Industry allocated project number	PHI allocated project number

SATI	CFPA	SAAPPA/SASPA	DFTS	Winetech
tarryn@satgi.co.za Tel: 021 863-0366	inmaak@mweb.co.za Tel: 021 872-1501	theresa@hortgro.co.za Tel: 021 882-8470	dappies@dtd.co.za Tel: 021 870 2900	andraga@winetech.co.za Tel: 021 276 0499
				X

FINAL REPORT
(2018)

1. PROGRAMME AND PROJECT LEADER INFORMATION

	Research Organisation Programme leader	ARC Research Team Manager	Project leader
Title, initials, surname	Prof B Fischer		Prof B Fischer
Present position	Professor, Division of Computer Science, University of Stellenbosch		Professor, Division of Computer Science, University of Stellenbosch
Organisation, department	Division of Computer Science General Engineering Building Banghoek Rd University of Stellenbosch		Division of Computer Science General Engineering Building Banghoek Rd University of Stellenbosch
Tel. / Cell no.	+27 (0)21 808-4232		+27 (0)21 808-4232
E-mail	b.fischer@cs.sun.ac.za		b.fischer@cs.sun.ac.za

2. PROJECT INFORMATION

Research Organisation Project number	
Project title	Analysis and Visualization of Merlot Tasting Notes
Short title	Analysis and Visualization of Merlot Tasting Notes

Fruit kind(s)	Wine		
Start date (mm/yyyy)	01/02/2018	**End date** (mm/yyyy)	31/03/2018

Key words	Sensory profiles, data mining

Approved by Research Organisation Programme leader (tick box) X

3. EXECUTIVE SUMMARY

Objectives & Rationale
Sensory data such as tasting notes and wine reviews capture the essence of “good” and “bad” wines – but only in aggregation: a single review may focus on specifics of an individual wine that is not characteristic for the varietal. It is thus necessary to analyse and aggregate these textual descriptions and to correlate them with (subjective) quality metrics such as ratings. However, in order to identify dominant characteristics that correlate with the different rating categories, it is necessary to aggregate along different dimensions (e.g., origin, vintage, …).

Methods
We used the ConceptCloud (http://conceptcloud.herokuapp.com/) data exploration tool to analyze the reviews of all single-varietal Merlots in Platter’s By Diners Club South AfricanWine Guides from 2003 to 2016. We extracted the relevant fields (primarily origin, vintage, star rating, and the free-text review) for each wine from the SQL database, and used the Stanford corenlp natural language processing system to extract meaningful phrases from the review texts. We manually cleaned these phrases before we built a formal context table and used the ConceptCloud system to explore this data set. In particular, we constructed and analyzed word clouds that show the distribution of vintages, origins, and taste description phrases within the different rating categories.

Key Results
We confirmed that ConceptCloud is stable enough to handle data sets such as the one analyzed here. We saw no obstacles to scaling this up to larger collections, e.g., the full set of wine reviews for all varietals. We identified several characteristic traits of both “good” resp. “bad” Merlots; these are detailed in Sections 5.3 and 5.4 of the attached technical report.

Conclusion/Discussion
We identified several smaller shortcomings with ConceptCloud’s original visualization and implemented some improvements already. We found that the natural language processing component, however, needs substantial improvements to extract more information (in particular, more consistent key phrases) from the reviews. We suggest to integrate the phrase extraction with an ontology, taxonomy, or controlled vocabulary to achieve this. We suggest to integrate more data sources into the underlying data set that was used to build the formal context table; in particular, we suggest to re-integrate barrelling information (which we purposefully excluded because it interferes with the taste descriptors), because the barrelling process is under immediate control of the wine maker. We could also integrate chemical analysis data, where available.

4. PROBLEM IDENTIFICATION AND OBJECTIVES

Merlot is a widely grown varietal in South Africa, but it is not considered a high-quality varietal: in the 2017 edition of Platter's By Diners Club South African Wine Guide (in the following denoted simply as "Platter's"), there are only two 5-Star Merlots or Merlot-driven blends, with about another dozen in the 4.5-Star rating category. Producers, tasters, and consumers are unsure and divided about the taste characteristics of "good" Merlots, and do not know what to expect from a Merlot.

The objective of this project was to identify, from tasting notes and wine reviews, the taste characteristics of "good" resp. "bad" single-varietal Merlots.

Objectives	Milestones (Significant event or stage in a project)	Target Date
Identify, from tasting notes and wine reviews, the taste characteristics of "good" resp. "bad" single-varietal Merlots	Data collection for ConceptCloud completed	28/02/2018
	Tool demonstration	31/03/2018
	Report finalized	31/03/2018

5. DETAILED REPORT

a. PERFORMANCE CHART (for the duration of the project)

Milestone	Target Date	Extension Date	Date completed
Data collection for ConceptCloud completed	28/02/2018		28/02/2018
Tool demonstration	31/03/2018		31/03/2018
Report finalized	31/03/2018		31/03/2018

b) WORKPLAN (MATERIALS AND METHODS)

Approach: We used formal concept analysis (FCA) [W82] as the underlying technology. FCA relies on formal contexts, which relate objects (i.e., wines) to attributes, as unifying data structure. FCA has developed efficient algorithms to convert the formal contexts into algebraic structures called concept lattices, which reveal hidden hierarchical structure in data, and which can be used as navigation structure to explore the fused data sets interactively, without predefined access paths.

We extracted different attributes for a given wine can from the data base used in the production of Platter's. We used lightweight natural language processing techniques to extract relevant taste characteristics from the tasting notes and wine reviews of a large cross-section of South African single-varietal Merlots, together with available meta-data (in particular, rating, price, and winery, vineyard and winemaker details), and construct a formal context table from this.

We then used the ConceptCloud tool (http://conceptcloud.herokuapp.com/) [GEF17] to visualize and interactively explore the data set. ConceptCloud constructs “clickable” tag clouds from the concept lattices. Tag (or word) clouds have been used in many web applications, and are helpful to summarize data sets, helping in data analysis and understanding, but typically provide only a static view on the data. ConceptCloud combines an intuitive tag cloud interface with the underlying concept lattice that provides a formal structure for navigation. This combination allows users to explore data sets serendipitously, without predefined search goals and along different navigation paths; it has been shown to be effective in other domains with similar data characteristics [DGF17].

We used ConceptCloud to identify (qualitatively) the most dominant characteristics that correlate with the different rating categories, as well as any other dimension that emerges from the data. We will demonstrate the tool and analysis to Winetech and interested industry, and documented our findings in a report.

Data Sources: We used data from Platter's as data source, which offers an almost complete coverage of South African wines. We analyzed data from 14 editions, due to the small number of highly-ranked Merlots.

References:

[DGF17] M Dunaiski, GJ Greene, B Fischer: Exploratory search of academic publication and citation data using interactive tag cloud visualizations. Scientometrics 110(3):1539-1571, 2017.
[GEF17] GJ Greene, M Esterhuizen, B Fischer: Visualizing and exploring software version control repositories using interactive tag clouds over formal concept lattices. Information and Software Technology, 87:223-241, 2017.
[W82] R Wille: Restructuring lattice theory: an approach based on hierarchies of concepts. Ordered Sets, Reidel, 445-470, 1982.

c) RESULTS AND DISCUSSION

On the technical side, this experiment confirmed that ConceptCloud is in principle stable enough to handle data sets such as the one analysed here. We saw no obstacles to scaling this up to larger collections, e.g., the full set of wine reviews for all varietals. We identified several smaller shortcomings with ConceptCloud’s original visualization and implemented some improvements already. We found that the natural language processing component, however, needs substantial improvements to extract more information (in particular, more consistent key phrases) from the reviews. We suggest to integrate the phrase extraction with an ontology, taxonomy, or controlled vocabulary to achieve this.

We identified from the reivews several characteristic traits of both “good” (i.e., highly rated) resp. “bad” (i.e., lower-rated) Merlots. Some highlights include:

- Herbal tastes are more indicative of lower-rated wines.
- Plummy tastes are common across all categories except the bottom end, but they are most dominant in the medium categories (3.5/3/2.5 Stars).
- Spicy tastes are common across all categories but are more common across top wines, in particular 4.5/5-Star wines; lower-rated wines are more likely to exhibit savory tastes instead.
- Fruit tastes are the descriptors most commonly associated with top wines, in particular 4.5/5-Star wines.
- Chocolate tastes are common across all categories except the very top end (4.5/5-Star wines). Dark chocolate is generally more correlated with higher ratings.

While we need to caution that the underlying data has not yet been fully cleaned and pre-processed, these conclusions have “passed muster” by an expert (P. van Zijl). We suggest to integrate more data sources into the underlying data set that was used to build the formal

context table; in particular, we suggest to re-integrate barrelling information (which we purposefully excluded because it interferes with the taste descriptors), because the barrelling process is under immediate control of the wine maker. We could also integrate chemical analysis data, where available.

In the following we detail some of the findings.

c.1) Data Sources and Initial Summary Statistics

The analysis presented in this report is based on the original (raw) data underlying the 2002 to 2017 editions of Platter's, although not all of the data was actually used, as outlined below.

Vintage	Total	Stars								no rating	average rating
		4.5/5	4	3.5	3	2.5	2	1.5	0.5/1		
2016	78	0	6	14	16	25	10	2	0	5	2.65
2015	141	0	26	34	31	30	10	2	0	8	2.94
2014	182	0	28	40	42	33	17	4	0	18	2.75
2013	187	2	27	40	54	24	32	3	0	5	2.92
2012	258	5	42	50	64	50	26	13	1	7	2.93
2011	290	9	41	58	80	49	32	10	1	10	2.93
2010	296	12	41	61	79	58	27	6	3	9	2.97
2009	275	12	56	58	67	56	15	3	0	8	3.11
2008	251	13	27	60	63	51	18	6	1	12	2.94
2007	291	10	41	59	74	53	22	6	2	24	2.84
2006	291	6	39	69	63	61	28	6	4	15	2.87
2005	245	6	48	63	66	29	14	7	4	8	3.07
2004	134	2	16	44	35	14	6	1	2	13	2.85
2003	65	1	9	15	18	7	2	4	0	7	2.68
Total	2982	78	447	665	752	540	259	73	18	149	2.92

Table 1: Vintage ratings across vintages for single-varietal Merlot wines

The data set contains a total of 4369 wines with vintages ranging from 2002 to 2017; however, the 2002 and 2017 vintages are represented only very sparsely (2002: 6 wines, 2017: 11 wines) and are therefore discarded from the data set; note that the 2017 vintage also contains mostly lower-rated wines, due to the lack of maturation, so keeping it would skew the analysis. 940 further wines are blends, which are not considered in this study, and thus also discarded from the data set. 149 further wines have no associated rating due to a variety of reasons (noted in the guides as "discontinued (D)", "not retasted (NR)" and "not tasted (NT)"). Since our goal is to find associations between traits and ratings, these are discarded from the data set as well. This leaves us with a total of 2833 wines as subject of the analysis.

In this set, the extreme rating categories are also extremely rare—for example, there are only three 5-star wines over all years together. We therefore group 4.5- and 5-star wines together, as well as 0.5- and 1-star wines. This gives us eight rating groups: 4.5/5, 4.0, 3.5, 3.0, 2.5, 2.0, 1.5, and 0.5/1. Table 1 shows the detailed breakdown of the ratings over the vintages

We note that the average rating is fairly stable, with most vintages' averages close to the long-term average of 2.92. Notable exceptions are 2009 and 2005 on the positive side, and 2014 and 2003 (although based on a smaller number of rated wines) on the negative side; note that the averages of the recent vintages may well improve by the late releases of well-rated wines with longer barrel maturation. We also note a steady downward trend in the number of wines per vintage, starting from 2010. The extent to which this is caused by late releases is unknown; presumably, this reflects (at least partially) a trend away from Merlot cultivation. (Note that

official statistics show that Merlot cultivation went from 6862.67 ha in 2006 down to 5557.74 ha in 2016.) The downward trend may also be due to a shift from single-varietal Merlot bottlings to Merlot-based blends that are not considered in this analysis.

c.2) Vintage Distribution across Different Rating Categories

In a first analysis, we use ConceptCloud to visualize the vintage distribution across the different rating categories. We open eight views (arranged in a 4x2 layout), where each view shows the vintage tags from the wines within the corresponding rating category only.

Here (and in the following analyses) we use four different normalization scaling mechanisms that emphasize different aspects of the data:

- *no normalization, default scaling*: each tag *i* is scaled between the given minimum and maximum font sizes f_{min} and f_{max}, according to its count t_i in relation to the minimum and maximum counts t_{min} and t_{max} in the whole data set:

  $$\text{size}(i) = \left\lceil \frac{(f_{max} - f_{min}) \cdot (t_i - t_{min})}{t_{max} - t_{min}} \right\rceil + f_{min} - 1$$

  This mechanism ensures that rare tags remain visible, but it "squishes" differences between common tags, and makes it hard to compare tags across the different views.
- *no normalization, linear scaling*: each tag *i* is scaled linearly between the given minimum and maximum font sizes f_{min} and f_{max}, according to its count t_i in relation to the size of the whole data set. This leads to a stronger pronunciation of the trends, only works when the tags are distributed relatively evenly across a fairly narrow range of occurrences--otherwise the common tags completely crowd out the rarer tags.
- *global normalization, default scaling*: each tag count t_i is first normalized against its global count over the whole data set, before the tag is scaled using the default scaling. This gives us an indication of where on the rating scale the majority of the tag's occurrences fall.
- *per-category normalization, linear scaling*: each tag count t_i is first normalized against its count in the corresponding category before the tag is scaled linearly. This can be seen as representing multiple histograms, one per rating category.

Figure 1 below shows the vintage distribution across the different rating categories, without normalization and using the default scaling; as predicted, differences between the large tags are difficult to discern. The linear scaling (see Figure 2) indeed leads to a stronger pronunciation of the trends. For example, we can now see the outstanding 2009 vintage indeed standing out in the 4.0 rating category.

In the globally normalized view (see Figure 3) we see, for example, that the (overwhelming) majority of the 2016 vintage is (still) rated at 2.5 Stars only (which is an indication that the better wines have not been rated yet), or that while there are in absolute numbers more wines rated at 4~Star wines than there are 2005 wines, their relative distribution is similar. Finally, the independent, per-category scaling (see Figure 4) shows, for example see the 2008--2010 vintages dominating the top category, as well as the 2011 and 2012 resp. 2005 and 2006 vintages standing out in the bottom brackets

Figure 1: Vintage distribution across the different rating categories (no normalization, default scaling)

Figure 2: Vintage distribution across the different rating categories (no normalization, linear scaling)

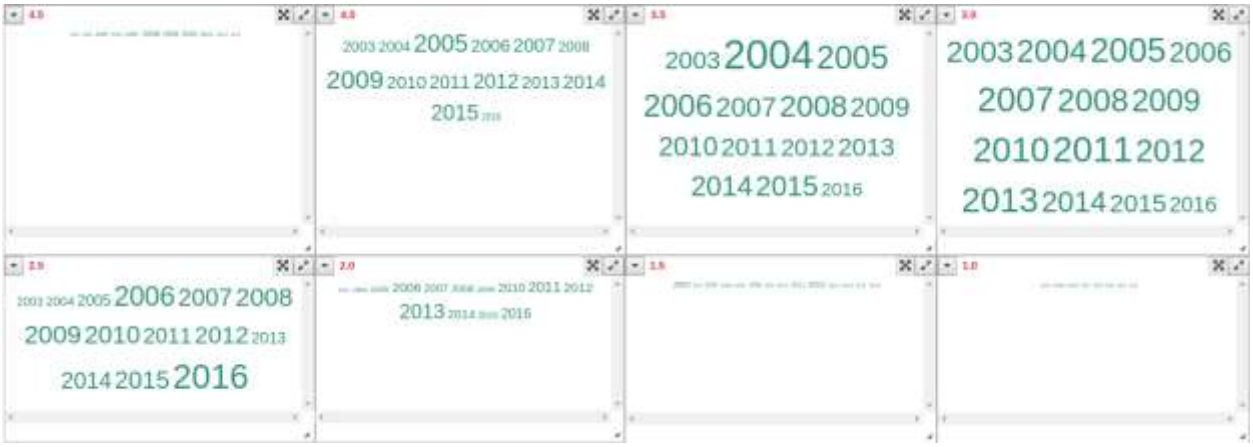

Figure 3: Vintage distribution across the different rating categories (global normalization, default scaling)

Figure 4: Vintage distribution across the different rating categories (per-category normalization, linear scaling)

c.3) Origin Distribution across Different Rating Categories

In a second analysis step, we then looked at the effect of the “terroir”, or more precisely, at the distribution of “wine of origin” tags across the different rating categories. We kept the 4x2 general layout and categorization we used in the vintage analysis.

Figure 5 shows this view using the default scaling. We see that the top category is dominated by wines from Stellenbosch, with other areas (Durbanville, Franschhoek, Paarl, and Simonsberg) also appearing prominently at the 4-Star rating category. Towards the middle rating categories (3.5/3/2.5 Stars), no clear trend is visible and wines from all regions are located here. We notice, however, that the generic origin “Western Cape” becomes more prominent in the lower rating categories. We also notice that Stellenbosch occurs prominently across all rating levels, but this is a consequence of it being one of the most common origin of Merlots.

In Figure 6, we therefore normalize each tag against its global count, as before. We can now see some interesting trends emerging, e.g., that the majority of wines from Banghoek and Groenekloof are rated at 4~Stars,

In Figure 7, we finally scale each view independently again. We now clearly see that Stellenbosch wines dominate all top rating categories, while the generic Western Cape the most common origin for lower-rated wines.

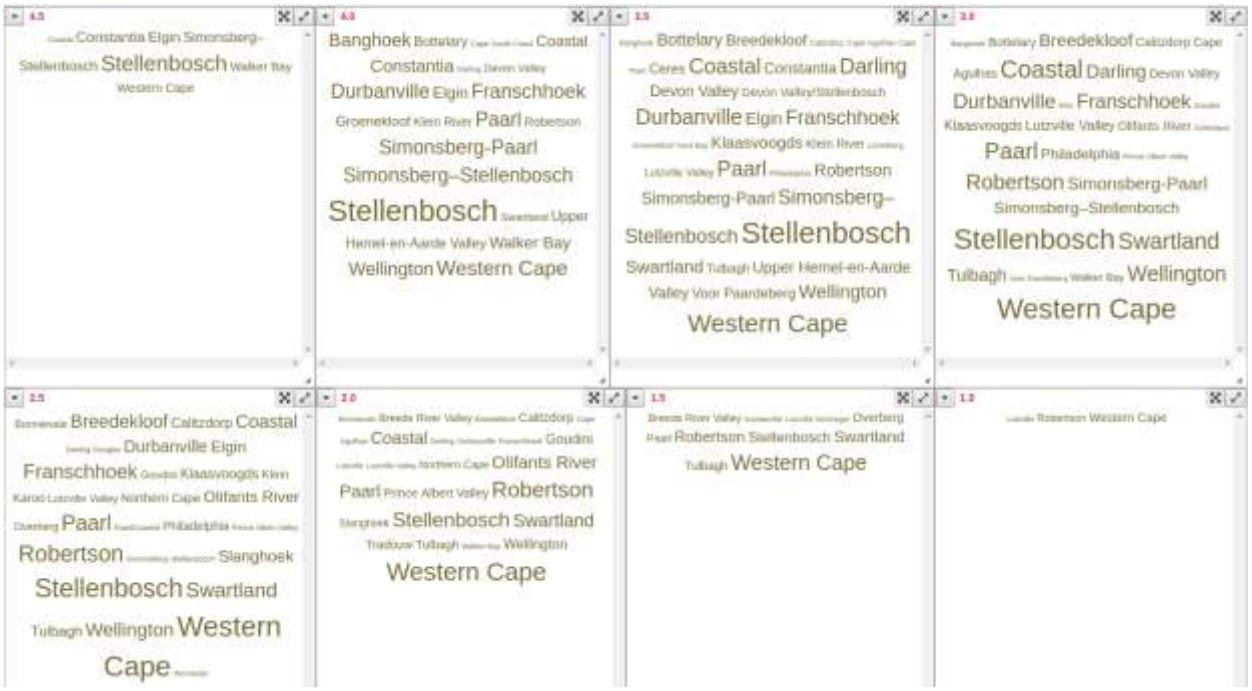

Figure 5: Origin distribution across the different rating categories (no normalization, default scaling)

Figure 6: Origin distribution across the different rating categories (global normalization, default scaling)

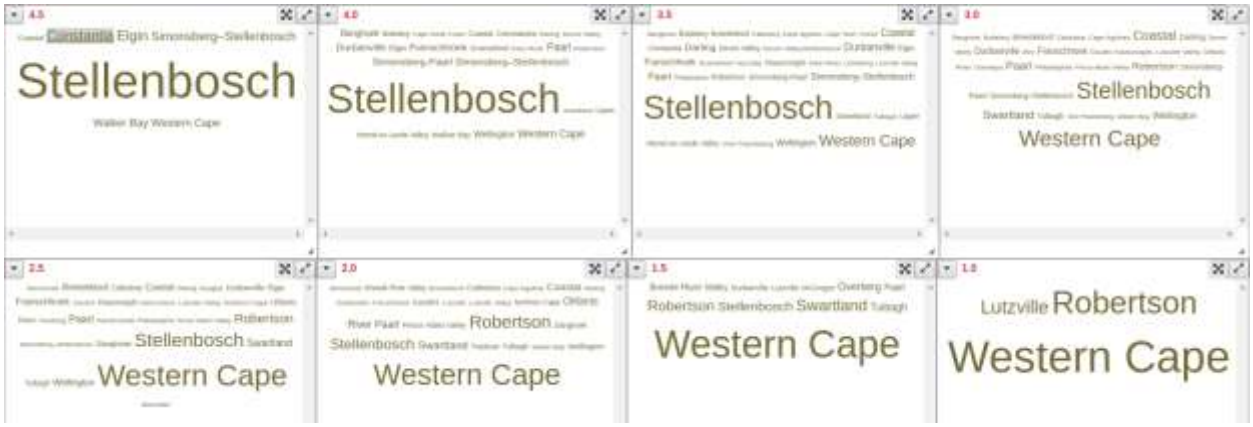

Figure 7: Origin distribution across the different rating categories (per-category normalization, linear scaling)

c.4) Taste Distribution across Different Rating Categories

Figures 8 and 9 show the 100 top phrases in each rating category using the default view, and using per-category normalization and linear scaling, respectively. While we need to caution that the underlying data has not yet been fully cleaned and pre-processed, we can already see a few interesting trends standing out from these two views:

- Herbal tastes are more indicative of lower-rated wines: herbal taste descriptors do occur even at 4- and 3.5-Star wines, but they are relatively much more common for 1.5- and 1-Star wines.
- Plummy tastes are common across all categories except the bottom end, but they are most dominant in the medium categories (3.5/3/2.5 Stars).
- Spicy tastes are common across all categories but are more common across top wines, in particular 4.5/5-Star wines; lower-rated wines are more likely to exhibit savory tastes instead.
- Fruit tastes are the descriptors most commonly associated with top wines, in particular 4.5/5-Star wines.
- Chocolate tastes are common across all categories except the very top end (4.5/5-Star wines). Dark chocolate is generally more correlated with higher ratings.

Figure 10 elaborates on the last point and shows the distribution of the phrases “dark chocolate” and “chocolate” across the rating categories. We can clearly see the higher prevalence of dark chocolate through the higher rating categories, and the absence of chocolate in the top-rated category. We can also see that the higher-rated occurrences of chocolate tend to correlate with spicy flavors (e.g., ginger), while the lower-rated occurrences tend to correlate with sweet tastes (e.g., ripe black cherry, sweet-spiced fruit cake, or raspberry)

Figure 8: Phrase distribution across the different rating categories (no normalization, default scaling)

Figure 9: Phrase distribution across the different rating categories (per-category normalization, linear scaling)

Figure 10: Dark chocolate vs. chocolate: distribution across the different rating categories (no normalization, default scaling)

c.5) Phrase Analysis

In a final analysis step, we picked a key phrase that is usually associated with good wines, in an attempt to see whether such phrases are already discriminative enough, and what other descriptors they imply. We thus created a single tag cloud that combines all rating categories, and includes phrases, ratings, origins, and vintages for all wines that are associated with the phrase “elegant” (see Figure 11).

We can first see that there is indeed an association between this phrase and good ratings--4.5 Star ratings are showing more prominently here than their overall numbers (see Figure 1) indicate. The vintage distribution shows the 2005 vintage over-represented; the reason for this is that the vintage was generally “muscular”, so that “elegant” wines were specifically called out. We also see some correlation with other highly-rated taste descriptors (e.g., fruit or plum), but the savoury taste is also surprisingly common here and, in particular, more common than spicy tastes.

Figure 11: Tag cloud for “elegant” (no normalization, default scaling)

d) CONCLUSIONS

Inferring general characteristics of a wine variety from a large collection of semi-structured data (specifically Platter’s reviews) appears to be feasible. We followed a semi-automatic approach in which humans drive the exploration of the data but are supported by an interactive data exploration tool (specifically ConceptCloud). The human does not need to be a wine expert to identify relevant characteristics, which may make the approach usable for the broader customer market.

The applied technology is stable enough for the task; the main shortcomings are in the natural language processing steps, in particular the phrase extraction. Here, more domain-knowledge must be exploited (e.g., in the form of a formal ontology) to get higher quality data.

6. ACCUMULATED OUTPUTS

a) TECHNOLOGY DEVELOPED, PRODUCTS AND PATENTS

We demonstrated the tool and analysis, and document our findings in a technical report (attached). The report will become property of the funders, but the project participants will be This document is confidential and any unauthorised disclosure is prohibited

allowed to use its results free of charge in their own work. The IP of the applied tool and the analyzed data sets will remain with the respective IP holders.

The primary outcome of the project is the set of taste characteristics of "good" resp. "bad" single-varietal Merlots. This can be of great of benefit beneficial for Merlot producers in terms of guiding the winemaking process to improve flavour and quality, thus lifting the standard across the board; it can also be used for marketing purposes.

b) SUGGESTIONS FOR TECHNOLOGY TRANSFER

An improved version of the tool could be developed into a customer product, for example a smartphone app. A refined analysis by domain experts can be used to develop a “tasting wheel” for Merlot wines.

c) HUMAN RESOURCES DEVELOPMENT/TRAINING

Student Name and Surname	**Student Nationality**	**Degree (e.g. MSc Agric, MComm)**	**Level of studies in final year of project**	**Total cost to industry throughout the project**
Honours students				n.a.
Masters Students				n.a.
PhD students				n.a.
Postdocs				n.a.
Support Personnel				n.a.

PERSONS PARTICIPATING IN THE PROJECT (Excluding students)

Initials & Surname	Highest Qualification	Degree/ Diploma registered for	Race (1)	Gender (2)	Institution & Department	Position (3)	Cost to Project R
B Fischer	PhD	n.a.	W	M	SU Division of Computer Science	PL	R0
P van Zijl		n.a.	W	M	Platter’s Wine Guide	Co	R0

(1)Race B = African, Coloured or Indian
W = White

(2)Gender F = Female
M = Male

(3)Position Co = Co-worker (other researcher at your institution)
Coll = Collaborator (participating researcher that does not receive funding for this project from industry)
PF = Post-doctoral fellow
PL = Project leader
RA = Research assistant
TA = Technical assistant/ technician

d) PUBLICATIONS (POPULAR, PRESS RELEASES, SEMI-SCIENTIFIC, SCIENTIFIC)

We are not planning any scientific publications during the period of the project, although it may be used as case study in other work of the proposer. The technical report describing the work in more detail is attached.

e) PRESENTATIONS/PAPERS DELIVERED

We are not planning any presentations beyond those to Winetech.

7. BUDGET

TOTAL COST SUMMARY OF THE PROJECT

TOTAL FUNDING REQUIRED FOR FOLLOWING YEAR	CFPA	DFTS	Deciduous	SATI	Winetech	THRIP	OTHER	TOTAL

Overheads (only if part of project cost)								n.a.
Research Personnel								n.a.
Research and Technical Assistance (directly linked to project)								R40,000
2018								R40,000
Bursaries								n.a.
Research materials and supplies - specify each item								n.a.
Research Equipment								n.a.
Local conferences (only specify separately for THRIP purposes)								n.a.
Capital items *								n.a.
Other								n.a.

* Industries will only fund capital items under exceptional circumstances

BUDGET MOTIVATION / DESCRIPTION/S

Budget Item	Motivation / detailed description of budget item
Operating Costs:	

Research Personnel	n.a
Research & Technical Assistance (directly linked to project)	R10000 is budgeted for Platter's to prepare the data used in the study; R30000 is budgeted for SU to help execute the study and present its results.
Bursaries	n.a.
Research Materials and Supplies	n.a.
Research Equipment	n.a.
Local accommodation and travel	n.a.

1. TOTAL COST IN REAL TERMS

YEAR	CFPA	DFTS	Deciduous	SATI	Winetech	THRIP	OTHER	TOTAL
2018					R40,000			R40,000

EVALUATION BY INDUSTRY

This section is for office use only

Project number	
Project name	
Name of Sub-Committee*	
Comments on project	

Committee's recommendation (Review panel in the case of PHI)	
• Accepted.	☐
• Accepted provisionally if the sub-committee's comments are also addressed. Resubmit this final report by_____	☐
• Unacceptable. Must resubmit final report.	☐

Chairperson_____ Date_____

*SUB-COMMITTEES

Winetech
Viticulture: Cultivation; Soil Science; Plant Biotechnology; Plant Protection; Plant Improvement;
Oenology: Vinification Technology; Bottling, Packaging and Distribution; Environmental Impact; Brandy and Distilling; Microbiology

Deciduous Fruit
Technical Advisory Committees: Post-Harvest; Crop Production; Crop Protection; Technology Transfer
Peer Work Groups: Post-Harvest; Horticulture; Soil Science; Breeding and Evaluation; Pathology; Entomology