pin. The server returns a set of objects containing at least an identifier and the geolocation. As a result, a Google Maps marker object is created and populated with the pre-configured attributes of that object. If desired and available, the user can populate a tool-tip text window of the marker with additional meta-data and links to external resources involving the object. In addition, the user needs to specify the attribute types present in the object. More specifically, the geolocation attribute must be of the format “lat, long”.

3.3.3.1 Map Driven Knowledge Discovery

A prevalent issue found in geolocation data sets is the limiting effect of pre-imposed borders in the data. These borders are typically political or geographical and the geolocation data collected is automatically sorted and grouped based on these boundaries, when in fact, the data seldom adheres to these boundaries. Crime data is a perfect example of this, where crimes are not restricted by political borders but perhaps by natural or socio-economic borders instead. Generally, any quantitative or statistical data collected is confined automatically to predefined borders. Any query on the data involving unique intersections of these borders becomes very complex or resource-intensive. However, the improvements made to the ConceptCloud architecture provide a solution to this problem by allowing the user to freely select naturally occurring biclusters to explore.

3.4 Graph Visualisation

3.4.1 Introduction

The geolocation aspect of data is not the only data that can be displayed more effectively using different visualisations. Data sets containing numeric fields are notoriously difficult for users to process when shown as raw values in a spreadsheet or table. Furthermore, they can induce a high cognitive load [35]. Graphs and charts have traditionally been used as an effective way to present numeric data in a manner that provides a more holistic view of the figures. As a result, the graphs and charts make it easier to identify trends and patterns in numeric data. Graphs play a pivotal role in data analysis. Furthermore, they work best as supplementary visualisations to the textual data [35].

After some experimentation with numeric data, it became clear that ConceptCloud's tag cloud is not well suited for this type of data. A reason for this is that numeric data, in general, does not work well as a tag since it is difficult for the user to discern which tags relate to which categories, especially when the data object contains multiple numeric fields. This could lead to unnecessary increases in cognitive load and can weaken the sensemaking process. ConceptCloud generates the tag cloud where the size of each tag displayed is based on the frequency of that tag within the data set. The problem with numeric data is that the odds of a number appearing multiple times in the data set is much lower than with textual fields. This problem can be solved by binning the numeric data into widely used and accepted categories or scales where possible. Generic categories ("low", "medium" and "high", for example) can also be used if no specific scale exists. This process allows for better tag cloud visualisation, albeit with less data accuracy than the original numeric values.

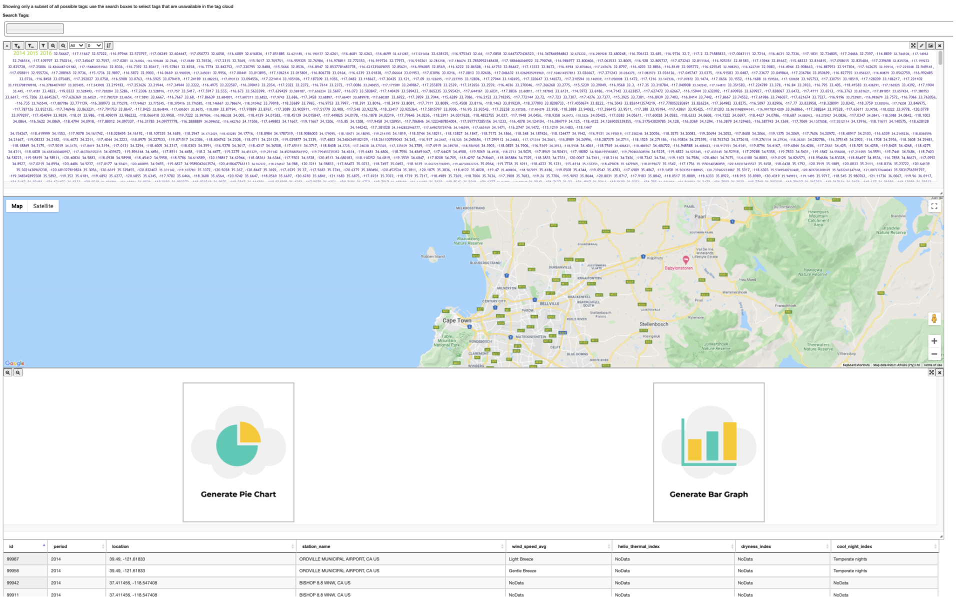
Instead of using graphs to visualise numeric data, once the data has been binned, a graph visualisation can be used to provide the user with insight into the composition of tag categories in the data, as well as the composition of tag categories in relation to specific tags. Implementing a graph viewer in ConceptCloud requires adherence to the same requirements applied to the map viewer stated in Section 3.2, in order to achieve an integrated and scalable system.

3.4.2 Abstraction of the Map Viewer

After implementing the map visualisation extension to ConceptCloud, additional visualisations were explored. One such visualisation is possible from abstracting the map viewer, essentially a 2D Cartesian plane, to support other metric space visualisations. In the map implementation, the count of the location attribute of the objects was used to roll up the objects i.e pin clusters. Instead of rolling up by a single attribute, it is possible to roll up by a set of attributes, or concept (a concept being the smallest object containing the object set). This results in new groups based on the collection of attributes. Further, this allows for the creation of graphs that have the ability to be "rolled up" or "drilled-down" to change the perspective and level at which the data is being viewed. This requires advanced graph visualisation tools which must be able to integrate with the existing ConceptCloud infrastructure and user interface.

3.4.3 Implementation

ConceptCloud's browser-based user interface allows for advanced graph visualisation through third-party Javascript libraries and the HTML5 canvas. Generally, these libraries can draw any type of graph and chart, trading flexibility and functionality for ease of implementation. To implement a graph viewer successfully, a library that allows for the necessary functionality and flexibility to draw and manipulate the graphs to the desired granular level while having the library handle the majority of the rendering and styling, is required.



Showing only a subset of all possible tags: use the search boxes to select tags that are unavailable in the tag cloud

Search Tags:

id	period	location	station_name	wind_speed_avg	helio_thermal_index	dryness_index	cool_night_index
99987	2014	39.49, -121.61833	OROVILLE MUNICIPAL AIRPORT, CA US	Light Breeze	NoData	NoData	Temperate nights
99956	2014	39.49, -121.61833	OROVILLE MUNICIPAL AIRPORT, CA US	Gentle Breeze	NoData	NoData	Temperate nights
99942	2014	37.411456, -118.547408	BISHOP 8.8 WNW, CA US	NoData	NoData	NoData	NoData
99911	2014	37.411456, -118.547408	BISHOP 8.8 WNW, CA US	NoData	NoData	NoData	NoData

Figure 3.7: Full User Interface With Map and Graph Viewer

The D3 library[3] was selected for this task since it is popular and satisfies the above criteria. The D3 library allows for low-level manipulation of the graph objects which is of key importance when attempting to implement graph refinement operations. For example, having to draw graphs with other graphs or combining the various graphs into a new graph. Figure 3.7 shows the full ConceptCloud user interface with both the map and graph viewers.

[3] https://d3js.org

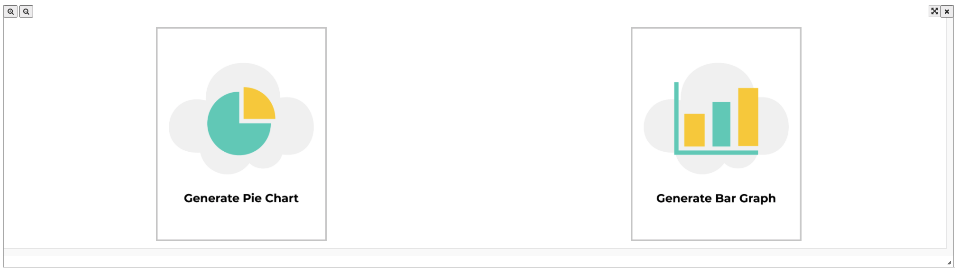


Figure 3.8: Initial Graph Viewer Window

The graph visualisation is triggered by the user from the front-end and as with the map viewer, can be done from either the tag cloud or new graph viewer window. The new viewer window is situated below the map viewer window on the user interface as can be seen in Figure 3.8. In this window, the user is presented with two options, the first for generating a pie chart and the second for generating a bar graph. Once one of the tiles is clicked, the user can select a tag category, or formal attribute, from the drop-down menu in the bottom left corner of the window, shown in Figure 3.9 to generate the graph.

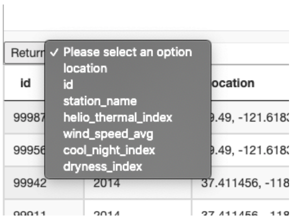


Figure 3.9: Graph Drop Down Menu

Once the category is selected, the graph is rendered in the graph viewer with the various tag categories broken down by count (occurrence in the data set). For a bar graph, the count is displayed on the y-axis in a linear scale and the attributes on the x-axis. The user may hover their cursor over each pie segment or bar to view the attribute and count of that segment via a tooltip popup window. Figure 3.10 shows the initial pie chart generated for a climate data set and Figure 3.11 shows the bar graph visualisation of this same data.

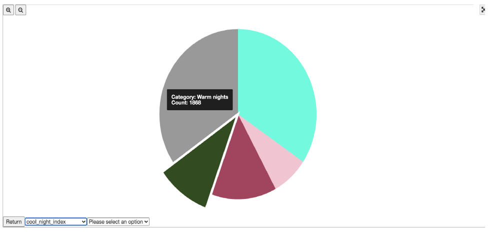


Figure 3.10: Initial Pie Chart

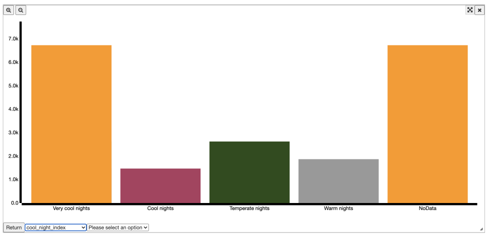


Figure 3.11: Initial Bar Graph

After rendering the graph, the user is given the option to refine the graph by selecting an additional tag category. This is achieved by selecting a specific attribute on the graph, either a pie segment or bar, and then selecting a second tag category from the corresponding drop-down menu. In the case of the pie chart, the entire graph is re-drawn based on the second tag category. This is shown in Figure 3.12.

However, when a second category is selected for the bar graph, a new stacked bar graph is generated. The stacked bar graph represents both the first tag category, in the full bar, and the second category attributes in the segments within each bar. This is shown in Figure 3.13.

The user can then switch between any first or second tag categories by changing their selection in either of the drop-down menus. When a pie segment or bar is clicked in the graph viewer, the focus concept is updated with this tag and both the tag cloud and map viewer is updated to maintain a consistent view of the lattice. At the bottom of the graph viewer there is a “Back” button which sets the viewer window back to the graph selection screen.

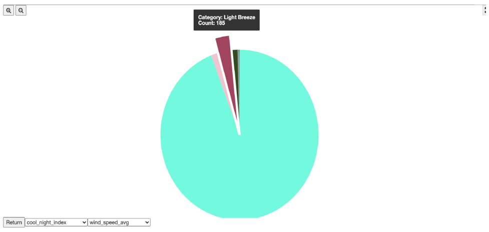


Figure 3.12: Drilled down pie chart

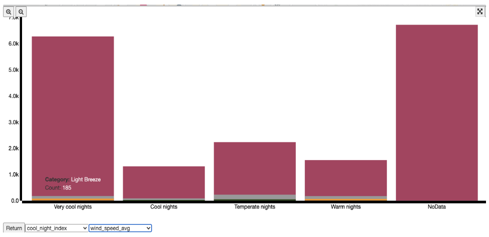


Figure 3.13: Drilled down bar graph

By having the drop down menu fixed at the bottom of the viewer window, the user is able to switch between different tag categories instantly on both the initial level as well as “drilled down” level in order to get a deeper understanding of the composition of the data set, both in its entirety as well as the different tag categories in terms of specific tags. Beyond that, the graph viewer provides the user with an additional visualisation of the same data visible in the tag cloud and map, which has been shown to improve data analysis and sense-making [31].

Alternatively, to generate a graph from the tag cloud the user may right-click any tag in the cloud and then select the desired graph from the specialised context menu seen in Figure 3.14. Based on the user's selection, the chosen graph will be rendered in the graph viewer window with the selected tag's tag category, similar to if it had been selected from the drop-down menu above. From here, the graph can be "drilled down" and explored in the same way as described above with the selected tag added to the focus concept.

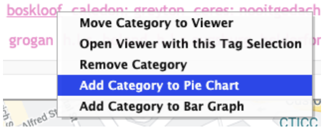


Figure 3.14: Generating graph from Tag Cloud

Both of the initial graphs are populated by the same server call, which expects a tag category and returns a count for each unique tag in that category. Since the grouping is done by the server before the data is returned to the browser, only a small amount of data is returned ensuring the call remains responsive as the size of the data set increases.

The server call that populates the "drilled down" graphs differ depending on whether the user selected a bar graph or pie chart. The refined pie chart's server call expects a primary and secondary tag category and a specific tag from the primary tag category, which it uses to get the counts for each of the tags in the secondary category grouped by tag. The D3 library then redraws the pie chart using these tags names and counts. The refined bar graph's server call is similar, but unlike the pie chart, it does not require a specific tag from the primary tag category. Instead it returns all secondary tag category tags in terms of the different values in the primary tag category.

In cases where the number of unique tag values in any tag category surpasses 15, the server call limits the returned tags to the 15 most frequently occurring tags in the data set and groups the remaining tags into a temporary "Other" category. This "Other" category is hidden by default but can be displayed if required.

Lastly, ways of generating a graph from the map viewer was investigated but proved difficult due to the already existing map functionality. As described in Section 3.3.1, left clicking a pin on the map is already bound to updating the focus concept of the lattice and the right click overwrites the existing Google Maps context menu in order to add the pin to the bicluster list. Therefore there is no intuitive way of adding the ability to generate a graph from the map without significantly changing how the map viewer functions.

3.5 Mobile Visualisation

3.5.1 Introduction

Mobile devices, more specifically smartphones and tablets, are becoming ever more powerful and capable of similar functionality as personal computers. These devices usually consist of a touch screen display which can be interacted with using a standardised set of gestures. These devices generally also include a range of other components such as GPS, accelerometer, camera and microphone, among others. This combination of components make these devices uniquely suited to certain applications but due to the limited screen size and a touch screen interface, these devices to display information to the user optimally by making use of symbolic icons, unique data visualisations and streamlined interaction and navigation. Touch screen user interfaces also allow for unique visualisations and user interactions. Gestures such as pinching, zooming, single and double tapping as well as swiping allow for an additional level of interaction not possible through traditional desktop interfaces.

3.5.2 Native and Hybrid Mobile Applications

Mobile devices (smartphones and tablets specifically) are becoming ever more powerful and capable of similar functionality as personal computers. These devices usually consist of a touch screen display that can be interacted with using a standardised set of gestures. These devices generally also include a range of other components such as GPS, accelerometer, camera, and a microphone. These components make these devices uniquely suited to specific applications. Due to the limited screen size and touch screen interface, these devices display information to the user optimally using symbolic icons, unique data visualisations, and streamlined interaction and navigation. Touch screen user interfaces also allow for unique visualisations and user interactions. Various gestures (such as pinching, zooming, single and double-tapping, and swiping) allow for an additional level of interaction not possible through traditional desktop interfaces.

The goal of mobile application development is to provide an application that can target multiple users as well as cater to numerous platforms and devices [23]. Platform vendors provide application developers with a set of applicable APIs, programming languages, Integrated Development Environments (IDEs), and application distribution markets (Application Stores) where the application can be bought and or downloaded [23].

Mobile applications generally take one of two forms. The first is a native application built for a specific platform or operating system in a dedicated programming language. For example, IOS applications use Objective C or

Swift [4] whereas Android applications use Java [5]. The latter, more modern type, is a hybrid web application for mobile devices which use modern browser languages such as HTML5, CSS and JavaScript.

Instead of making use of device specific APIs like native applications [24], hybrid applications allow for the wrapping of CSS, HTML, and JavaScript code depending on the platform. Furthermore, it extends the features of HTML and JavaScript to access the components of the device, something not possible with pure web-based applications.

These "Hybrid" applications are named as such since they are not truly native applications (since layout rendering is done via Web views as opposed to the platform's native user interface framework), nor are they purely web-based since they have access to native device functionality [24]. These hybrid applications are packaged as native mobile applications for distribution through application distribution markets. Figure 3.15 showcases the benefits and costs of the different forms of mobile applications.

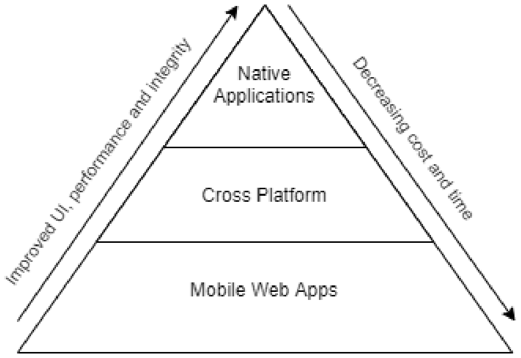


Figure 3.15: A comparison of the different developmental approaches

Developers are faced with two options for mobile application development. Producing one platform-specific application at a time is the first option [23]. This approach supports the use of an application on different platforms, yet, it is time-consuming as developers are required to familiarise themselves with the various IDEs [23]. The second option requires the developers to separate into teams where they develop each specific platform in parallel development [23].

Figure 3.16 represents both the traditional life cycle on the left as well as the cross-platform life cycle on the right. The cross-platform cycle improves the software development life cycle as it facilitates the structure of developing an app once and deploying it across a variety of platforms [23].

[4] https://developer.apple.com/swift

[5] https://developer.android.com

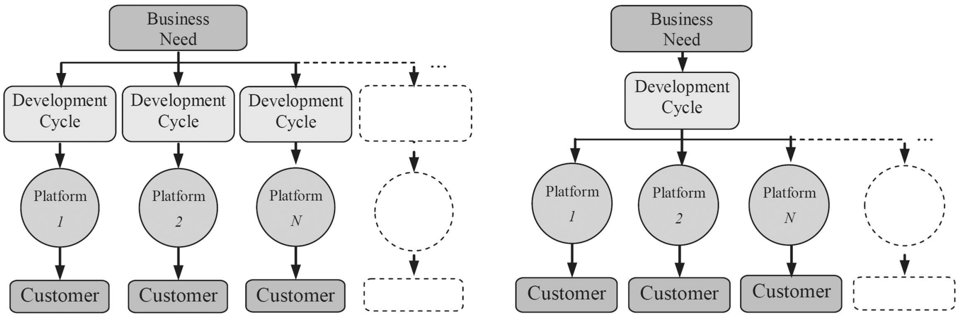


Figure 3.16: Traditional software development vs Cross-Platform development

There are six main tools available when developing hybrid cross-platform applications, namely: Apache Cordova; Appcelerator Titanium; Rhodes; Canappi mdsl; mobl and DragonRAD [23]. The Thesis uses Apache Cordova[6] since it is the most popular open-source mobile application development framework which allows for the development of hybrid applications.

Wrappers provide access to the mobile device's native features such as GPS and the camera. Cordova is a wrapper that provides such features as it loads web applications into native applications that then allow access to the device's native functionality [23]. Furthermore, Cordova applications are hybrid since they are neither "purely native" nor "purely web-based". With hybrid applications, CSS3 and HTML5 is used for rendering the user interface and JavaScript for powering the logic of the application [3]. The HTML and JavaScript components provide access to underlying mobile hardware, although the extent of this access varies depending on the device and browser used. Functionality is especially limited in older versions of the Android operating system. To overcome these limitations, Cordova embeds the HTML code inside a native "WebView" on the device, using a foreign function interface (a mechanism by which a program written in one programming language can make use of services written in another) to access the native resources of the mobile hardware [23]. Figure 3.17 shows the architecture of Cordova and how it interacts with the mobile application's operating system through the HTML APIs.

[6] https://cordova.apache.org/

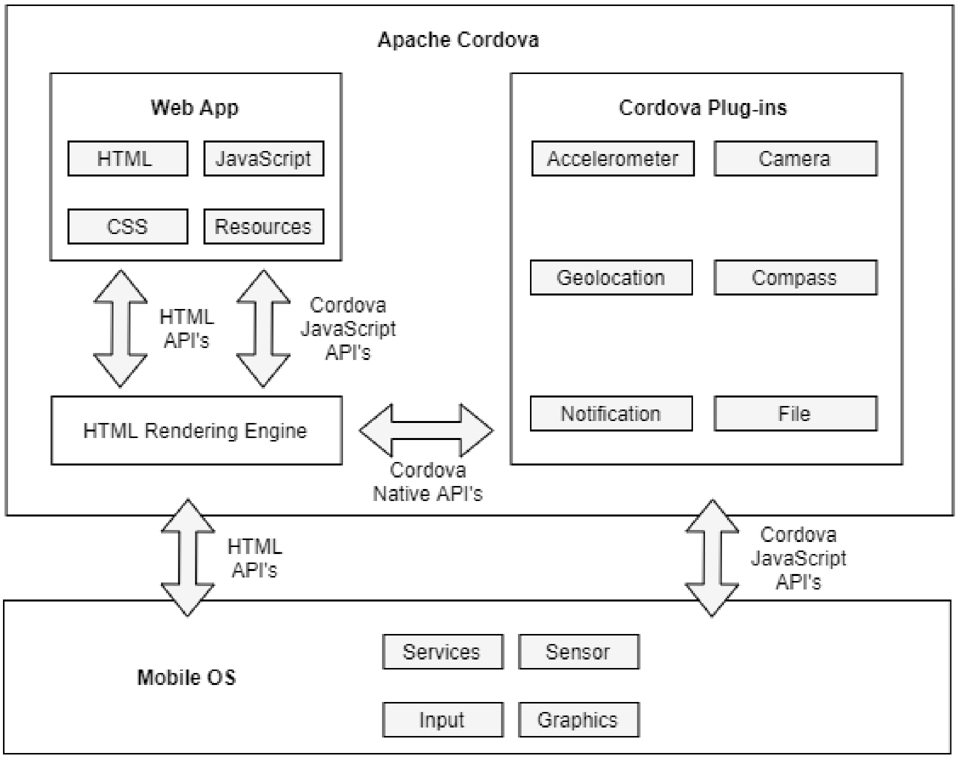


Figure 3.17: PhoneGap Tool Architecture

These hybrid applications are naturally not quite as adept at handling memory-intensive tasks when compared to native applications. However, they allow for faster development and for JavaScript libraries to be used. These are also operating system agnostic, meaning the same application can run on any mobile operating system with no additional effort.

3.5.3 ConceptCloud Mobile

ConceptCloud's server architecture allows for the addition of an externally accessible, representational state transfer (REST) application programming interface (API) to enable cross-origin resource sharing (CORS) of its powerful lattice generation and processing capabilities. This addition is achieved by configuring the Play! Server to return lattice data in JSON format.

This ability of the ConceptCloud server to support an externally accessible REST API and hybrid mobile application allows for the unique opportunity to build a mobile application that implements ConceptCloud's capabilities. As a starting point, a mobile application was built that mimics the functionality of ConceptCloud with an updated, more appropriate user interface and visualisation. It is not easy to utilise the current ConceptCloud user interface on a mobile device. As a result, developing a mobile version of ConceptCloud presents unique visualisation challenges.

3.5.3.1 Mobile User Interface

Figure 3.18 shows the main user interface of ConceptCloud Mobile. The user can interact with the touch screen interface by making use of standard touch screen gestures such as *pinch*, *zoom* and *tap*. Below are the features and elements found on all screens of the application:

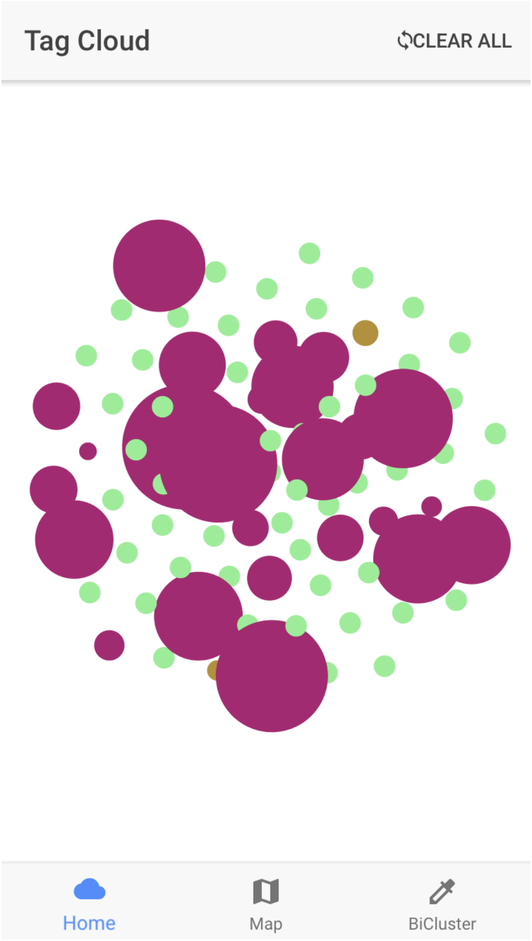


Figure 3.18: ConceptCloud Mobile

- **Page Title:** At the top of the screen is the page title indicating which screen of the application the user is currently viewing.
- **Clear all:** At the top of the *Tag Cloud* and *Bicluster* screen there is a *Clear All* button that will return the user to the top of the lattice as well as return all tags.
- **Navigation tabs:** At the bottom of the screen are three tabs, namely: *Tag Cloud (Home)*, *Map* and *Bicluster* which allow the user to navigate between the different screens of the application.

3.5.3.2 Tag Cloud

Figure 3.18 shows the mobile version of the Tag Cloud. The floating bubbles represent the tag cloud instead of text since the larger surface area of the bubbles allows for easier navigation on a touch screen interface. Similar to ConceptCloud, the bubble colour denotes the different categories of tags, and their size denotes the tag's frequency in the data set. The user can navigate the tag cloud by *pinching* the screen to zoom and *swiping* to change which section of the tag cloud is visible. When a bubble is *tapped*, it appears in the top left of the screen and the tag cloud is regenerated with the new focus concept (See Figure 3.19 and Figure 3.20). To ensure bi-directional navigation between the tag cloud and map is maintained, this action also updates the map screen and the pins displayed similar to how these viewers function in the browser based version of ConceptCloud.

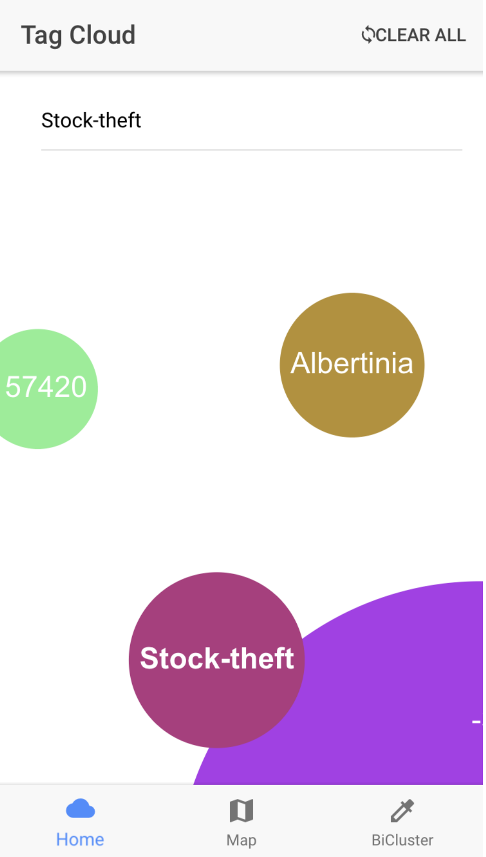


Figure 3.19: Filtered Mobile Tag Cloud

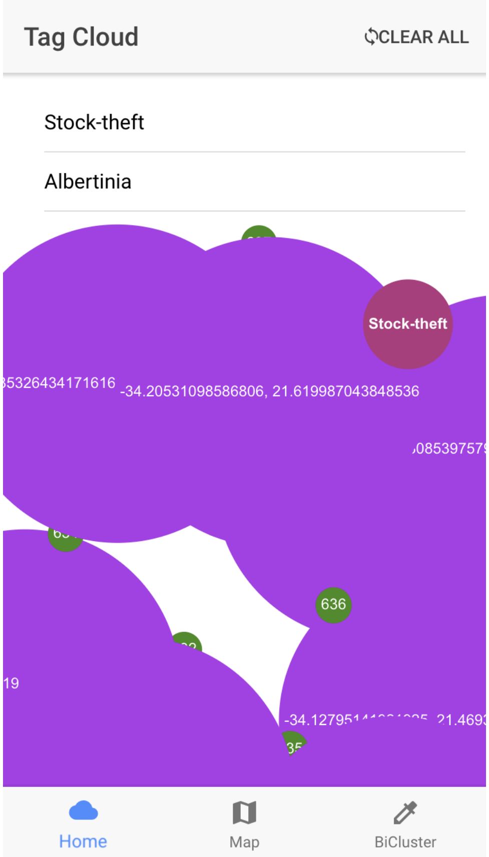


Figure 3.20: Filtered Mobile Tag Cloud 2

3.5.3.3 Map

Figure 3.21 shows the mobile map screen. The map screen is accessed by *tapping* on the map icon in the navigation bar at the bottom of the screen. The mobile map screen functions similar to the map viewer in ConceptCloud. This view also makes use of the Google Maps API, albeit within a mobile-friendly Ionic wrapper. Similar to the browser version of ConceptCloud, pins are drawn on the map based on the geolocation aspect of the data objects. Furthermore, the pins are clustered automatically to prevent the pins from overlapping one another. The user can navigate the map in the same manner as the Tag Cloud, that is, by *pinching* the screen to zoom in or out and *swiping* to change which section of the map is visible. The users may then *tap* any pin to update the tag cloud to display the object and attributes associated with that specific pin.

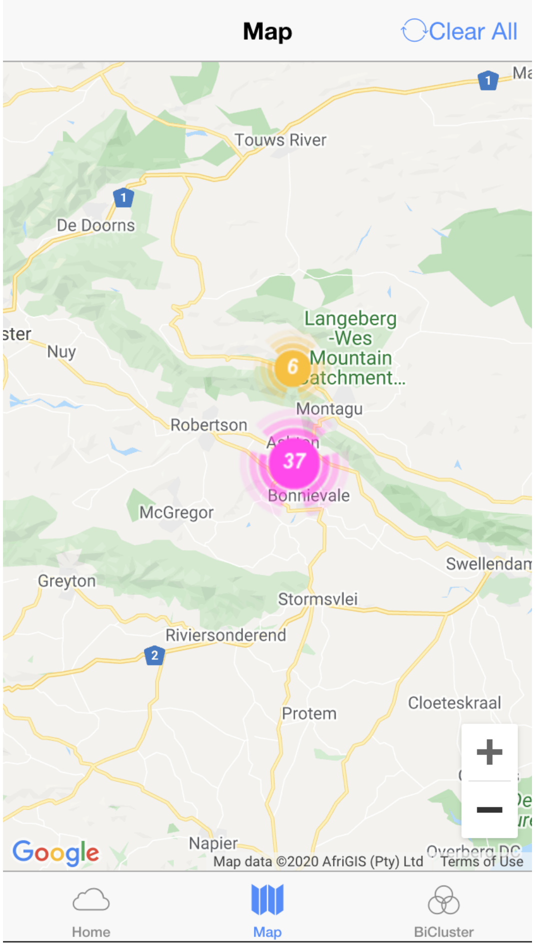


Figure 3.21: Mobile Map Screen

3.5.3.4 Bicluster

Figure 3.22 shows the Bicluster Tag Cloud screen that is accessed by *tapping* on the Bicluster icon in the navigation bar at the bottom of the screen. The mobile bicluster functions in a similar manner to the bicluster functionality in ConceptCloud. Individual map pins or pin clusters can be *tapped* and *held* to add them to the Bicluster lattice, similar to right-clicking them in the browser version of ConceptCloud. From there, the Bicluster cloud can be navigated and explored in the same way as the main Tag Cloud.

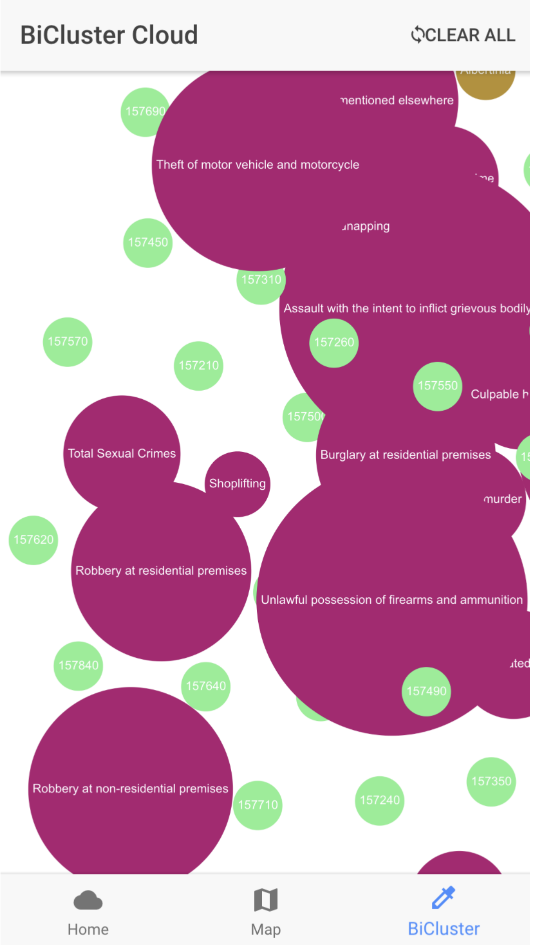

BiCluster Cloud
CLEAR ALL
mentioned elsewhere
157690
Theft of motor vehicle and motorcycle
napping
157450
157310
Assault with the intent to inflict grievous bodily
157570
157260
157210
157550
Culpable h
15750
Total Sexual Crimes
Shoplifting
Burglary at residential premises
murder
157620
Robbery at residential premises
Unlawful possession of firearms and ammunition
157840
157640
157490
Robbery at non-residential premises
157710
157240
157350
Home
Map
BiCluster

Figure 3.22: Mobile Bicluster

3.5.4 Server Implementation

In order to replicate the functionality found in the browser version of ConceptCloud, the server architecture was expanded. The expansion was achieved by developing REST API endpoints that accepted HTTP requests from external sources and returned data in the more agnostic and modern JSON format, as opposed to Play! framework "Views". This allows external developers to also make use of these methods to develop their own applications. The newly added API endpoints are:

3.5.4.1 API Endpoints:

- **GET/mobileTags:** This API call serves the same role as the method used for the main Tag Cloud but instead of returning the tags as a Play! framework view, it returns the tags as JSON objects instead.

- **GET/mobileFilteredTags:** This method requires a Tag object, selected by the user, and returns all Tags, also in JSON format, based on the user's new position in the lattice.

- **GET/mapPinsInBounds:** This method is the same one used in the Map Implementation since the Google Maps API, like the mobile application, require the geolocation data in JSON format to calculate the location of pins and render them on the map.

- **POST/mobileBiClusterTags:** This method accepts a JSON object of all tags required in the Bicluster Tag Cloud and returns all Tags in the newly generated lattice as a JSON object.

3.6 Chapter Summary

Chapter 3 discussed the extensions used to fully integrate new visualisations into ConceptCloud's existing user interface by ensuring the extensions adhered to key requirements. The key requirements and the methods used to ensure each requirement was met is summarised below:

- The extension should be clear and readable for a user with moderate levels of data literacy.

- The visualisations must remain an accurate representation of the data where opportunity for misinterpretation of the data by the user is minimised.

- The new extension must aid in data analysis by providing cognitive support and reducing the cognitive load experienced by the user as far as possible.

- The extension must aid in data exploration and knowledge discovery by providing different visualisations of the same data.

- The extension must allow the new visualisation as well as the *Tag Cloud* to drive exploration. Thus a pivotal function of this tool is *to ensure the data exploration works in a bidirectional manner*. In other words, to update the new visualisation when changes are made in one of the visualisations and vice versa.

- The extension must ensure that *the visualisation and exploration tool be highly scalable to process large data sets*, since the rate at which such data is being generated is ever increasing.

These extensions included a map viewer for visualising semi-structured data with geolocation aspects, a graph viewer for visualising the occurrences of attributes in the semi-structured data set, and a mobile application that mimics the functionality of the desktop application. These extensions were added seamlessly and comprehensively into the existing ConceptCloud infrastructure and user interface. To conclude, the latest ConceptCloud architecture was expanded by including a mobile architecture extension. The Mobile architecture extension included an external API controller as well as various mobile applications.

Chapter 4

Evaluation

4.1 Overview

This chapter describes the evaluation conducted on the functionality of the newly implemented data visualisations made to ConceptCloud described in Chapter 3. The evaluation takes the form of two case studies conducted on two publicly available data sets in the viticulture domain. These data sets contain textual, numeric, and geolocation aspects. This chapter begins by providing a background discussion on viticulture and the climatic indexes used by viticulturists. Moreover, a discussion on the data preparation and transformation of both sets for ConceptCloud is provided. Thereafter, a discussion of the evaluation and case studies is provided as well the results obtained from it using the newly developed data visualisations of ConceptCloud.

4.2 Introduction

ConceptCloud is a versatile tool that applies to any domain as long as the data is of a suitable format, namely JSON. The two data sets utilised in this chapter both fall in the viticulture domain, and in particular, contains data from California in the United States of America. A state that is widely known and highly regarded for its wine production [7]. Consequently, viticulturists from all over the world study the atmospheric conditions in which these grapes are grown and the characteristics of the wines produced there. This makes the data sets good candidates for example case studies viticulturist might conduct using a tool such as ConceptCloud.

As discussed in Section 3.4, ConceptCloud is capable of processing and operating on numeric data. However, the data exploration of the main tag cloud visualisation is greatly improved when displaying textual data, preferably short phrases. Thus, when utilising a numeric data set, it is beneficial to categorise the numeric data using a widely known and accepted scale or index.

The first data set used in the evaluation is an example of a purely numeric data set, namely climatic measurements, that was binned into indices, described in Section 4.4 for improved tag cloud visualisation. The second data set used in the evaluation, namely wine reviews, contains more textual fields and required no data binning. It does however contain large amounts of free text which again, can be processed by ConceptCloud as is, but is not conducive to effective tag cloud visualisations. ConceptCloud has some Natural Language Processing (NLP) capabilities, which was used to extract key-phrases from the free text. This process is described in Sections 4.5.6 and 4.5.7.

Further, the second data set lacks an explicit geolocation aspect which is required to make use of the map viewer. To resolve this, the data set was combined with the first, resulting in a diverse and complex data set with geolocation aspects used in the final evaluation. This process is described further in Section 4.5.8.

4.3 Viticulture

Viticulture or “winegrowing” is a branch of horticulture and involves the cultivation and harvesting of grapes. The duties of the viticulturist include monitoring and controlling pests and diseases, fertilisation, irrigation, canopy management, monitoring of fruit development and characteristics, deciding on optimal harvest windows and managing vine pruning during winter months. Viticulturists generally work closely with winemakers since vineyard management is the first step in crafting wine that is of a high-quality [41].

4.4 Climatic Indices

The climate found in different grape-growing regions contributes significantly to the variety of viticultural products, wine quality, and wine type. In previous studies, researchers have concluded that the temperature, for example, affects both the quality and the composition of grapes [17]. Consequently, viticulturists have developed several indices to formalise these effects and to group and compare different wine-growing regions over the world.

Some of the first indices, such as The Thermal Index of Winkler, focused only on temperature [6]. However, researchers later discovered that multiple climatic factors play a role in grape development and composition. Furthermore, researchers found that these climatic factors also influence one another. Jackson and Cherry, (1988) discovered that the ripening capacity of grapes in regions with high rainfall is significantly lower compared to the predictions of climatic thermal indices. Thus, requiring the development of new indices.

The new indices include various climatic factors for a given region, namely the day and night temperature, the air temperature, and the water level in the soil. These factors influence the coloration, aroma, ripening quality, and overall quality of the grapes. As a result, the factors directly impact the characteristics and quality of the wine [54].

Lastly, the Géoviticulture Multi-criteria Climatic Classification (MCC) System was developed in 2004 by Tonietto and Carbonneau [54]. The system was developed to identify weather conditions in grape-growing regions more simply. Furthermore, the system consists of three indexes, namely the Heliothermal Index, Cool Night Index, and Dryness Index. Various researchers collaborated with the World Meteorological Organization (W.M.O.), members of the Agricultural Meteorology Commission of various wine-producing countries, and scientific institutions to obtain the necessary climate data for grape-growing regions. These climatic indices are strongly related to the characteristics and qualitative potential of grapes/viticultural products including the acidity, colour, and type of the wines. Furthermore, these indices are used to characterise a region's climatic potential.

4.4.1 Heliothermal Index

The Heliothermal Index of Huglin, developed by Huglin in 1978 [37], calculates the heliothermal potential of a region. The index can be used worldwide and exhibits a good relation to the potential sugar content of the grape.

It was chosen for the Géoviticulture MCC System above other heliothermal indices for a variety of reasons. Firstly, the index calculates the heliothermal potential based on a period which is nearest to the average cycle of the grapevine. Secondly, it provides a more accurate estimation of the sugar potential according to varieties compared to the classic temperature sums. The index therefore provides qualitative information as well. Thirdly, when the index is combined with the cool night index,it allows for a good discrimination of the region climate with regards to cool night conditions during the ripening period of the grape and the heliothermal conditions during the vegetative cycle.

The index is calculated from monthly climatic means using the below equation:

$$HI = \sum_{Mi}^{Mf} \frac{[(T - 10) + (T_x - 10)]}{2} d \tag{4.0}$$

- In the Northern Hemisphere, T is the mean air temperature (°C), $\mathrm{T_x}$ is the maximum air temperature (°C), d is the length of day coefficient, ranging from 1.02 to 1.06 between 40° and 50° of latitude for the period of 1 April to 30 September.

- In the Southern Hemisphere the index is equally calculated based on the 6-month period from 1 October to 31 March.

For this index, the "helio" component is a result of the "d" coefficient of the thermal component which expresses the mean day length in relation to the latitude. Secondly, it is due to the calculation of the thermal component being estimated over the mean period when most of metabolisms in the plant is active.

4.4.2 Cool Night Index

The Cool Night Index (CI) represents the minimum average air temperature of the last month of the growing period and varies from values of 12° C (very cool nights) to values greater than 18° C (warm nights). Night coolness is a climatic factor that has significant influence on grape and wine colour as well as aromas [45]. The Cool night index is used in order to improve the evaluation of the qualitative potentials of wine-growing regions, especially with regards to secondary metabolites. Depending on the hemisphere, the cool night index (CI) is calculated as follows:

$$CI = T_y \tag{4.0}$$

- Northern Hemisphere: CI= minimum air temperature in the month of September (mean of minima), in ° C.
- Southern Hemisphere: CI= minimum air temperature in the month of March (mean of minima), in ° C.

4.4.3 Dryness Index

The Dryness Index (DI) indicates the potential water availability in soil, which is related to the level of dryness in a region, and varies from values greater than 150mm (humid) to values of -100mm (very dry). Essentially, this index allows for the characterisation of the water component of a climate. This is an important climatic factor to consider seeing as it influences grape ripening and wine quality [40]. The dryness index is an adaptation of the potential water balance of the soil index of Riou [53], which was developed specifically for vineyard use.

The Dryness Index is calculated using the following formula [54]:

$$W = \sum_{Mi}^{Mf} Wo + P - T_v - E_s \tag{4.0}$$

Where:

- W is the estimate of soil water reserve at the end of a chosen period
- W_0 is the initial useful soil water reserve

- P is the precipitation
- T_v is the potential transpiration in the vineyard
- E_s is the direct evaporation from the soil

Similar to the HI Index, this calculation is based on a 6 month period which is suitable for most vineyards in the world. The Index is calculated month by month using the monthly values for all the variables mentioned above. The Dryness Index is the value of W calculated at the end of the 6 month period.

The values for T_v and E_s are calculated monthly using the following formulas:

$$T_v = ETPk \tag{4.0}$$

Where:

- ETP is the potential evapotranspiration (monthly total calculated using the Penman method [49])
- k is the coefficient of radiation absorption by vine-plant

$$E_s = \frac{ETP}{N}(1-k)JPm \tag{4.0}$$

Where:

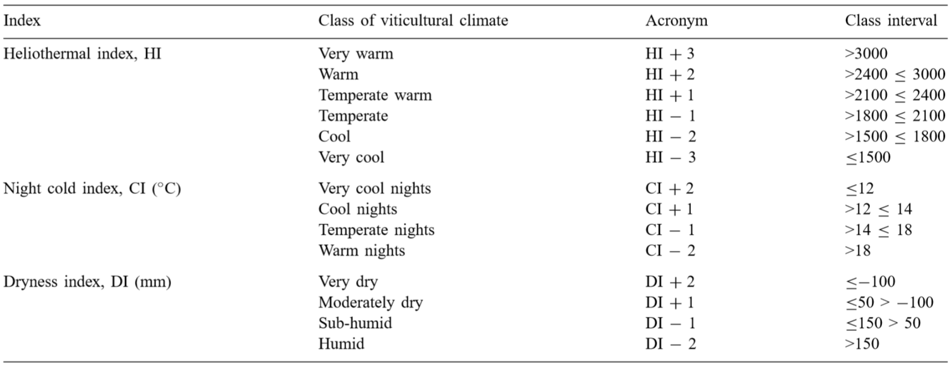
- N is the number of days in the month
- JPm is the number of days of effective evaporation from the soil per month (rainfall per month in mm/5)

Values used for k differ depending on the hemisphere the calculation is based on. The different values for k are as follows:

- In the Northern Hemisphere: k is equal to 0.1 for April, 0.3 for May and 0.5 for the months June to September.
- In the Southern Hemisphere: k is equal to 0.1 for October, 0.3 for November and 0.5 for the months from December to March.

4.4.4 A Multicriteria Climatic Classification System (Géoviticulture MCC System)

As discussed above, The Géoviticulture Multicriteria Climatic Classification System is a climatic classification system for grape-growing regions based on the integration of the different classes of the three climatic indices, namely DI, HI and CI. A multicriteria classification system is used in order to allow the grouping of grape-growing regions with similar viticultural climates. This is useful since the different classes of viticultural climate represents both climatic differences as well as the responses of the vine and grapes to the climatic factors. Different classes of viticultural climate is proposed for the heliothermal index, cool night index and the dryness index. The proposed classification system of the different viticultural climates as well as the interpretation of each class is presented in Figure 4.1.



Index	Class of viticultural climate	Acronym	Class interval
Heliothermal index, HI	Very warm	HI + 3	>3000
	Warm	HI + 2	>2400 ≤ 3000
	Temperate warm	HI + 1	>2100 ≤ 2400
	Temperate	HI − 1	>1800 ≤ 2100
	Cool	HI − 2	>1500 ≤ 1800
	Very cool	HI − 3	≤1500
Night cold index, CI (°C)	Very cool nights	CI + 2	≤12
	Cool nights	CI + 1	>12 ≤ 14
	Temperate nights	CI − 1	>14 ≤ 18
	Warm nights	CI − 2	>18
Dryness index, DI (mm)	Very dry	DI + 2	≤−100
	Moderately dry	DI + 1	≤50 > −100
	Sub-humid	DI − 1	≤150 > 50
	Humid	DI − 2	>150

Figure 4.1: Classes of viticultural climate for the dryness index, heliothermal index and cool night index of the grape-growing regions

This system makes use the following definitions to describe climates and regions [54]:

- **Viticultural climate:** this refers to the climate of some locality, grape-growing region or vineyard described by the set of three viticultural climatic indices. The viticultural climate is thus a more specific description of the climate than its more general definition. This climate can also change annually and distinctions are made between the mean viticultural climate and the range of viticultural climate.

- **Climatic group:** this refers to the group assigned to a locality, grape-growing region or vineyard based on its viticultural climate. Usually, the climatic group includes a large share of the (inter-annual) ranges of viticultural climates.

- **Viticultural climate with intra-annual variability:** this relates to regions that, under natural climate conditions, change viticultural climate class as a result of the time of the year at which grapes can be produced (This definition is specifically used for regions where it is possible to have more than one grape harvest a year).

The Geoviticulture MCC System is only applicable to grape-growing regions, since it is applied once the climatic criteria limiting viticulture (risk of frost, excessive humidity, etc.) are taken into account. Figure 4.2 shows the full process of converting climatic variables into the MCC system.

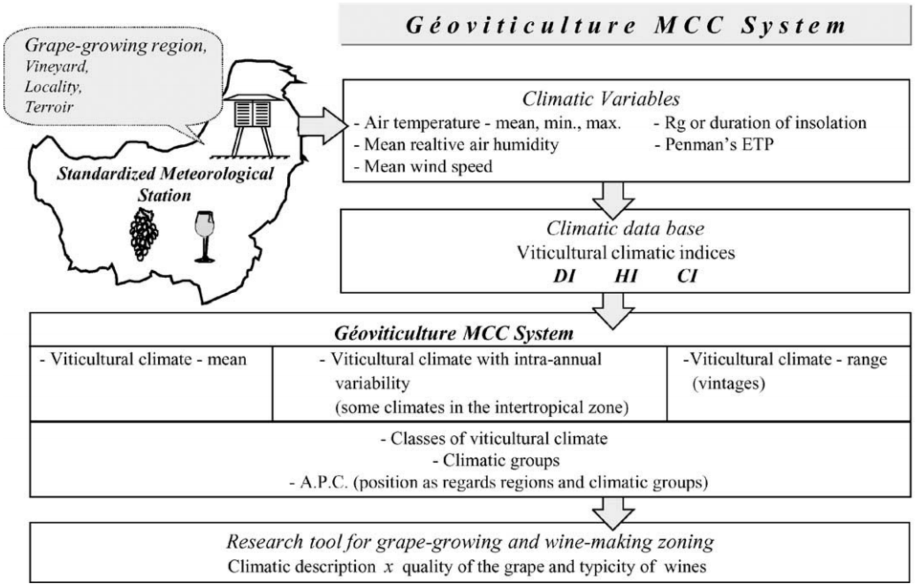

Géoviticulture MCC System
Grape-growing region,
Vineyard,
Locality,
Terroir
Standardized Meteorological Station
Climatic Variables
- Air temperature - mean, min., max.
- Mean realtive air humidity
- Mean wind speed
- Rg or duration of insolation
- Penman's ETP
Climatic data base
Viticultural climatic indices
DI HI CI
Géoviticulture MCC System
- Viticultural climate - mean
- Viticultural climate with intra-annual variability (some climates in the intertropical zone)
-Viticultural climate - range (vintages)
- Classes of viticultural climate
- Climatic groups
- A.P.C. (position as regards regions and climatic groups)
Research tool for grape-growing and wine-making zoning
Climatic description x quality of the grape and typicity of wines

Figure 4.2: The Géoviticulture MCC System [54]

4.5 Discussion

4.5.1 Climatic Data Set

The first data set used in the evaluation is publicly available and was collected from *The National Centers for Environmental Information*'s portal. This is part of the National Oceanic and Atmospheric Administration [25], which in turn is an American scientific and regulatory agency within the United States Department of Commerce.

The data set specifically is from the "GHCN (Global Historical Climatology Network)- Daily database", which contains records from various sources which are merged and reviewed for quality assurance. The database contains more than 40 different meteorological measurements such as wind speed average, soil temperature, evaporation, precipitation, temperature maximum, temperature minimum and a wide array of other elements.

When downloading the data, one can select from a number of different atmospheric and climatic measurements at any weather station in the country, some of which extend as far back as the nineteenth century. Further, the measurements are updated daily where possible and are usually available within a few days of observation. Furthermore, when downloading the data, it is possible to include geographic location, station name as well as other flags within the data request.

The final data set of atmospheric measurements were chosen based on their ability to satisfy the below criteria:

- It should be possible to categorise or index any numeric field in the set into some formalised textual scale. For atmospheric measurements to be categorised into the climatic indices described in Section 4.4, it is further required that: 1) the measurement be from stations in a grape growing region (California in this case), as well as that 2) the measurements span a certain period (For the northern hemisphere this is period is 1 April to 30 September).
- The data must contain some explicit geolocation aspect, preferably latitude and longitudinal co-ordinates, in order to evaluate the map viewer.
- The total data set must contain at least 15 000 unique objects in order to evaluate tool scalability.

As a result the data set contains indicators measured at various weather stations and reservoirs across the entire state of California, measured daily from April 2014 up until the end of September 2016. The data set contains 520631 unique sets of measurements at 1086 different stations. The data contains all the measurements required to calculate the climatic indices described in Section 4.4 above. These include measurements such as minimum and maximum temperature, rainfall and average wind speed among others. The designations used in the data set link to the following climate measurements:

- AWND: Average daily wind speed (meters per second).
- DATE: Year, month and day of record.
- ELEVATION: elevation above mean sea level (tenths of meters).
- EVAP: Evaporation of water from evaporation pan (mm).
- LATITUDE: latitude coordinate (decimated degrees w/northern hemisphere values /> 0).
- LONGITUDE: longitude coordinate (decimated degrees w/western hemisphere values /< 0).
- NAME: Weather Station Common Name (city or airport).
- PRCP: Precipitation (mm).
- TAVG: Average temperature (Celsius).
- TMAX: Maximum temperature (Celsius).

- TMIN: Minimum temperature (Celsius).
- STATION: Station identification code.

Figure 4.3 shows the initial data object before any data categorisation was done.

```
{
   "AWND": 2.1,
   "DATE": "2014/06/20",
   "ELEVATION": 57.9,
   "EVAP": "",
   "LATITUDE": 39.49,
   "LONGITUDE": "-121.61833",
   "NAME": "OROVILLE MUNICIPAL AIRPORT, CA US",
   "PRCP": 0.0,
   "TAVG": 22.3,
   "TMAX": 35.6,
   "TMIN": 16.1,
   "STATION": "USW00093210"
},
```

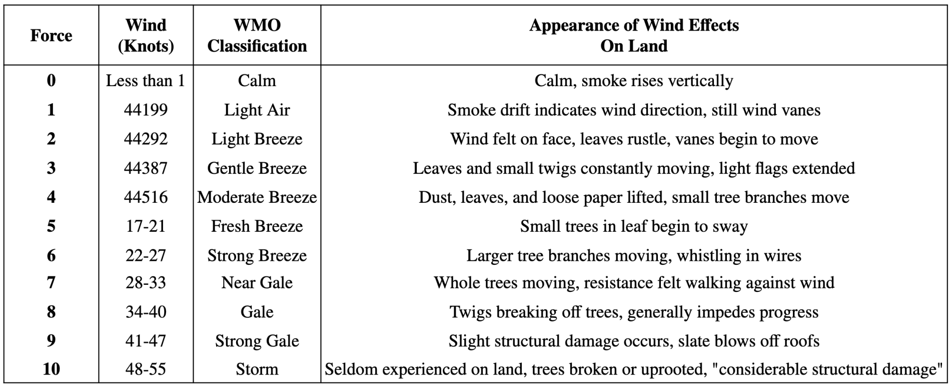
Figure 4.3: Initial JSON object of climatic measurements

4.5.2 Data Transformation

The data set was collected as a comma separated value (.csv) file of daily measurements and was loaded into a custom data processing script which processed the individual daily station measurements and calculated the climatic indices, namely the Heliothermal Index and Cool night Index. This was done for each station for the grape growing seasons of 2014, 2015 and 2016. The Dryness Index could unfortunately not be calculated due to a lack of potential evapotranspiration measurements in the data set, and as a result has been assigned a value of “NoData” instead.

This process resulted in the daily measurements being summarised into a period index where each unique station, and its climatic indices, appears in the processed data set a maximum of three times (once for each growing period the indices could be calculated) data allowing.

The only numeric field in the data set that is not used in any climatic index calculation, is AWND (Average daily wind speed). This climatic measurement was averaged per station for each of the grape growing periods and categorised into the Beaufort Wind Scale. This scale has historically been used to assign a textual description to wind speed measured in knots [1]. Figure 4.4 shows this scale as well as the categories and descriptions these measurements relate to.



Force	Wind (Knots)	WMO Classification	Appearance of Wind Effects On Land
0	Less than 1	Calm	Calm, smoke rises vertically
1	44199	Light Air	Smoke drift indicates wind direction, still wind vanes
2	44292	Light Breeze	Wind felt on face, leaves rustle, vanes begin to move
3	44387	Gentle Breeze	Leaves and small twigs constantly moving, light flags extended
4	44516	Moderate Breeze	Dust, leaves, and loose paper lifted, small tree branches move
5	17-21	Fresh Breeze	Small trees in leaf begin to sway
6	22-27	Strong Breeze	Larger tree branches moving, whistling in wires
7	28-33	Near Gale	Whole trees moving, resistance felt walking against wind
8	34-40	Gale	Twigs breaking off trees, generally impedes progress
9	41-47	Strong Gale	Slight structural damage occurs, slate blows off roofs
10	48-55	Storm	Seldom experienced on land, trees broken or uprooted, "considerable structural damage"

Figure 4.4: The Beaufort Wind Scale

As a final step in the data transformation script, the data was transformed to a JSON object since this is the format ConceptCloud requires. The newly transformed JSON object can be seen in Figure 4.5.

```
{
    "period": "2014",
    "location": "37.4681, -120.1106",
    "station_name": "OROVILLE MUNICIPAL AIRPORT, CA US",
    "wind_speed_avg": "Light Breeze",
    "helio_thermal_index": "Very warm",
    "dryness_index": "NoData",
    "cool_night_index": "Warm nights"
    "avg wind speed": "Light breeze"
},
```

Figure 4.5: Transformed JSON Object

4.5.3 Case Study 1: Climatic Data

The goal of this case study is to evaluate the new additions made to ConceptCloud and determine how they facilitate and improve the data exploration and knowledge discovery aspects of the software artefact through a case study. The case study done on the publicly available data set serves as a demonstration and it is not the primary goal of this evaluation to conduct any deep data exploration and mining, but rather demonstrate how ConceptCloud enables and promotes the exploration of data and knowledge discovery. It should be viewed as as an example of how a domain expert might utilise the tool to gain a better understanding of their data set, identify any correlations therein and determine which parts of the data can be investigated further.

This case study is presented alongside a series of screenshots of the different viewers in ConceptCloud's user interface. It is recommended that they be viewed in colour as the software tool makes use of uniquely coloured tags and clusters to indicate the different attribute types and cluster sizes in the tag viewers, as well as pie segments and bars.

4.5.3.1 California Climate Overview

As stated in Section 4.2, the state of California is widely known and highly regarded for its wine production [7]. It attracts the attention of viticulture researchers from all over the world to study the atmospheric conditions in which these grapes are grown and the characteristics possessed by the wines grown in this climate.

California is unique in its climate and displays five major climate types depending on latitude, elevation, and proximity to the coast. These are Desert, Cool interior, Highland, Steppe and a small section of Mediterranean climate [42].

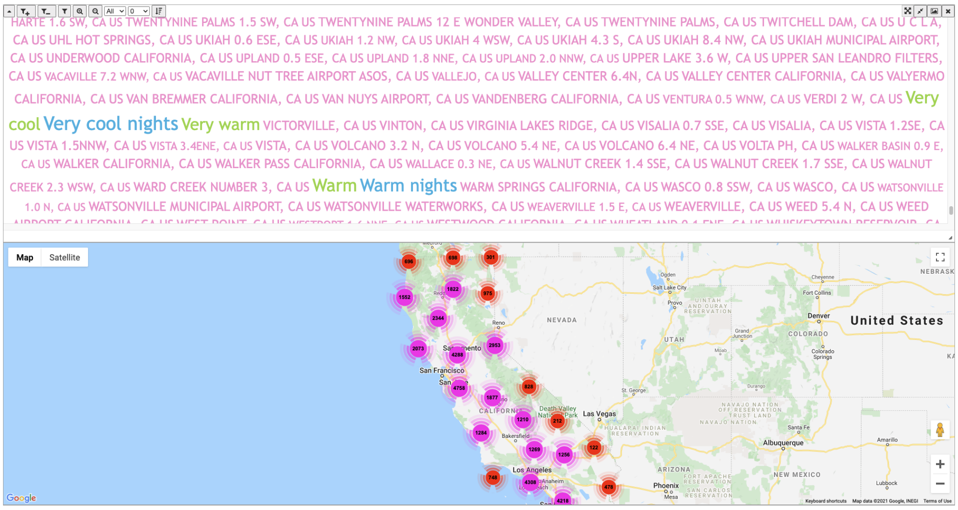
This Mediterranean climate in the state has three variations, namely a cool summer and cool winter climate found along the coast and the western slopes of the Sierra Nevada foothills. The second is similar to the first, also found along the coast, but with frequent summer fog. The final type of Mediterranean climate found in California has hotter, dry summers and cooler, rainy winters [42]. This climate is found in the Central Valley. Most of the rainfall occurs in the winter and the ocean moderates temperature extremes.

Rainfall in the state ranges from more than 4300 mm in the northwest to only traces in the southeastern desert [47]. Further, in the higher eastern deserts of California, summer temperatures are more moderate. Winter temperatures in the Sierra Nevada can drop to near freezing [47]. In Los Angeles, the average annual temperature is around 18 degrees Celsius and annual precipitation is roughly 350mm. San Francisco on the other hand is on average cooler with a annual temperate average of 14 degrees Celsius, and also on average receives more rainfall, namely 508 mm annual average [47].

4.5.3.2 Initial View

After completing the data transformation process described in Section 4.5.2, a single JSON file was generated by the custom data processing script. This JSON file was then loaded into ConceptCloud through the standard data configuration script and process discussed in Section 2.5.1.1, resulting in the view shown in Figure 4.6, after loading the user interface.

Upon first viewing the data set in ConceptCloud's user interface, the user is able to easily determine the different attribute categories by looking at the colour of the tags. It is also easy to determine which tags appear most frequently in the data set by identifying the largest tags. From Figure 4.6, we can see that the tag "Very cool nights", which belongs to the 'Cool night index' tag category and is typed in blue, appears as one of the largest tags in the cloud.

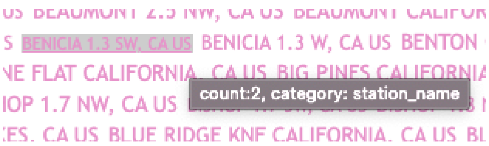


HARTE 1.6 SW, CA US TWENTYNINE PALMS 1.5 SW, CA US TWENTYNINE PALMS 12 E WONDER VALLEY, CA US TWENTYNINE PALMS, CA US TWITCHELL DAM, CA US U C L A, CA US UHL HOT SPRINGS, CA US UKIAH 0.6 ESE, CA US UKIAH 1.2 NW, CA US UKIAH 4 WSW, CA US UKIAH 4.3 S, CA US UKIAH 8.4 NW, CA US UKIAH MUNICIPAL AIRPORT, CA US UNDERWOOD CALIFORNIA, CA US UPLAND 0.5 ESE, CA US UPLAND 1.8 NNE, CA US UPLAND 2.0 NNW, CA US UPPER LAKE 3.6 W, CA US UPPER SAN LEANDRO FILTERS, CA US VACAVILLE 7.2 WNW, CA US VACAVILLE NUT TREE AIRPORT ASOS, CA US VALLEJO, CA US VALLEY CENTER 6.4N, CA US VALLEY CENTER CALIFORNIA, CA US VALYERMO CALIFORNIA, CA US VAN BREMMER CALIFORNIA, CA US VAN NUYS AIRPORT, CA US VANDENBERG CALIFORNIA, CA US VENTURA 0.5 WNW, CA US VERDI 2 W, CA US Very cool Very cool nights Very warm VICTORVILLE, CA US VINTON, CA US VIRGINIA LAKES RIDGE, CA US VISALIA 0.7 SSE, CA US VISALIA, CA US VISTA 1.2SE, CA US VISTA 1.5NNW, CA US VISTA 3.4ENE, CA US VISTA, CA US VOLCANO 3.2 N, CA US VOLCANO 5.4 NE, CA US VOLCANO 6.4 NE, CA US VOLTA PH, CA US WALKER BASIN 0.9 E, CA US WALKER CALIFORNIA, CA US WALKER PASS CALIFORNIA, CA US WALLACE 0.3 NE, CA US WALNUT CREEK 1.4 SSE, CA US WALNUT CREEK 1.7 SSE, CA US WALNUT CREEK 2.3 WSW, CA US WARD CREEK NUMBER 3, CA US Warm Warm nights WARM SPRINGS CALIFORNIA, CA US WASCO 0.8 SSW, CA US WASCO, CA US WATSONVILLE 1.0 N, CA US WATSONVILLE MUNICIPAL AIRPORT, CA US WATSONVILLE WATERWORKS, CA US WEAVERVILLE 1.5 E, CA US WEAVERVILLE, CA US WEED 5.4 N, CA US WEED

Figure 4.6: Initial tag cloud and map viewer

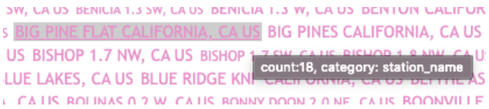
Another aspect of the data that is immediately obvious from looking at the tag cloud, is that the 'Station name' tags, shown in pink text, all appear to be roughly the same size. This indicates that measurements from these stations are relatively consistent in the data set, with no single station making up the bulk of the measurements.

There are a few stations which do appear smaller that the other tags such as "BENICIA 1.3 SW, CA US". This can be confirmed by hovering over the tag and observing the tag count in the tool tip pop up. Figure 4.7 shows an example of this, and confirms that "BENICIA 1.3 SW, CA US" only appears twice in the data set while most other stations with the mode tag size, appear on average eighteen times.



count:2, category: station_name

Figure 4.7: BENICIA 1.3 SW, CA US Tooltip



count:18, category: station_name

Figure 4.8: BIG PINE FLAT, CA US Tooltip

By looking at the initial visualisation of the map viewer, it can be seen that the pin clusters for the various station measurements are relatively evenly spread across the state of California. By looking at the different coloured clusters, it is clear that the majority of stations appear in the centre and bottom of the state (bigger clusters appear in purple while smaller clusters appear in red and even smaller clusters in yellow), with the highest concentration of station measurements appearing around the three big cities, namely San Francisco, Los Angeles and San Diego.

In order to get a clearer view of the station distributions near the cities, the user can zoom in on one of the clusters, breaking up the purple clusters into red and yellow clusters, to observe a lower level breakdown of the pins. In the case of San Francisco, Figure 4.9 shows that the concentration of stations are higher around the highly developed areas as opposed to the rural areas.

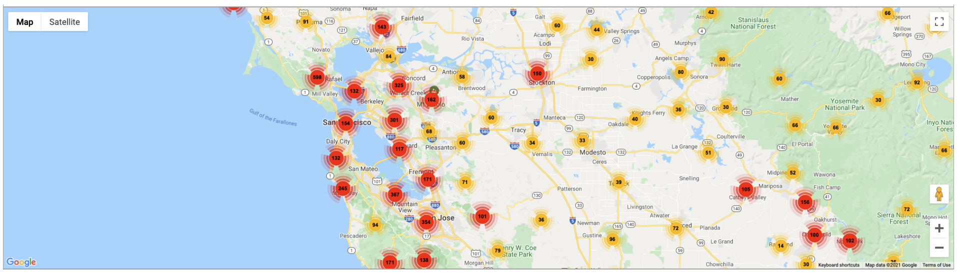
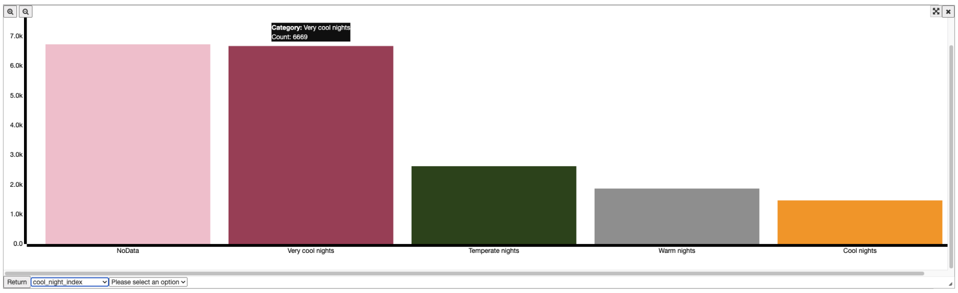


Figure 4.9: Map Viewer view of San Francisco

4.5.3.3 Tag and Map Exploration

Since “Very cool nights” is the most frequently occurring tag in the data set, it provides a natural starting point for further knowledge discovery. This tag belongs to the “Cool night index” category. In order to get a high level view of the distribution of this category, the user can generate a graph in the graph viewer by selecting the “Cool night index” category from the first drop down and visualising the composition of tags in this category. Figure 4.10 shows a bar graph of this tag category and the maroon bar with the tool tip indicates that the “Very cool nights” tag occurs in the data set 6669 times. By considering the graph as a whole, it is clear that this tag does indeed make up the majority of the data set (The pink bar on the far left relates to instances where the index could not be calculated due to lack of data and can be ignored).



Category: Very cool nights
Count: 6669

cool_night_index
Please select an option

Figure 4.10: Bar graph of the 'Cool night index' tag category

Once the first graph is generated, the user can easily get a view of the composition of other tag categories in relation to the "Cool night index" tag category. This can be achieved by selecting a category from the second drop down menu. This will update the focus concept and update both the tag cloud and map viewer to the new location in the concept lattice.

The same data can be visualised in a pie chart. Figure 4.11 shows the "drilled down" pie chart for the "Wind speed average" tag category for the "Very cool night" tag. The largest cyan segment once again relates to the "No Data" category and should be ignored. The actual largest segment is the "Light Breeze" tag, shown in maroon, with 185 objects that share both the "Very cool night" and "Light Breeze" attributes.

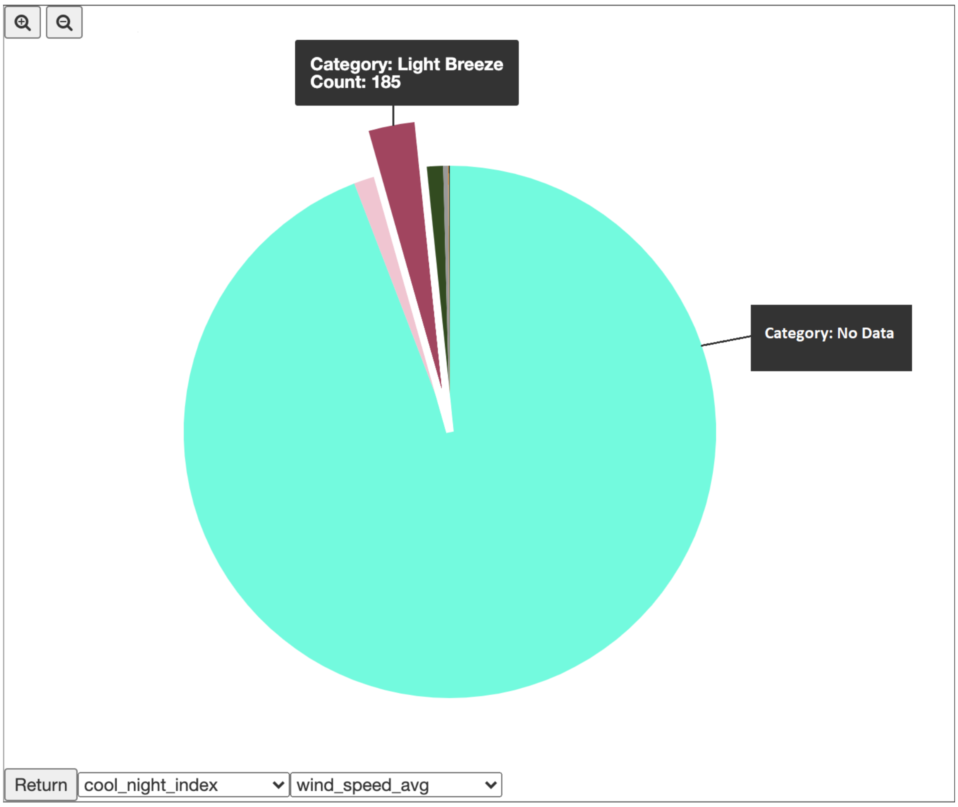


Figure 4.11: Drilled down pie chart

With "Very cool night" now being the focus concept, the tag cloud is updated to show the tags visible in Figure 4.12. The station names have been removed for clarity. By looking at the remaining tags, it can seen that the viticultural class of "Very cool nights" are most frequently associated with the period of "2016" and the "Gentle Breeze" wind speeds among others.

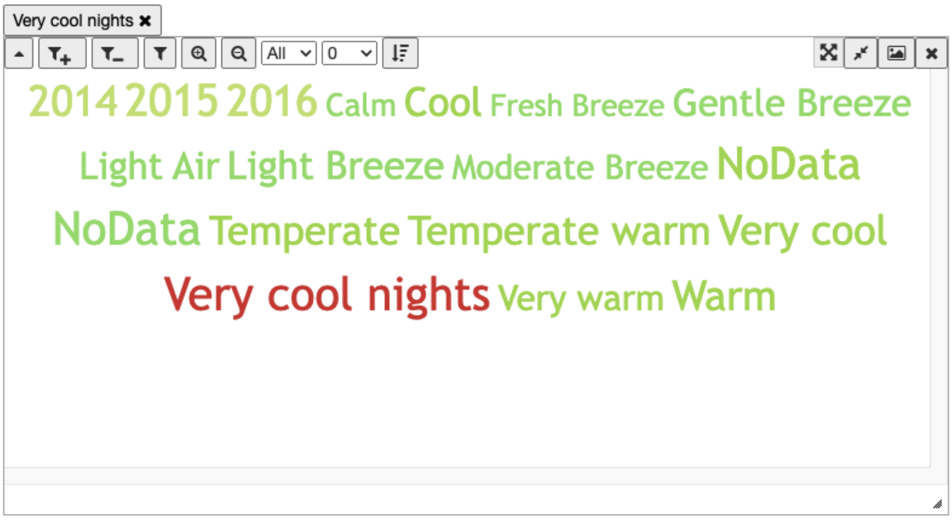

Very cool nights ✖
All
0
2014 2015 2016 Calm Cool Fresh Breeze Gentle Breeze
Light Air Light Breeze Moderate Breeze NoData
NoData Temperate Temperate warm Very cool
Very cool nights Very warm Warm

Figure 4.12: Tag cloud with 'Very cool night' as the focus concept

4.5.3.4 Temporal Filter

Since the data set contains measurements taken from the year 2014 to 2016, displayed in the "Period" tag category, the user is able to use these tags as a temporal filter to compare how the measurements and indices differed over the years. This can be accomplished by selecting a tag belonging to the "Period" tag category, appearing in maroon text in the tag cloud, which will update the focus concept and only display objects containing this attribute in the tag cloud and map viewer. Figures 4.13 and Figure 4.14 shows a comparison of the tag clouds having the "2014" and "2016" tags as the focus concept respectively. As can be seen, there are no major differences with regards to the climatic indices, displayed in blue and green, but it does appear that the 2016 grape growing period had on average, higher wind speeds than 2014. This can be derived from the "Light Air" and "Light Breeze" tags appearing larger in Figure 4.14 than in Figure 4.13.

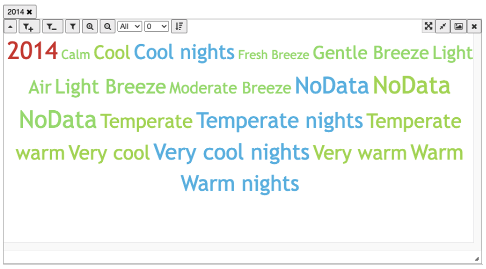

2014 ✖
All
0
2014 Calm Cool Cool nights Fresh Breeze Gentle Breeze Light
Air Light Breeze Moderate Breeze NoData NoData
NoData Temperate Temperate nights Temperate
warm Very cool Very cool nights Very warm Warm
Warm nights

Figure 4.13: Tag cloud with "2014" as the focus concept

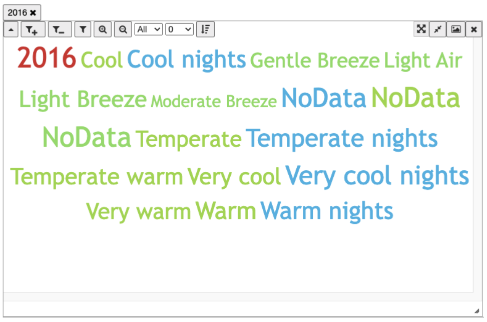

2016 ✖
All
0
2016 Cool Cool nights Gentle Breeze Light Air
Light Breeze Moderate Breeze NoData NoData
NoData Temperate Temperate nights
Temperate warm Very cool Very cool nights
Very warm Warm Warm nights

Figure 4.14: Tag cloud with "2016" as the focus concept

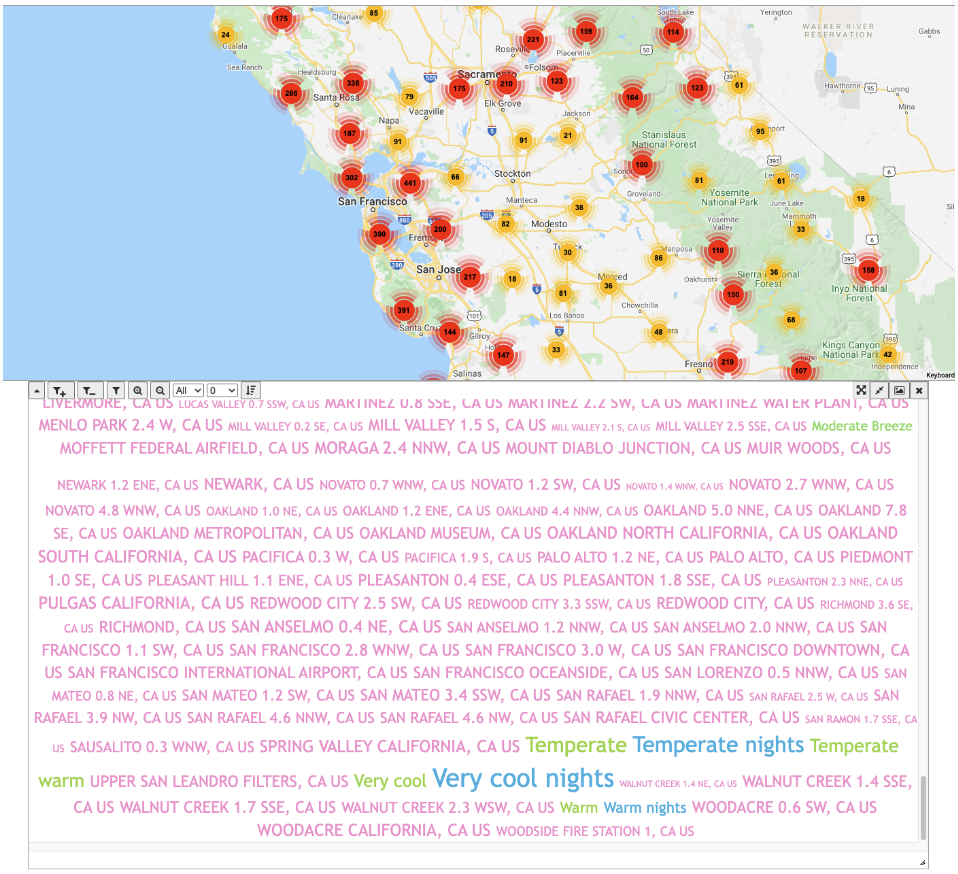
In a similar fashion, the graph viewer can also be used with this temporal filter to explore the data. This can be achieved by selecting the "Period" tag category in the first drop down menu and then visualising the occurrence of different tag categories for that period by changing the selection in the second drop down menu.

4.5.3.5 Bicluster Functionality

In order to get a better view of the climate indicators from stations in certain areas, the user can make use of the biclustering tool on the map viewer. The user can generate this bicluster tag cloud by right-clicking either a single or multiple pin clusters as well as individual pins on the map.

As mentioned in Section 3.3, a prevalent issue found in geolocation data sets is the limiting effect of pre-imposed borders on the data, usually political or socio-economic in nature. The biclustering functionality in Conceptcloud circumvents these restrictions. Figure 4.15 and Figure 4.16 shows a comparison of the tag clouds, appearing below the map viewer, generated from pin clusters around San Francisco and Los Angeles respectively.

In order to generate the San Francisco bicluster, the map was zoomed in on the city to a relatively low level, until the pin clusters grouped into small enough clusters around the city borders (As discussed in Section 3.3.1, the pins are re-clustered automatically based on the current zoom level of the map). Right clicking all of the pin clusters in the city as well as the clusters touching the city borders and then clicking the "Generate Bicluster" button to the left of the map viewer resulted in the tag cloud seen in Figure 4.15



LIVERMORE, CA US LUCAS VALLEY 0.7 SSW, CA US MARTINEZ 0.8 SSE, CA US MARTINEZ 2.2 SW, CA US MARTINEZ WATER PLANT, CA US MENLO PARK 2.4 W, CA US MILL VALLEY 0.2 SE, CA US MILL VALLEY 1.5 S, CA US MILL VALLEY 2.1 S, CA US MILL VALLEY 2.5 SSE, CA US Moderate Breeze MOFFETT FEDERAL AIRFIELD, CA US MORAGA 2.4 NNW, CA US MOUNT DIABLO JUNCTION, CA US MUIR WOODS, CA US

NEWARK 1.2 ENE, CA US NEWARK, CA US NOVATO 0.7 WNW, CA US NOVATO 1.2 SW, CA US NOVATO 1.4 WNW, CA US NOVATO 2.7 WNW, CA US NOVATO 4.8 WNW, CA US OAKLAND 1.0 NE, CA US OAKLAND 1.2 ENE, CA US OAKLAND 4.4 NNW, CA US OAKLAND 5.0 NNE, CA US OAKLAND 7.8 SE, CA US OAKLAND METROPOLITAN, CA US OAKLAND MUSEUM, CA US OAKLAND NORTH CALIFORNIA, CA US OAKLAND SOUTH CALIFORNIA, CA US PACIFICA 0.3 W, CA US PACIFICA 1.9 S, CA US PALO ALTO 1.2 NE, CA US PALO ALTO, CA US PIEDMONT 1.0 SE, CA US PLEASANT HILL 1.1 ENE, CA US PLEASANTON 0.4 ESE, CA US PLEASANTON 1.8 SSE, CA US PLEASANTON 2.3 NNE, CA US PULGAS CALIFORNIA, CA US REDWOOD CITY 2.5 SW, CA US REDWOOD CITY 3.3 SSW, CA US REDWOOD CITY, CA US RICHMOND 3.6 SE, CA US RICHMOND, CA US SAN ANSELMO 0.4 NE, CA US SAN ANSELMO 1.2 NNW, CA US SAN ANSELMO 2.0 NNW, CA US SAN FRANCISCO 1.1 SW, CA US SAN FRANCISCO 2.8 WNW, CA US SAN FRANCISCO 3.0 W, CA US SAN FRANCISCO DOWNTOWN, CA US SAN FRANCISCO INTERNATIONAL AIRPORT, CA US SAN FRANCISCO OCEANSIDE, CA US SAN LORENZO 0.5 NNW, CA US SAN MATEO 0.8 NE, CA US SAN MATEO 1.2 SW, CA US SAN MATEO 3.4 SSW, CA US SAN RAFAEL 1.9 NNW, CA US SAN RAFAEL 2.5 W, CA US SAN RAFAEL 3.9 NW, CA US SAN RAFAEL 4.6 NNW, CA US SAN RAFAEL 4.6 NW, CA US SAN RAFAEL CIVIC CENTER, CA US SAN RAMON 1.7 SSE, CA US SAUSALITO 0.3 WNW, CA US SPRING VALLEY CALIFORNIA, CA US Temperate Temperate nights Temperate warm UPPER SAN LEANDRO FILTERS, CA US Very cool Very cool nights WALNUT CREEK 1.4 NE, CA US WALNUT CREEK 1.4 SSE, CA US WALNUT CREEK 1.7 SSE, CA US WALNUT CREEK 2.3 WSW, CA US Warm Warm nights WOODACRE 0.6 SW, CA US WOODACRE CALIFORNIA, CA US WOODSIDE FIRE STATION 1, CA US

Figure 4.15: San Francisco Bicluster

The San Francisco bicluster was generated from 1333 unique objects/pins, and similar to the data set as a whole, "Very cool nights", appearing in blue text, is still the most frequently occurring tag. "2016" had the most measurements compared to the other periods. "Light breeze" was the most frequently occurring average wind speed and "Temperate" was the modal Heliothermal Index class.

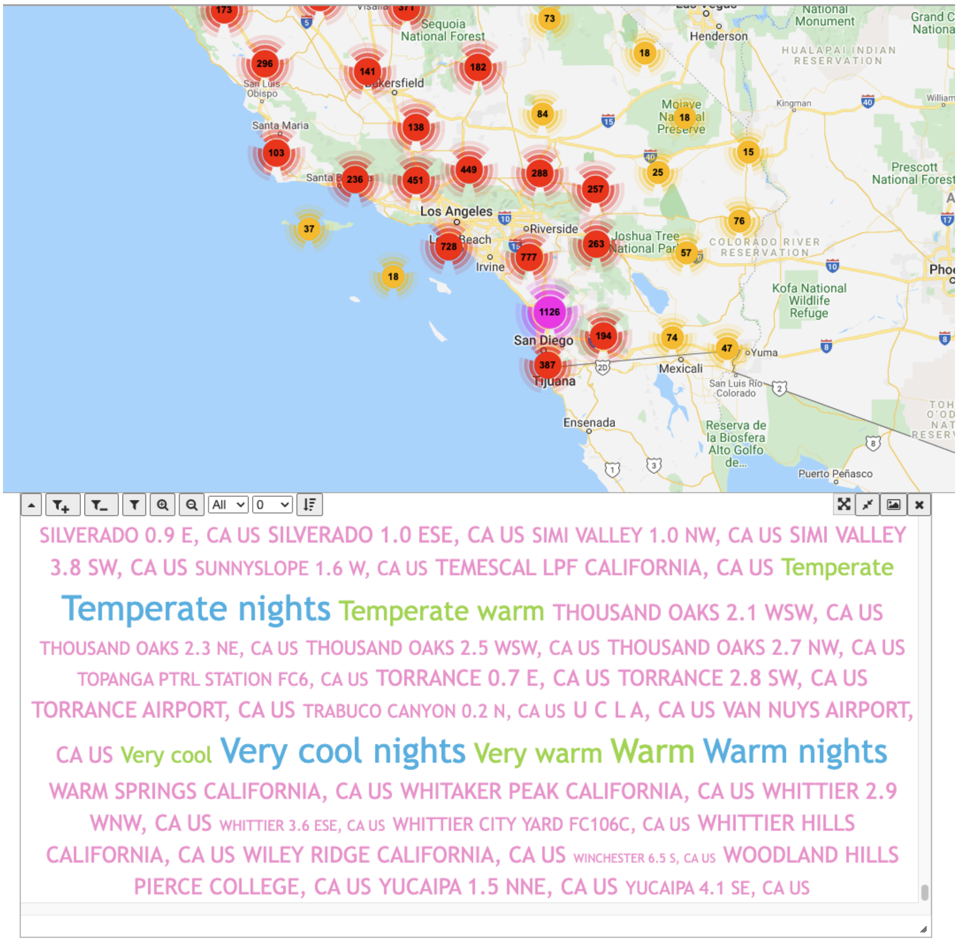

Sequoia National Forest
Henderson
San Luis Obispo
Bakersfield
Santa Maria
Mojave National Preserve
Kingman
Los Angeles
Riverside
Long Beach
Irvine
Joshua Tree National Park
COLORADO RIVER RESERVATION
Kofa National Wildlife Refuge
San Diego
Tijuana
Mexicali
Yuma
Ensenada
Puerto Peñasco
173
296
141
182
73
18
138
84
103
236
451
449
288
257
25
15
37
76
728
777
263
57
18
1126
194
74
47
387
All
0
SILVERADO 0.9 E, CA US SILVERADO 1.0 ESE, CA US SIMI VALLEY 1.0 NW, CA US SIMI VALLEY 3.8 SW, CA US SUNNYSLOPE 1.6 W, CA US TEMESCAL LPF CALIFORNIA, CA US Temperate
Temperate nights Temperate warm THOUSAND OAKS 2.1 WSW, CA US
THOUSAND OAKS 2.3 NE, CA US THOUSAND OAKS 2.5 WSW, CA US THOUSAND OAKS 2.7 NW, CA US
TOPANGA PTRL STATION FC6, CA US TORRANCE 0.7 E, CA US TORRANCE 2.8 SW, CA US
TORRANCE AIRPORT, CA US TRABUCO CANYON 0.2 N, CA US U C L A, CA US VAN NUYS AIRPORT, CA US Very cool Very cool nights Very warm Warm Warm nights
WARM SPRINGS CALIFORNIA, CA US WHITAKER PEAK CALIFORNIA, CA US WHITTIER 2.9 WNW, CA US WHITTIER 3.6 ESE, CA US WHITTIER CITY YARD FC106C, CA US WHITTIER HILLS CALIFORNIA, CA US WILEY RIDGE CALIFORNIA, CA US WINCHESTER 6.5 S, CA US WOODLAND HILLS PIERCE COLLEGE, CA US YUCAIPA 1.5 NNE, CA US YUCAIPA 4.1 SE, CA US

Figure 4.16: Los Angeles Bicluster

As for the Los Angeles bicluster, it was generated in the same way as the San Francisco bicluster, from 1974 pins appearing in and around the city. Interestingly, when considering the Cool night Index class, "Temperate nights' was the most frequently occurring, as well as "Warm" for the Heliothermal Index. This indicates a slightly warmer climate when compared to San Francisco.

4.5.4 Wine Review Data Set

The second data set used in the evaluation, which is also publicly available, was downloaded from Kaggle [1]. This web portal allows users to share large data sets for machine learning exercises and other statistical analyses. The data set

[1] www.kaggle.com

consists of wine reviews published on the Wine Enthusiast Magazine's website [2], a popular publication among wine drinkers.

These reviews were collected during the week of 22 November 2017 and contains 129975 wine reviews from various countries. The data describes information about the wine such as the review score, variety, winery, description and country of origin among others. The full data object consists of the ten fields below:

- Points: The number of points WineEnthusiast rated the wine on a scale of 1-100 (WineEnthusiast only publishes reviews for wines that achieve a score of 80 or higher).
- Title: The title of the wine review, which often contains the vintage of the wine.
- Variety: The type of grapes used to make the wine (ie. Sauvignon Blanc).
- Description: A sommelier's description of the wine's taste, smell, look and feel.
- Country: The wine's country of origin.
- Province: The province or state that the wine is from.
- Region 1: The wine growing area in a province or state (ie Napa).
- Region 2: The more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley).
- Winery: The winery that made the wine.
- Designation: The vineyard within the winery where the grapes that made the wine are from.
- Price: The cost for a bottle of the wine in USD ($).
- Taster Name: The name of the wine reviewer.
- Taster Twitter Handle: Twitter handle for the wine reviewer.

4.5.5 Data Cleaning

An important step when processing any data, especially large amounts of uncleaned data from publicly available sources, is to perform a list of standardised checks in order to ensure the data is of good quality [19]. Good quality data is valid, accurate, complete, consistent and uniform. Data completeness, accuracy, consistency and uniformity is relatively simple to check but difficult to

[2] www.winemag.com

correct through the data cleaning process and must be achieved earlier in the data capturing process. Other factors such as validity and completeness can be assessed and rectified as required [19].

As a first step in checking data validity, the data was checked for duplicate observations and all duplicates were removed. Further, a check was done to ensure that all variables were of the correct type ie. numeric data is of type number, categorical data adheres to the expected categories and so forth. During this check, data fields of the wrong type were corrected. The data set also contained a number of non-ASCII characters in the "Descriptions", "Varietals" and "Wineries" which had to be stripped before continuing with any further data transformation. The next step was a check for data completeness, where again there were multiple records found with missing fields. These fields were assigned a "NoData" value. Data accuracy and consistency and uniformity were all at an acceptable level.

The final step in the data transformation process involves the "Description" field in the data object which consists of multiple sentences (See Figure 4.17). It can be used as is, but becomes much more effective in the tag cloud visualisation when split into smaller phrases. This can be accomplished by utilising ConceptCloud's key-phrase extraction functionality. Key phrase extraction is one of many natural language processing techniques used to identify important concepts in strings of text. Other techniques such as filtering of common stop word, stemming, n-gramming and lemmatisation all play a role in determining important phrases.

```
{
    "id": 1,
    "Region1": "Napa Valley",
    "Region2": "Napa",
    "Designation": "Martha’s Vineyard",
    "Points": 96.0,
    "Price": 235.0,
    "Variety": "Cabernet Sauvignon",
    "Winery": "Heitz",
    "Description": "This tremendous 100% varietal wine
                    hails from Oakville and was aged
                    over three years in oak. Juicy
                    red-cherry fruit and a compelling
                    hint of caramel greet the palate,
                    framed by elegant, fine tannins
                    and a subtle minty tone in the
                    background. Balanced and
                    rewarding from start to finish,
                    it has years ahead of it to
                    develop further nuance.
                    Enjoy 20222030."
}
```

Figure 4.17: Wine review JSON object with “Description” attribute

4.5.6 Natural Language Processing

When processing natural language, there are a number of challenges to consider depending on the data in question. The first of these relate to ambiguity found in natural language. Allen 2003, identifies five types of ambiguity.

These are:

- Simple lexical ambiguity which refers to words that are both nouns and verbs, i.e. “duck” being either an animal or the action of bending down.

- Structural or syntactic ambiguity, which refers to cases where it is unclear what information pertains to which subject in a sentence, i.e.“I saw a man with a telescope”.

- Semantic ambiguity, referring to words with multiple definitions, i.e. the word “Go” which, depending on context, has more than 10 definitions in any dictionary.

- Pragmatic ambiguity which refers to sentences that can be interpreted as either a question or a request, i.e. "Can you lift that rock?"
- Referential ambiguity, which refers to cases where it is not clear who or what the subject of the sentence is, i.e. "Jack met Sam at the station. He was feeling ill". In this case it is unclear who was feeling ill.

The interaction of these various ambiguities result in a complex interpretation process for the natural language processing system [5]. In order to avoid these ambiguities, some type of processing needs to be done which normalises and simplifies the text. This can be achieved through stemming or lemmatisation. Stemming is the process of reducing inflected, or derived words to their stem or root form, even if the stem itself is not a valid root. Lemmatisation on the other hand, is the process of grouping together the different inflected forms of a word so they can be analysed as a single item [11]. Lemmatisation as opposed to stemming, always normalises words to a valid root, namely the dictionary form of the word. Any natural language processing system will have to make use of these techniques in order to simplify input and remove

When considering how to identify important phrases in a piece of text, a paper by Brown et al. [12], provides insights into assigning words to classes according to statistical behaviour. Using statistical algorithms, the researchers assigned words to classes based on how frequently they occur together with other words. This results in classes that are either syntactically based groupings or semantically based groupings [12].

Brown et al. 1992, differentiate between two types of sticky pairs of words, basically 2-grams, one type which occur next to each other but are non-symmetric, and semantically sticky pairs which are symmetric and occur near each other only [12]. An example of a normal sticky pair is "Humpty Dumpty", but not "Dumpty Humpty as it is non-symmetric. "Dumpty Humpty" would be a new sticky pair, if it appeared in the text. Normal non-symmetric sticky pairs are useful for determining important phrases or key words in text.

Semantic pairs are more useful in creating classes of grouped together words having the same morphological stem, such as performance, performed, perform, performs, and performing, and others containing words that are related semantically but have different stems, such as attorney, counsel, trial, court, and judge [12]. This method of statistically analysing pairs of words to form classes and to determine sticky pairs, is a very effective method for determining the most important words and phrases in a piece of text.

4.5.7 Key-phrase Extraction

In order to split the text in the "Description" field into useful key phrases, ConceptCloud's key phrase extraction was used. The tool makes use of the

Stanford CoreNLP Java library [3] for natural language processing. This process consists of firstly splitting the text into individual sentences and removing any punctuation, escape characters and symbols that may interfere with language processing done later. The sentences are then further split into key-phrases or n-grams on which tokenisation, named-entity recognition and lemmatisation is performed. Although stemming is faster, it does not always reduce words into proper roots, for example "having" is stemmed to "hav". Lemmatisation considers the part of speech of a word when performing normalisation, which generally results in a better root [11]. Finally these phrases are added to each data object as additional attributes which will appear as tags in the tag cloud and can be explored the same as any other tag. The key-phrases extracted from the wine review shown in Figure 4.17 is shown in Figure 4.18.

```
{"Key-phrases": "compelling hint , further nuance ,
three years , caramel , year , subtle minty tone ,
start to finish , elegant , fine tannins ,
compelling hint of caramel ,
three years in oak , background ,
oak , Juicy red-cherry fruit , palate"}
```

Figure 4.18: Key-phrases extracted from the "Description" attribute in Figure 4.17

4.5.8 Combining Data Sets

As discussed in Section 4.2, in order to have a diverse and complex data set for the final evaluation, the wine review data set was combined with the climatic data set described in Section 4.5.1 above. This was achieved by adding a "County" field to both data sets. For the climatic data, the county was determined by the coordinates of the weather station. The climatic data was then averaged for all stations within a county and the climatic indices calculated using these new averaged measurements.

For the wine review data set, the "Region 1" and "Region 2" fields were used to determine the county of the wine. The region attributes describe the American Viticultural Area (AVA) of the wine in question. An AVA is a designated wine grape-growing region in the United States, providing an official appellation for the mutual benefit of wineries and consumers. As often the regions in which wine is produced can have a profound effect on the characteristics of the wine, it is vital for both winemakers as well as customers to know the geographic locations and its environment's potential impact. As it is required for a wine with an AVA label to be produced with 85% or more

[3] https://stanfordnlp.github.io/CoreNLP/

grapes grown within the AVA, often prices and customers can be influenced by different AVAs [13]. Figure 4.19 shows the various AVAs in California, with the darker shades of purple indicating overlapping AVAs. By combining the wine reviews with the climate data, only wine reviews for Californian wines could be used, where the region values were precise enough to assign the wine to a county. This meant that a total of 24028 wine objects were included in the final dataset.

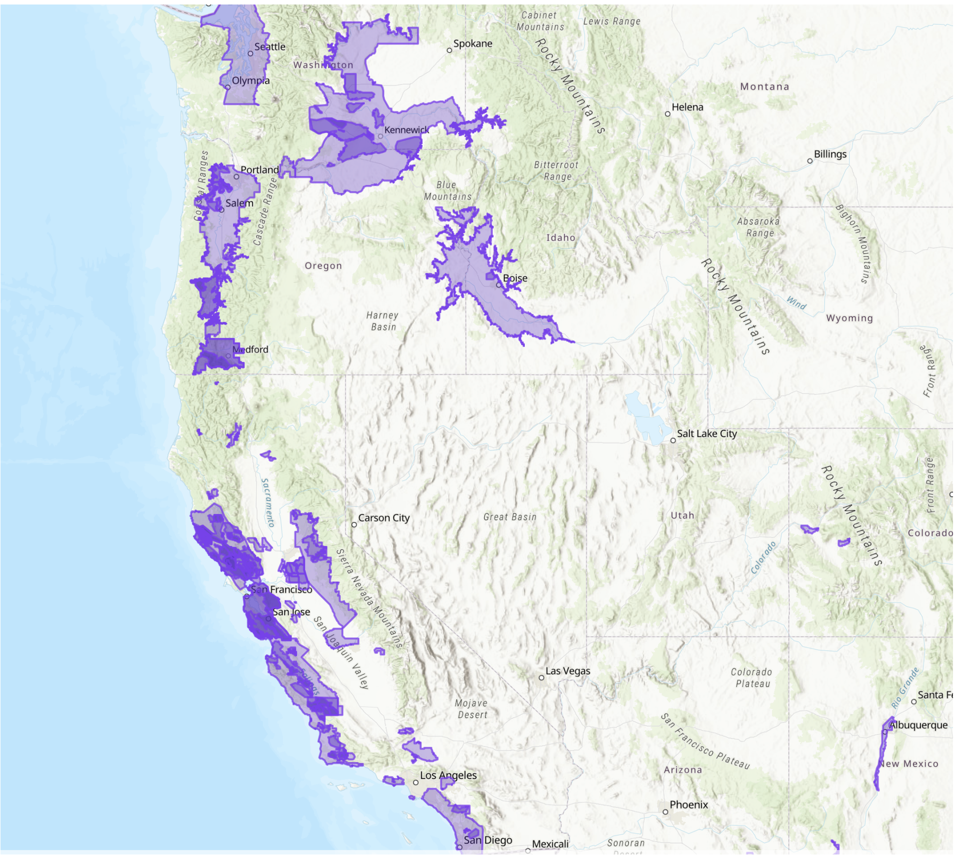


Figure 4.19: California AVAs

Since AVAs are defined by certain climatic factors, it would have been ideal to join the data sets on a common AVA instead of county but the reason for not doing so is twofold. Firstly, the size of the AVA's vary greatly when compared to counties, with the biggest AVA in California being the *Central Coast* AVA, spanning around 100,000 acres (400 km2) and includes portions of 6 counties [46]. Secondly, as there is overlap between AVAs as well as multiple levels of nesting, in many cases it is difficult to assign a single AVA to a wine. The state of California alone contains 139 AVAs as opposed to only 58 counties. Thus, the county was chosen as the joining attribute between the two data sets. It was also chosen as the geolocation aspect, with the centre coordinates of the county, based on US census data [4], being used in the final data object.

The final data object, shown in Figure 4.20, has numeric, textual and geolocation aspects which can be fully explored and mined by ConceptCloud in order to gain insight into the interplay between climatic factors and wine characteristics. The combined data object now contains the following fields:

```
{
    {
        "id": 2,
        "Location": "38.532574 -122.945194",
        "County": "sonoma",
        "CountyArea": 4578.96,
        "HI2014": "Temperate warm",
        "HI2015": "Temperate warm",
        "HI2016": "Temperate warm",
        "CI2014": "Very cool nights",
        "CI2015": "Very cool nights",
        "CI2016": "Very cool nights",
        "DI2014": "Moderately dry",
        "DI2015": "Moderately dry",
        "DI2016": "Moderately dry",
        "Region1": "Knights Valley",
        "Region2": "Sonoma",
        "Description1": "toasted hazelnut",
        "Description2": "pear compote",
        "Description3": "orange peel",
        "Designation": "Special Selected Late Harvest",
        "Points": 96.0,
        "Price": 90.0,
        "Variety": "Sauvignon Blanc",
        "Winery": "Macauley"
    }
},
```

Figure 4.20: Combined JSON Object

[4] https://census.gov

4.5.9 Limitations of the Data Set

Unlike the previous set, the wine becomes the formal object in this data set. This means that the geolocation of each wine had to be approximated based on the centre coordinates of its county, which in turn had to be approximated based on its region attributes. This is not ideal since counties are not determined by natural borders but rather by political legislature, and the climate in any one county can differs greatly depending on the latitude, topography, elevation, nearby bodies of water, ocean currents and prevailing winds. This approximation also means the climatic indices have to be averaged across counties per grape growing period. This results in each wine having an attribute per year for the three climatic indices. The county surface area is included as an attribute to give an indication of how accurate the climate indices potentially are.

Further, there is unfortunately no consistent way to determine the vintage of the wines, but sometimes it is included in the title. The data set was created in November 2017, so it can be assumed that these wines all have a vintage of at least 2017 or earlier. The climatic indices are available for the grape growing regions of 2014 through to 2016, which might not directly relate to the period some of the grapes used to produce these wines were grown, but should provide a relatively accurate indicator of the climate at the time. Further, we also lose the temporal filtering possible in the previous data set, since the wine data object lacks a "Vintage" attribute.

Finally, it is important to note that this data set is biased towards high scoring wines due to the original wine review data set's limitation on review scores below 80. This means the scores and varietals in the data set might not be representative of all wines produced in the state.

4.5.10 Case Study 2: Wine Review Data

It should be re-iterated that this evaluation is for demonstration purposes and the primary goal is not to gain new insight from the data but rather determine how well the tool facilitates data exploration and knowledge discovery. This evaluation is again presented alongside a series of screenshots of the different viewers in ConceptCloud's user interface.

4.5.10.1 California Wine Overview

California is responsible for 90% of all wine produced in North America, and there are over 1200 wineries in the state [46]. The majority of vines are grown ranging from Mendocino County to the southwestern tip of Riverside County. There are over 100 different grape varieties grown in California, with the four leading varieties in descending order being [46]:

- Cabernet Sauvignon

- Chardonnay
- Merlot
- Pinot noir

California's warm climate allows many wineries to make use of very ripe fruits in the wine making process, which promotes the fruit flavours as opposed to the earthy and mineral tones found in European wines. This also creates the opportunity for higher alcohol levels in the wines [46].

4.5.10.2 Initial View

After completing the data transformation processes described in Sections 4.5.5, 4.5.7 and 4.5.8, a single JSON file was generated by the custom data processing script. This JSON file was then loaded into ConceptCloud through the standard data configuration script and process discussed in Section 2.5.1.1, resulting in the view shown in Figure 4.21, after loading the user interface.

From the initial cloud, seen in Figure 4.21, it is clear the data set contains a wide variety of wines ranging from "Bordeaux-style Red Blend" to "Zinfandel". Of these, the tags for "Chardonnay", "Cabernet Sauvignon" and "Pinot Noir" appear to be the largest, which is in line with expectation. The "Merlot", although present, is not as large as the other three tags mentioned. This view gives the user an immediate view of the varietals that appear most frequently in the set. This might be due to the fact that California produces more of these varietals than any other, or that these varietals are generally reviewed higher than other varietals produced in the region.

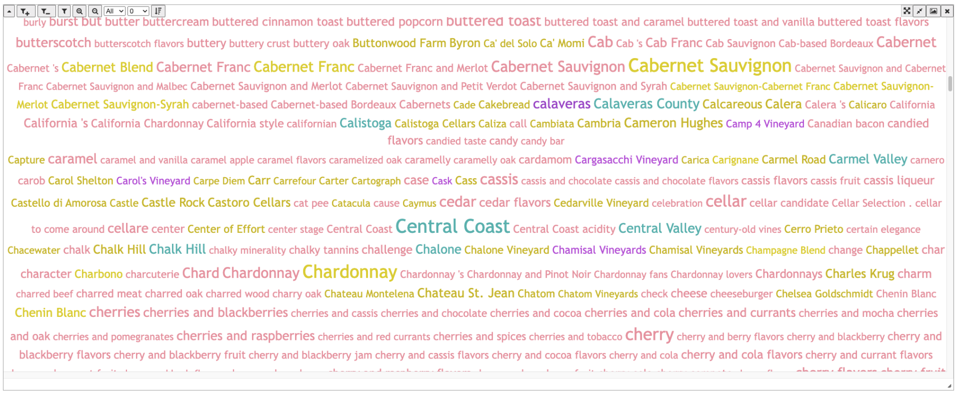

All
0
burly burst but butter buttercream buttered cinnamon toast buttered popcorn buttered toast buttered toast and caramel buttered toast and vanilla buttered toast flavors
butterscotch butterscotch flavors buttery buttery crust buttery oak Buttonwood Farm Byron Ca' del Solo Ca' Momi Cab Cab 's Cab Franc Cab Sauvignon Cab-based Bordeaux Cabernet
Cabernet 's Cabernet Blend Cabernet Franc Cabernet Franc Cabernet Franc and Merlot Cabernet Sauvignon Cabernet Sauvignon Cabernet Sauvignon and Cabernet
Franc Cabernet Sauvignon and Malbec Cabernet Sauvignon and Merlot Cabernet Sauvignon and Petit Verdot Cabernet Sauvignon and Syrah Cabernet Sauvignon-Cabernet Franc Cabernet Sauvignon-
Merlot Cabernet Sauvignon-Syrah cabernet-based Cabernet-based Bordeaux Cabernets Cade Cakebread calaveras Calaveras County Calcareous Calera Calera 's Calicaro California
California 's California Chardonnay California style californian Calistoga Calistoga Cellars Caliza call Cambiata Cambria Cameron Hughes Camp 4 Vineyard Canadian bacon candied
flavors candied taste candy candy bar
Capture caramel caramel and vanilla caramel apple caramel flavors caramelized oak caramelly caramelly oak cardamom Cargasacchi Vineyard Carica Carignane Carmel Road Carmel Valley carnero
carob Carol Shelton Carol's Vineyard Carpe Diem Carr Carrefour Carter Cartograph case Cask Cass cassis cassis and chocolate cassis and chocolate flavors cassis flavors cassis fruit cassis liqueur
Castello di Amorosa Castle Castle Rock Castoro Cellars cat pee Catacula cause Caymus cedar cedar flavors Cedarville Vineyard celebration cellar cellar candidate Cellar Selection . cellar
to come around cellare center Center of Effort center stage Central Coast Central Coast Central Coast acidity Central Valley century-old vines Cerro Prieto certain elegance
Chacewater chalk Chalk Hill Chalk Hill chalky minerality chalky tannins challenge Chalone Chalone Vineyard Chamisal Vineyards Chamisal Vineyards Champagne Blend change Chappellet char
character Charbono charcuterie Chard Chardonnay Chardonnay Chardonnay 's Chardonnay and Pinot Noir Chardonnay fans Chardonnay lovers Chardonnays Charles Krug charm
charred beef charred meat charred oak charred wood charry oak Chateau Montelena Chateau St. Jean Chatom Chatom Vineyards check cheese cheeseburger Chelsea Goldschmidt Chenin Blanc
Chenin Blanc cherries cherries and blackberries cherries and cassis cherries and chocolate cherries and cocoa cherries and cola cherries and currants cherries and mocha cherries
and oak cherries and pomegranates cherries and raspberries cherries and red currants cherries and spices cherries and tobacco cherry cherry and berry flavors cherry and blackberry cherry and
blackberry flavors cherry and blackberry fruit cherry and blackberry jam cherry and cassis flavors cherry and cocoa flavors cherry and cola cherry and cola flavors cherry and currant flavors

Figure 4.21: Initial tag cloud of the wine review data set

As for the most frequently occurring counties, "Sonoma" and "Napa" appear to be the largest. The "Central Coast" region also appears very largely in the tag cloud. This is expected as the Central Coast AVA is the largest in the state, spanning 400 km2 and producing 75% of the state's wine [46]. These tags also provides a clear indication of where the majority of the map pins will appear. This is confirmed by looking at the initial map view, shown in Figure 4.22, where most of the markers do indeed appear in the centre of the state, around the Napa Valley and Central Coast regions. The purple cluster is the largest and consists of 14282 pins.

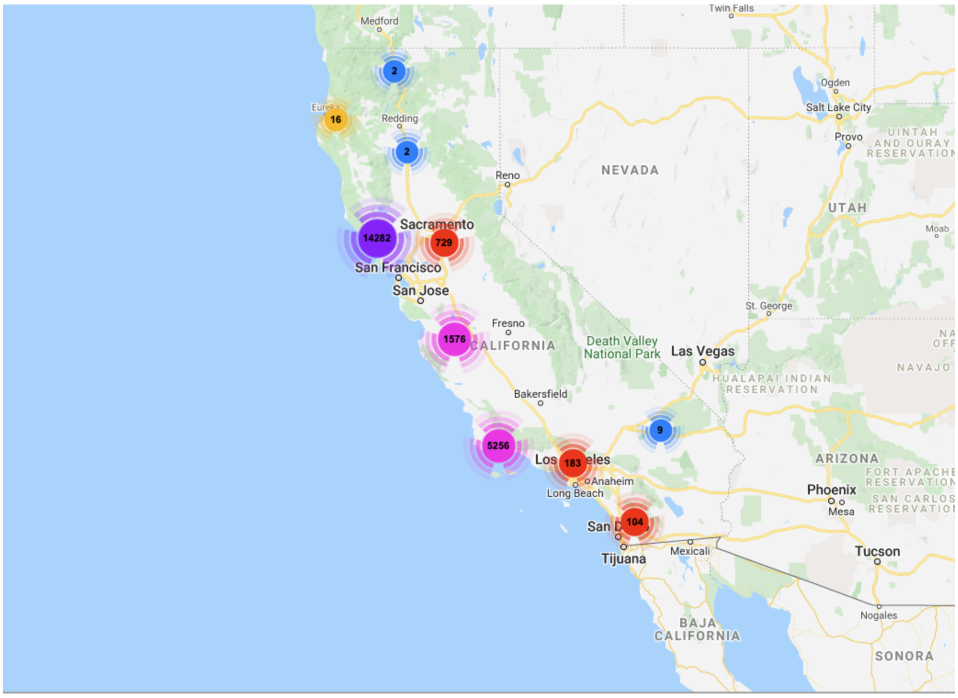
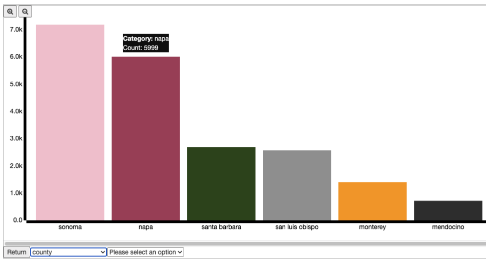


Figure 4.22: Initial map of the wine review data set

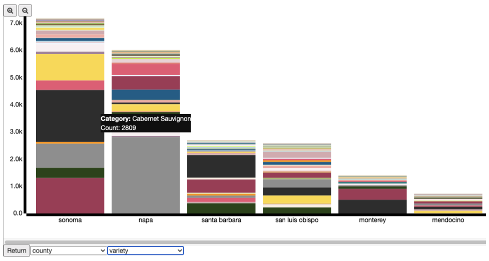
Again, as with the first data set, utilising the graph viewer can give the user a high level view of the composition of the data. By selecting "County" in the first drop down menu, the user is provided with the visualisation seen in Figure 4.23. Here it can be seen that most of the wine reviews fall within Sonoma, the pink bar, and Napa county, the maroon bar. This view can be drilled down further by selecting "Variety" in the second drop down list and generating the view shown in in Figure 4.24. This allows the user to quickly determine that most of the reviews in Sonoma are for the Pinot Noir varietal, the charcoal segment, and for Napa it is Cabernet Sauvignon, the grey segment .



Category: napa
Count: 5999

county

Figure 4.23: Bar graph view of the "County" tag category



Category: Cabernet Sauvignon
Count: 2809

county
variety

Figure 4.24: Drilled down bar graph view of the "Variety" tag category in terms of "County"

4.5.10.3 Tag and Map Exploration

With some insight gained into the composition of the data set, data exploration can now be conducted. By clicking the maroon "Chardonnay" tag, we get the view displayed in Figure 4.25. The map view provides us with the location distribution of this varietal. By generating the pie chart for this varietal, drilled down by points, we get three different perspectives on this data attribute allowing for very low level examination of this varietal in the data set. The pie chart can also be drilled down by any other tag category as desired.

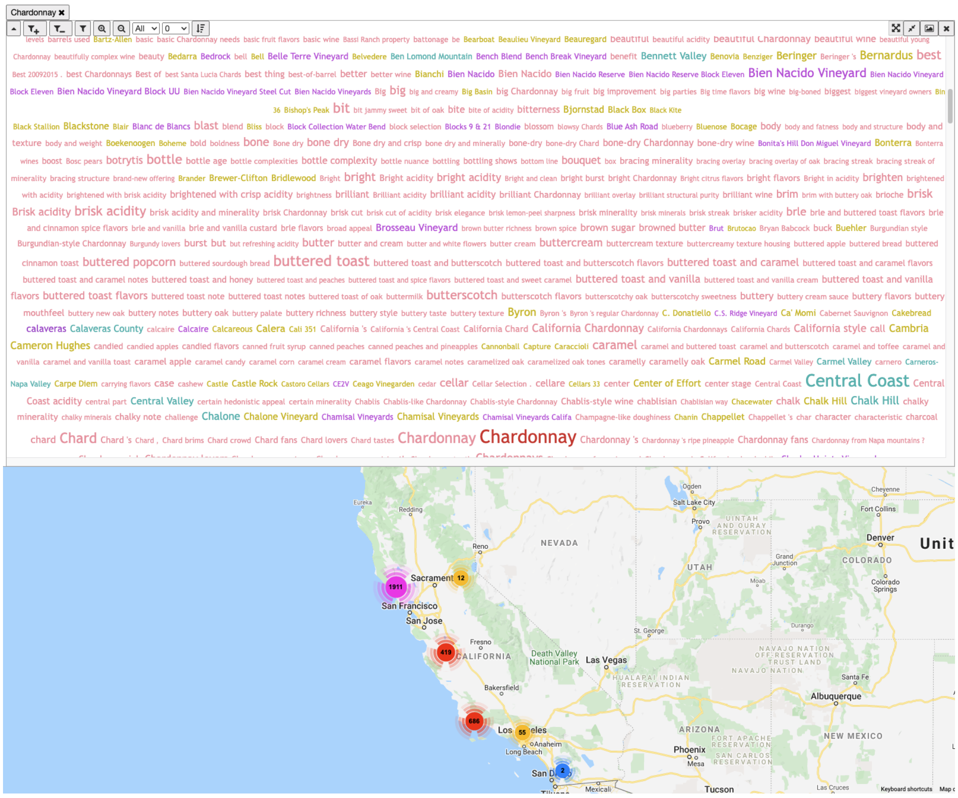

Chardonnay
All
0
Bartz-Allen basic basic Chardonnay needs basic fruit flavors basic wine Bassi Ranch property battonage be Bearboat Beaulieu Vineyard Beauregard beautiful beautiful acidity beautiful Chardonnay beautiful wine beautiful young
Chardonnay beautifully complex wine beauty Bedarra Bedrock bell Bell Belle Terre Vineyard Belvedere Ben Lomond Mountain Bench Blend Bench Break Vineyard benefit Bennett Valley Benovia Benziger Beringer Beringer 's Bernardus best
Best 20092015 . best Chardonnays Best of best Santa Lucia Chards best thing best-of-barrel better better wine Bianchi Bien Nacido Bien Nacido Bien Nacido Reserve Bien Nacido Reserve Block Eleven Bien Nacido Vineyard Bien Nacido Vineyard
Block Eleven Bien Nacido Vineyard Block UU Bien Nacido Vineyard Steel Cut Bien Nacido Vineyards Big big big and creamy Big Basin big Chardonnay big fruit big improvement big parties Big time flavors big wine big-boned biggest biggest vineyard owners Bin
36 Bishop's Peak bit bit jammy sweet bit of oak bite bite of acidity bitterness Bjornstad Black Box Black Kite
Black Stallion Blackstone Blair Blanc de Blancs blast blend Bliss block Block Collection Water Bend block selection Blocks 9 & 21 Blondie blossom blowsy Chards Blue Ash Road blueberry Bluenose Bocage body body and fatness body and structure body and
texture body and weight Boekenoogen Boheme bold boldness bone Bone dry bone dry Bone dry and crisp bone dry and minerally bone-dry bone-dry Chard bone-dry Chardonnay bone-dry wine Bonita's Hill Don Miguel Vineyard Bonterra Bonterra
wines boost Bosc pears botrytis bottle bottle age bottle complexities bottle complexity bottle nuance bottling bottling shows bottom line bouquet box bracing minerality bracing overlay bracing overlay of oak bracing streak bracing streak of
minerality bracing structure brand-new offering Brander Brewer-Clifton Bridlewood Bright bright Bright acidity bright acidity Bright and clean bright burst bright Chardonnay Bright citrus flavors bright flavors Bright in acidity brighten brightened
with acidity brightened with brisk acidity brightened with crisp acidity brightness brilliant Brilliant acidity brilliant acidity brilliant Chardonnay brilliant overlay brilliant structural purity brilliant wine brim brim with buttery oak brioche brisk
Brisk acidity brisk acidity brisk acidity and minerality brisk Chardonnay brisk cut brisk cut of acidity brisk elegance brisk lemon-peel sharpness brisk minerality brisk minerals brisk streak brisker acidity brle brle and buttered toast flavors brle
and cinnamon spice flavors brle and vanilla brle and vanilla custard brle flavors broad appeal Brosseau Vineyard brown butter richness brown spice brown sugar browned butter Brut Brutocao Bryan Babcock buck Buehler Burgundian style
Burgundian-style Chardonnay Burgundy lovers burst but but refreshing acidity butter butter and cream butter and white flowers butter cream buttercream buttercream texture buttercreamy texture housing buttered apple buttered bread buttered
cinnamon toast buttered popcorn buttered sourdough bread buttered toast buttered toast and butterscotch buttered toast and butterscotch flavors buttered toast and caramel buttered toast and caramel flavors
buttered toast and caramel notes buttered toast and honey buttered toast and peaches buttered toast and spice flavors buttered toast and sweet caramel buttered toast and vanilla buttered toast and vanilla cream buttered toast and vanilla
flavors buttered toast flavors buttered toast note buttered toast notes buttered toast of oak buttermilk butterscotch butterscotch flavors butterscotchy oak butterscotchy sweetness buttery buttery cream sauce buttery flavors buttery
mouthfeel buttery new oak buttery notes buttery oak buttery palate buttery richness buttery style buttery taste buttery texture Byron Byron 's Byron 's regular Chardonnay C. Donatiello C.S. Ridge Vineyard Ca' Momi Cabernet Sauvignon Cakebread
calaveras Calaveras County calcaire Calcaire Calcareous Calera Cali 351 California 's California 's Central Coast California Chard California Chardonnay California Chardonnays California Chards California style call Cambria
Cameron Hughes candied candied apples candied flavors canned fruit syrup canned peaches canned peaches and pineapples Cannonball Capture Caraccioli caramel caramel and buttered toast caramel and butterscotch caramel and toffee caramel and
vanilla caramel and vanilla toast caramel apple caramel candy caramel corn caramel cream caramel flavors caramel notes caramelized oak caramelized oak tones caramelly caramelly oak Carmel Road Carmel Valley Carmel Valley carneros Carneros-
Napa Valley Carpe Diem carrying flavors case cashew Castle Castle Rock Castoro Cellars CE2V Ceago Vinegarden cedar cellar Cellar Selection . cellare Cellars 33 center Center of Effort center stage Central Coast Central Coast Central
Coast acidity central part Central Valley certain hedonistic appeal certain minerality Chablis Chablis-like Chardonnay Chablis-style Chardonnay Chablis-style wine chablisian Chablisian way Chacewater chalk Chalk Hill Chalk Hill chalky
minerality chalky minerals chalky note challenge Chalone Chalone Vineyard Chamisal Vineyards Chamisal Vineyards Chamisal Vineyards Califa Champagne-like doughiness Chanin Chappellet Chappellet 's char character characteristic charcoal
chard Chard Chard 's Chard , Chard brims Chard crowd Chard fans Chard lovers Chard tastes Chardonnay Chardonnay Chardonnay 's Chardonnay 's ripe pineapple Chardonnay fans Chardonnay from Napa mountains ?
Eureka
Redding
Reno
NEVADA
Sacramento
12
1911
San Francisco
San Jose
Fresno
419
CALIFORNIA
Death Valley National Park
Las Vegas
Bakersfield
686
Los Angeles
55
Anaheim
Long Beach
San Diego
2
Tijuana
Mexicali
Ogden
Salt Lake City
Provo
UINTAH AND OURAY RESERVATION
UTAH
Grand Junction
Moab
St. George
Cheyenne
Fort Collins
Denver
United
COLORADO
Colorado Springs
Durango
NAVAJO NATION OFF-RESERVATION TRUST LAND
NAVAJO NATION
Santa Fe
HUALAPAI INDIAN RESERVATION
Albuquerque
ARIZONA
FORT APACHE RESERVATION
Phoenix
Mesa
SAN CARLOS RESERVATION
NEW MEXICO
Tucson
Las Cruces
Keyboard shortcuts
Map

Figure 4.25: View of the"Chardonnay" tag

From the tag cloud, it can be seen that the phrases "Buttered toast" and "Buttered popcorn" are associated with this wine, among others. In terms of climate, it seems the "Very Cool Night" climate class for the Cool night Index and "Warm" for the Heliothermal Index are most commonly associated with this varietal. This holds true for all periods, from 2014 to 2016 and could indicate that highly reviewed Chardonnays are grown in areas with overall warm climates but having cool nights over the grape growing period. The "Russian River Valley" region seems to be the most popular region for growing highly rated Chardonnays. The map viewer shows a large cluster of pins north of San Francisco, which is where the Russian River Valley AVA is located. The graph viewer in Figure 4.26 shows that Chardonnays in the data set most frequently fall between the range of 82 to 94 points.

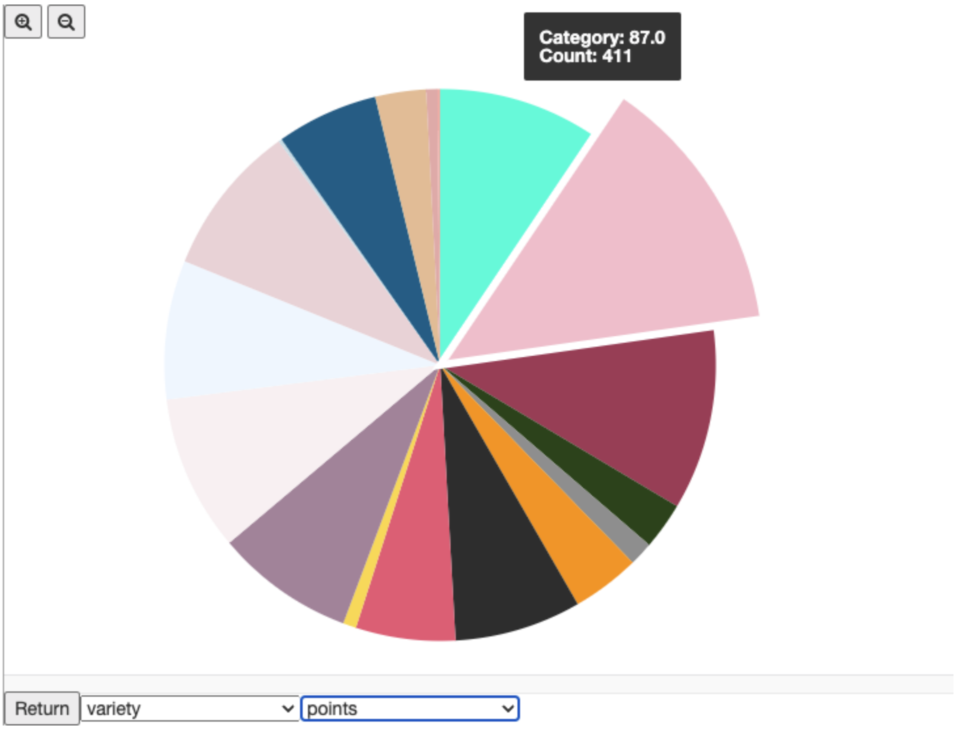
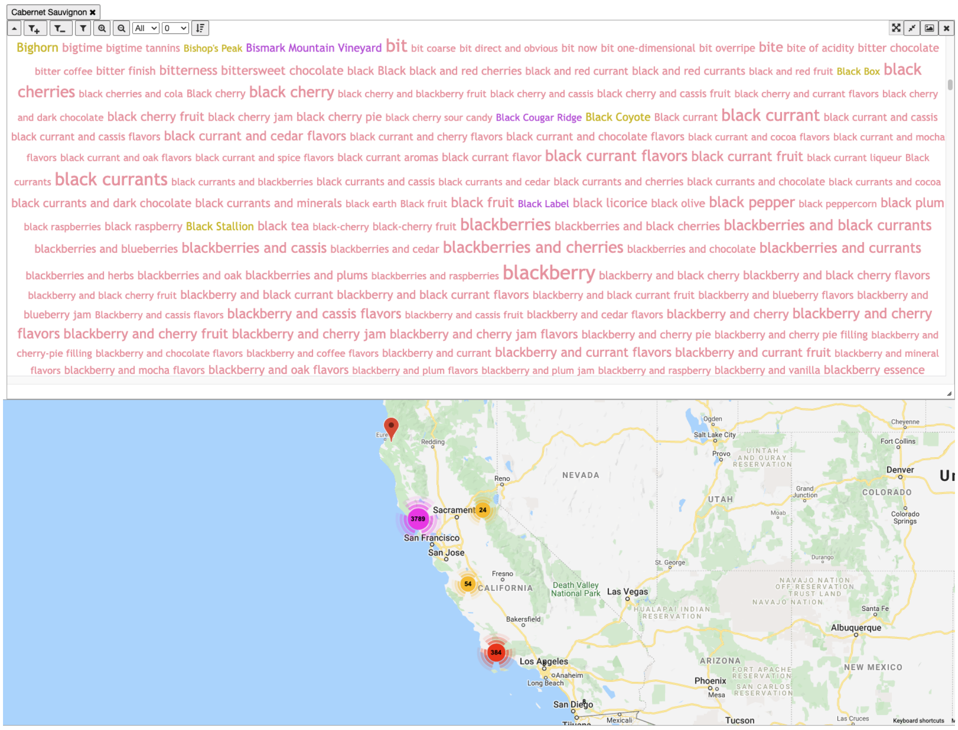


Figure 4.26: Pie chart of the“Chardonnay” varietal tag, drilled down by the “Points” category

Investigating another varietal, namely Cabernet Sauvignon, we are presented with the tag cloud shown in Figure 4.27, and it can be seen that the key-phrases “blackberry” and “blackcurrants”, shown in salmon coloured text, are frequently associated with highly rated wines of this varietal. This gives us an idea of the characteristics a highly reviewed Cabernet Sauvignon might possess. The “Napa valley”, and “Alexander Valley” region tags are the most prominent for this variety, and this is again supported by the map view showing large pin clusters to the north of San Francisco. This is also supported by what was seen in the graph viewer in Figure 4.24. As with Chardonnay, “Very cool night” was again the most frequently occurring Cool night Index class, but unlike Chardonnay, the “Temperate Warm” class was the modal Heliothermal Index.



Bighorn bigtime bigtime tannins Bishop's Peak Bismark Mountain Vineyard bit bit coarse bit direct and obvious bit now bit one-dimensional bit overripe bite bite of acidity bitter chocolate bitter coffee bitter finish bitterness bittersweet chocolate black Black black and red cherries black and red currant black and red currants black and red fruit Black Box black cherries black cherries and cola Black cherry black cherry black cherry and blackberry fruit black cherry and cassis black cherry and cassis fruit black cherry and currant flavors black cherry and dark chocolate black cherry fruit black cherry jam black cherry pie black cherry sour candy Black Cougar Ridge Black Coyote Black currant black currant black currant and cassis black currant and cassis flavors black currant and cedar flavors black currant and cherry flavors black currant and chocolate flavors black currant and cocoa flavors black currant and mocha flavors black currant and oak flavors black currant and spice flavors black currant aromas black currant flavor black currant flavors black currant fruit black currant liqueur Black currants black currants black currants and blackberries black currants and cassis black currants and cedar black currants and cherries black currants and chocolate black currants and cocoa black currants and dark chocolate black currants and minerals black earth Black fruit black fruit Black Label black licorice black olive black pepper black peppercorn black plum black raspberries black raspberry Black Stallion black tea black-cherry black-cherry fruit blackberries blackberries and black cherries blackberries and black currants blackberries and blueberries blackberries and cassis blackberries and cedar blackberries and cherries blackberries and chocolate blackberries and currants blackberries and herbs blackberries and oak blackberries and plums blackberries and raspberries blackberry blackberry and black cherry blackberry and black cherry flavors blackberry and black cherry fruit blackberry and black currant blackberry and black currant flavors blackberry and black currant fruit blackberry and blueberry flavors blackberry and blueberry jam Blackberry and cassis flavors blackberry and cassis flavors blackberry and cassis fruit blackberry and cedar flavors blackberry and cherry blackberry and cherry flavors blackberry and cherry fruit blackberry and cherry jam blackberry and cherry jam flavors blackberry and cherry pie blackberry and cherry pie filling blackberry and cherry-pie filling blackberry and chocolate flavors blackberry and coffee flavors blackberry and currant blackberry and currant flavors blackberry and currant fruit blackberry and mineral flavors blackberry and mocha flavors blackberry and oak flavors blackberry and plum flavors blackberry and plum jam blackberry and raspberry blackberry and vanilla blackberry essence

Figure 4.27: View of the "Cabernet Sauvignon" tag

Further, using the tag cloud we are able to determine which common attributes are possessed by the most expensive Cabernet Sauvignons, as well as what review scores they achieved. By having "Cabernet Sauvignon" as our focus concept, we can then click the "350" tag which is the highest price tag for this varietal. The tag cloud and map viewer is then updated to display the view seen in Figure 4.28. There are 4 wine reviews that share these attributes. Of these, they all share a very high rating of 94 points and are grown in Napa county. Upon further investigation, it is found that these wines are all produced by Lokoya winery, an exclusive winery in Napa valley, renowned for their Cabernet Sauvignons [5].

[5] https://www.lokoya.com/the-cellar

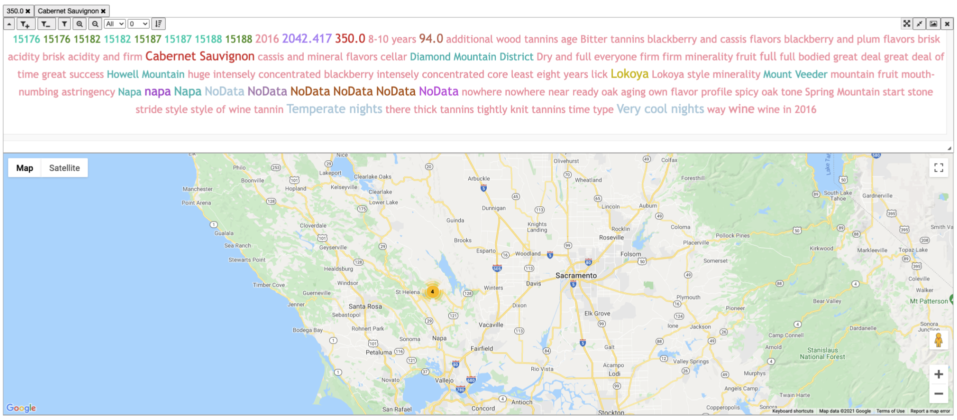

350.0 ✖
Cabernet Sauvignon ✖
All
0
15176 15176 15182 15182 15187 15187 15188 15188 2016 2042.417 350.0 8-10 years 94.0 additional wood tannins age Bitter tannins blackberry and cassis flavors blackberry and plum flavors brisk acidity brisk acidity and firm Cabernet Sauvignon cassis and mineral flavors cellar Diamond Mountain District Dry and full everyone firm firm minerality fruit full full bodied great deal great deal of time great success Howell Mountain huge intensely concentrated blackberry intensely concentrated core least eight years lick Lokoya Lokoya style minerality Mount Veeder mountain fruit mouth-numbing astringency Napa napa Napa NoData NoData NoData NoData NoData NoData nowhere nowhere near ready oak aging own flavor profile spicy oak tone Spring Mountain start stone stride style style of wine tannin Temperate nights there thick tannins tightly knit tannins time type Very cool nights way wine wine in 2016
Map
Satellite
Santa Rosa
Napa
Sacramento
St Helena
Vacaville
Vallejo
Stockton
Google

Figure 4.28: View of the most expensive Cabernet Sauvignons

4.5.10.4 Highest Reviewed Wines

By clicking the "99" tag belonging to the points tag category, we are able to get a view of the common attributes shared between exceptionally highly rated wines, shown in Figure 4.29. Doing so, shows that there is a correlation between high ratings and red varietals such as Cabernet Blend, Cabernet Sauvignon and Pinot Noir, appearing as gold text in the tag cloud. There also seems to be a correlation between these red wines and key phrases involving fruity flavours such as "raspberry" and "red currant". "Acidity" and "strong tannins" also appear among the phrases for these wines, which is in line with expectations around Californian wines. Further, the price of these wines are relatively high, ranging from 75 to 290 USD. Using the graph viewer it is possible to see that roughly only 10% of wines are more expensive than 75 USD, and only 2 are 290 USD in price.

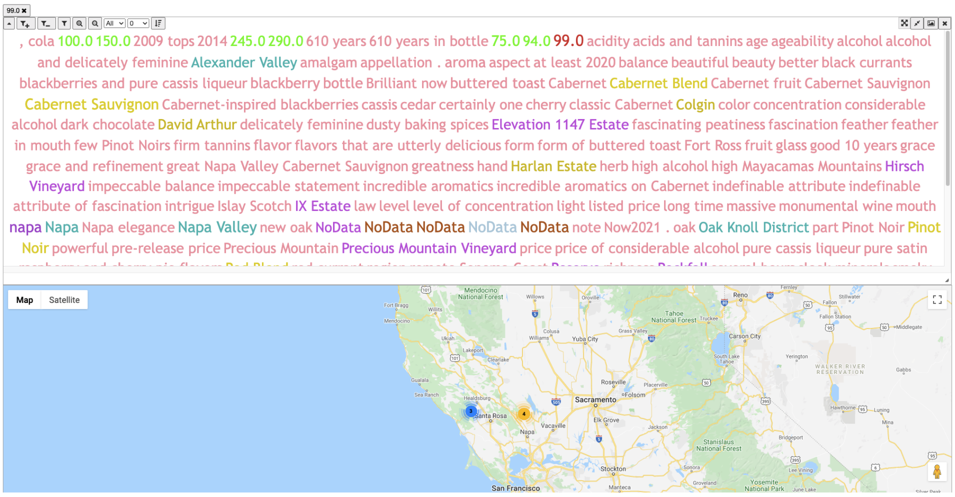
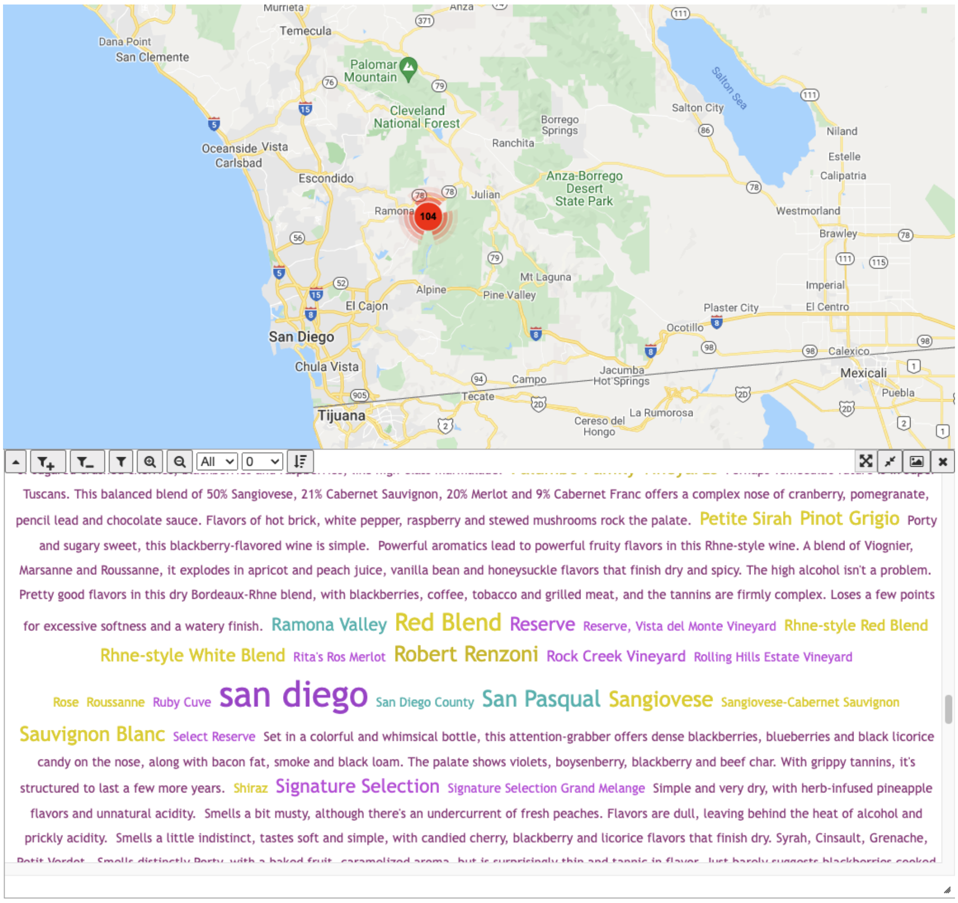

99.0 ✖
All
0
, cola 100.0 150.0 2009 tops 2014 245.0 290.0 610 years 610 years in bottle 75.0 94.0 99.0 acidity acids and tannins age ageability alcohol alcohol and delicately feminine Alexander Valley amalgam appellation . aroma aspect at least 2020 balance beautiful beauty better black currants blackberries and pure cassis liqueur blackberry bottle Brilliant now buttered toast Cabernet Cabernet Blend Cabernet fruit Cabernet Sauvignon Cabernet Sauvignon Cabernet-inspired blackberries cassis cedar certainly one cherry classic Cabernet Colgin color concentration considerable alcohol dark chocolate David Arthur delicately feminine dusty baking spices Elevation 1147 Estate fascinating peatiness fascination feather feather in mouth few Pinot Noirs firm tannins flavor flavors that are utterly delicious form form of buttered toast Fort Ross fruit glass good 10 years grace grace and refinement great Napa Valley Cabernet Sauvignon greatness hand Harlan Estate herb high alcohol high Mayacamas Mountains Hirsch Vineyard impeccable balance impeccable statement incredible aromatics incredible aromatics on Cabernet indefinable attribute indefinable attribute of fascination intrigue Islay Scotch IX Estate law level level of concentration light listed price long time massive monumental wine mouth napa Napa Napa elegance Napa Valley new oak NoData NoData NoData NoData NoData note Now2021 . oak Oak Knoll District part Pinot Noir Pinot Noir powerful pre-release price Precious Mountain Precious Mountain Vineyard price price of considerable alcohol pure cassis liqueur pure satin
Map
Satellite
Reno
Sacramento
Santa Rosa
Napa
Carson City
Stockton
San Francisco

Figure 4.29: Highest rated wines

4.5.10.5 Biclusters Exploration

Similarly to what was done during the evaluation on the climate data set, pin clusters in the map viewer can be used to generate new tag cloud visualisations. These in turn can be explored as per normal. To get a view of the wines produced in the southern parts of the state, a bicluster of 104 objects was used to generate the tag cloud shown in Figure 4.30.



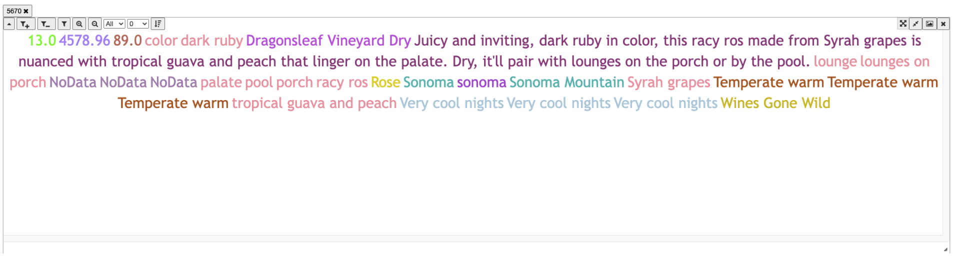
Tuscans. This balanced blend of 50% Sangiovese, 21% Cabernet Sauvignon, 20% Merlot and 9% Cabernet Franc offers a complex nose of cranberry, pomegranate, pencil lead and chocolate sauce. Flavors of hot brick, white pepper, raspberry and stewed mushrooms rock the palate. Petite Sirah Pinot Grigio Porty and sugary sweet, this blackberry-flavored wine is simple. Powerful aromatics lead to powerful fruity flavors in this Rhne-style wine. A blend of Viognier, Marsanne and Roussanne, it explodes in apricot and peach juice, vanilla bean and honeysuckle flavors that finish dry and spicy. The high alcohol isn't a problem. Pretty good flavors in this dry Bordeaux-Rhne blend, with blackberries, coffee, tobacco and grilled meat, and the tannins are firmly complex. Loses a few points for excessive softness and a watery finish. Ramona Valley Red Blend Reserve Reserve, Vista del Monte Vineyard Rhne-style Red Blend Rhne-style White Blend Rita's Ros Merlot Robert Renzoni Rock Creek Vineyard Rolling Hills Estate Vineyard Rose Roussanne Ruby Cuve san diego San Diego County San Pasqual Sangiovese Sangiovese-Cabernet Sauvignon Sauvignon Blanc Select Reserve Set in a colorful and whimsical bottle, this attention-grabber offers dense blackberries, blueberries and black licorice candy on the nose, along with bacon fat, smoke and black loam. The palate shows violets, boysenberry, blackberry and beef char. With grippy tannins, it's structured to last a few more years. Shiraz Signature Selection Signature Selection Grand Melange Simple and very dry, with herb-infused pineapple flavors and unnatural acidity. Smells a bit musty, although there's an undercurrent of fresh peaches. Flavors are dull, leaving behind the heat of alcohol and prickly acidity. Smells a little indistinct, tastes soft and simple, with candied cherry, blackberry and licorice flavors that finish dry. Syrah, Cinsault, Grenache, Petit Verdot. Smells distinctly Porty, with a baked fruit, caramelized aroma, but is surprisingly thin and tannic in flavor. Just barely suggests blackberries cooked

Figure 4.30: Tag Cloud generated from the bottom most pin cluster

The tag cloud for this bilcuster contains mostly red varietals such as Cabernet Sauvignon and Merlot. The majority of review scores seem to be in the low to mid 80s and the climatic indicators point to a definite warmer climate as opposed to the ones found towards Napa and Sonoma county.

4.5.10.6 Individual Map Pins

Finally, to see all of the attributes associated with a particular wine review, a single map pin can be clicked. This provides the user with the view shown in Figure 4.31, showing all of the wine review objects' formal attributes. In this case it is a Rose grown in the Dragonsleaf Vineyard in Sonoma county. The wine has notes of tropical guava and peach, and a score of 89 points, costing 13 USD.



13.0 4578.96 89.0 color dark ruby Dragonsleaf Vineyard Dry Juicy and inviting, dark ruby in color, this racy ros made from Syrah grapes is nuanced with tropical guava and peach that linger on the palate. Dry, it'll pair with lounges on the porch or by the pool. lounge lounges on porch NoData NoData NoData palate pool porch racy ros Rose Sonoma sonoma Sonoma Mountain Syrah grapes Temperate warm Temperate warm Temperate warm tropical guava and peach Very cool nights Very cool nights Very cool nights Wines Gone Wild

Figure 4.31: Tag cloud after a map pin was clicked

4.6 Evaluation Learnings

4.6.1 Viewer Extensions

As was demonstrated by the evaluation, much can be learned from the data sets just from the initial visualisations shown in the tag cloud and map viewer. Having multiple visualisations of the same data also assists the user in analysing the data, providing them insight into the composition of the data. From there, the data exploration process can be initiated from any of the three viewers, with the tag cloud still being the primary visualisation used for navigating the data. The graph viewer is very effective for the early stages of the data exploration process, since it is able to instantly give the user an idea of the concentration of tags in the data set, especially the volume of tags with regards to another tag.

As for the map viewer, it proved to be a very effective visualisation of the geolocation data as opposed to having the lat and long coordinates appear as text in the tag cloud only. The map viewer together with the bicluster functionality, makes the map viewer not just a useful data visualisation, but also supports powerful data exploration from the map.

In terms of data analysis and sensemaking, the visualisations are clear and easily readable. The viewers contain all the information a user requires to interpret the visualisation and provides methods of exploring the data further, while not incurring unnecessary cognitive load on the user.

As mentioned in Section 2.3.5, the visualisation design process is an iterative one of refinement through evaluation and feedback. This is especially true of dynamic and reactive visualisations as the ones described in this thesis. As such, further changes and improvements can be made to these extensions through further evaluations and applying the tool to other data sets.

4.6.2 Data Sets

During the evaluation, it once again become obvious that purely numeric data does not result in an effective visualisation in ConceptCloud, but as demonstrated, this problem can be bypassed by performing binning or categorising the data.

Since ConceptCloud is a data exploration tool, the type and quality of data used with the tool directly impacts how effective the data analysis and exploration process is going to be. Effective knowledge discovery can not be conducted on poor quality data. Thus it is important to make use of high quality data as far as possible and be aware of any biases within the data, such as the ones described in Section 4.5.9. As was the case with the example data sets used, some data cleaning and transformation processes might be required before the data is loaded into ConceptCloud. When analysing the data, it is beneficial to formulate questions about the data to investigate. Domain experts will know which correlations to look out for and know which aspects of the data is worth exploring further. This gives direction and purpose to the exploratory search process.

The evaluation also showed that ConceptCloud is capable of processing very large data sets. The server architecture processes the data effectively and efficiently, and the tag cloud displayed the top 5000 tags without issue. The map viewer makes use of pin clustering to avoid the map from becoming overcrowded and provides a clear visualisation even when all pins are displayed on the map. The graph viewer also caters for very large data sets by limiting the number of pie segments or bars. This ensures that the visualisation remains clear and readable by the user. Some detail is sacrificed for less frequently occurring tags, but the benefits of maintaining a readable visualisation outweighs this negative.

4.6.3 Conclusion

In conclusion, ConceptCloud is a data exploration and knowledge discovery tool that works best on data sets that contain textual information as opposed to numeric. This evaluation proved that ConceptCloud is effective in providing a user with a low level and multi faceted view of the data, and allows for exploratory search through any of the three viewer windows. This functionality is useful when the user is not familiar with the data set, and as a result, ConceptCloud is most useful during the initial stages of data exploration where the user is still becoming familiar with the details and composition of the data. The tool is very effective in identifying correlations in the data and other aspects that the user would like to explore further, perhaps using a different method of analysis.

4.7 Chapter Summary

Chapter 4 provided an evaluation on the functionality of the newly implemented data visualisations made to ConceptCloud's existing architecture. This was achieved by using two data sets that fall within the viticulture domain. The first data set was obtained from the *National Centers for Environmental Information's Portal* and is a numeric data set that consisted of different meteorological measurements. The second data set consisted of wine reviews that were posted on the *Wine Enthusiast Magazine's* website, a popular and highly regarded publication among wine enthusiasts. The data sets both contained data pertaining to the wine regions of California, United States of America, thus making it the area of focus for the evaluation. Both data sets were used to demonstrate ConceptCloud's ability in enabling and promoting data exploration and knowledge discovery. From the evaluation, it was found that having multiple visualisations of the same data assisted the user in analysing the data and provided them with insight into the composition of the data. The graph viewer provided the user with an understanding of the concentration of tags in the data set, especially the volume of tags with regards to another tag. Secondly, the map viewer was very effective for displaying geolocation data and together with the bicluster functionality, provided support for powerful data exploration. Overall, the evaluation reinforced ConceptCloud's ability in effectively providing a user with a low level and multi faceted view of the data.

Chapter 5

Conclusion

5.1 Thesis Summary

This thesis aimed to improve and extend ConceptCloud by achieving the following:

- Expanding on the visualisation and exploration capabilities of the tool when handling large semi-structured data sets.
- Providing the user with additional data visualisations and exploration methods.
- Extending the tool's functionality to other applications and devices.

The above objectives were achieved through the development of specialised map and graph viewer windows which were integrated seamlessly with the existing tag cloud visualisation. Further, the map and graph viewers provided additional visualisations of the data to assist the user with data exploration and knowledge discovery.

Furthermore, a mobile application was created that mimicked the functionality found in the browser-based version of the tool, albeit with an updated user interface for mobile touch screens and reduced screen sizes. In ConceptCloud, this was accomplished through the implementation of a REST API made possible by its modern web-based architecture.

ConceptCloud incorporated the aforementioned extensions into its existing architecture. The benefits of data analysis and knowledge discovery were maximised for the user by following a user-centric approach.

The following subsections provided a summary of the major topics covered in this thesis.

5.1.1 Background

Chapter 2 provided the reader with a theoretical background of Formal Concept Analysis (FCA). This section provided the reader with a brief overview of FCA fundamentals and examples of the formal context and concept lattice. This was followed by background theory on data visualisation, analysis, and sensemaking before several formal concept analysis data visualisation and exploration tools were introduced and explored. The initial version of ConceptCloud's architecture was introduced and its functionality was discussed.

5.1.2 Map Visualisation

The first extension made to ConceptCloud was a fully integrated map-based viewer window that allowed for the visualisation and exploration of semi-structured data with some geolocation aspects. Chapter 3 discussed the implementation of the Google Maps API and integration with the existing tag cloud functionality. Furthermore, Chapter 3 discussed the development of the biclustering functionality that allows the user to generate a new tag cloud and lattice, based on any cluster or individual map pins on the map.

5.1.3 Graph Visualisation

The second extension made to ConceptCloud, an integrated graph viewer, was implemented to assist the user in gaining a better understanding of the composition of their data set, especially in terms of specific tags. Furthermore, the graph viewer provided additional supporting visualisations of the data.

5.1.4 Mobile Visualisation

Last, a discussion of the extensions made to ConceptCloud's server architecture with a REST API and the development of a mobile application that mimics the browser user interface, with an updated tag cloud visualisation, was provided.

5.1.5 Evaluation

Chapter 4 evaluated the extensions made to ConceptCloud using two large semi-structured data sets, namely a climate data set and a wine review data set. The two data sets were explored and interrogated using a combination of all three data visualisations. The map and graph viewer provided additional visualisations of the data present in the tag cloud and supported the explorative search process. Furthermore, the map and graph viewer provided the user with methods for navigating the concept lattice beyond the tag cloud.

5.2 Future Work

When considering future work involving ConceptCloud, firstly additional visualisations for the user interface can be developed for other aspects of semi-structured data, e.g. an image viewer for photo data. Secondly, even though ConceptCloud is a modern web application, it can benefit from being ported to a newer version of the Play! framework or a different modern framework entirely such as Microsoft's .NET core framework. Together with this, opportunities for refactoring and optimisation can be explored to improve its ability to process very large data sets even further.

As for ConceptCloud Mobile, there is potential for expanding the functionality offered by the application. One such feature could involve utilising the user's current location with the geolocation aspect of the data to change how data is displayed to the user. The size of the tags displayed to the user can vary depending on the object's proximity to the user's current location. Moreover, further integration with the Google Maps API can allow the map viewer to calculate and indicate the fastest route to various objects in the data set.

Furthermore, it is recommended that further research be conducted on extending and refining the functionality of the map and graph viewer by integrating further with the Google Maps API beyond pins and pin clustering. The API allows for a range of visualisations that include shapefiles, legends, and heat maps. For the graph viewer, support for more advanced 3D graphs as well as improved lattice navigation between the graph and tag cloud can be added. Furthermore, adding the graph viewer functionality to the bicluster lattice could assist users with more advanced and in-depth data exploration.

5.3 Limitations of Visualisations in ConceptCloud

When processing large data sets, it is important to consider performance limitations imposed by third-party libraries that could negatively affect the performance of ConceptCloud as a whole. Problems may arise if the library does not allow extensive modification or low-level manipulation to resolve the issue.

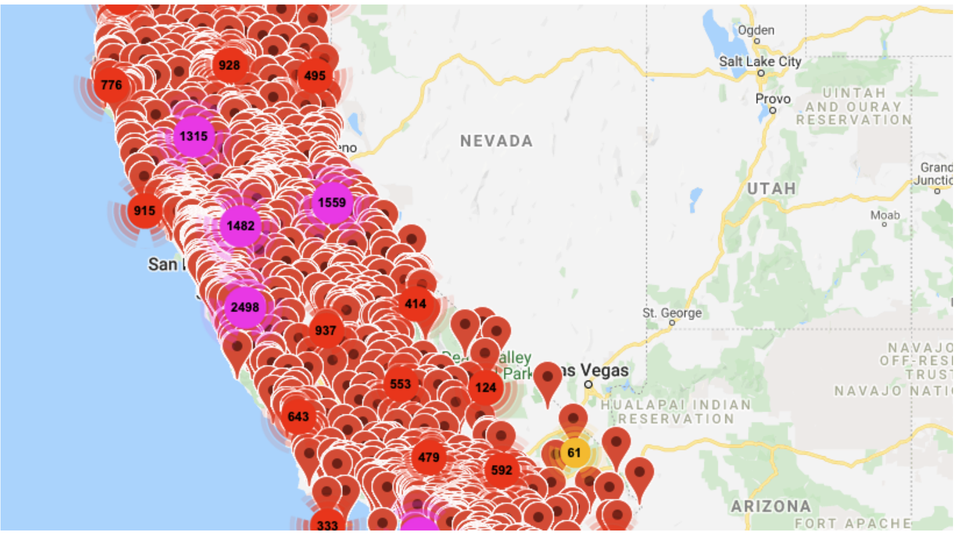


Figure 5.1: Map pins being added to the map after the focus concept is reset

An example of this involves the Google Maps API utilised in the map viewer, which adds marker objects to the map individually before they are clustered. Figure 5.1 shows this process when the focus concept is reset to the top of the lattice in ConceptCloud. Generally, this is not a problem, but during initial experimentation with large data sets, the process was highly memory intensive and resulted in an unresponsive web browser when exceeding 250 000 markers. Due to the highly granular developer functionality provided by this library, this limitation was resolved through a combination of front-end and back-end optimisations described in Chapter 3.

5.4 Conclusion

The primary goals of this thesis were to extend the functionality and visualisation capabilities of Conceptcloud and ensure these extensions adhered to the following requirements:

- The extension must be clear and readable for a user with moderate levels of data literacy.
- The visualisations must remain an accurate representation of the data where an opportunity for misinterpretation of the data by the user is minimised.
- The new extension must aid in data analysis by providing cognitive support and reducing the cognitive load experienced by the user as far as possible.

- The extension must aid in data exploration and knowledge discovery by providing different visualisations of the same data.

- The extension must allow the new visualisation as well as the *Tag Cloud* to drive exploration. Thus a pivotal function of this tool is *to ensure the data exploration works in a bidirectional manner*. In other words, to update the new visualisation when changes are made in one of the visualisations and vice versa.

- The extension must ensure that *the visualisation and exploration tool be highly scalable to process large data sets* since the rate at which such data is being generated is ever increasing.

5.4.1 Visualisation Extension

Despite the capability of handling visualisation for general data sets, it is vital to acknowledge the drawbacks of the default tag-cloud visualisation tool that ConceptCloud provides, specifically when working with specialised data. Support for specialised viewers was added to ConceptCloud to account for these shortcomings. Thereafter the graph viewer was implemented together with the REST API and mobile application.

5.4.2 Architecture and Mobile Extension

Although the mobile application did not form part of the evaluation, it served as an example of what is possible through the REST API. There is still much to explore regarding extension of the the REST API and more advanced mobile applications with powerful lattice processing capabilities.

5.4.3 Evaluation

The newly developed viewers were successfully utilised to conduct case studies on the two data sets described in Chapter 4. The evaluation showcased the flexibility and scalability of ConceptCloud, and the novel ways data can be grouped and explored, especially through the bicluster in the map viewer.

5.4.4 Comparison with existing Tools

As discussed in Chapter 2, none of the other FCA data visualisation tools provide specialised visualisations for specific data types. This now makes ConceptCloud a novel and unique tool for this purpose. Further, due to its approach to lattice generation and visualisation, ConceptCloud remains the top choice for visualising very large and diverse data sets.

5.4.5 Final Conclusions

This thesis achieved the goals set out for it and lays the foundations for further research in data visualisation. The extensions made to ConceptCloud serve as a starting point for further, more advanced extensions, especially in the field of mobile applications. The methods of efficiently visualising large data sets dynamically could potentially be adapted upon to be used for future work involving visualisations. Ultimately, ConceptCloud is now more flexible and feature rich in terms of its data exploration and knowledge discovery functionality than its previous iteration.

Bibliography

[1] 1805. Beaufort wind scale. (1805). https://www.spc.noaa.gov/faq/tornado/beaufort.html

[2] 2019. History of the budget 1937-1950gg. (2019). http://budgetapps.artzub.com/minfin/

[3] 2020. Overview. (2020). https://cordova.apache.org/docs/en/latest/guide/overview/index.html

[4] 2020. Welcome to the ToscanaJ Suite. (2020). http://toscanaj.sourceforge.net/

[5] James F. Allen. 2003. Natural language processing. *Encyclopedia of Computer Science* (2003), 1218–1222.

[6] Maynard Amerine and Albert Winkler. 1944. Composition and quality of musts and wines of California grapes. *Hilgardia* 15, 6 (1944), 493–675.

[7] Christopher J Bargmann. 2003. Geology and wine 7. Geology and wine production in the coastal region, western Cape Province, South Africa. *Geoscience Canada* (2003).

[8] Peter Becker, Joachim Hereth, and Gerd Stumme. 2002. ToscanaJ: An open source tool for qualitative data analysis. In *Advances in Formal Concept Analysis for Knowledge Discovery in Databases. Proc. Workshop FCAKDD of the 15th European Conference on Artificial Intelligence (ECAI 2002). Lyon, France*, Vol. 426.

[9] Joshua Berndt. 2020. *Scaling the ConceptCloud browser to very large semi-structured data sets: architecture and data completion.* Ph.D. Dissertation. Stellenbosch: Stellenbosch University.

[10] Joshua Berndt, Bernd Fischer, and Arina Britz. 2018. Scaling the ConceptCloud browser to large semi-structured data sets. (2018).

[11] Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural language processing with Python.* (1st ed.). O'Reilly Media Inc., Sebastopol.

[12] Peter. F. Brown, Peter. V. DeSouza, Robert. L. Mercer, Vincent. J. Della Pietra, and Jenifer. C. Lai. 1992. Class-Based n-gram models of natural language. *Computational Linguistics* 18, 4 (1992).

[13] Tim Bugher. 2020. TTB: Wine: Appellations of origin. (2020). `https://www.ttb.gov/appellations-of-origin`

[14] Claudio Carpineto, Giovanni Romano, and Fondazione Ugo Bordoni. 2004. Exploiting the potential of concept lattices for information retrieval with CREDO. *J. UCS* 10, 8 (2004), 985–1013.

[15] Stephen Cass. 2015. The 2015 top ten programming languages. *IEEE Spectrum, July* 20 (2015).

[16] Min Chen, David Ebert, Hans Hagen, Robert S Laramee, Robert Van Liere, Kwan-Liu Ma, William Ribarsky, Gerik Scheuermann, and Deborah Silver. 2008. Data, information, and knowledge in visualization. *IEEE Computer Graphics and Applications* 29, 1 (2008), 12–19.

[17] BG Coombe. 1986. Influence of temperature on composition and quality of grapes. In *Symposium on Grapevine Canopy and Vigor Management, XXII IHC 206*. 23–36.

[18] Big Data. 2013. for better or worse: 90% of world's data generated over last two years. *SCIENCE DAILY, May* 22, 3 (2013).

[19] Marco Di Zio, Nadežda Fursova, Tjalling Gelsema, Sarah Gießing, Ugo Guarnera, Jūratė Petrauskienė, L Quensel-von Kalben, Mauro Scanu, KO ten Bosch, Mark van der Loo, and Katrin Walsdorfer. 2016. Methodology for data validation 1.0. *Essnet Validat Foundation, Brussels, Belgium* (2016), 1–76.

[20] Tiaan du Toit, Joshua Berndt, Katarina Britz, and Bernd Fischer. 2019. ConceptCloud 2.0: Visualisation and exploration of geolocation-rich semi-structured data sets. (2019).

[21] Marcel Dunaiski, Gillian J Greene, and Bernd Fischer. 2017. Exploratory search of academic publication and citation data using interactive tag cloud visualizations. *Scientometrics* 110, 3 (2017), 1539–1571.

[22] Ecma International. 2013. The JSON data interchange format. Standard ECMA-404. (Oct. 2013).

[23] Wafaa S El-Kassas, Bassem A Abdullah, Ahmed H Yousef, and Ayman M Wahba. 2017. Taxonomy of cross-platform mobile applications development approaches. *Ain Shams Engineering Journal* 8, 2 (2017), 163–190.

[24] Jose Fermoso. 2009. PhoneGap seeks to bridge the gap between mobile app platforms. (Apr 2009). `https://gigaom.com/2009/04/05/phonegap-seeks-to-bridge-the-gap-between-mobile-app-platforms/`

[25] National Centers for Environmental Information (NCEI). 2021. Climate data online. (2021). `https://www.ncdc.noaa.gov/cdo-web`

[26] Bernhard Ganter and Rudolf Wille. 2012. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media.

[27] Robert Godin, Claude Pichet, and Jan Gecsei. 1989. Design of a browsing interface for information retrieval. In *Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval*. 32–39.

[28] Gillian J Greene, Marvin Esterhuizen, and Bernd Fischer. 2017. Visualizing and exploring software version control repositories using interactive tag clouds over formal concept lattices. *Information and Software Technology* 87 (2017), 223–241.

[29] Gillian J Greene and Bernd Fischer. 2015. Interactive tag cloud visualization of software version control repositories. In *Software Visualization (VISSOFT), 2015 IEEE 3rd Working Conference on*. IEEE, 56–65.

[30] Gillian J. Greene and Bernd Fischer. 2016. Single-focus broadening navigation in concept lattices. In *CDUD@CLA*.

[31] Garrett Grolemund and Hadley Wickham. 2014. A cognitive interpretation of data analysis. *International Statistical Review* 82, 2 (2014), 184–204.

[32] Terry Halpin and Tony Morgan. 2010. *Information modeling and relational databases*. Morgan Kaufmann.

[33] Melanie R. Herrmann, Duncan P. Brumby, and Tadj Oreszczyn. 2018. Watts your usage? A field study of householders' literacy for residential electricity data. *Energy Efficiency* 11, 7 (01 Oct 2018), 1703–1719. `DOI: http://dx.doi.org/10.1007/s12053-017-9555-y`

[34] Alan R Hevner. 2007. A three cycle view of design science research. *Scandinavian Journal of Information Systems* 19, 2 (2007), 87–92. `DOI: http://dx.doi.org/sjis/vol19/iss2/4`

[35] Weidong Huang, Peter Eades, and Seok-Hee Hong. 2009. Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization* 8, 3 (2009), 139–152.

[36] Peter J Huber. 2012. *Data analysis: what can be learned from the past 50 years*. Vol. 874. John Wiley & Sons.

[37] M Huglin. 1978. Nouveau mode d'évaluation des possibilités héliothermiques d'un milieu viticole. (1978).

[38] Dmitry I. Ignatov. 2017. Introduction to formal concept analysis and its applications in information retrieval and related fields. *CoRR* abs/1703.02819 (2017). `http://arxiv.org/abs/1703.02819`

[39] Juhani Iivari. 2007. A paradigmatic analysis of information systems as a design science. *Scandinavian Journal of Information Systems* 19, 2 (2007), 39–64. `DOI:http://dx.doi.org/10.1.1.218.2636`

[40] DI Jackson and NJ Cherry. 1988. Prediction of a district's grape-ripening capacity using a latitude-temperature index (LTI). *American Journal of Enology and Viticulture* 39, 1 (1988), 19–28.

[41] Hugh Johnson. 1989. *Vintage: The story of wine*. Simon and Schuster.

[42] Eric Kauffman. 2003. Climate and topography. *Atlas of the Biodiversity of California* (2003), 12–15.

[43] Mehdi Kaytoue, Sergei O Kuznetsov, Juraj Macko, Wagner Meira, and Amedeo Napoli. 2011. Mining biclusters of similar values with triadic concept analysis. *arXiv preprint arXiv:1111.3270* (2011).

[44] Slava Kisilevich, Florian Mansmann, and Daniel Keim. 2010. P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research #38; Application (COM.Geo '10)*. ACM, New York, NY, USA, Article 38, 4 pages.

[45] W Mark Kliewer and Rodrigo E Torres. 1972. Effect of controlled day and night temperatures on grape coloration. *American Journal of Enology and Viticulture* 23, 2 (1972), 71–77.

[46] Karen MacNeil. 2015. *The wine bible*. Workman Publishing.

[47] Neil Morgan and Gregory Lewis McNamee. 2021. California. (2021). `https://www.britannica.com/place/California-state`

[48] Alexey A Neznanov, Dmitry A Ilvovsky, and Sergei O Kuznetsov. 2013. FCART: A new FCA-based system for data analysis and knowledge discovery. *Contributions to the 11th International Conference on Formal Concept Analysis* (2013), 65–78.

[49] Howard Latimer Penman. 1948. Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 193, 1032 (1948), 120–145.

[50] Donna J Peuquet and Menno-Jan Kraak. 2002. Geobrowsing: creative thinking and knowledge discovery using geographic visualization. *Information Visualization* 1, 1 (2002), 80–91.

[51] Jonas Poelmans, Dmitry I. Ignatov, Sergei O. Kuznetsov, and Guido Dedene. 2013. Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications* 40, 16 (2013), 6538 – 6560. `DOI:http://dx.doi.org/https://doi.org/10.1016/j.eswa.2013.05.009`

[52] Uta Priss. 2006. Formal concept analysis in information science. *Annual Review of Information Science and Technology* 40, 1 (2006), 521–543. `DOI:http://dx.doi.org/10.1002/aris.1440400120`

[53] Ch Riou, N Becker, V Sotes Ruiz, V Gomez-Miguel, A Carbonneau, M Panagiotou, A Calo, A Costacurta, R de CASTRO, and A Pinto. 1994. Le déterminisme climatique de la maturation du raisin: application au zonage de la teneur en sucre dans la communauté européenne. Luxembourg. *Publications officielles de la CEE* (1994).

[54] Jorge Tonietto and Alain Carbonneau. 2004. A multicriteria climatic classification system for grape-growing regions worldwide. *Agricultural and Forest Meteorology* 124, 1-2 (2004), 81–97.

[55] Ingo Wassink, Olga Kulyk, Betsy van Dijk, Gerrit van der Veer, and Paul van der Vet. 2009. Applying a user-centered approach to interactive visualisation design. *Trends in Interactive Visualization* (2009), 175–199.