Data/text mining techniques in modelling the climate effects on crops

Subana Shanmuganathan PhD
Geoinformatics Research Centre (GRC)
School of Computer and Mathematical Sciences (SCMS)
Auckland University of Technology (AUT), New Zealand
December 2013

from data mining *and* knowledge discovery (in databases) *to* big data analytics *and* knowledge extraction; *for* applications in science

GRC research projects

overview

background

(what's DM, KD (KDD), big data, data analytics and data science)

- constraints/ challenges
- initiatives: techniques, algorithms needed
- "hot" topics

where are we today ?

big data analytics / data science applications

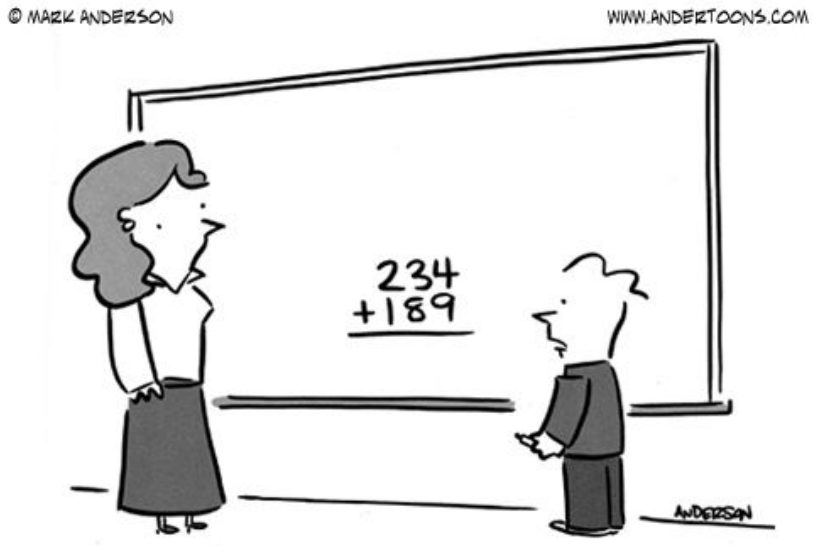
- in enterprises
- disciplines/ problem domains
- new approaches

examples of DM (GRC research)

- climate change effects on grapevine phenology and wine quality
- multi-sensor data analysis
- spatial data mining

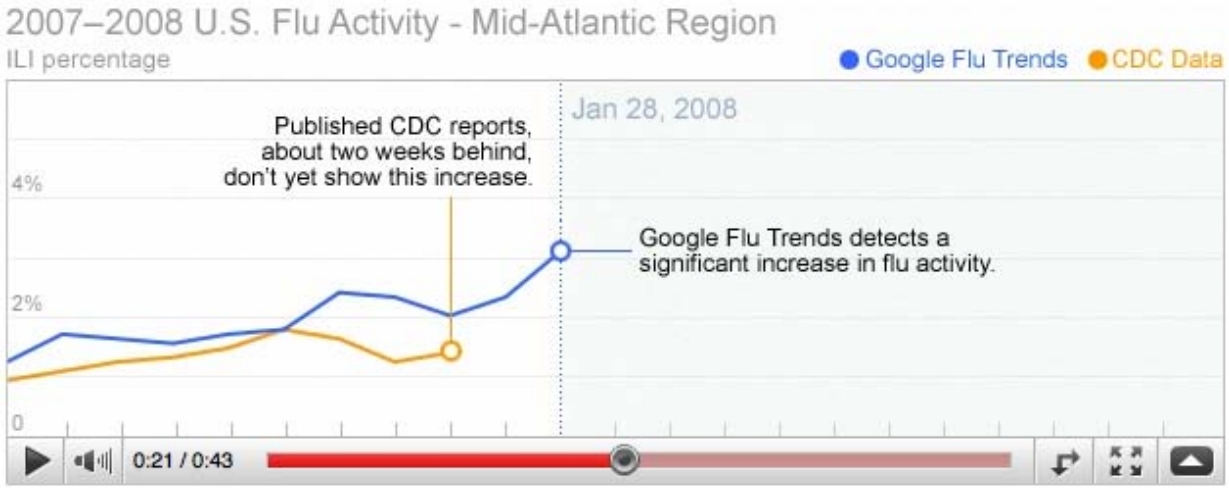
Is this the place to learn about mining?

http://cadeh.com/



"Does this count as big data?"

http://magnus-notitia.blogspot.co.nz/2013/02/big-data-is-dead-whats-next.html



"Google was able to spot trends in the Swine Flu epidemic roughly two weeks before the Center for Disease Control by analyzing searches that people were making in different regions of the country."
http://radar.oreilly.com/2010/06/what-is-data-science.html

http://www.benitaepstein.com/science%20cartoons/science.html

What's DM & KD(D)

DM : Data mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data (Fayyad et al., 1996).

KD(D) : Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996).

Big data: Big data is a buzzword, or catch-phrase, used to describe **massive volume of both structured and unstructured data** that is so large that it's difficult to process using **traditional database and software techniques** (www.webopedia.com/).

Big data refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from **instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources** available today and in the future.

Big data analytics and Data science

Big data analytics : Big data analytics is the process of examining large amounts of data of a variety of types (big data: 3 Vs : **volume, variety, velocity,** and lately added **veracity**) to uncover hidden patterns, unknown correlations and other useful information (**actionable knowledge**). Such information can provide competitive advantages over rival organisations and result in business benefits, such as more effective marketing and increased revenue

http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

http://www.theserverside.com/feature/Handling-the-four-Vs-of-big-data-volume-velocity-variety-and-veracity

Data science : Data Science deals with the whole process of gathering data, pre-processing them and finally making sense out of them, producing what can be called as **data products**. Large volumes of **noisy** and **unstructured** data generated in our daily lives, from social media to search terms on Google cannot be analysed using traditional data mining and warehousing strategies with such large and dynamic data sources.

The need for far more advanced ….for scientific research is historically significant

Recent NSF big data initiatives

Core Techniques and Technologies or fundamental advances in following relevant disciplines:

- computer science
- computational science
- statistics and
- mathematics

deals with research challenges relating to 3 themes:

- data collection and management (DCM)
- analytics
- collaborative environments

NSF: National Science Foundation, USA

http://www.nsf.gov/events/event_summ.jsp?cntn_id=124058&org=CISE http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767
http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf

Data collection & management

DCM relates to

- types of research being conducted in data management, information retrieval at an expanded scale

need new

- techniques for storing data, new ways of doing I/O
- news architectures to deal with heterogeneous data types
- methods for archiving, indexing and recovering data
- dealing with streaming data
- dealing data already in some complex form
- using various techniques for retrieving the amount of bandwidth for retrieving information from remote data sources
- various techniques for automated annotation
- discovering data sources
- Languages and tools for programming data related obligations

Data analytics

focusses on

- scalable data mining (DM), machine learning (ML) algorithms, statistical inference techniques for dealing with extremely massive and high dimensional, highly heterogeneous and dynamic data sets
- new techniques for predictive modelling such data
- algorithms, programming languages and data structures to deal…
- data-driven simulations techniques for mixing data with simulation, simulation with formal models, information extraction from unstructured and multi modal types of data
- scalable techniques for data visualisation in real time
- techniques for dealing with memory problems that constraints with having to do with big data analytics

Collaborative environments

add science to this:

- data analytics and interpretation become highly interdisciplinary requiring collaborations that
- need techniques for representations, new modelling techniques, and tools that allow for collaborations across individuals looking at complex data sets or across disciplines using multiple representations that make sense within the respective disciplines
- these are Foundational aspects of an effective cyber infrastructure and basic problems have to be addressed in the respective disciplines

Topics

Biomedical applications

- Various techniques for analysing structural and functional correlatives, interactions and networks, various protein interaction networks, network of neurons
- Example analysis on social media for understanding local, original and national health
- New techniques for mining literature and other types of data to get an understanding of the biomedical research landscape, techniques for analysing multiple clinical research data sets
- Predictive modelling in biology are related to human health and treating disease
- Clinical science to generate hypotheses using already available background knowledge
- New techniques for disseminating scientific knowledge beyond traditional publications. Methods to link the publications to data sets, simulations → one can actually replicate the study reported in the publication

Artemis platform

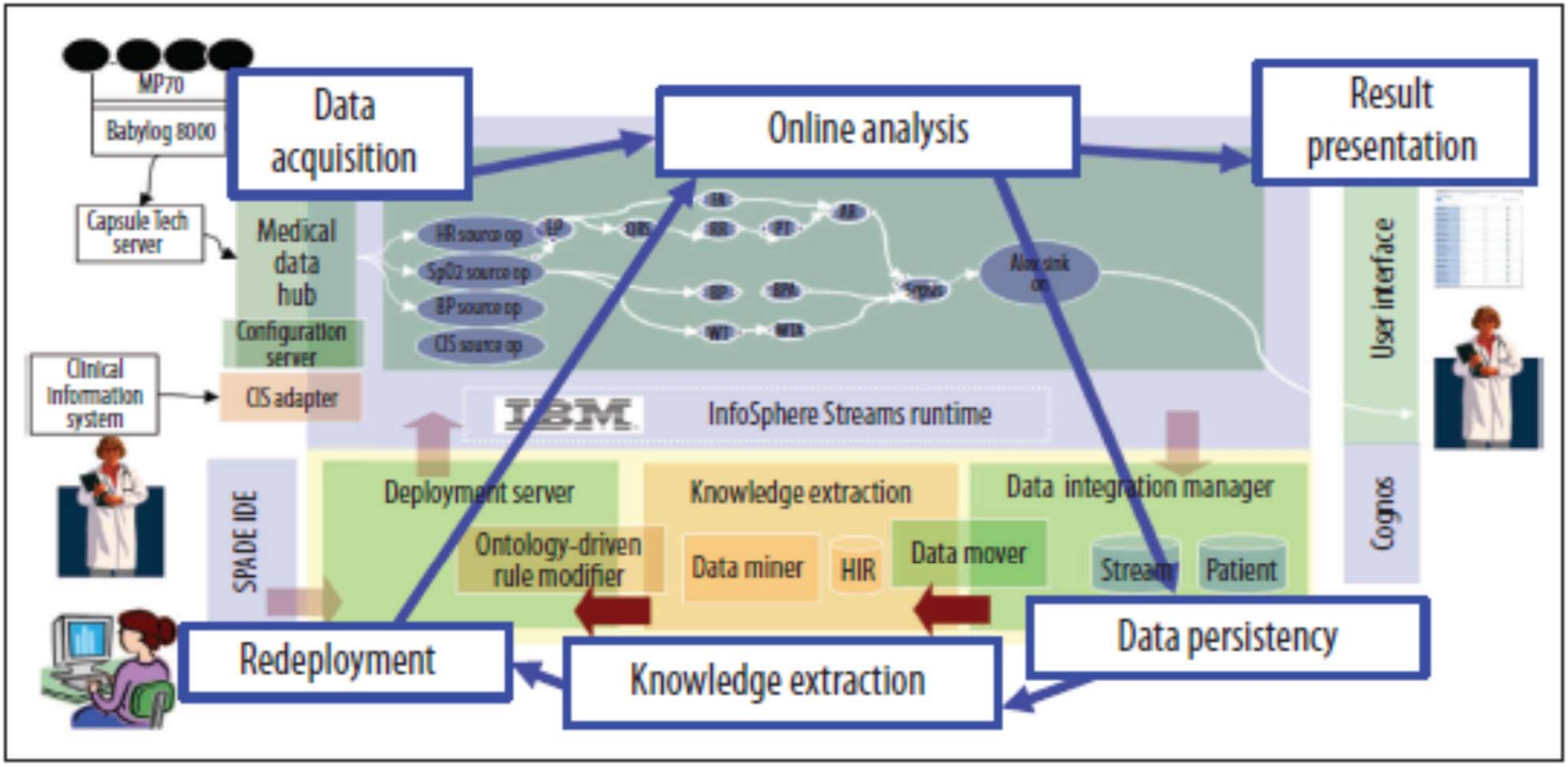
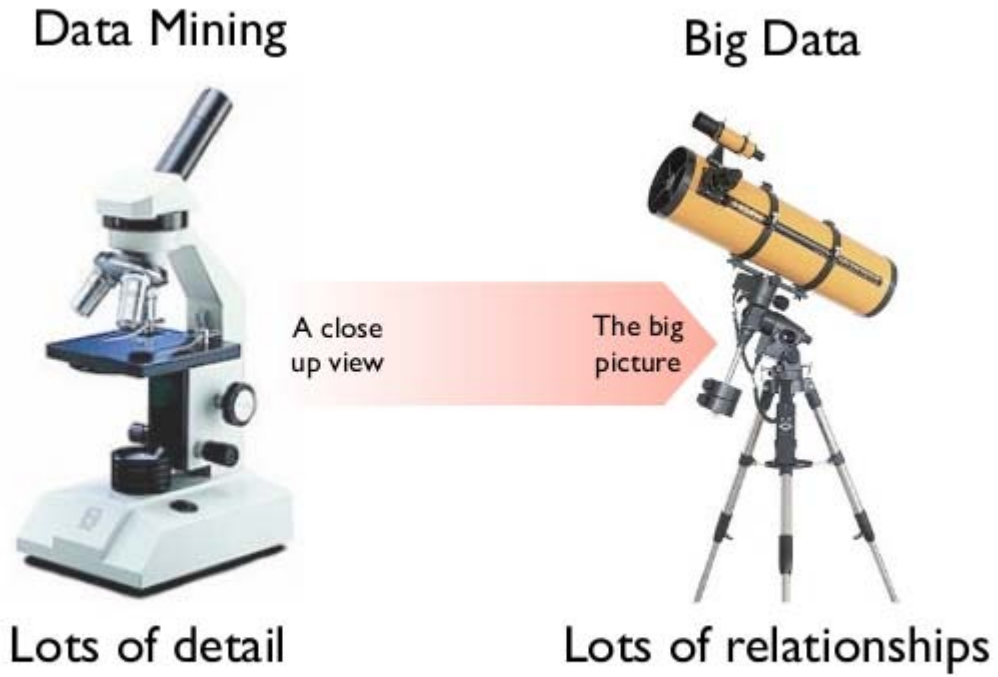


Figure 2. Artemis platform. Artemis enables concurrent diagnoses of multiple patients through real-time analysis of multiple data streams.

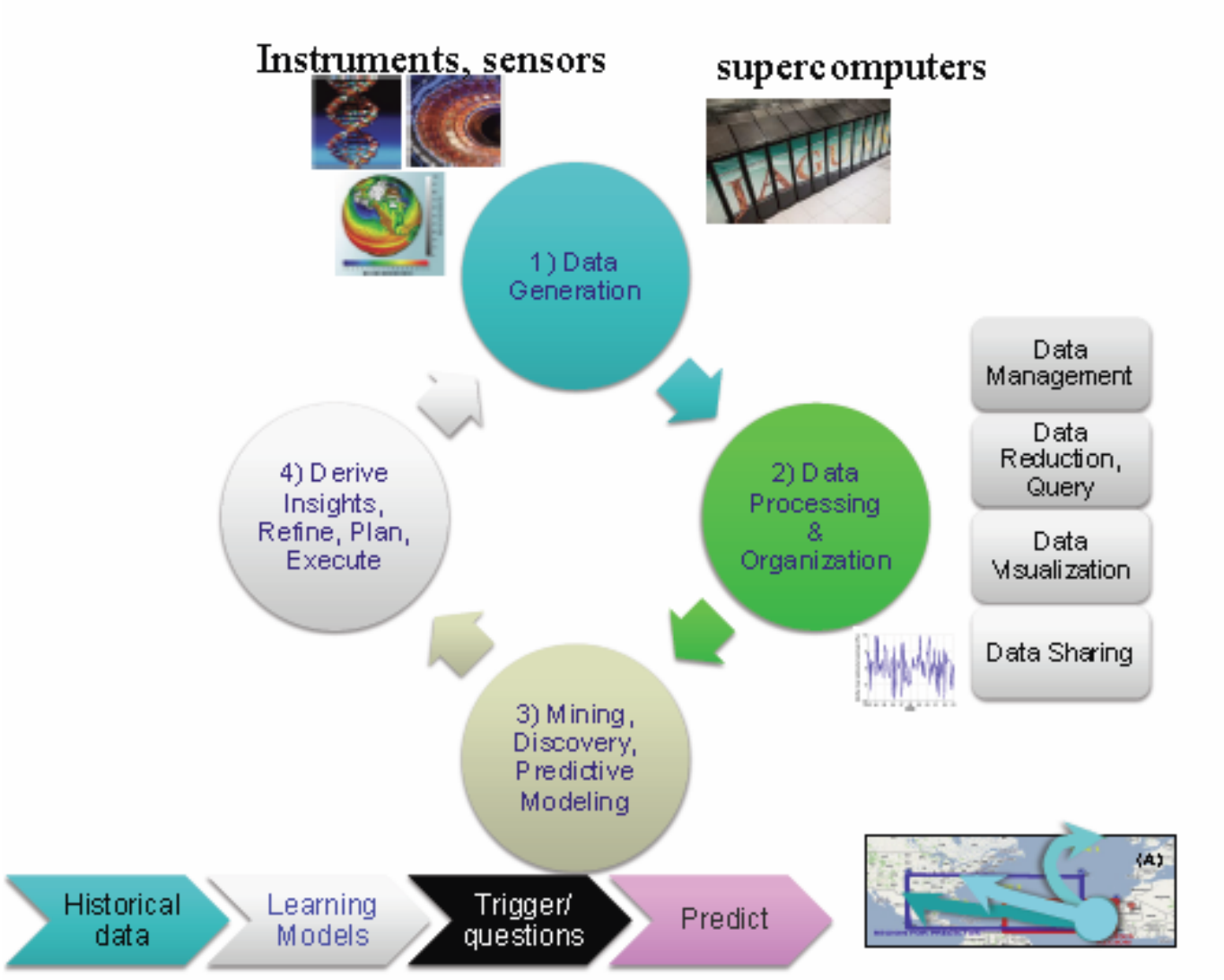
McGregor, C 2013 **Big Data in Neonatal Intensive Care.** ComputervVolume:46 Issue:6 54-59

big data and data mining



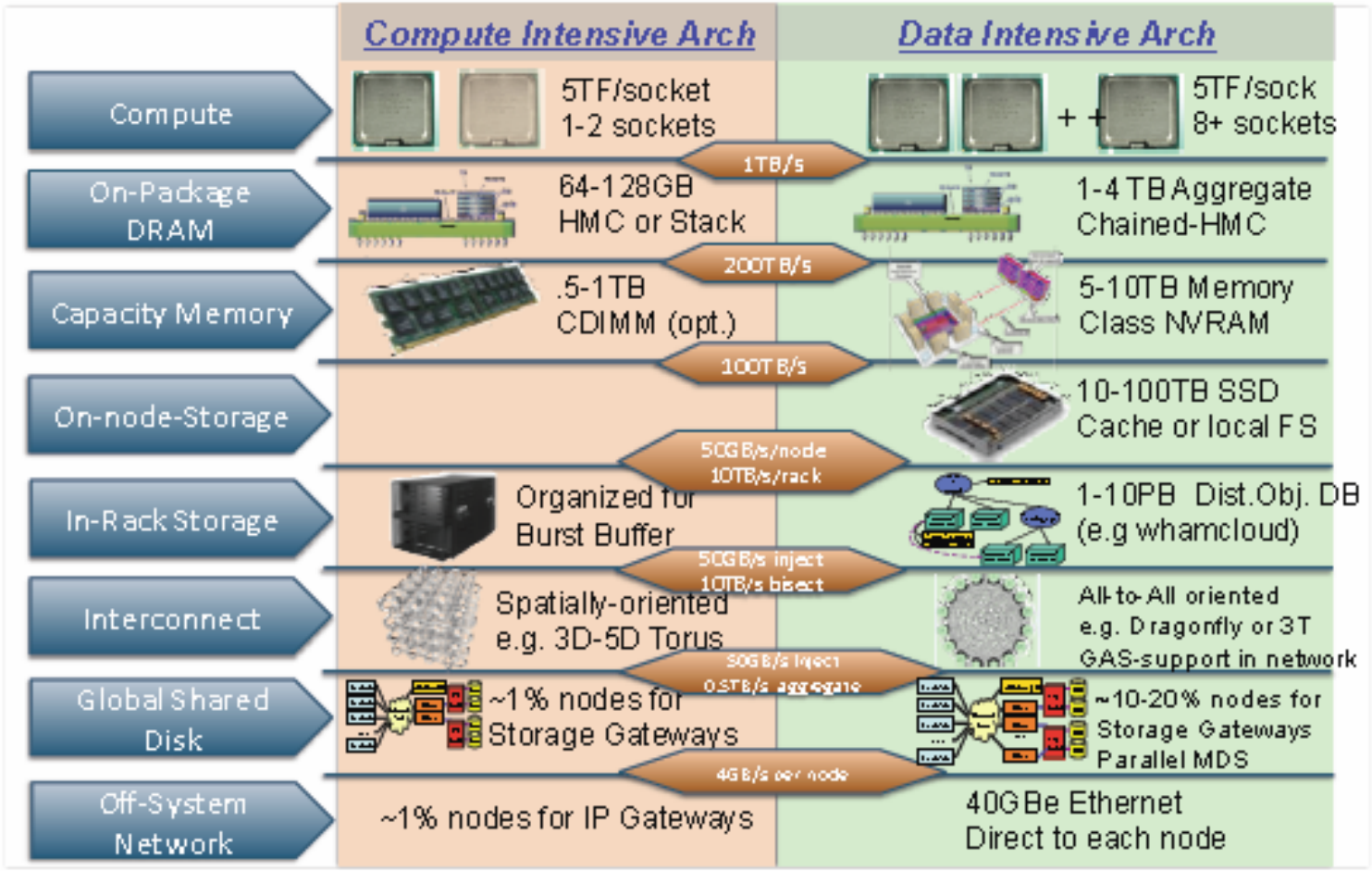
Thursday, 31 January 13

Peter Cochrane, Cochrane Associates on Jan 31, 2013
http://www.slideshare.net/PeterCochrane/big-data-v-data-mining



A knowledge-discovery life-cycle for Big Data

Synergistic Challenges in Data-Intensive Science and Exascale Computing, Summary Report of the **Advanced Scientic Computing Advisory Committee (ASCAC) Subcommittee**, Office of Science, US Department of Energy, 30, March 2013 page 6



Strawman compute-intensive vs. data-intensive computer architectures in the 2017 timeframe

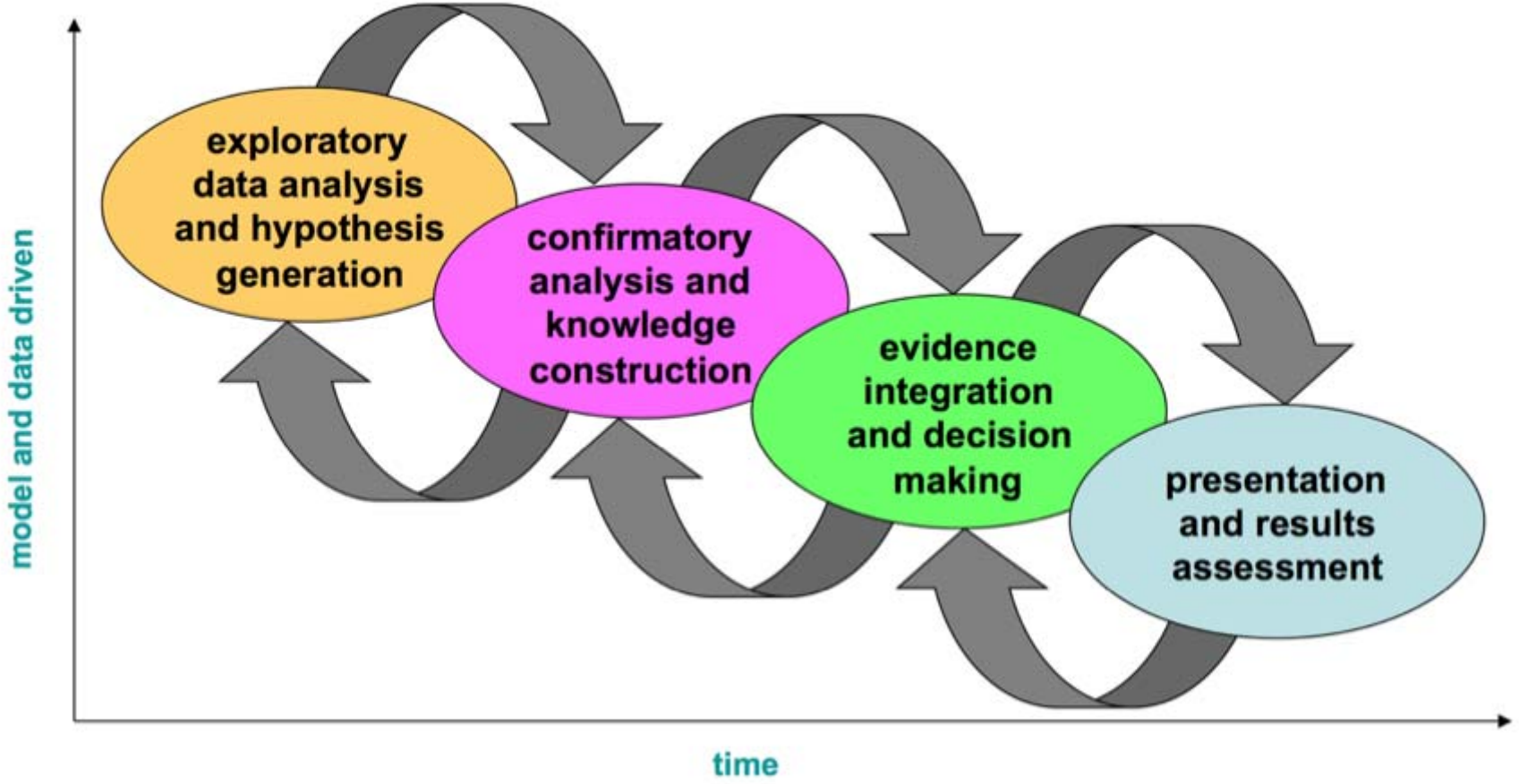
Synergistic Challenges in Data-Intensive Science and Exascale Computing, Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee, Office of Science, US Department of Energy, 30, March 2013 page 8

Parameter	Benchmarks				
	SPECINT	SPECFP	MediaBench	TPC-H	MineBench
Data References	0.81	0.55	0.56	0.48	1.10
Bus Accesses	0.030	0.034	0.002	0.010	0.037
Instruction Decodes	1.17	1.02	1.28	1.08	0.78
Resource Related Stalls	0.66	1.04	0.14	0.69	0.43
ALU Instructions	0.25	0.29	0.27	0.30	0.31
L1 Misses	0.023	0.008	0.010	0.029	0.016
L2 Misses	0.0030	0.0030	0.0004	0.0020	0.0060
Branches	0.13	0.03	0.16	0.11	0.14
Branch Mispredictions	0.0090	0.0008	0.0160	0.0006	0.0060

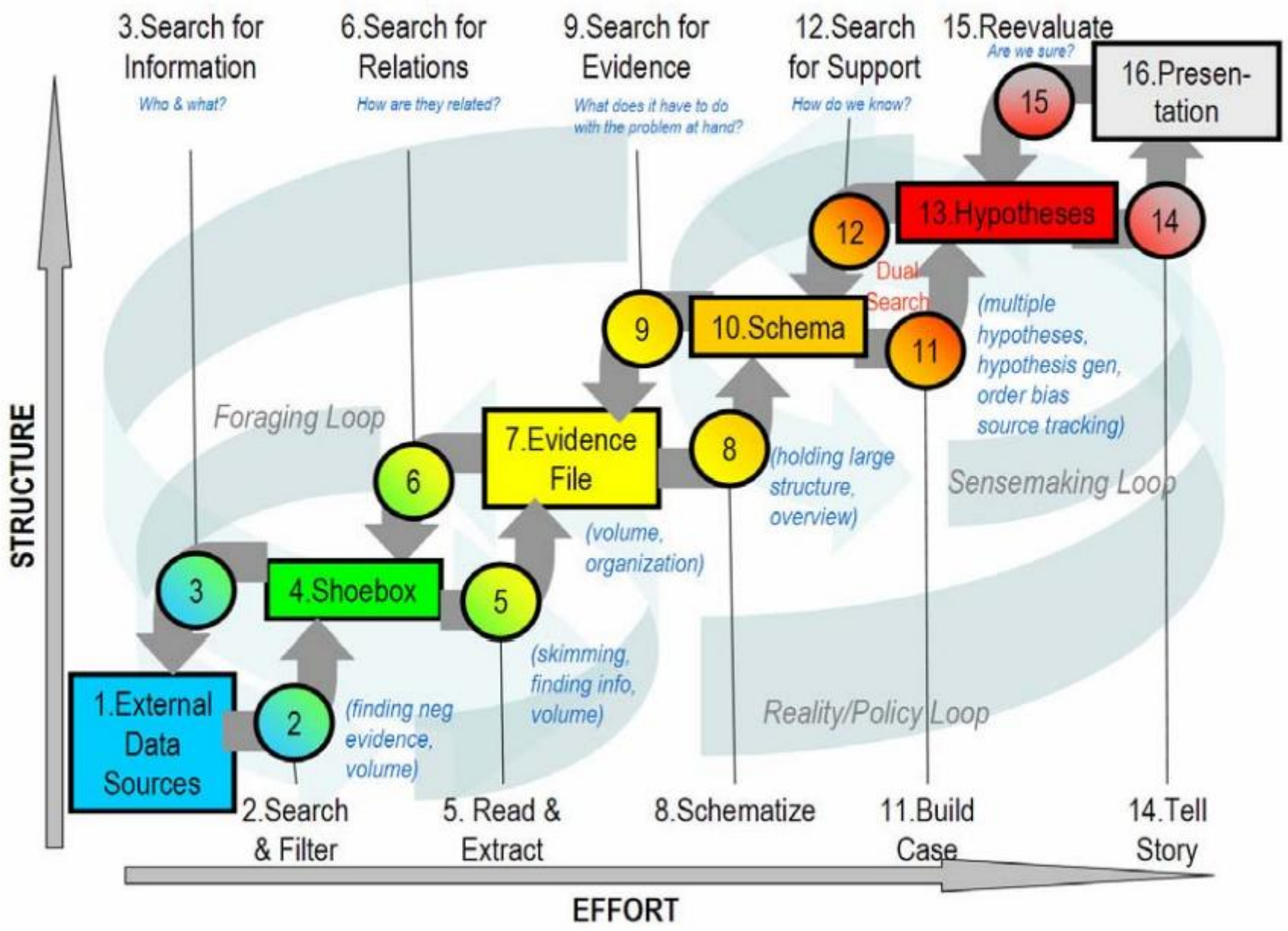
A comparison of selected performance parameters for deferent benchmarks with data analytics and mining workloads

Berkin Ozisikyilmaz, Ramanathan Narayanan, Joseph Zambreno, Gokhan Memik, and Alok N. Choudhary. An architectural characterization study of data mining and bioinformatics workloads. In IISWC, pages 61-70, 2006.

Generalized Scientific Workflow



Synergistic Challenges in Data-Intensive Science and Exascale Computing, Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee, Office of Science, US Department of Energy, 30, March 2013 page24



http://www.rosebt.com/1/post/2012/12/data-science-for-better-business-decisions-formula.html

Data Generation Phase (scenarios)	Overview	Transactional / in situ processing require-ments	Storage for Post processing	Sharing and distribution	Visualization
1) Design Oriented Exascale Simulations e.g., combustion, CFD	Generation of data from simulations.	(1) Data reduction for post processing (2) feature detection & tracking (3) advanced analytics	Reduced data	Low (Only the producer or a few scientists may analyze data in the future)	In-situ, inter-action, feature display, uncer-tainty, visual debugging
2) Discovery Oriented Exascale Simulations (2) e.g., climate, cosmology	Integration of data generated from simulation and observations	(1) Data reduction for post processing (2) time series (3) statistics (4) advanced analytics	(1) Raw data (2) Well organized (DB) (3) Enabled for queries	High (A large number of scientists, geographically distributed)	InfoVis and SciVis, pattern detection, correlation, clustering, ensemble vis, uncertainty
3) Large centralized instruments e.g., LHC	Data generation from large devices Extremely high rates Centralized, coordinated, controlled access	(1) HW/SW for high-rate data processing (2) derived data (3) metadata (4) Extensive queries	(1) Raw data (2) Different forms of derived data (3) Lots of distributed copies	High (A large number of scientists, geographically distributed), different sets defined by queries and other parameters	Custom user interfaces en-abling query visual analysis, trajectory vis/analysis, user driven data triage/-summarization
4) Smaller distributed instruments e.g., field work, sensors, biology	Data generation from massive numbers of distributed devices, sensors	(1) Local processing and derivations (2) Local analytics (3) Integration of massive data (possibly at an exascale level system, data centers)	(1) Raw data (2) Derived data and subsets (3) Distributed copies	High (A large number of scientists, geographically distributed)	InfoVis, high dimensional vis, large-scale graphs, patterns, clustering, scalability

Table 3.1: Data-generation requirements for different domains

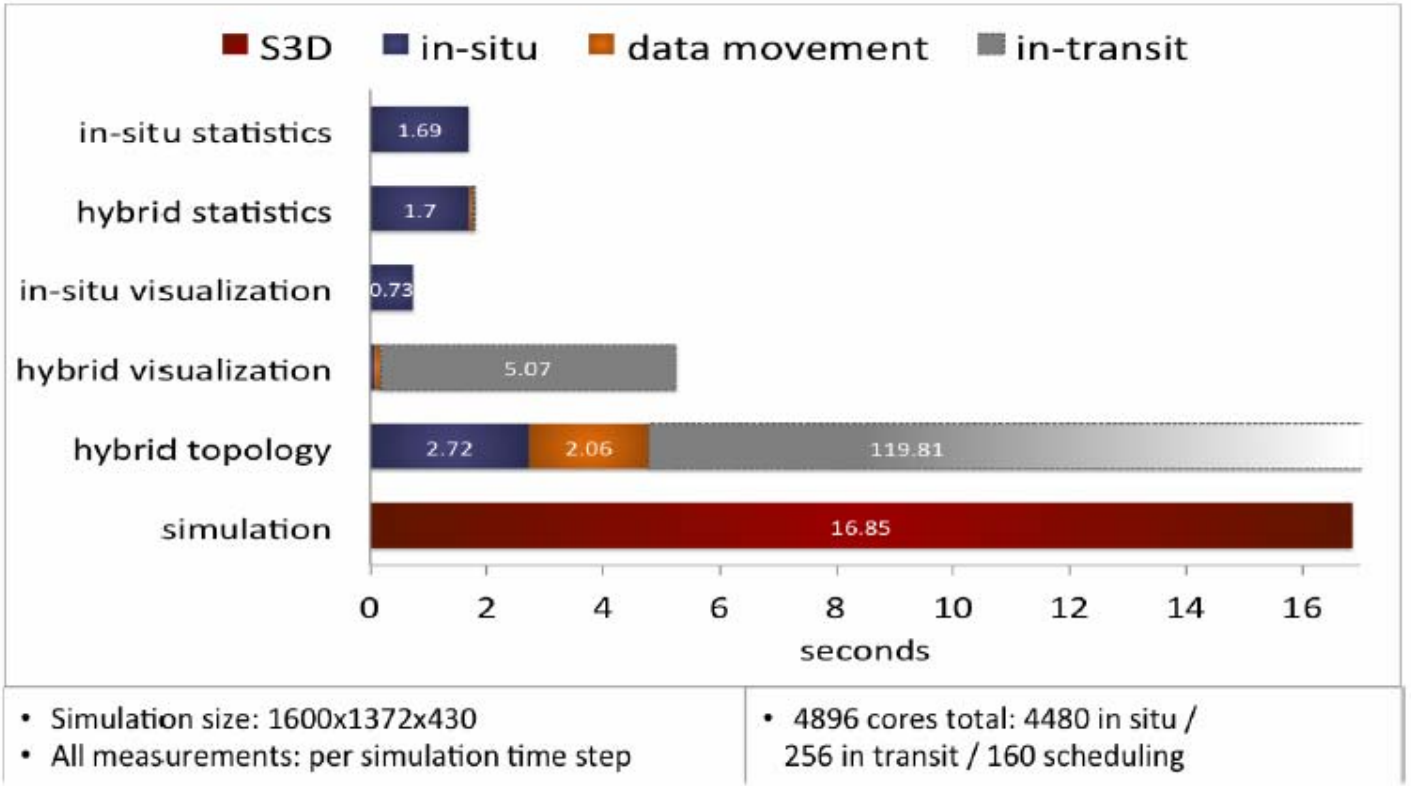


Figure 4.3: The timing breakdown (in seconds) for in-situ, in-transit, and data movement for the simulation and the various analytics algorithms using 4896 cores on Jaguar, the Cray XK6 at Oak Ridge National Laboratory's National Center for Computational Sciences. 4480 cores were used for the simulation and in-situ processing, 256 cores were used for in-transit processing and 160 cores were used for task scheduling and data movement. The simulation grid size was 1600x1372x430 and all measurements are per simulation time step.

where are we today?

preparing for post Moore Era : race for next level performance/ Exascale systems

- need it by the end of this decade/early next
- to achieve progress in the simulation of
  - societal impacts of weather, environmental change
  - continued certification of nuclear stockpile
  - combustion simulaiton
  - national security
  - to begin to understand brain functioning

DRAPA launched UHPC research aimed @ achieving petascale performance in a single rack system consuming only 57KW

DRAPA: Defence Advanced Research Projects Agency
UHPC: Ubiquitous High Performance Computing
Exascale Research: Preparing for the Post Moore Era Marc Snir, William Gropp and PeterKogge 6/19/11
https://www.ideals.illinois.edu/bitstream/handle/2142/25469/Exascale%20Research.pdf?sequence=2

Achieving Exascale X1000 in 2015 ?

Not just high end, floating point intensive, supercomputers ("exoflops" machines) but across the board 3 classes 1) data center-sized systems 2) departmental-sized systems and 3) embedded systems

Identified challenges

- Energy and Power Challenge
- Memory and Storage Challenge
- Concurrency and Locality Challenge
- Resiliency Challenge

More research reqd in Co-development and optimization of Exascale

- Hardware Technologies and Architectures
- Architectures and Programming Models
- Algorithms, Applications, Tools, and Run-times
- Development of a deep understanding of how to architect Resilient Exascale Systems

Suggested 3 phased research agenda

- A System Architecture Exploration Phase
- A Technology Demonstration Phase
- A Scalability Slice Prototyping Phase

ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems
http://users.ece.gatech.edu/mrichard/ExascaleComputingStudyReports/exascale_final_report_100208.pdf

current CMOS won't last

CMOS technology is slowing down

- Stein's Law… something cannot go on forever, forecasts a feature size of 7.5 nm by 2024
- will plateau next decade and no alternative is ready yet i.e., spintronics, Rapid Single Flux Quantum (RSFQ) Logic (requires cryogenic cooling)

Why we need exascale

http://www.bbc.co.uk/news/health-24428162

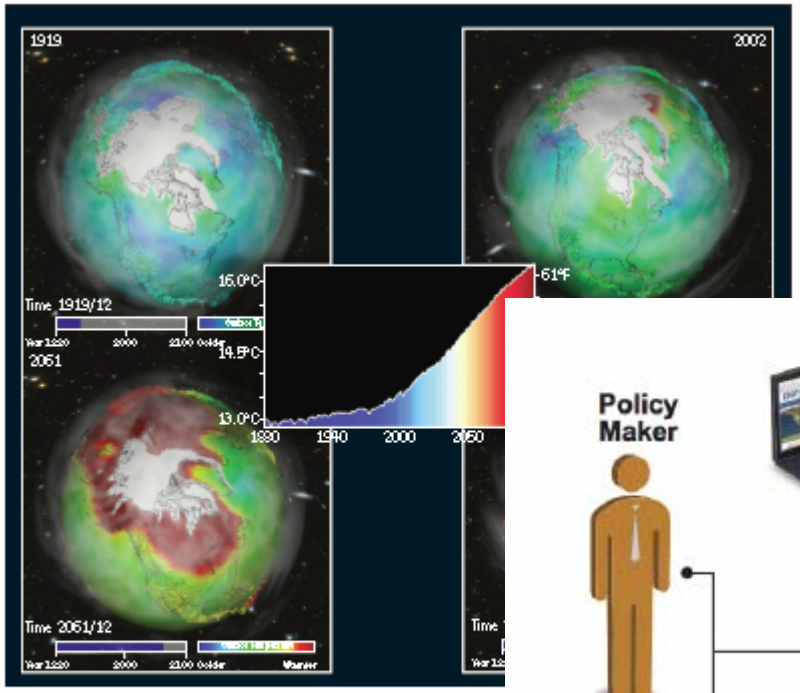


Figure 3.2: The surface of the earth is warming. Four maps
the trend, with average surface temperature for 1919 (top le
and 2099 (bottom right) shown. The global average, as a f
The computations were based on observed temperature (p
model (i.e., multi-model ensemble) assuming a continued h

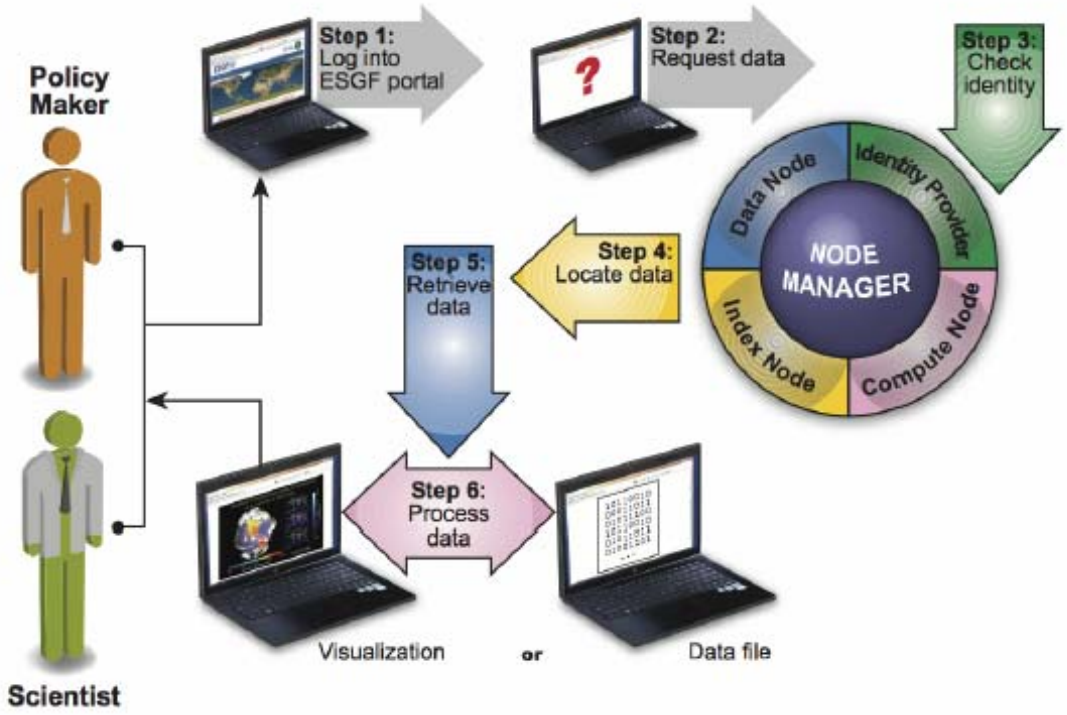


Figure 4.2: Federated data access in climate workflow

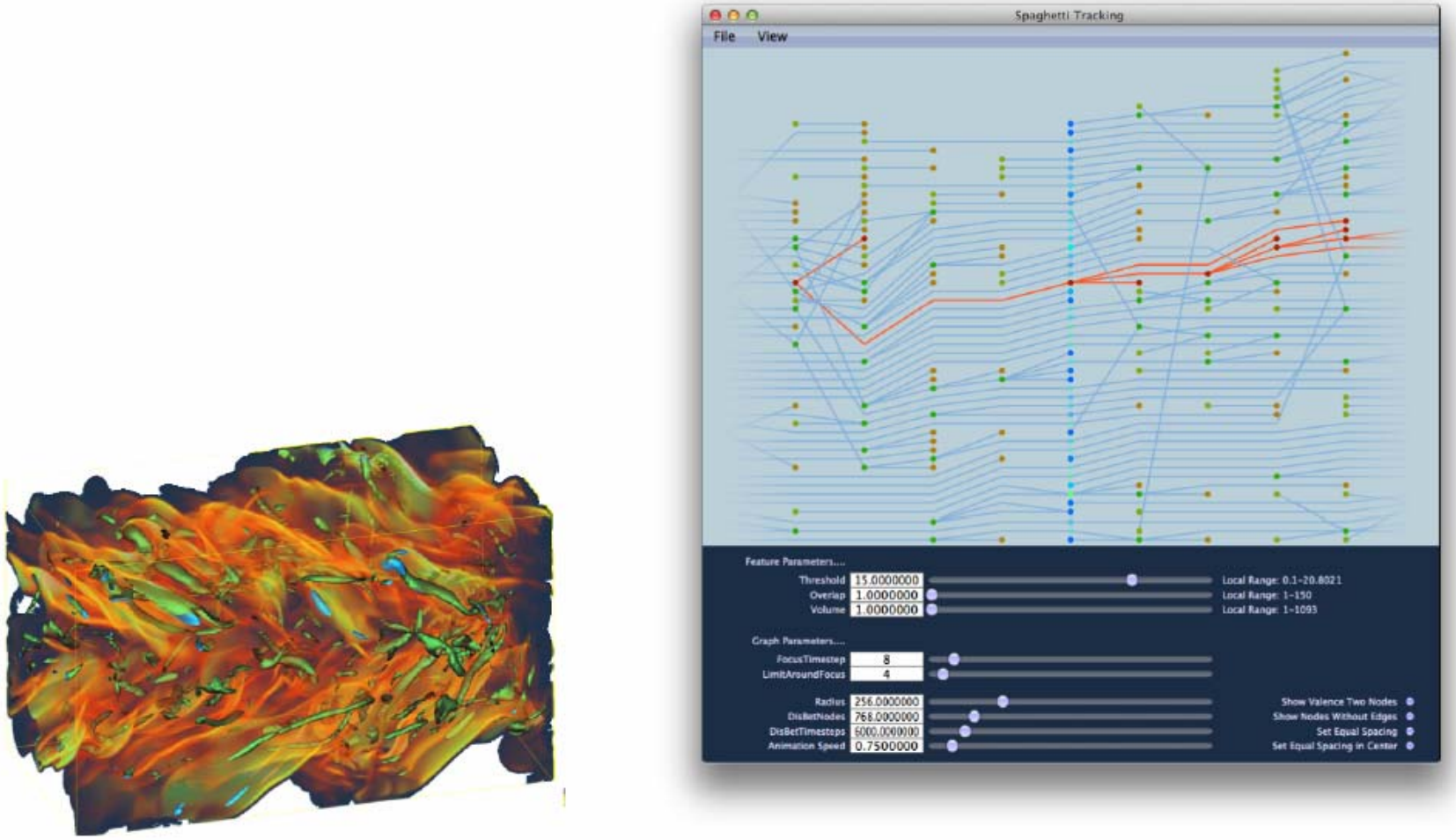


Figure 4.6: The visualization on the left shows features for a single time step within a large-scale combustion simulation while the figure on the right shows a graph that tracks features over time. Combustion scientists can follow specific events over time. Simulation by Jackie Chen, Sandia National Laboratories; Visualization by the Scientific Computing and Imaging (SCI) Institute.

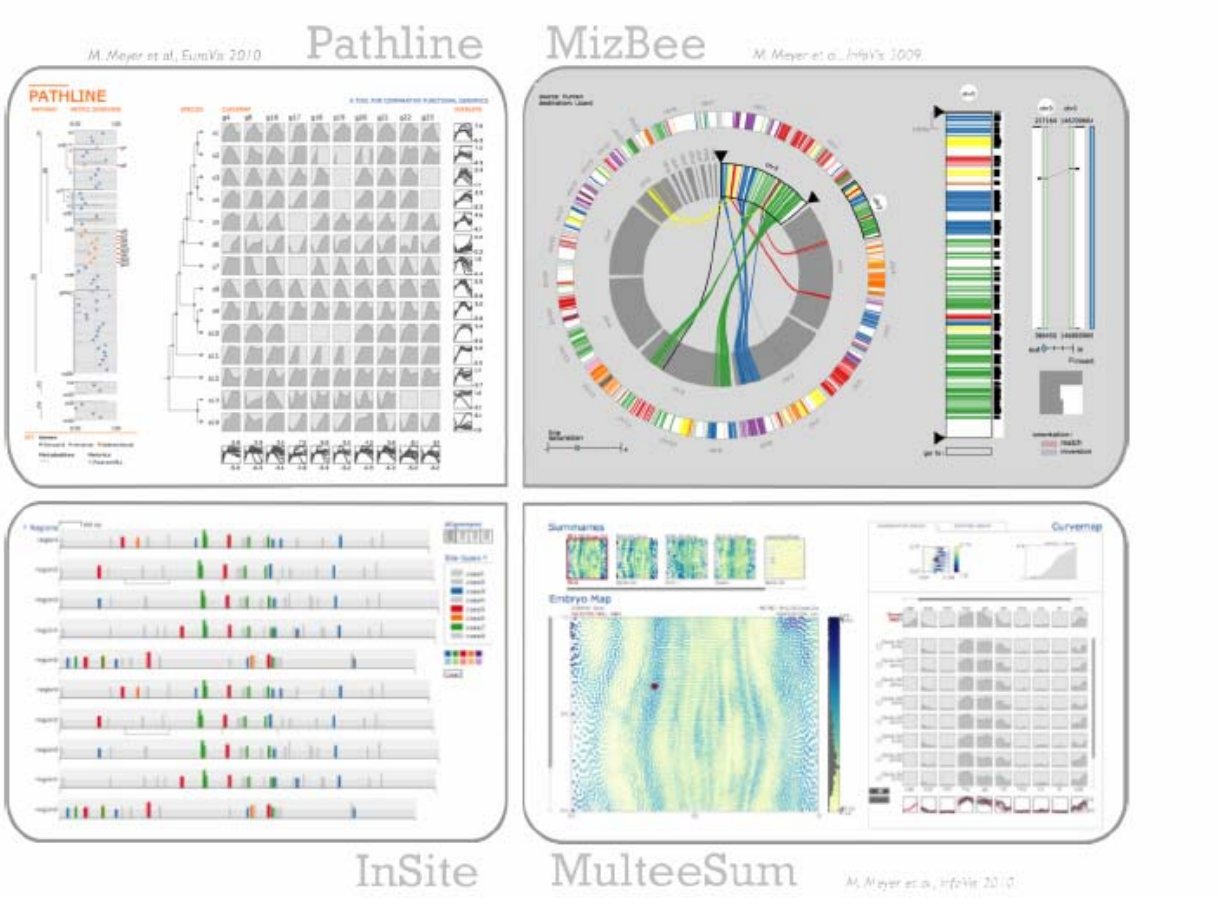
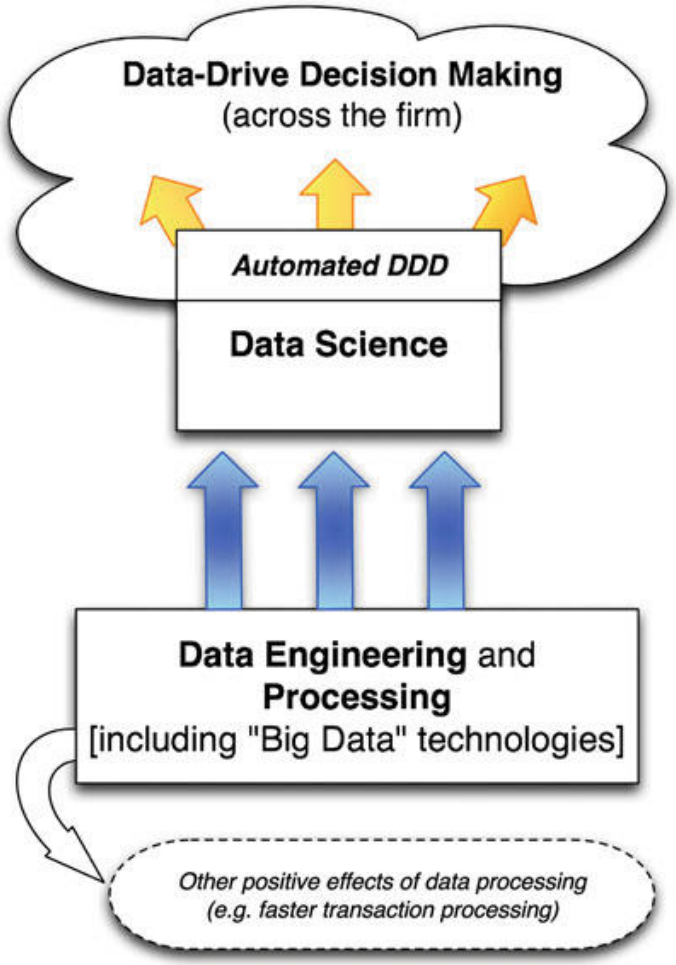
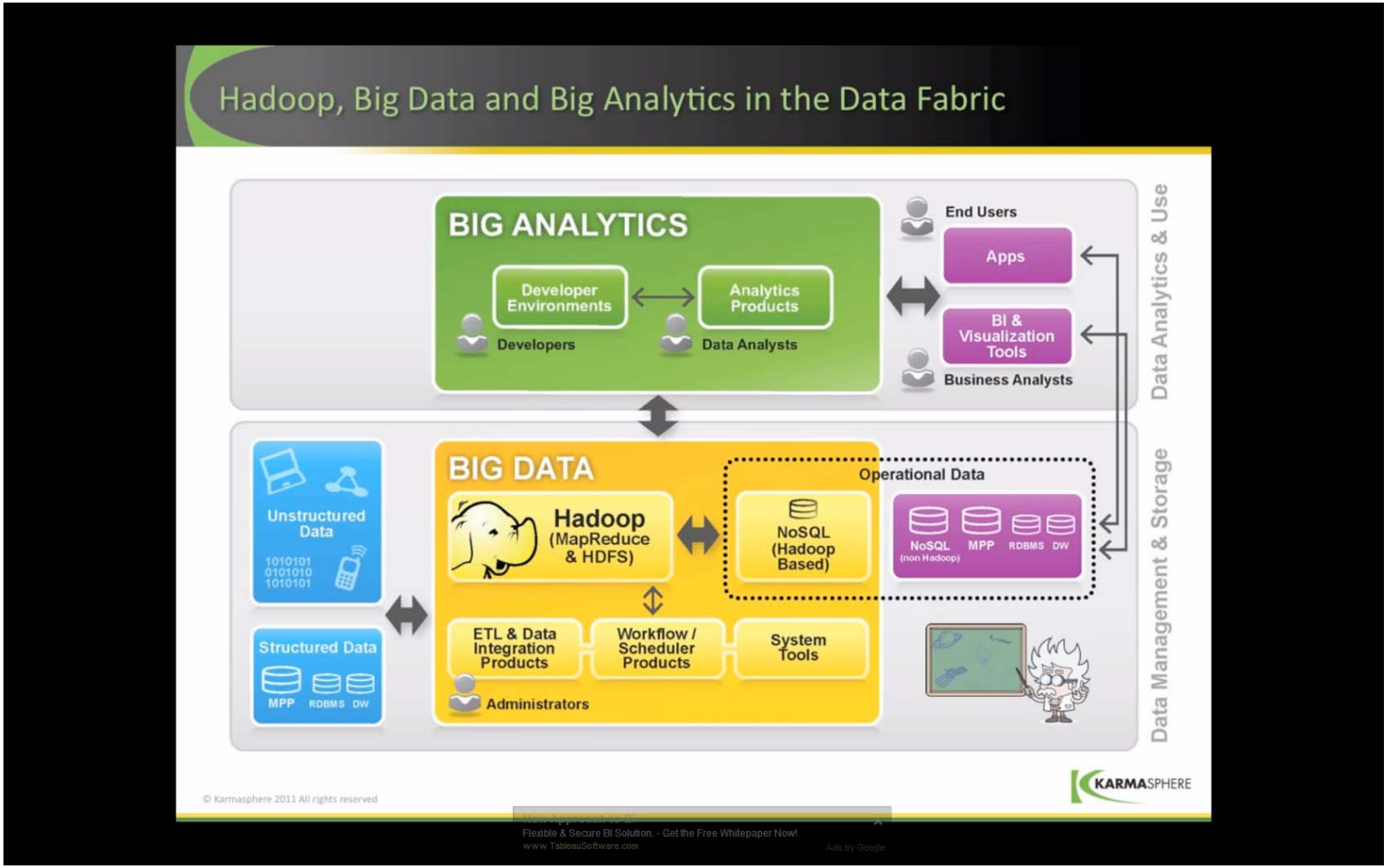


Figure 4.8: The rise of the fields of computational biology and bioinformatics has brought significant advances in algorithms for processing biological datasets. However, deciphering raw data and computational results through visual representations is too often done as the final step in a complex research process, with tools that are rarely specific to the task. Hence, there is a significant opportunity to enhance biological data analysis through a thoughtful and principled investigation of visualization.These four tools are examples of custom, interactive visualization tools designed in collaborations with biologists – they have all been deployed in biological research labs and led to a variety of scientific insights [14].

Data science

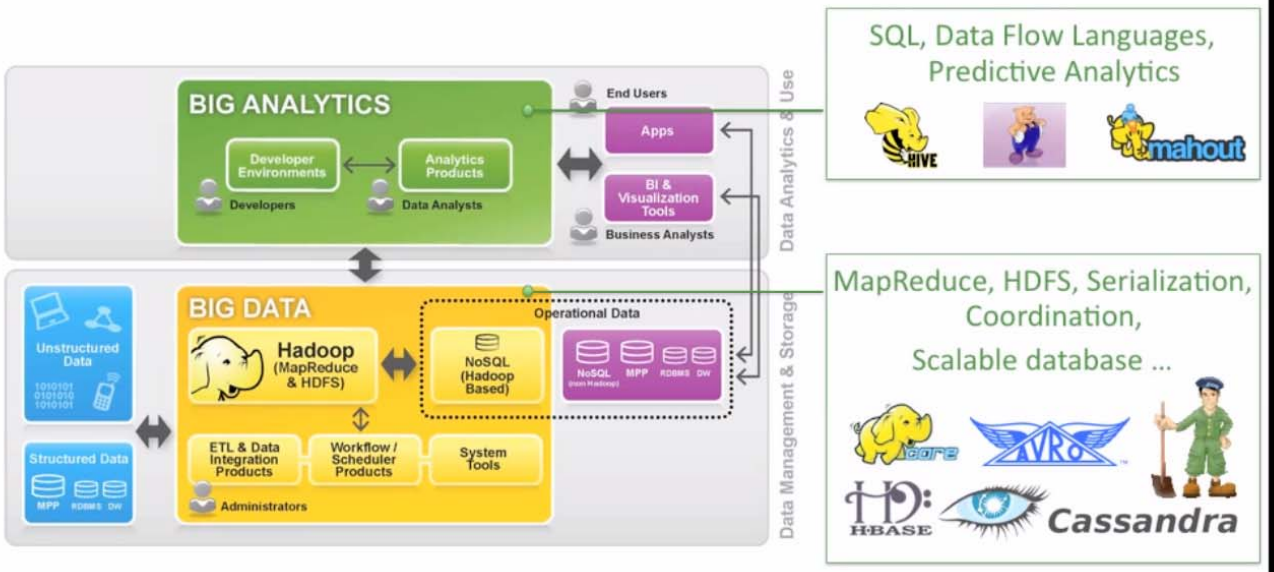


DOI: 10.1089/big.2013.1508 MARY ANN LIEBERT, INC. VOL. 1 NO. 1 MARCH 2013 BIG DATA



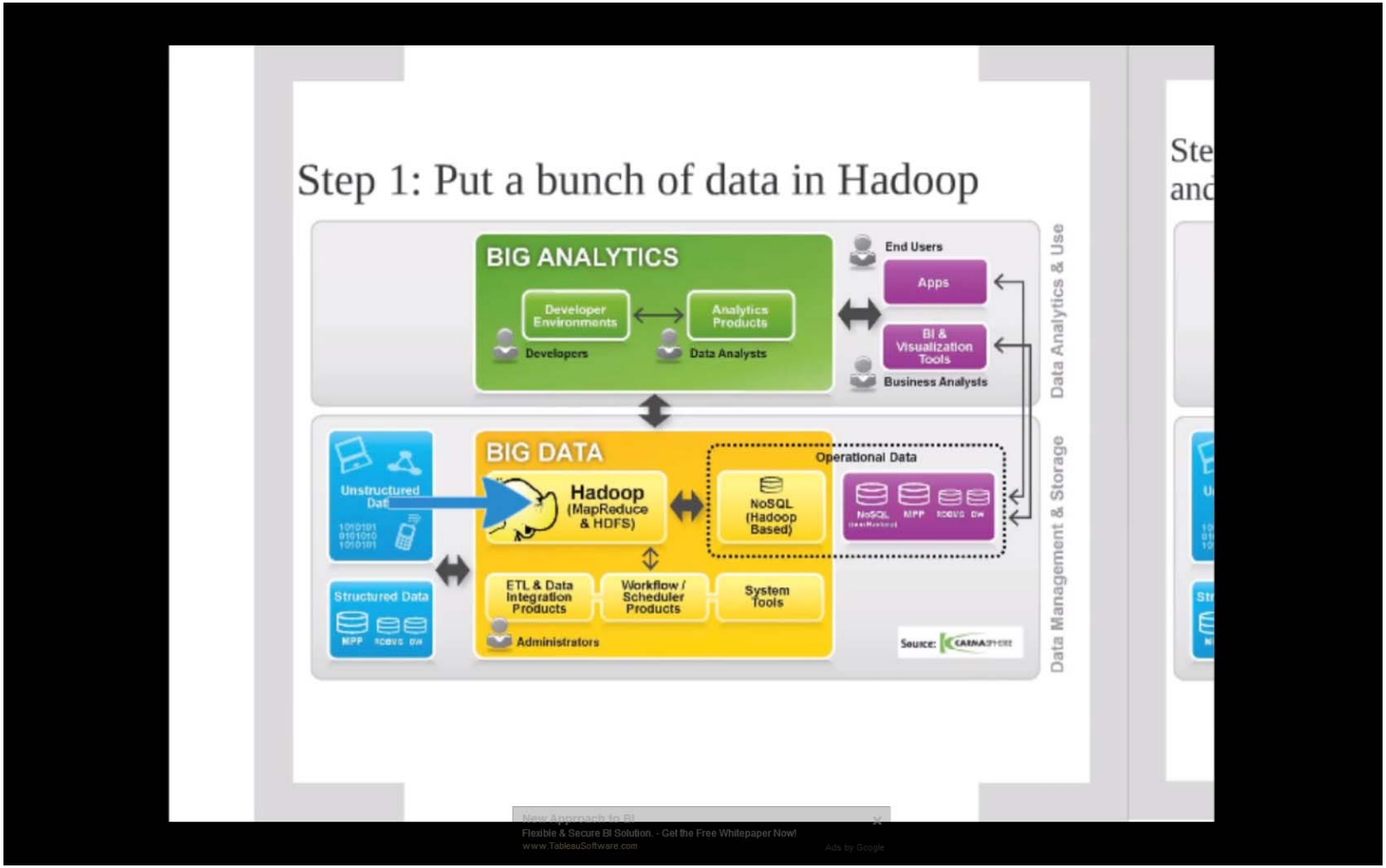
http://www.youtube.com/watch?v=duC4PDOBFwU#t=16

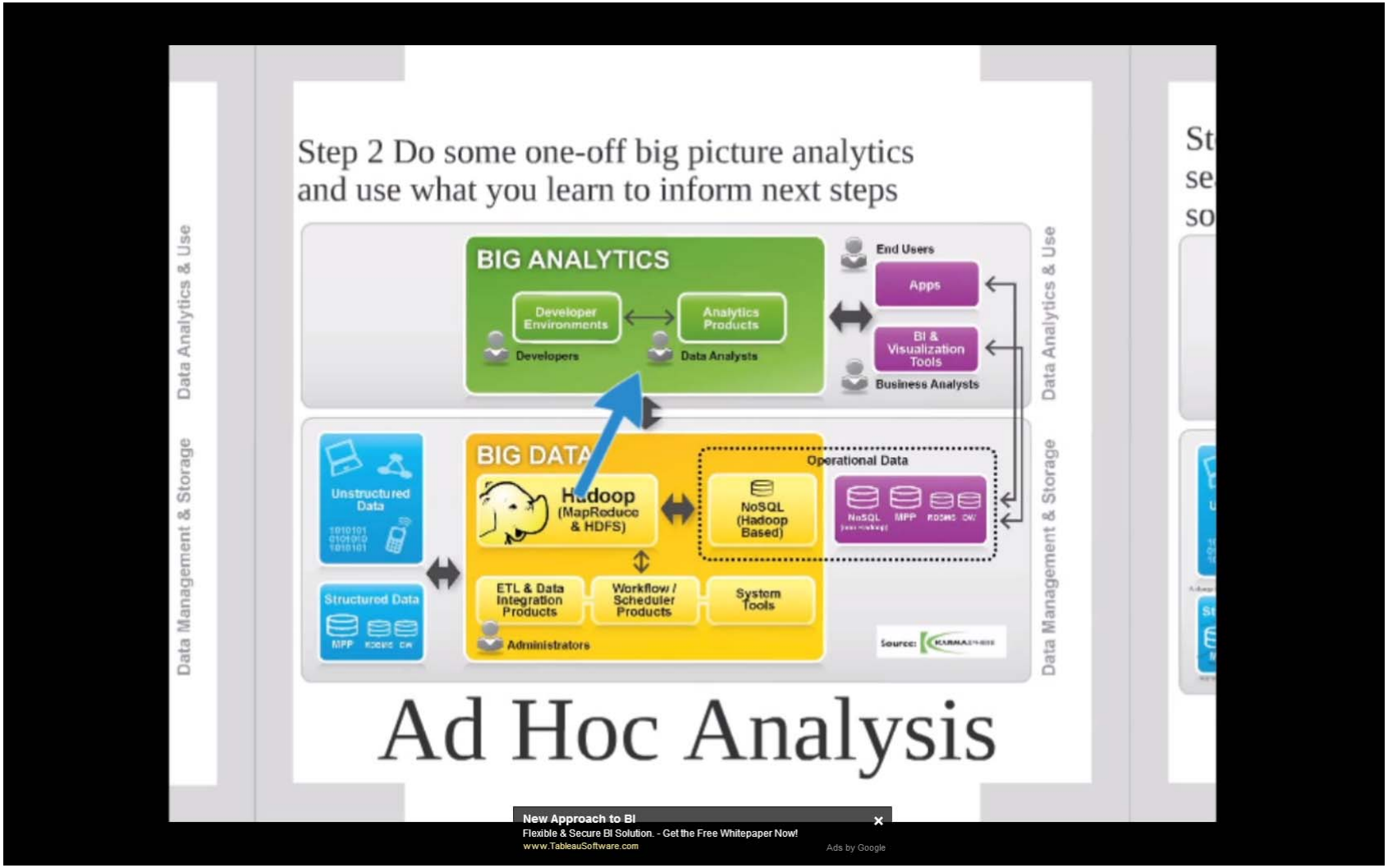
Open Source Apache Hadoop

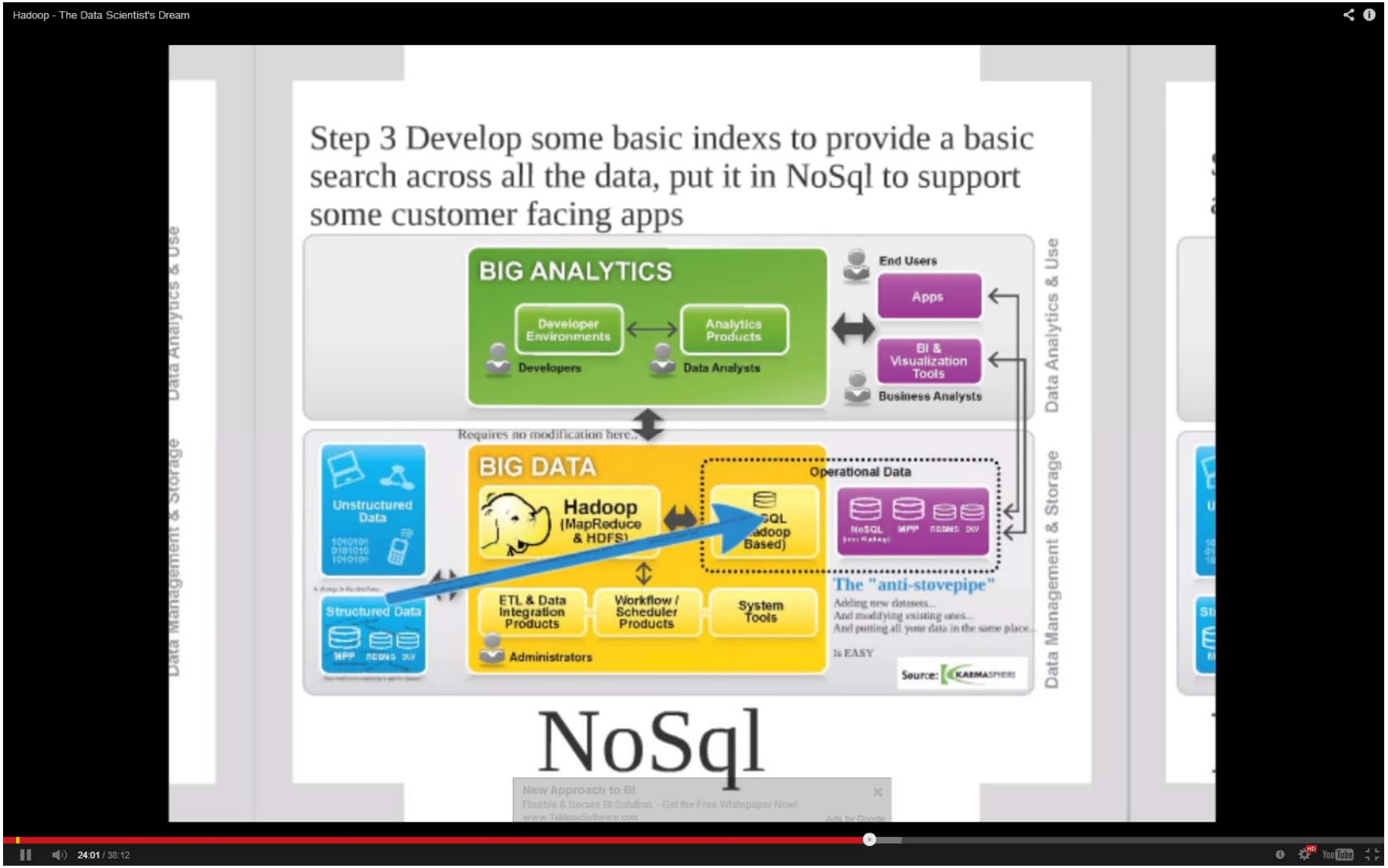


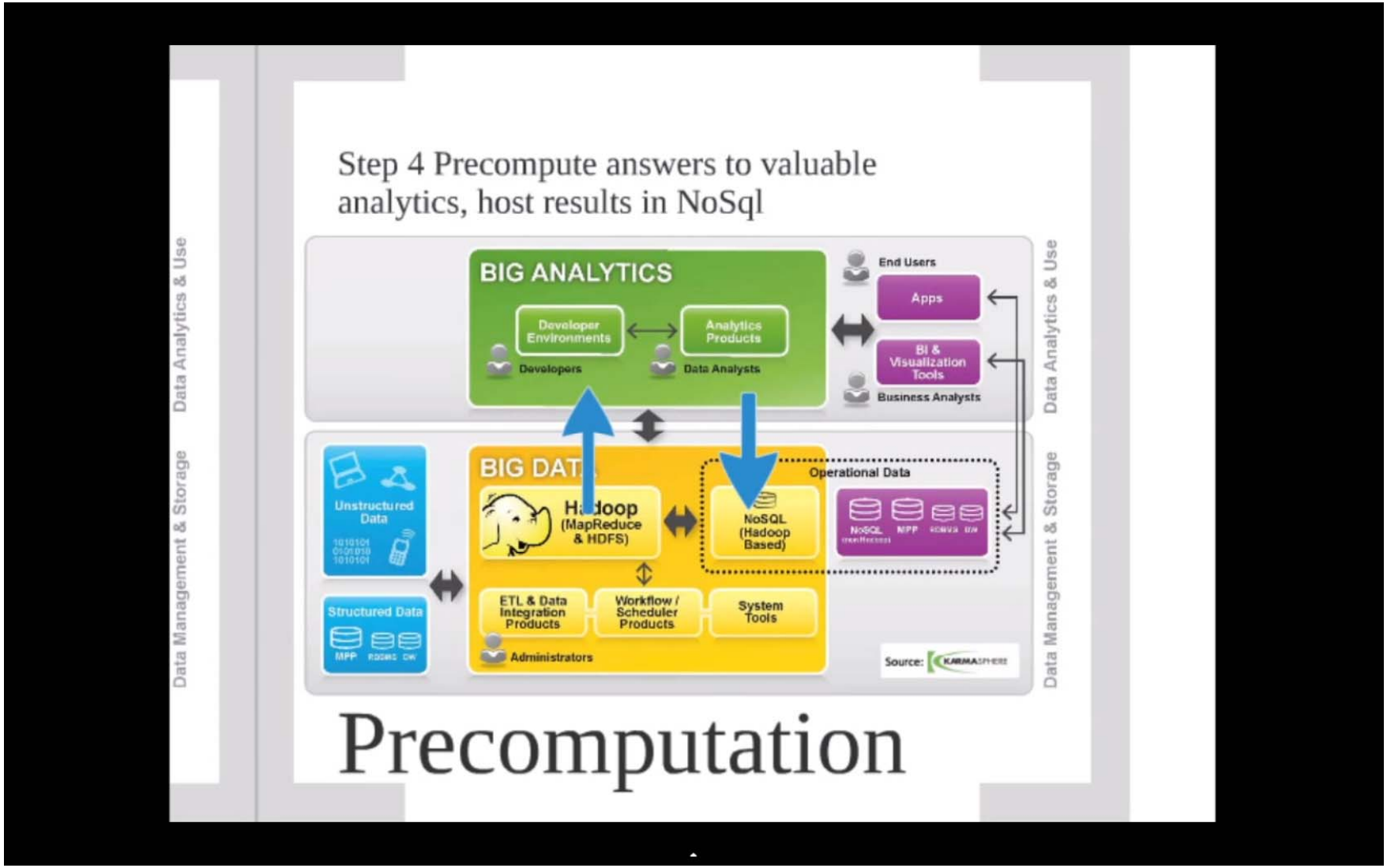
- Hadoop is an Apache Software Foundation collection of projects

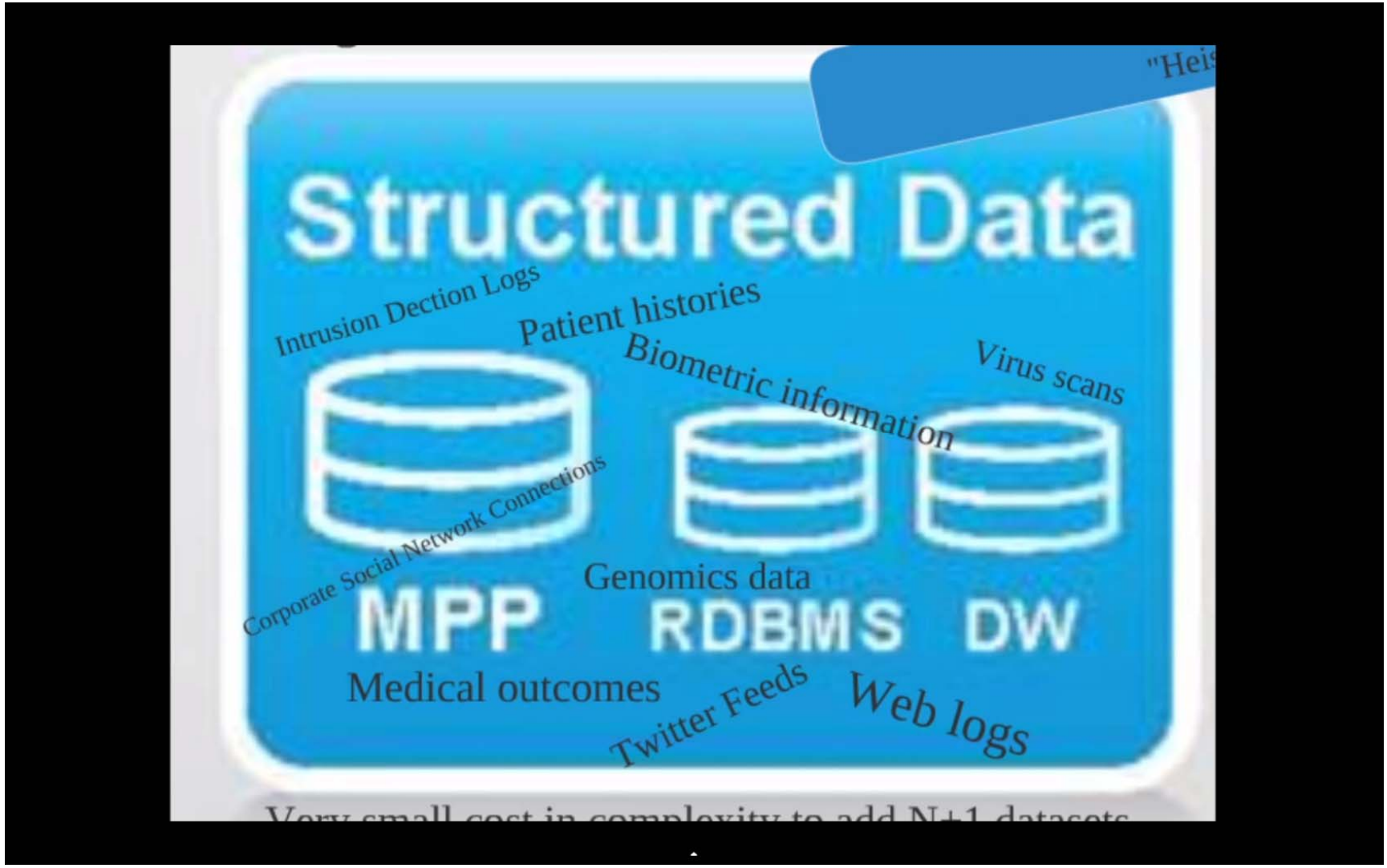
KARMASPHERE


Step 1: Put a bunch of data in Hadoop
BIG ANALYTICS
Developer Environments
Analytics Products
Developers
Data Analysts
End Users
Apps
BI & Visualization Tools
Business Analysts
Data Analytics & Use
BIG DATA
Unstructured Dat
Hadoop (MapReduce & HDFS)
Operational Data
NoSQL (Hadoop Based)
Structured Data
ETL & Data Integration Products
Workflow / Scheduler Products
System Tools
Administrators
Source:
Data Management & Storage
Flexible & Secure BI Solution. - Get the Free Whitepaper Now!
www.TableauSoftware.com
Ads by Google


Step 2 Do some one-off big picture analytics and use what you learn to inform next steps
Data Analytics & Use
Data Management & Storage
BIG ANALYTICS
Developer Environments
Analytics Products
Developers
Data Analysts
End Users
Apps
BI & Visualization Tools
Business Analysts
Unstructured Data
Structured Data
MPP
BIG DATA
Hadoop (MapReduce & HDFS)
ETL & Data Integration Products
Workflow / Scheduler Products
System Tools
Administrators
Operational Data
NoSQL (Hadoop Based)
NoSQL
MPP
Source:
Ad Hoc Analysis
New Approach to BI
Flexible & Secure BI Solution. - Get the Free Whitepaper Now!
www.TableauSoftware.com
Ads by Google


Hadoop - The Data Scientist's Dream
Step 3 Develop some basic indexs to provide a basic search across all the data, put it in NoSql to support some customer facing apps
Data Management & Storage
Data Analytics & Use
BIG ANALYTICS
Developer Environments
Analytics Products
Developers
Data Analysts
End Users
Apps
BI & Visualization Tools
Business Analysts
Requires no modification here...
BIG DATA
Unstructured Data
Structured Data
Hadoop (MapReduce & HDFS)
Operational Data
NoSQL
ETL & Data Integration Products
Workflow / Scheduler Products
System Tools
Administrators
The "anti-stovepipe"
Adding new datasets...
And modifying existing ones...
And putting all your data in the same place...
Is EASY
Source:
NoSql
New Approach to BI
Flexible & Secure BI Solution. - Get the Free Whitepaper Now!
24:01 / 38:12


Step 4 Precompute answers to valuable analytics, host results in NoSql
Data Analytics & Use
Data Management & Storage
BIG ANALYTICS
Developer Environments
Analytics Products
Developers
Data Analysts
End Users
Apps
BI & Visualization Tools
Business Analysts
Unstructured Data
Structured Data
MPP
Hadoop (MapReduce & HDFS)
NoSQL (Hadoop Based)
Operational Data
ETL & Data Integration Products
Workflow / Scheduler Products
System Tools
Administrators
Source:
Precomputation


Structured Data
Intrusion Dection Logs
Patient histories
Biometric information
Virus scans
Corporate Social Network Connections
Genomics data
MPP
RDBMS
DW
Medical outcomes
Twitter Feeds
Web logs


Scema-less format requires very little pre-processing
"Heisenberg principal" applied to data
ed Data
Virus scans
data
ETL &
Integ
Prod


(MapReduce
Hadoop - The Data Scientist's Dream
(non Hadoop)
The "anti-
Adding new datas
And modifying ex
And putting all yo
Is EASY
Hadoop - The Data Scientist's Dream
Network Defense
Anomalous Web Traffic Detection Using Cloud
Press Esc to exit full screen mode.
Cool User Interface
Hadoop - The Data Scientist's Dream
Press Esc to exit full screen mode.
Dicoveries quickly integrated into operational use through appropriate use of NoSql linked to a user interface
Analysis: discovers new
rns and threat vectors
omalous

Traditional Vs modern day fraud detection

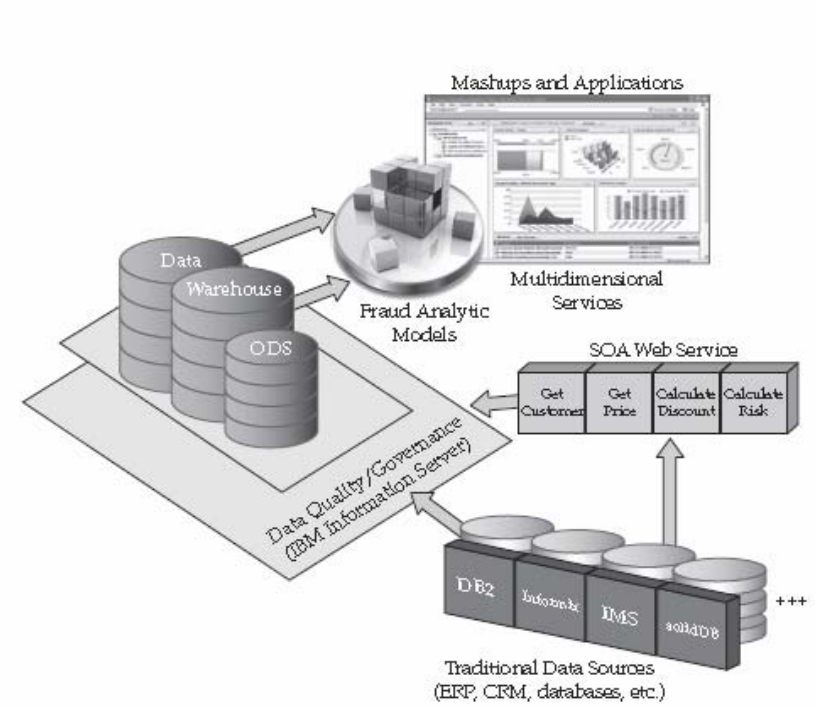


Figure 2-1 *Traditional fraud detection patterns use approximately 20 percent of available data.*

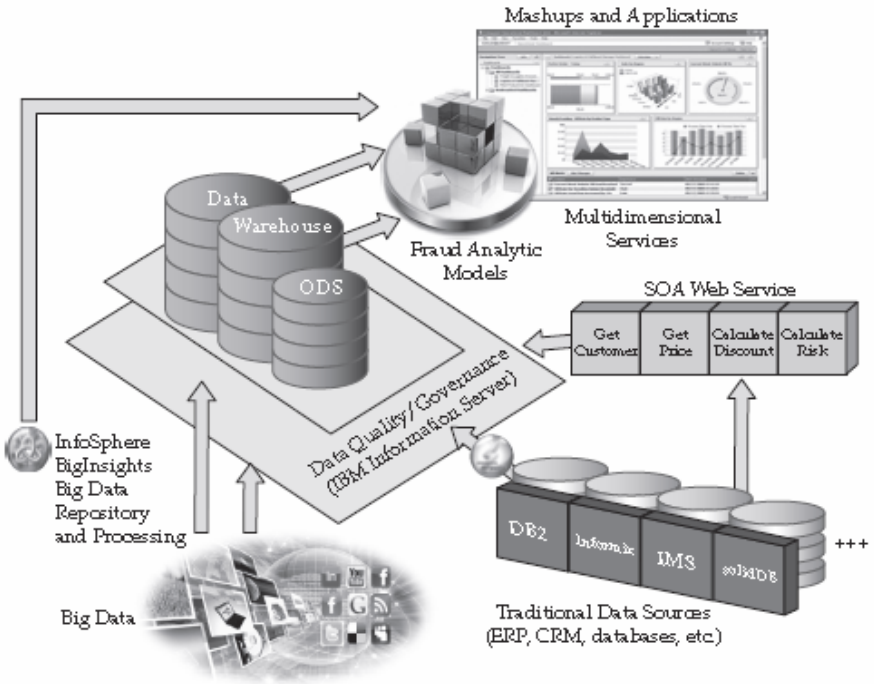


Figure 2-2 *A modern-day fraud detection ecosystem synergizes a Big Data platform with traditional processes.*

Understanding Big Data: ***Analytics for Enterprise Class Hadoop and Streaming Data: page 22-23***

https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=500016891&S_CPM=is_bdebook1_hdfs

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection(1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access(1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing &Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining(Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

http://www.thearling.com/text/dmwhite/dmwhite.htm

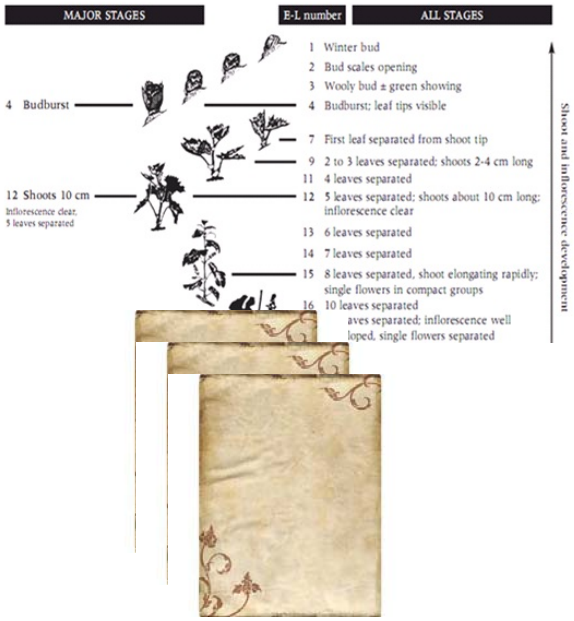
GRC research projects

1. climate change effects on grapevine phenology and wine quality
2. multi-sensor data analysis
3. Pixel clustering for spatial data mining

Climate change effects on grapevine phenology and wine quality

Precise and structured

Precise and structured



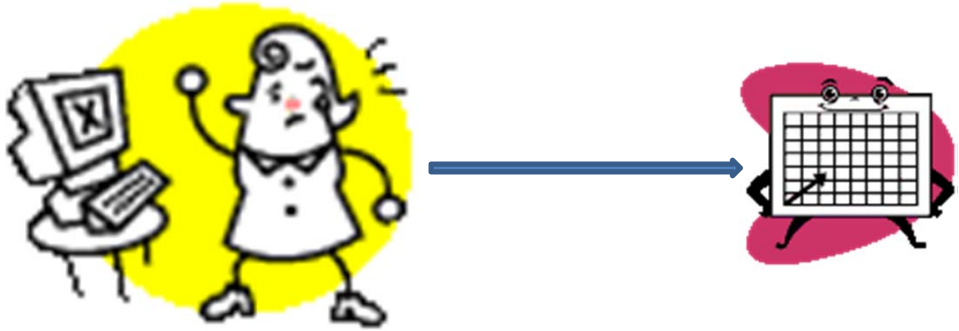
1000 years old diary

Video, text, ratings
Audio, web

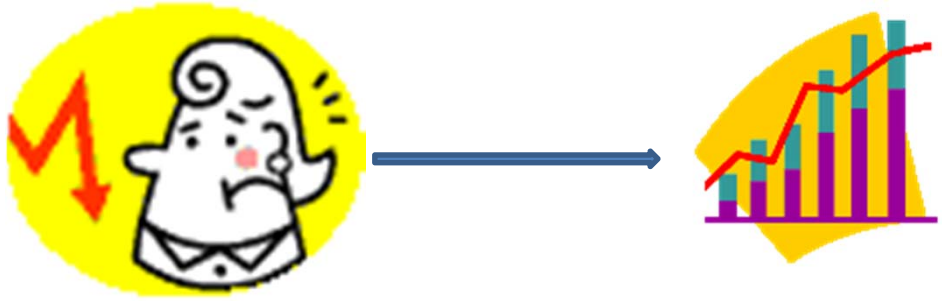
sommelier comments…

Issues and solutions

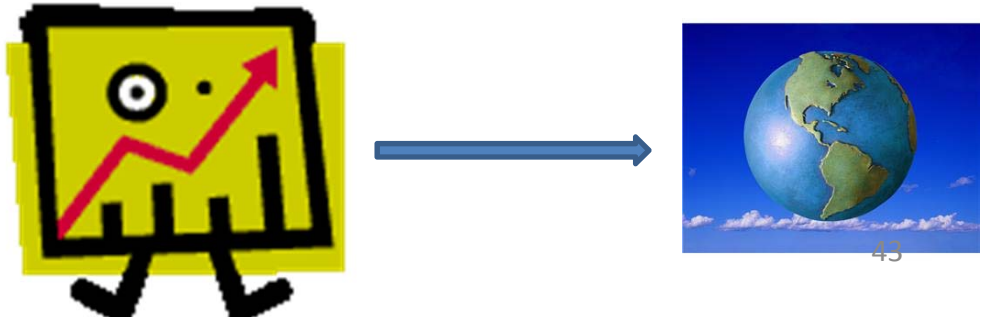
Extract data from web portal



Data/text pre-processing

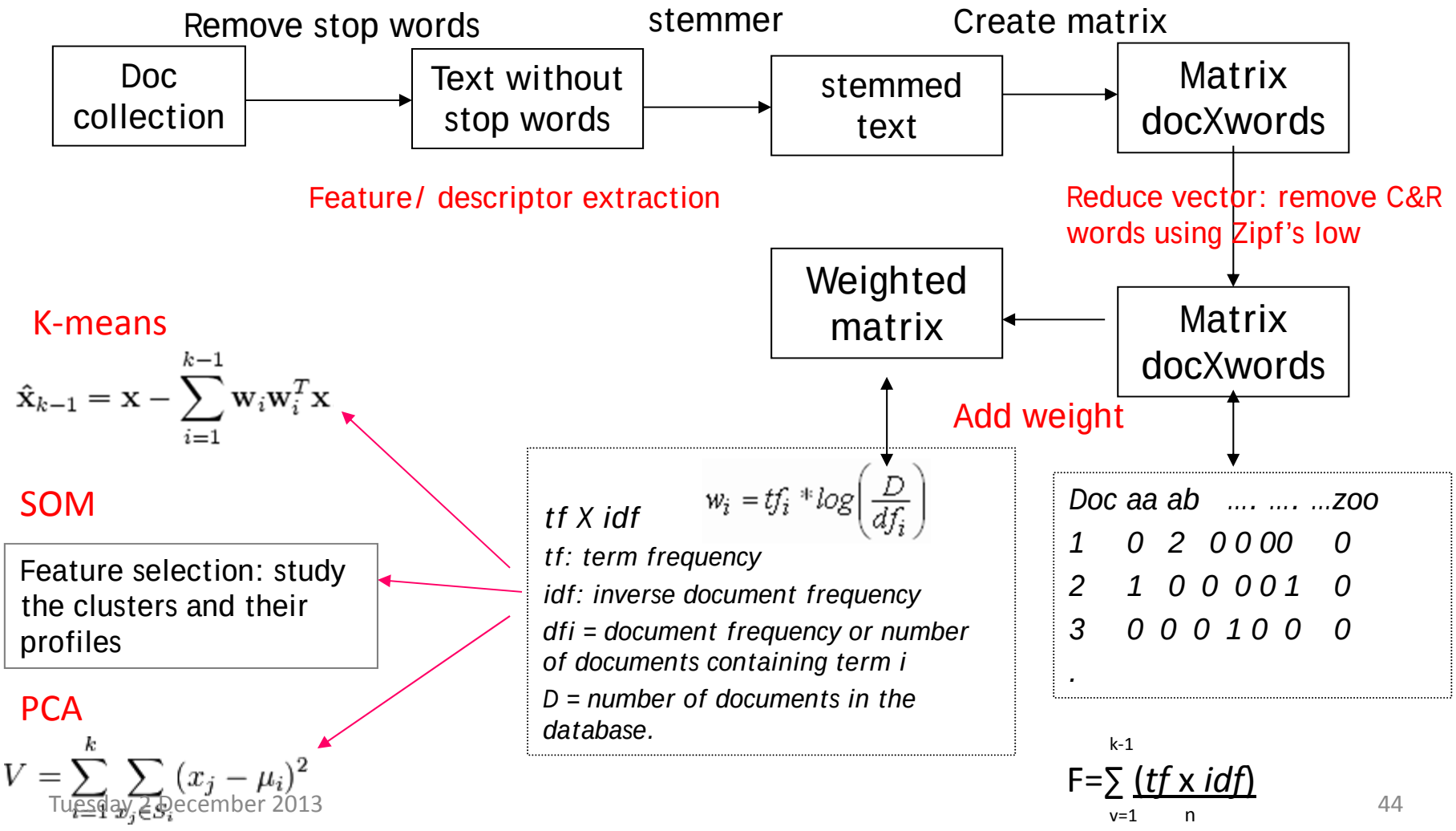


data Dimension reduction

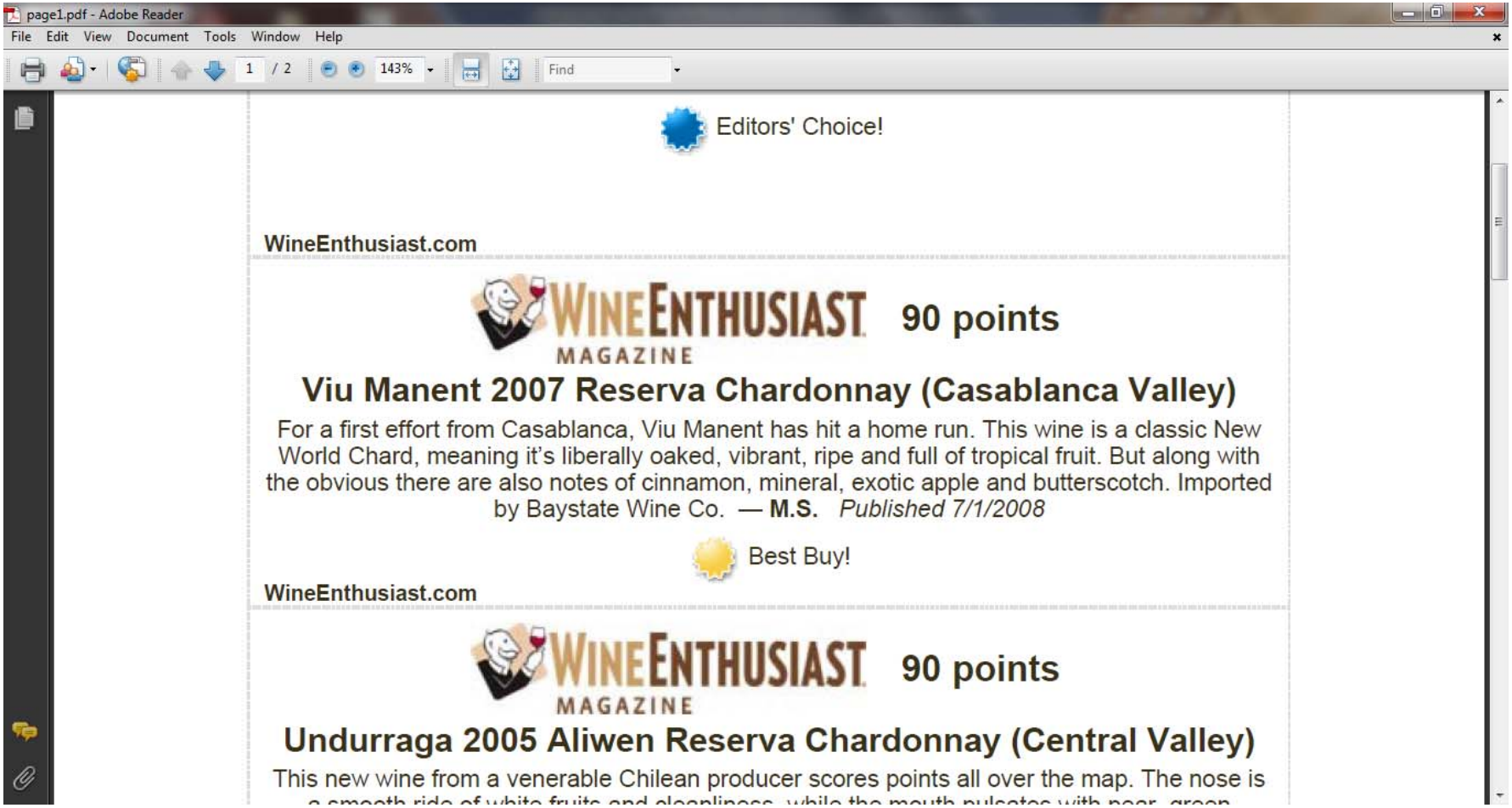


Analysis

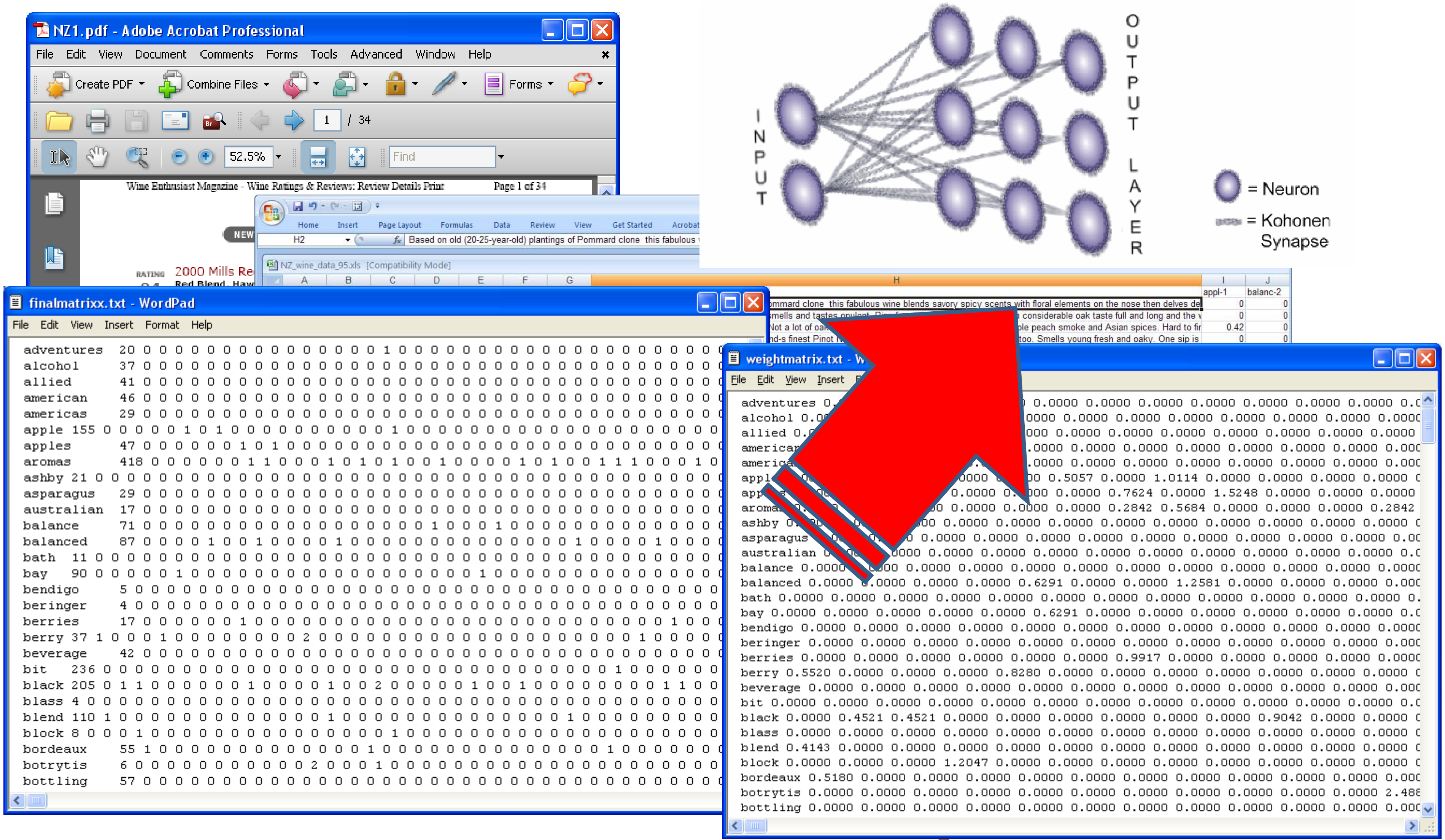
Text mining-vector space model



Text mining : Sommelier comments



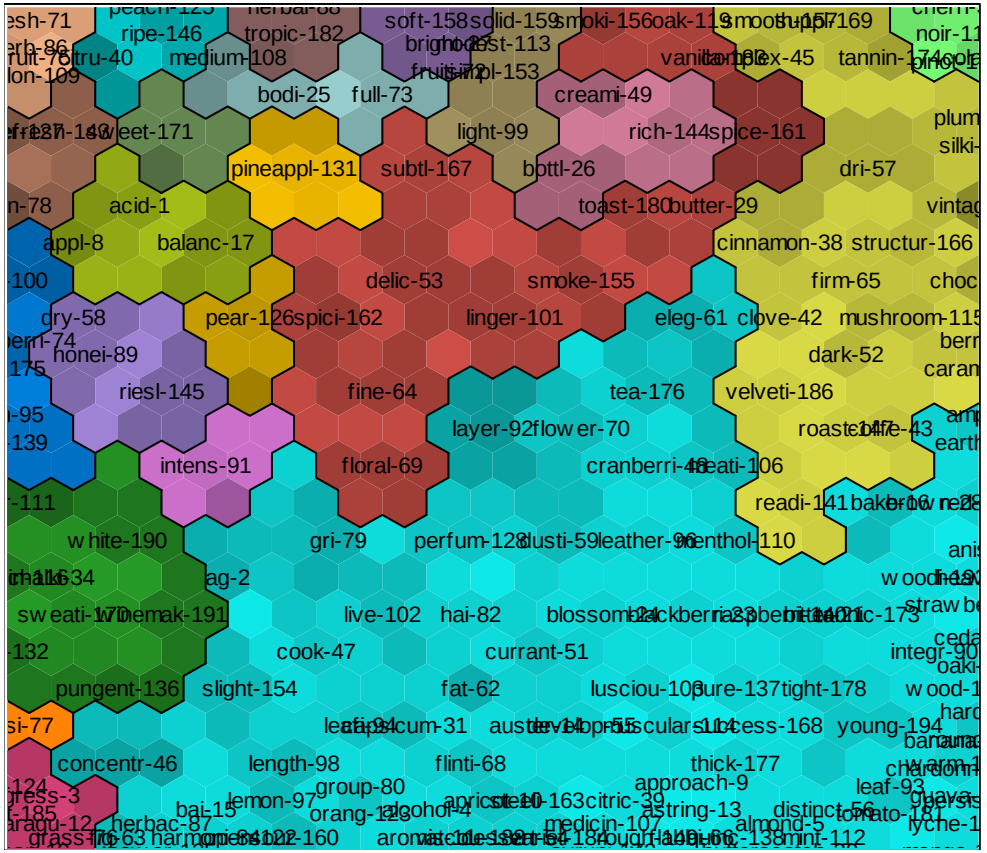
Web text mining wine comments

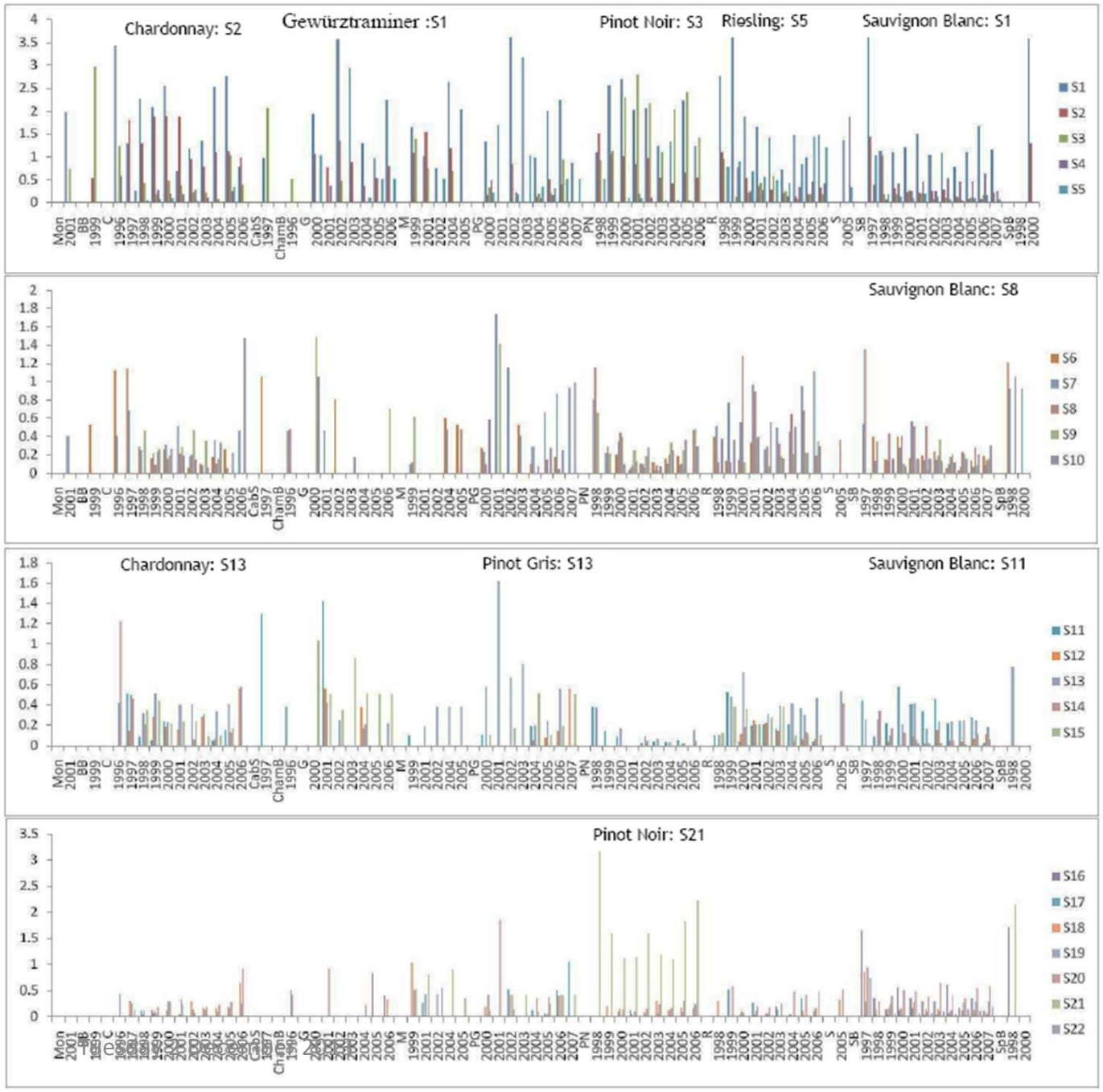


websom results

The WEBSOM approach to Marlborough wine styles

WEBSOM of **195** descriptors extracted from sommelier comments provided for **253** Marlborough vintages of styles produced from this famous wine region. The different segments in the SOM show the descriptors used to state the features (S1-S22) of the vintages by sommeliers.





S1 (Sauvignon Blanc): eleg-61, tea-176, layer-92, flower-70, ampl-6 earthi-60, cranberri-48, meati-106, bake-16, brown-28, red-142, gri-79, perfum-128, dusti-59, leather-96, menthol-110, ag-2, woodi-193, anis-7 heavi-85 strawberri-164, live-102, hai-82, blossom-24, blackberri-23, raspberri-140, bitter-21, tannic-173, cook-47, currant-51, integr-90, cedar-33 oaki-120, slight-154, fat-62, lusciou-103, pure-137, tight-178, wood-192, leafi-94, capsicum-31, auster-14, develop-55, muscular-114, success-168, young-194, hard-83 round-150, concentr-46, length-98, flinti-68, thick-177, warm-189, lemon-97, group-80 orang-123, apricot-10, steeli-163, citric-39, leaf-93 tomato-181, persist-129, fig-63, herbac-87 power-135, bai-15 harmoni-84 sharp-151, open-122, sour-160, alcohol-4 aromat-11 strong-165, viscou-188, dessert-54, variet-184, medicin-107 syrupi-172, rough-149, approach-9 astring-13 flabbi-66 fleshi-67 zesti195, quinc-138, almond-5 butterscotch-30, distinct-56 mint-112 nutti-118, banana-1

S2 (Chardonnay): smoki-156, oak-119, vanilla-183, spice-161, subtl-167, toast-180, butter-29, delic-53, smoke-155, spici-162, linger-101, fine-64, floral-69

S3 (Pinot Noir): smooth-157, suppl-169, complex-45, tannin-174, plum-134 silki-152, dri-57, vintag-187, cinnamon-38, structur-166, firm-65, chocol-37, clove-42, mushroom-115, dark-52, berri-20 caramel-32, velveti-186, roast-147, coffe-43, readi-141

S19: herbal-88 tropic-182

S20: clean-41 crisp-50 fresh-71 herb-86 melon-109, grapefruit-75 **S21**: black-22 cherri-36 noir-117 pinot-133, cola-44

S22: grassi-77

S 2	butter-29 delic-53 fine-64 floral-69 linger-101 oak-119 smoke-155 smoki-156 spice-161 spici-162 subtl-167 toast-180 vanilla-183 - **Chardonnay**
S 3	berri-20 caramel-32 chocol-37 cinnamon-38 clove-42 coffe-43 complex-45 dark-52 dri-57 firm-65 mushroom-115 plum-134 readi-141 roast-147 silki-152 smooth-157 structur-166 suppl-169 tannin-174 velveti-186 vintag-187 **– Pinot Noir**
S 4	chalki-34 miner-111 nectarin-116 pink-132 pungent-136 sweati-170 white-190 winemak-191
S 5	dry-58 honei-89 riesl-145
S 6	bottl-26 creami-49 rich-144
S 7	acid-1 appl-8 balanc-17
S 8	gooseberri-74 lean-95 lime-100 raci-139 tart-175
S 9	light-99 modest-113 simpl-153 solid-159
S 10	bodi-25 full-73 medium-108
S 11	green-78 pepper-127 refresh-143
S 12	pineappl-131
S 13	citru-40 peach-125 ripe-146
S 14	intens-91
S 15	pear-126
S 17	sweet-171
S 18	bright-27 fruiti-72 soft-158
S 19	herbal-88 tropic-182
S 20	clean-41 crisp-50 fresh-71 grapefruit-75 herb-86 melon-109
S 21	black-22 cherri-36 cola-44 noir-117 pinot-133
S 22	grassi-77

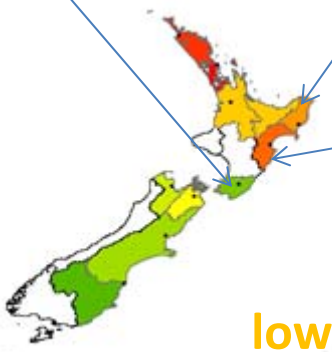
NZ Chardonnay

Waipara
toast-8 <= 0.26
| citru-3 <= 0: med (8.0/2.0)
| citru-3 > 0: high (2.0/1.0)
toast-8 > 0.26: high (3.0)

Training: 76%
Cross validation : 38%

Gisborne
sweet-19 <= 0
| spice-18 <= 0
| | appl-1 <= 0.27: med (28.0/7.0)
| | appl-1 > 0.27: high (2.0)
| spice-18 > 0: high (3.0/2.0)
sweet-19 > 0
| vanilla-23 <= 0: med (3.0)
| vanilla-23 > 0: low (3.0)

Training: 76%
Cross validation : 46%



Hawke's Bay
lime-19 <= 0
| ripe-28 <= 0.23
| | orang-23 <= 0
| | | creami-9 <= 0
| | | | honei-17 <= 0
| | | | | intens-18 <= 0: med (19.0/3.0)
| | | | | intens-18 > 0: high (5.0/1.0)
| | | | honei-17 > 0: high (2.0)
| | | creami-9 > 0: high (2.0)
| | orang-23 > 0: high (3.0)
| ripe-28 > 0.23: med (8.0/1.0)
lime-19 > 0: med (6.0/1.0)

Training: 86%
Cross validation : 48%

low <80 medium (med) >79 and <90 high >89 (100 point)

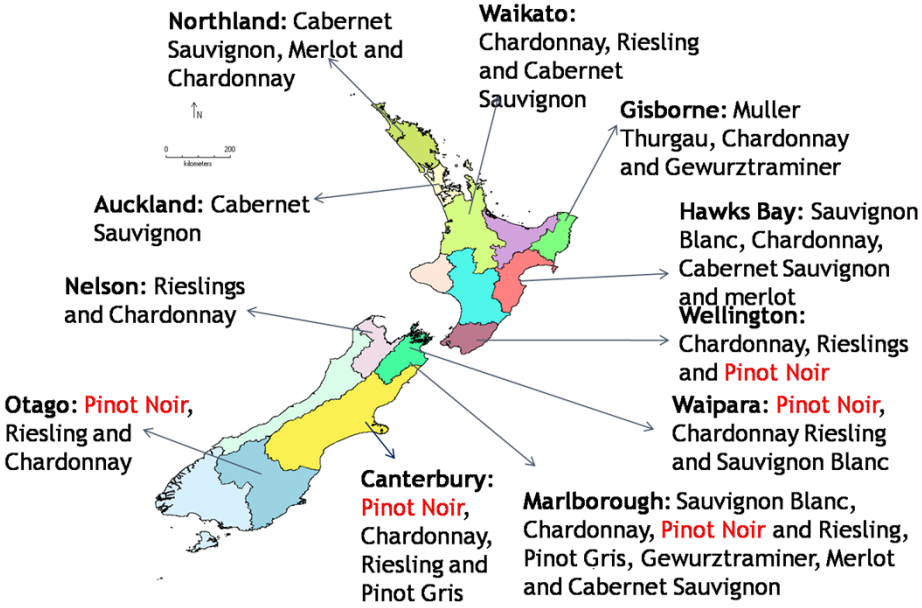
Chardonnay

region	low	med	high
Waipara			toast, citrus
Gisborne	vanilla		apple. spice
Hawke's Bay			Intense, honey, creamy, orange

Climate Vs Viticulture

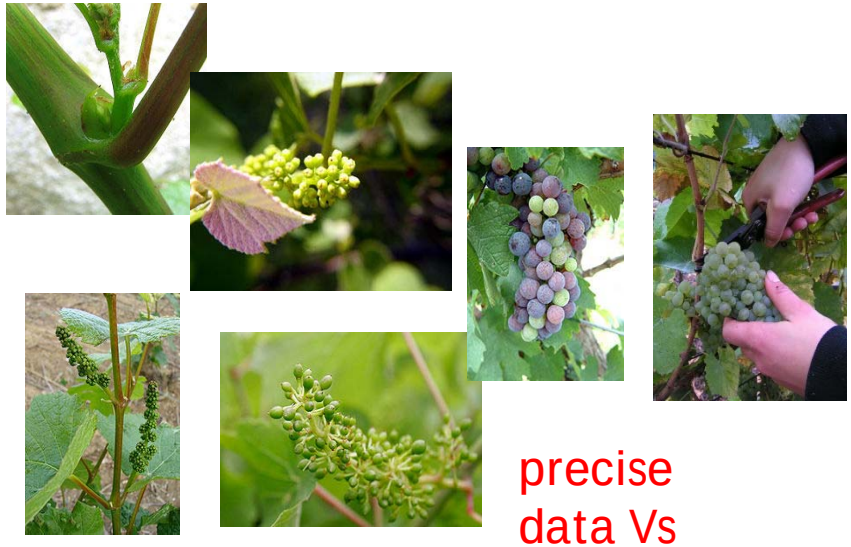
Base climate

- 3-5 decade average
- Used for Grapevine variety selection



Annual (year to year)

- Determines vintage quality
- Responsible for phenology



precise data Vs subjective wine quality

Spatiotemporal scales & climate change modelling

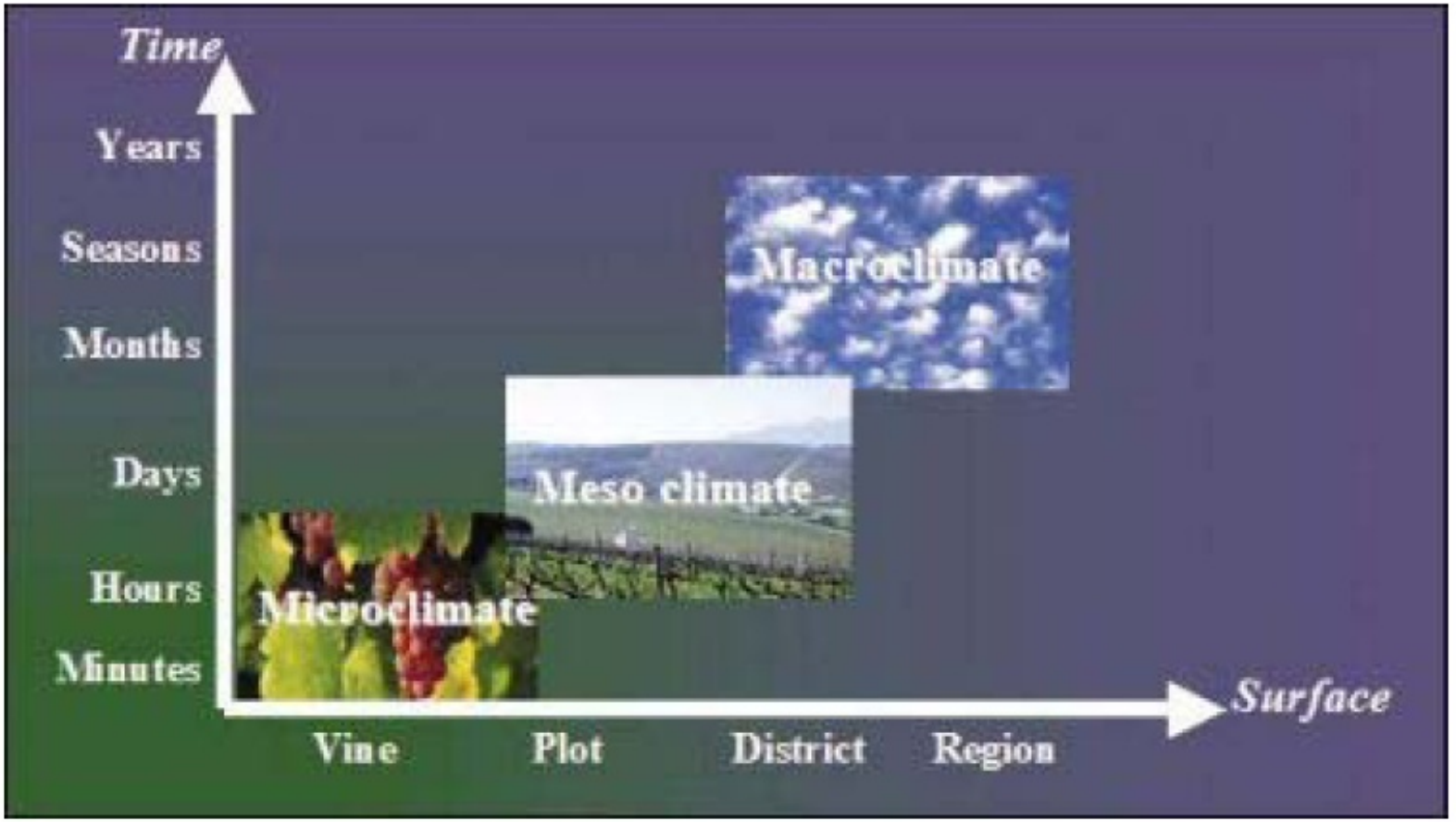
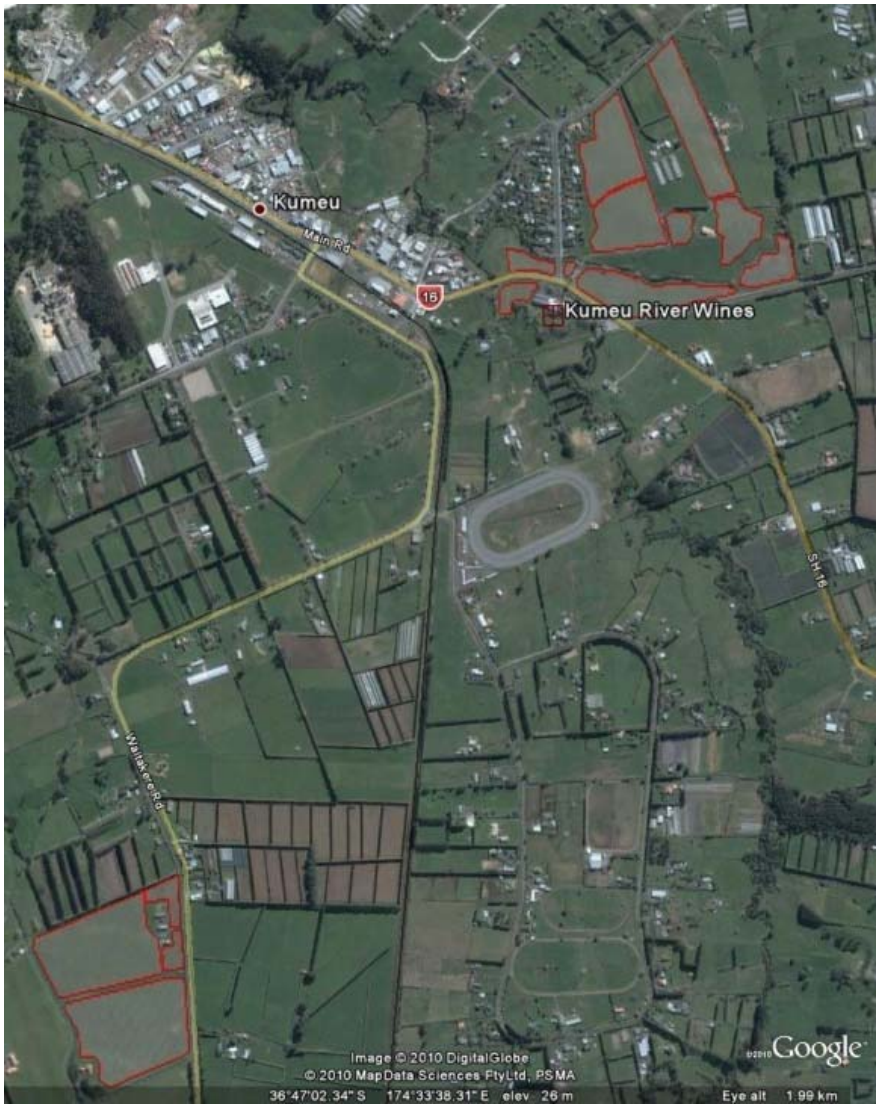


Fig. 1 Climatic scales related to surface and time.

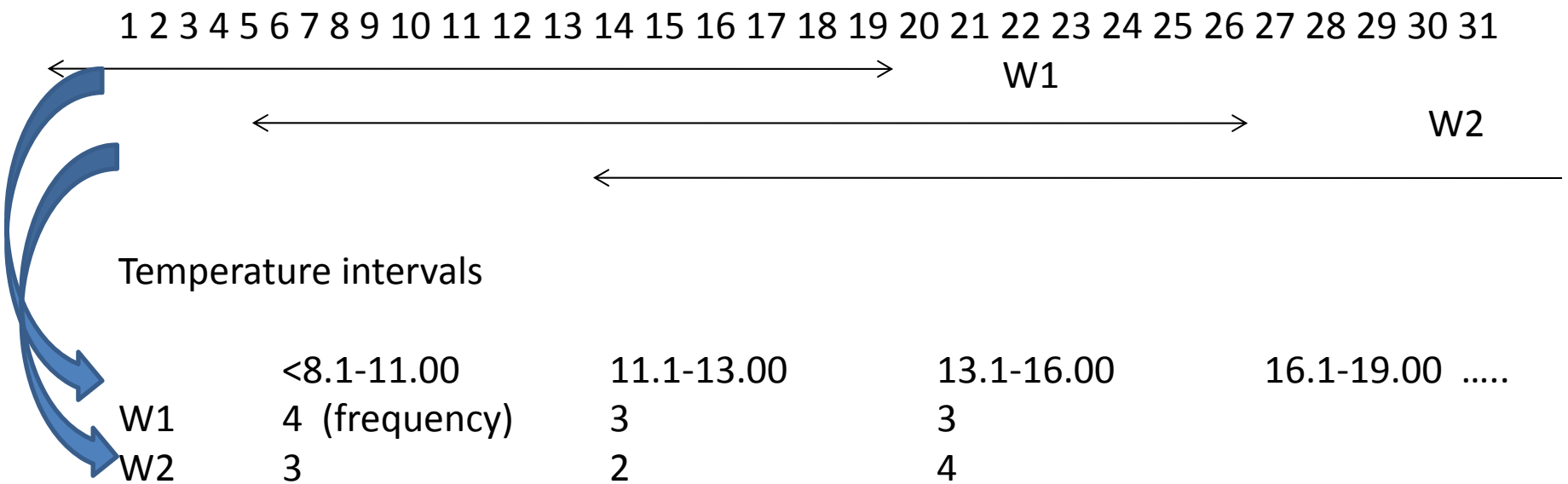
Bonnardot V, Carey VA & Strydom J, 2000, **Weather stations: Applications for viticulture, Wynboer (**incorporated in **WineLand,** magazine of the SA wine producers. **A Technical Guide for Wine Producers** wynboer.co.za/…/0405weather.php3. page 1



Kumeu River Wines, New Zealand

Based on annual yield high (1) / med (2) / low (3) yield years are determined

Daily extreme weather matrix



Temperature intervals

	<8.1-11.00	11.1-13.00	13.1-16.00	16.1-19.00 …..
W1	4 (frequency)	3	3	
W2	3	2	4	

χ^2 test: Daily max temp freq matrix

weekNo	moving3weekW	8.1-11	11.1-14	14.1-17	17.1-20	20.1-23	23.1-26	26.1-29	29.1-32
1	(1997 4 29-5 19)	0	0	0	13	8	0	0	0
2	(1997 5 6-5 26)	0	0	0	15	6	0	0	0
3	(1997 5 13-6 2)	0	1	0	19	1	0	0	0
4	(1997 5 20-6 9)	0	3	4	13	1	0	0	0
5	(1997 5 27-6 16)	0	5	8	8	0	0	0	0
6	(1997 6 3-6 23)	0	5	14	2	0	0	0	0
7	(1997 6 10-6 30)	0	7	13	1	0	0	0	0
8	(1997 6 17-7 7)	0	7	14	0	0	0	0	0
9	(1997 6 24-7 14)	0	7	13	1	0	0	0	0
10	(1997 7 1-7 21)	0	7	13	1	0	0	0	0
11	(1997 7 8-7 28)	0	6	14	1	0	0	0	0
12	(1997 7 15-8 4)	1	6	14	0	0	0	0	0
13	(1997 7 22-8 11)	1	4	15	1	0	0	0	0
14	(1997 7 29-8 18)	1	3	14	3	0	0	0	0
15	(1997 8 5-8 25)	1	3	12	5	0	0	0	0
16	(1997 8 12-9 1)	1	1	13	6	0	0	0	0
17	(1997 8 19-9 8)	1	3	11	6	0	0	0	0
18	(1997 8 26-9 15)	0	3	13	5	0	0	0	0
19	(1997 9 2-9 22)	0	3	9	9	0	0	0	0
20	(1997 9 9-9 29)	0	2	8	11	0	0	0	0
21	(1997 9 16-10 6)	0	1	5	15	0	0	0	0
22	(1997 9 23-10 13)	0	1	6	11	3	0	0	0
23	(1997 9 30-10 20)	0	0	4	10	7	0	0	0
24	(1997 10 7-10 27)	0	0	5	9	7	0	0	0
25	(1997 10 14-11 3)	0	0	3	9	8	1	0	0
26	(1997 10 21-11 10)	0	0	3	8	9	1	0	0
27	(1997 10 28-11 17)	0	0	1	7	12	1	0	0
28	(1997 11 4-11 24)	0	0	1	9	11	0	0	0
29	(1997 11 11-12 1)	0	0	1	7	9	4	0	0
30	(1997 11 18-12 8)	0	0	0	4	11	6	0	0
31	(1997 11 25-12 15)	0	0	0	2	10	8	1	0
32	(1997 12 2-12 22)	0	0	0	2	7	6	6	0
33	(1997 12 9-12 29)	0	0	0	2	3	6	9	1
34	(1998 12 16-1 5)	0	0	0	0	4	4	12	1
35	(1998 12 23-1 12)	0	0	0	1	6	3	10	1
36	(1998 12 30-1 19)	0	0	0	1	8	3	9	0
37	(1998 1 6-1 26)	0	0	0	1	5	8	7	0
38	(1998 1 13-2 2)	0	0	0	0	3	7	8	3
39	(1998 1 20-2 9)	0	0	0	0	0	5	6	8
40	(1998 1 27-2 16)	0	0	0	0	1	0	8	10
41	(1998 2 3-2 23)	0	0	0	0	1	2	7	9
42	(1998 2 10-3 2)	0	0	0	0	3	6	8	4
43	(1998 2 17-3 9)	0	0	0	0	2	9	8	2
44	(1998 2 24-3 16)	0	0	0	0	3	9	9	0
45	(1998 3 3-3 23)	0	0	0	0	2	11	8	0

χ^2 test on daily ext max T

week	<23	23.1-26	>26	rate
31	17.33	2.67	1.00	1 low yield
31	11.67	8.00	1.33	3 high yield
32	17.67	3.00	0.33	1
32	8.67	9.00	3.33	3
33	16.00	4.67	0.33	1
33	8.33	8.33	4.33	3
34	11.33	7.00	2.67	1
34	7.67	7.67	5.67	3
35	6.00	9.33	5.67	1
35	5.33	9.33	6.33	3
36	4.00	9.67	7.33	1
36	4.67	8.67	7.67	3
37	5.00	8.00	8.00	1
37	3.67	8.67	8.67	3
38	5.00	6.00	10.00	1
38	5.33	7.00	8.67	3

χ^2 test results daily ext max T

week No.	<23	23.1-26	>26	chi square rate	p-value
31	11.67	**8.00**	1.33	8.000	0.005
32	17.67	3.00	0.33	9.228	0.002
32	8.67	**9.00**	**3.33**	7.364	0.007
33	16.00	4.67	0.33	7.247	0.007
33	8.33	8.33	**4.33**	10.286	0.001
44	5.33	8.00	7.67	6.125	0.013
45	4.33	7.33	9.33	14.235	0.000
45	**9.00**	10.00	2.00	4.900	0.027

Low yield

High yield

Pinot Gris

PG	Wk	<23°C	23-26 °C	>26 °C	X^2	*p-value*	h/l
13N-4De	30	13	*7.5*	0.5	*3.86*	*0.050*	*low*
4-24Dec	33	*12*	7.5	0.5	*6.82*	*0.009*	*low*
1-20Jan	37	*6*	7	8	*4.00*	*0.046*	*low*
12-22Feb	**43**	**6**	7.5	7.5	**4.00**	**0.046**	**high**
19F-11M	**44**	**9.5**	7.5	4	**5.54**	**0.190**	**high**
		3.5	7	*10.5*	*5.83*	*0.016*	*low*
26F-18M	**45**	**11.5**	8	1.5	**4.24**	**0.400**	**high**
				9.5	*11.64*	*0.001*	*low*

early Feb mid March (berry ripening) → < 23 ºC produces high yield and >26 ºC leads to low yield

Chardonnay

Char	wk	<23 °C	23-26 °C	>26 °C	X^2	*p-value*	h / l
11-31No	**31**	18.0	**13.0**	0.5	**7.11**	**0.008**	**high**
18N-7De	**32**	15.0	**13.0**	3.5	**6.08**	**0.014**	**high**
2-22Dec	**34**	**16.0**	10.5	5.0	**4.59**	**0.032**	**high**
9-29Dec	**35**	**12.0**	15.5	4.0	**5.77**	**0.016**	**high**
16D-5Ja	**36**	**9.0**	15.5	7.0	**7.35**	**0.007**	**high**
20J-9Feb	41	*4.5*	6.5	9.0	*6.40*	*0.011*	*low*
		0.5	11.0	**20.0**	**8.35**	**0.004**	**high**
27J-6Feb	42	*4.5*	9.0	6.5	*6.40*	*0.011*	*low*
		0.5	13.0	**18.0**	**10.8**	**0.001**	**high**
3-23 Feb	**43**	2.5	**17.5**	11.5	**5.45**	**0.020**	**high**
17F-9 Ma	45	3.0	6.5	*13.5*	*4.12*	*0.042*	*low*
		9.0	**15.5**	7.0	**6.0**	**0.014**	**high**
					7.35	**0.007**	**high**

early Feb early March (berry ripening) → < 26ºC produces high yield

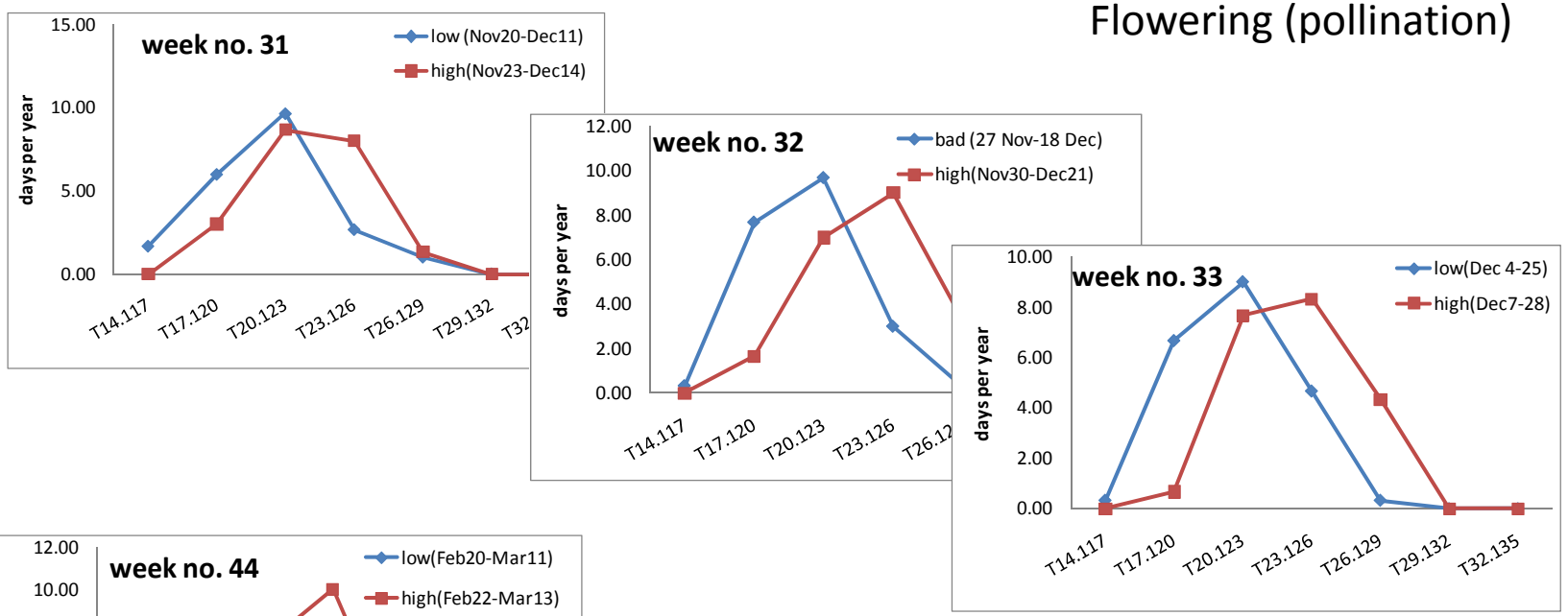
Pinot Noir

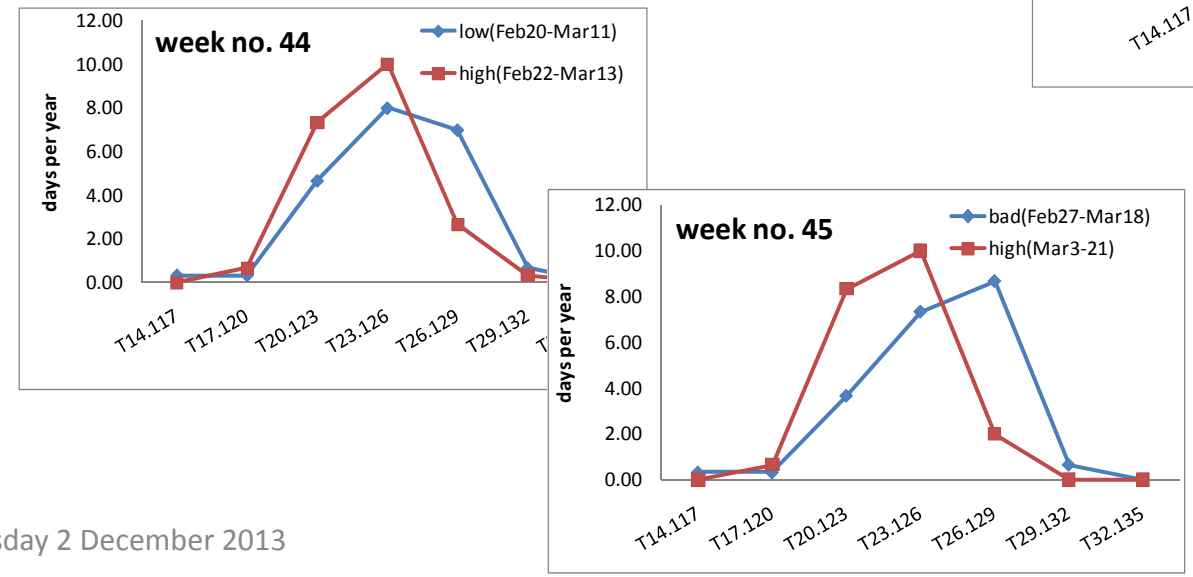
PN	wk	<23 ºC	23-26 ºC	>26 ºC	X^2	p-value	h / l
12N-1Dec	**32**	17	**3.5**	0.5	**4.50**	**0.034**	**high**
19N-8Dec	**33**	13	**7.5**	0.5	**9.94**	**0.002**	**high**
26N-15De	34	*17.5*	2.0	1.5	*4.74*	*0.029*	*low*
		9.5	**8.5**	3.0	**8.05**	**0.005**	**high**
3-23 Dec	35	*16.0*	4.0	1.0	*12.90*	*0.000*	*low*
	35	4.5	**12.0**	**4.5**	**8.00**	**0.005**	**high**
					4.46	**0.035**	**high**
17D-6Jan	37	*11.5*	8.0	1.5	*4.24*	*0.400*	*low*
		5.5	9.5	**6.0**	**5.40**	**0.020**	**high**
24D-13Ja	**38**	5.0	8.5	**7.5**	**6.37**	**0.012**	**high**
31D-20Ja	39	*7*	9	5	*5.56*	*0.018*	*low*
		2	8	**11**	**4.50**	**0.034**	**high**
4-24Feb	**44**	**4**	6.5	10.5	**5.44**	**0.020**	**high**
11F-3Mar	**45**	**6**	7.5	7.5	**9.32**	**0.002**	**high**

early Feb early March (berry ripening) → < 23 ºC produces high yield

χ^2 test results/ graphs daily ext max T

Flowering (pollination)

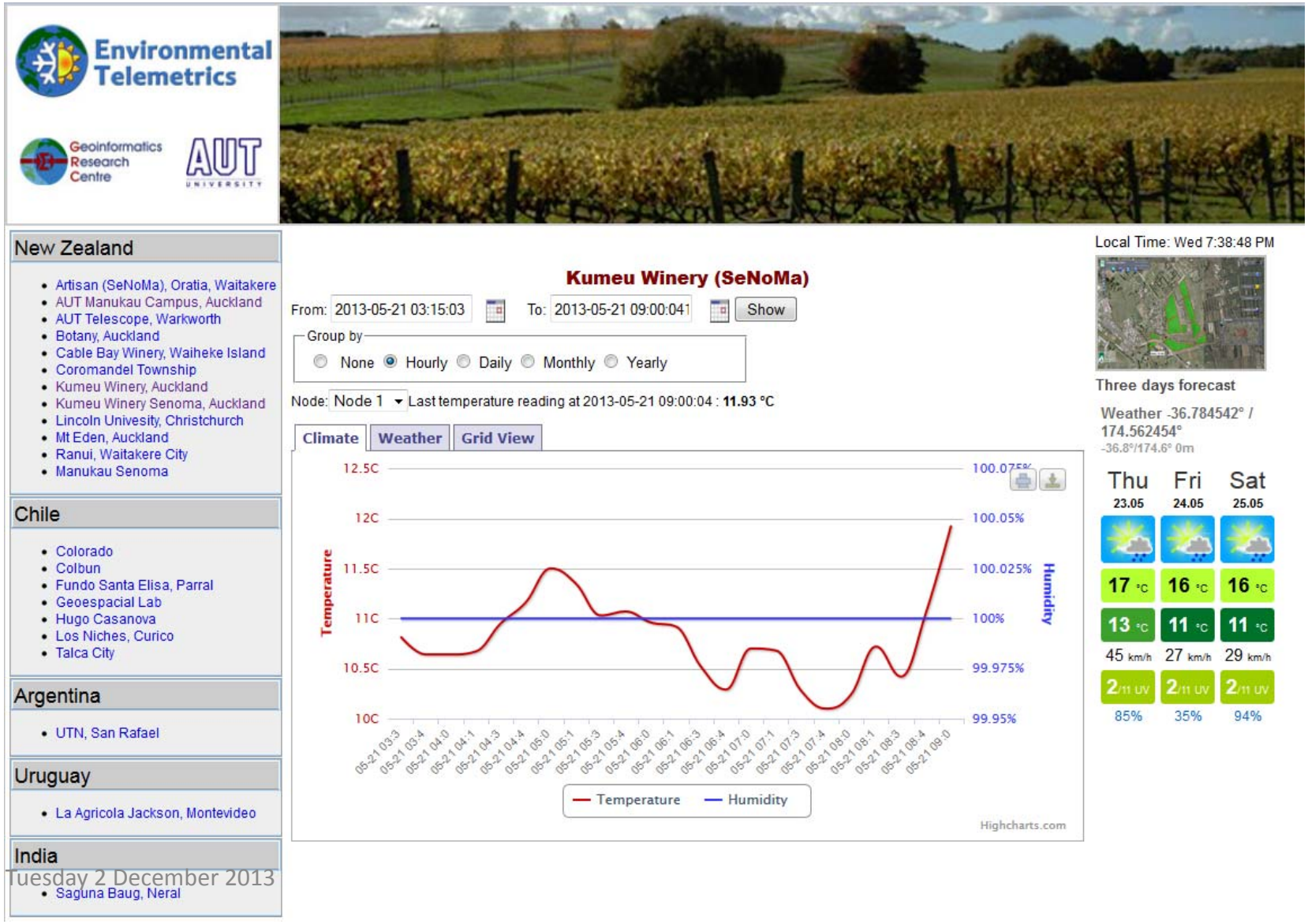




Berry ripening

MULTI SENSOR DATA

live web display



The methodology

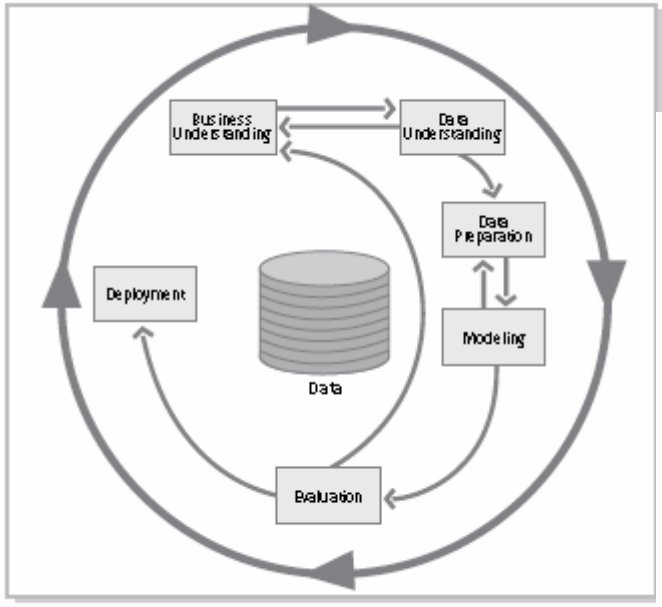


Figure 2: Phases of the CRISP-DM reference model

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria* Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*	Collect Initial Data *Initial Data Collection Report* Describe Data *Data Description Report* Explore Data *Data Exploration Report* Verify Data Quality *Data Quality Report*	Select Data *Rationale for Inclusion/ Exclusion* Clean Data *Data Cleaning Report* Construct Data *Derived Attributes* *Generated Records* Integrate Data *Merged Data* Format Data *Reformatted Data* *Dataset* *Dataset Description*	Select Modeling Techniques *Modeling Technique* *Modeling Assumptions* Generate Test Design *Test Design* Build Model *Parameter Settings* *Models* *Model Descriptions* Assess Model *Model Assessment* *Revised Parameter Settings*	Evaluate Results *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* Review Process *Review of Process* Determine Next Steps *List of Possible Actions* *Decision*	Plan Deployment *Deployment Plan* Plan Monitoring and Maintenance *Monitoring and Maintenance Plan* Produce Final Report *Final Report* *Final Presentation* Review Project *Experience Documentation*

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

Chapman, Pete , et al., et al. (2000) CRISP-DM 1.0 (Cross Industry Standard Process for Data Mining), Step-by-step data mining guide. SPSS Inc. CRISPMWP-1104, 2000. pp73. Page 10 & page 12

The multi-sensor data

1. id
2. DateTime
3. NodeId
4. Pressure_Rel
5. Ind_Temp
6. Ind_Hum
7. Out_Temp
8. Out_Hum
9. Dewp
10. Windc
11. Winds
12. Wind_Dir
13. Gust
14. Rain_Rate
15. Act_rain
16. Rain_Today
17. Pressure_Abs
18. VinyardId
19. Rain_Total
20. Heat_Indx
21. High_Gust

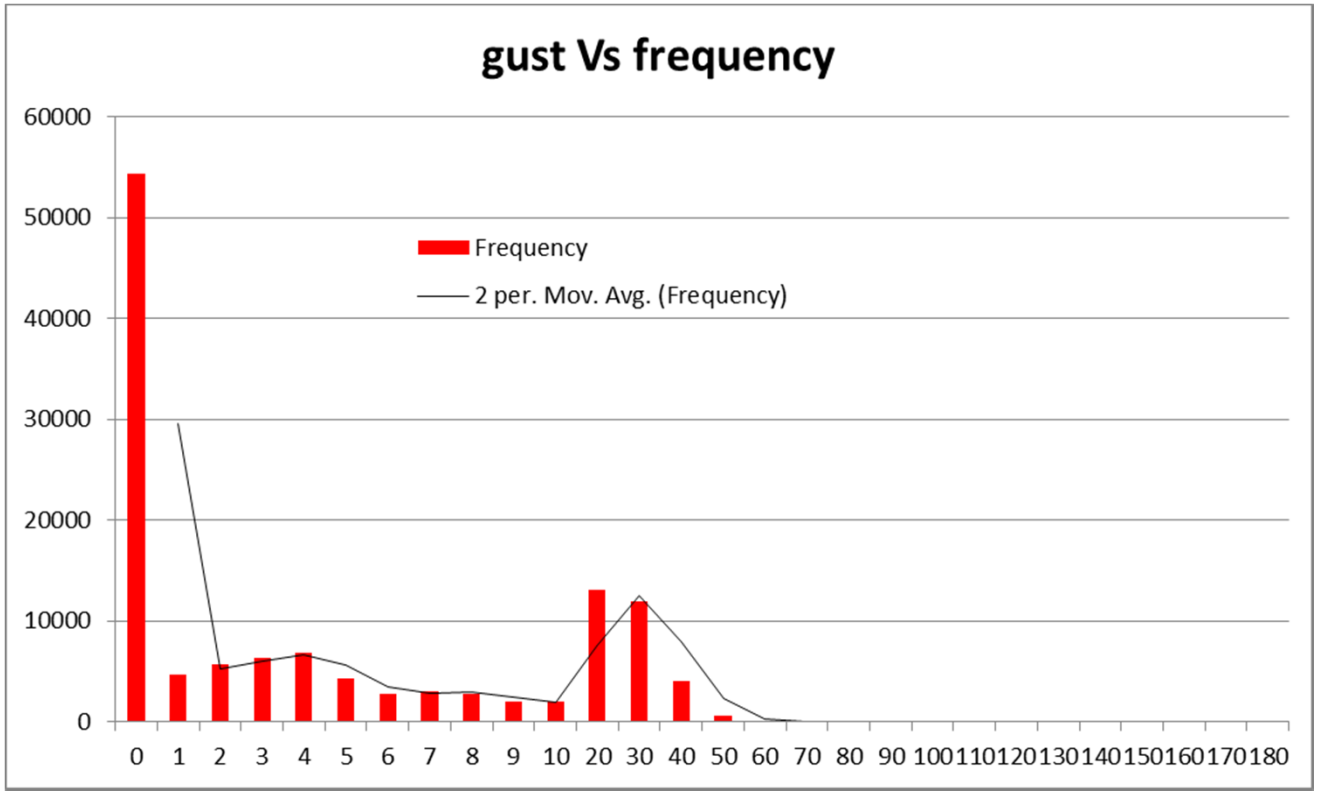
Record high temperature 32.8°C
Record low temperature -8.9°C
Record high gust 172.2 km/h
Record high average 172.2 km/h
Record daily rain 82.6 mm
Record low wind chill -11.5°C
Record high barometer 1035.6 hPa
Record low barometer 977.9 hPa

From
http://www.binoscope.co.nz/Kumeu.htm

1. id
2. Date Time
3. Pressure_Rel
4. Out_Temp
5. Out_Hum
6. Dewp
7. Windc
8. Winds
9. Wind_Dir
10. Gust
11. class
12. Heat_Indx

13. E_code

Data distribution



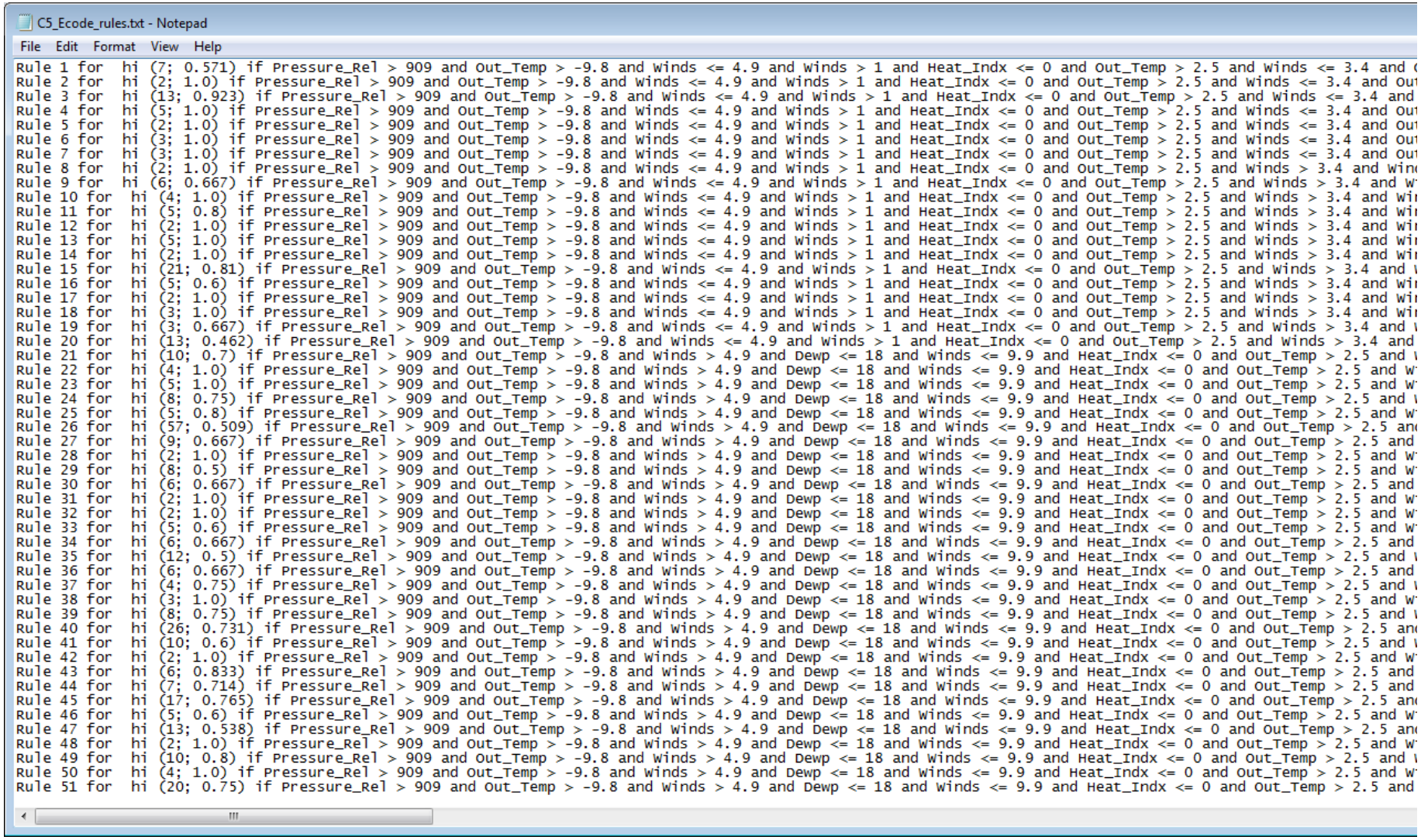
gust classes <1,"no", <5,"low", <10,"med", <20,"high", >20,"very high"

Data mining

- C5.0
- C&RT (classification and regression trees-B)
- CAHID (Chi-squared Automatic Interaction) Detector
- ANN
- Regression
- PCA

Sw: SPSS clementine

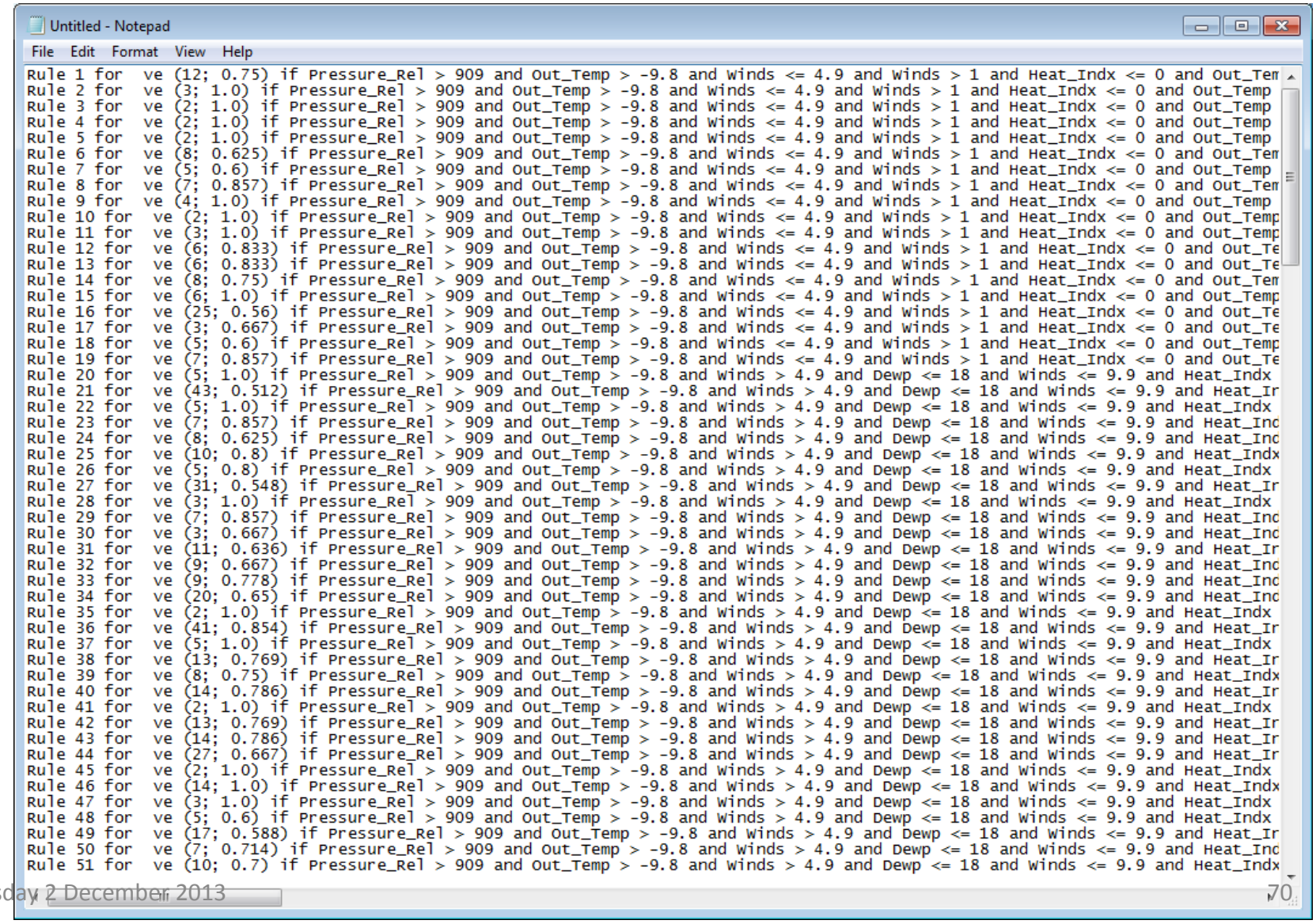
C5.0 for high gust


C5_Ecode_rules.txt - Notepad

File Edit Format View Help

```
Rule 1 for  hi (7; 0.571) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds <= 3.4 and
Rule 2 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds <= 3.4 and Ou
Rule 3 for  hi (13; 0.923) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds <= 3.4 and
Rule 4 for  hi (5; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds <= 3.4 and Ou
Rule 5 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds <= 3.4 and Ou
Rule 6 for  hi (3; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds <= 3.4 and Ou
Rule 7 for  hi (3; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds <= 3.4 and Ou
Rule 8 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and Win
Rule 9 for  hi (6; 0.667) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and W
Rule 10 for  hi (4; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and Wi
Rule 11 for  hi (5; 0.8) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and Wi
Rule 12 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and Wi
Rule 13 for  hi (5; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and Wi
Rule 14 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and Wi
Rule 15 for  hi (21; 0.81) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and
Rule 16 for  hi (5; 0.6) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and Wi
Rule 17 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and Wi
Rule 18 for  hi (3; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and Wi
Rule 19 for  hi (3; 0.667) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and
Rule 20 for  hi (13; 0.462) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds <= 4.9 and Winds > 1 and Heat_Indx <= 0 and Out_Temp > 2.5 and Winds > 3.4 and
Rule 21 for  hi (10; 0.7) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 22 for  hi (4; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 23 for  hi (5; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 24 for  hi (8; 0.75) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 25 for  hi (5; 0.8) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 26 for  hi (57; 0.509) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 an
Rule 27 for  hi (9; 0.667) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 28 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 29 for  hi (8; 0.5) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 30 for  hi (6; 0.667) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 31 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 32 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 33 for  hi (5; 0.6) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 34 for  hi (6; 0.667) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 35 for  hi (12; 0.5) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 36 for  hi (6; 0.667) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 37 for  hi (4; 0.75) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 38 for  hi (3; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 39 for  hi (8; 0.75) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 40 for  hi (26; 0.731) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 an
Rule 41 for  hi (10; 0.6) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 42 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 43 for  hi (6; 0.833) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 44 for  hi (7; 0.714) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 45 for  hi (17; 0.765) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 an
Rule 46 for  hi (5; 0.6) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 47 for  hi (13; 0.538) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 an
Rule 48 for  hi (2; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 49 for  hi (10; 0.8) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
Rule 50 for  hi (4; 1.0) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and W
Rule 51 for  hi (20; 0.75) if Pressure_Rel > 909 and Out_Temp > -9.8 and Winds > 4.9 and Dewp <= 18 and Winds <= 9.9 and Heat_Indx <= 0 and Out_Temp > 2.5 and
```

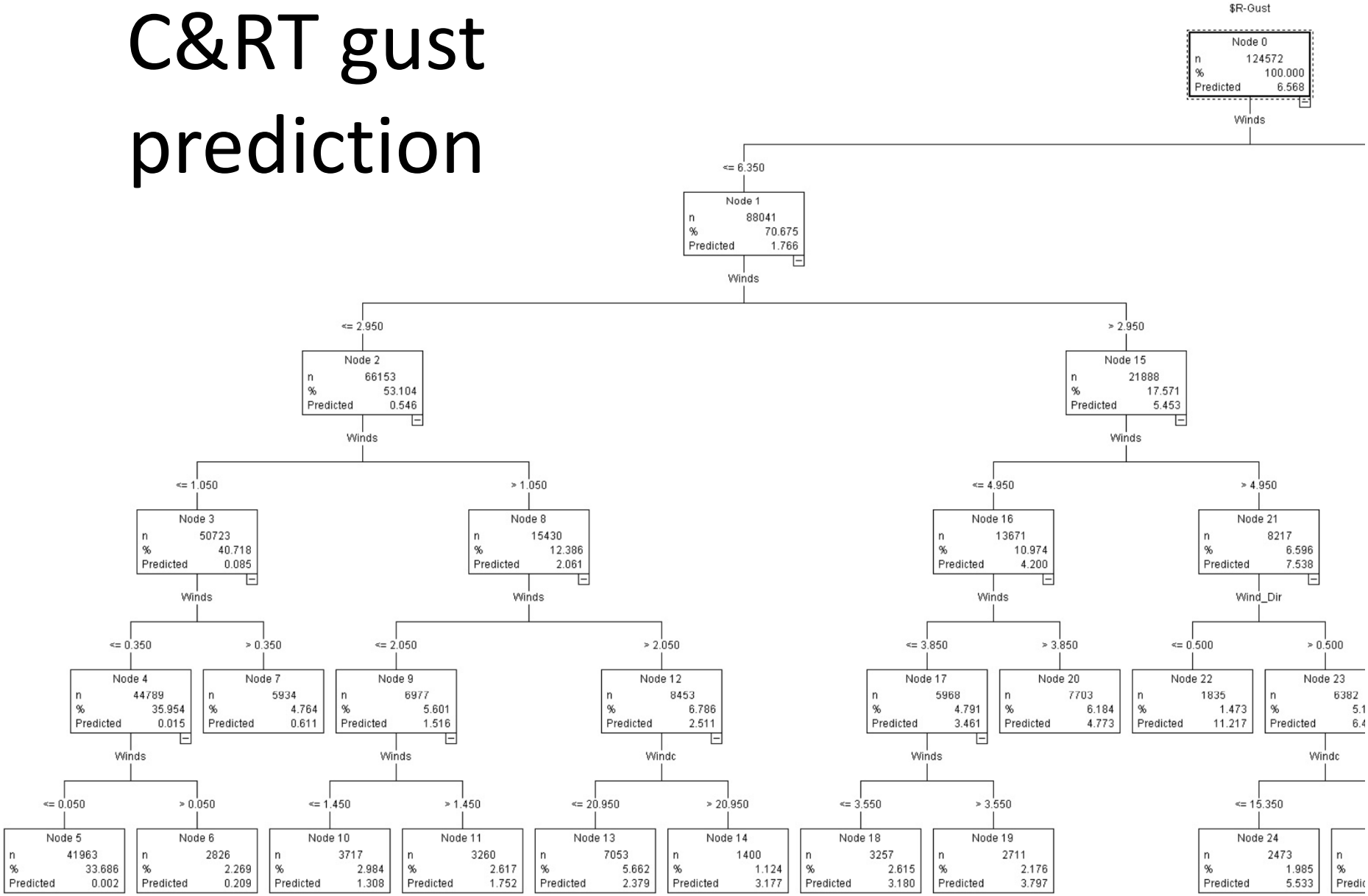
C5.0 for very high gust



C5.0 rule for high gust

Rule 118 for high gust (8; 0.75)
if Pressure_Rel > 909
and Out_Temp > -9.8
and Winds > 4.9
and Dewp <= 18
and Winds <= 9.9
and Heat_Indx <= 0
and Out_Temp > 2.5
and Winds > 7.3
and Dewp > 8.3
and Out_Temp > 14.9
and Out_Hum <= 90
and Pressure_Rel <= 1018.9
and Winds <= 8.8
and Wind_Dir <= 242
and Out_Hum > 45
and Winds > 8
and Pressure_Rel > 997.7
and Out_Temp <= 24.9
and Out_Hum > 69
and Dewp <= 16.4
and Wind_Dir <= 135
and Out_Hum > 70
and Winds <= 8.3
and Wind_Dir <= 67
and Wind_Dir <= 22
and Out_Hum <= 72
and Pressure_Rel > 1003.1
and Pressure_Rel <= 1010.5
then high gust

C&RT gust prediction



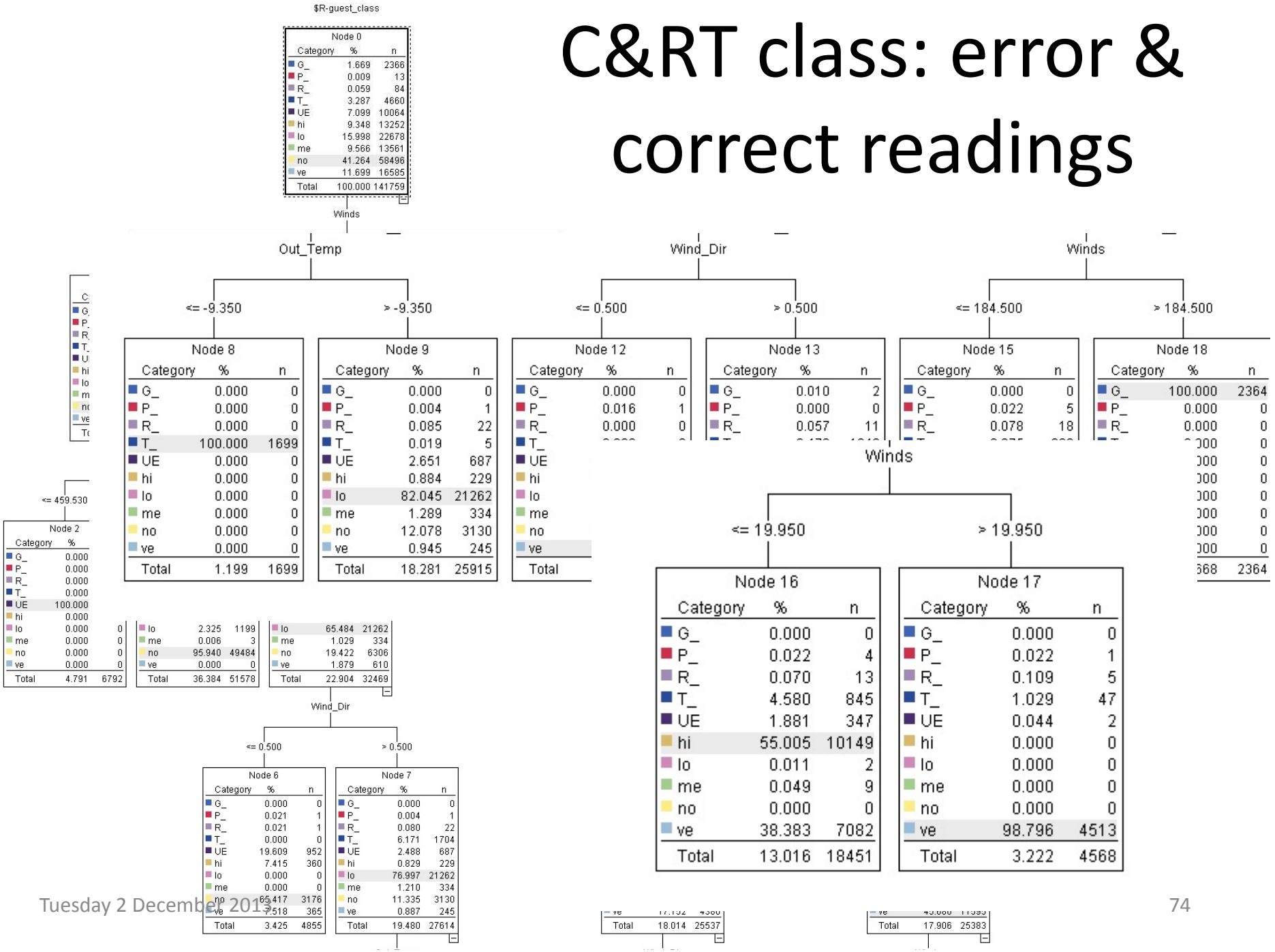
C&RT rules for Gust prediction

Gust_CRT_model.txt - Notepad

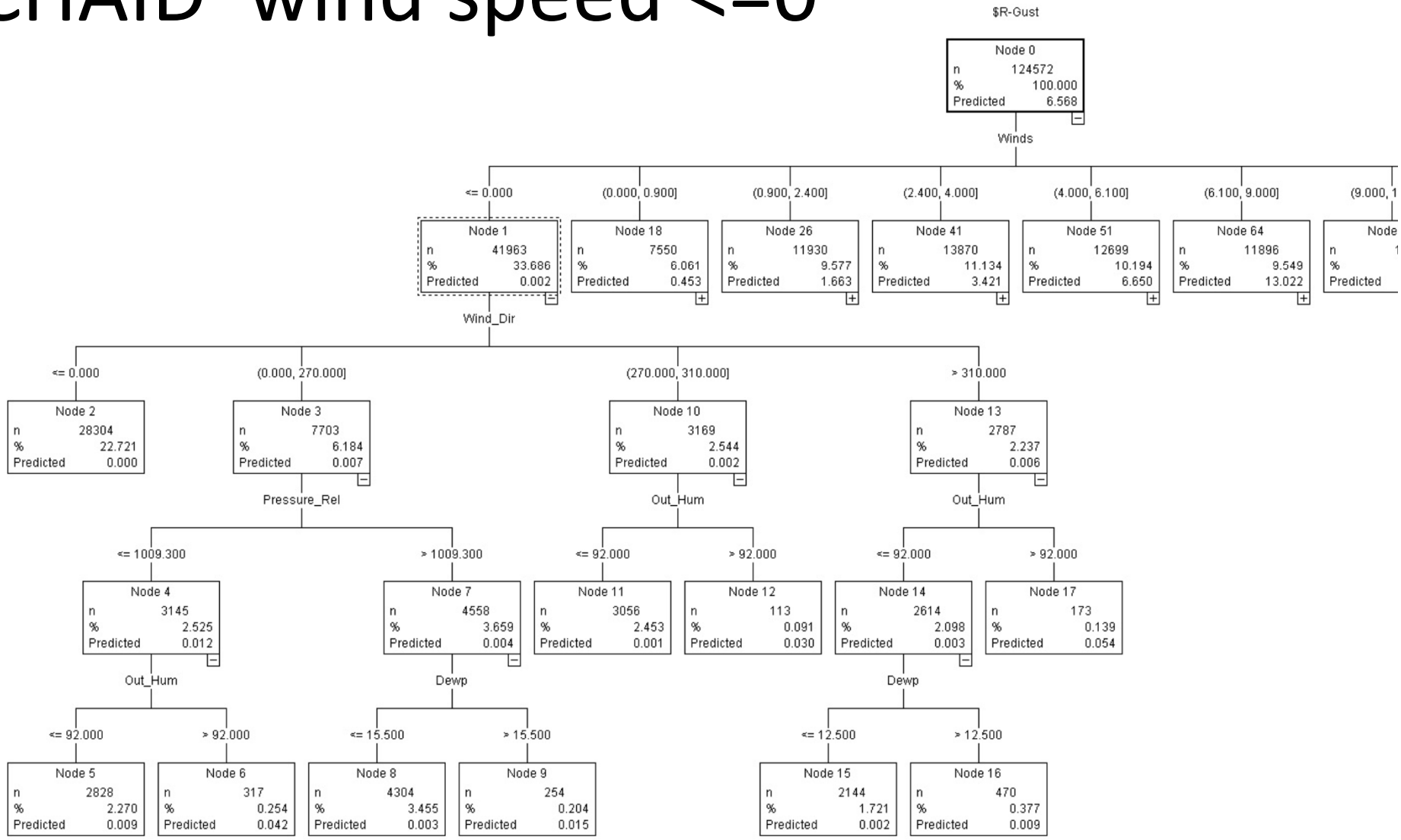
File Edit Format View Help

```
Winds <= 6.3500 [ Ave: 1.766, Effect: -4.802 ] (88,041)
        Winds <= 2.9500 [ Ave: 0.546, Effect: -1.22 ] (66,153)
                Winds <= 1.0500 [ Ave: 0.085, Effect: -0.461 ] (50,723)
                        Winds <= 0.3500 [ Ave: 0.015, Effect: -0.07 ] (44,789)
                                Winds <= 0.0500 [ Ave: 0.002, Effect: -0.013 ] => 0.002 (41,963)
                                Winds > 0.0500 [ Ave: 0.209, Effect: 0.194 ] => 0.209 (2,826)
                        Winds > 0.3500 [ Ave: 0.611, Effect: 0.526 ] => 0.611 (5,934)
                Winds > 1.0500 [ Ave: 2.061, Effect: 1.515 ] (15,430)
                        Winds <= 2.0500 [ Ave: 1.516, Effect: -0.545 ] (6,977)
                                Winds <= 1.4500 [ Ave: 1.308, Effect: -0.208 ] => 1.308 (3,717)
                                Winds > 1.4500 [ Ave: 1.752, Effect: 0.237 ] => 1.752 (3,260)
                        Winds > 2.0500 [ Ave: 2.511, Effect: 0.45 ] (8,453)
                                Windc <= 20.9500 [ Ave: 2.379, Effect: -0.132 ] => 2.379 (7,053)
                                Windc > 20.9500 [ Ave: 3.177, Effect: 0.666 ] => 3.177 (1,400)
        Winds > 2.9500 [ Ave: 5.453, Effect: 3.688 ] (21,888)
                Winds <= 4.9500 [ Ave: 4.2, Effect: -1.253 ] (13,671)
                        Winds <= 3.8500 [ Ave: 3.461, Effect: -0.739 ] (5,968)
                                Winds <= 3.5500 [ Ave: 3.18, Effect: -0.28 ] => 3.18 (3,257)
                                Winds > 3.5500 [ Ave: 3.797, Effect: 0.337 ] => 3.797 (2,711)
                        Winds > 3.8500 [ Ave: 4.773, Effect: 0.573 ] => 4.773 (7,703)
                Winds > 4.9500 [ Ave: 7.538, Effect: 2.085 ] (8,217)
                        Wind_Dir <= 0.500 [ Ave: 11.217, Effect: 3.679 ] => 11.217 (1,835)
                        Wind_Dir > 0.500 [ Ave: 6.481, Effect: -1.058 ] (6,382)
                                Windc <= 15.3500 [ Ave: 5.533, Effect: -0.947 ] => 5.533 (2,473)
                                Windc > 15.3500 [ Ave: 7.08, Effect: 0.599 ] => 7.08 (3,909)
Winds > 6.3500 [ Ave: 18.142, Effect: 11.574 ] (36,531)
        Winds <= 17.0500 [ Ave: 16.094, Effect: -2.048 ] (29,226)
                Wind_Dir <= 0.500 [ Ave: 22.322, Effect: 6.228 ] (7,753)
                        Winds <= 8.4000 [ Ave: 17.918, Effect: -4.404 ] => 17.918 (2,434)
                        Winds > 8.4000 [ Ave: 24.337, Effect: 2.015 ] => 24.337 (5,319)
                Wind_Dir > 0.500 [ Ave: 13.846, Effect: -2.249 ] (21,473)
                        Winds <= 10.2500 [ Ave: 11.309, Effect: -2.537 ] (11,758)
                                Pressure_Rel <= 1004.1500 [ Ave: 14.72, Effect: 3.411 ] => 14.72 (3,325)
                                Pressure_Rel > 1004.1500 [ Ave: 9.964, Effect: -1.345 ] => 9.964 (8,433)
                        Winds > 10.2500 [ Ave: 16.916, Effect: 3.071 ] (9,715)
                                Out_Temp <= 15.4500 [ Ave: 14.33, Effect: -2.586 ] => 14.33 (3,317)
                                Out_Temp > 15.4500 [ Ave: 18.257, Effect: 1.341 ] => 18.257 (6,398)
        Winds > 17.0500 [ Ave: 26.337, Effect: 8.194 ] => 26.337 (7,305)
```

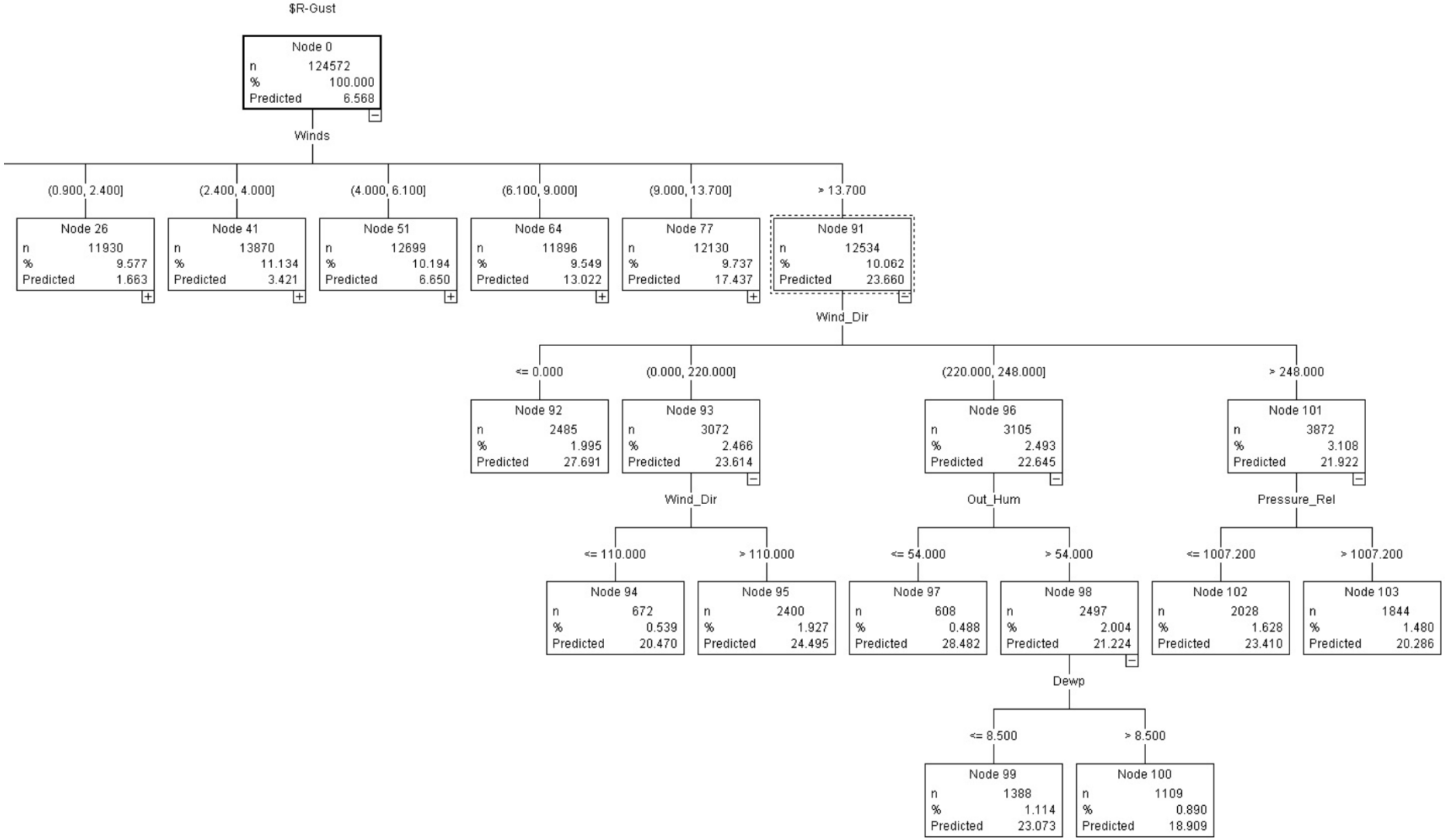
C&RT class: error & correct readings



CHAID wind speed <=0



Wind speed >13.7



ANN predict gust class

Analysis

Estimated accuracy: 83.652
Input Layer: 8 neurons
Hidden Layer 1: 3 neurons
Output Layer: 10 neurons
Relative Importance of Inputs

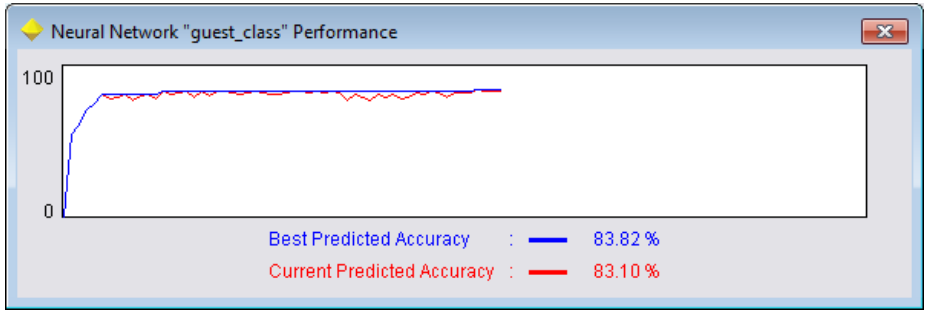
Out_Temp	0.340085
Heat_Indx	0.2521
Winds	0.224171
Dewp	0.218885
Windc	0.186994
Out_Hum	0.113052
Pressure_Rel	0.0187872
Wind_Dir	0.00938279
Out_Temp	0.340085

Heat_Indx	0.2521
Winds	0.224171
Dewp	0.218885
Windc	0.186994
Out_Hum	0.113052
Pressure_Rel	0.0187872
Wind_Dir	0.00938279

Fields
Target
guest_class
Inputs
Dewp, Heat_Indx, Out_Hum, Out_Temp, Pressure_Rel, Wind_Dir, Windc, Winds

Build Settings
Use partitioned data: false
Method: Quick
Stop on: Default
Set random seed: false
Prevent overtraining: true
Sample %: 50.0
Optimize: Memory

Training Summary
Model type: Neural net
Stream: Stream1
User: sshanmug
Date built: 23/05/13 01:56
Application: Clementine 10.1
Elapsed time for model build: 0 hours, 0 mins, 31 secs



Regression

Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed	Method
1	Winds, Pressure_Rel, Out_Hum, Wind_Dir, Windc, Out_Temp, Dewp(b)	.	Enter
a. Dependent Variable: Gust			
b. All requested variables entered.			

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000(a)	.999	.999	5.46882
a. Predictors: (Constant), Winds, Pressure_Rel, Out_Hum, Wind_Dir, Windc, Out_Temp, Dewp				

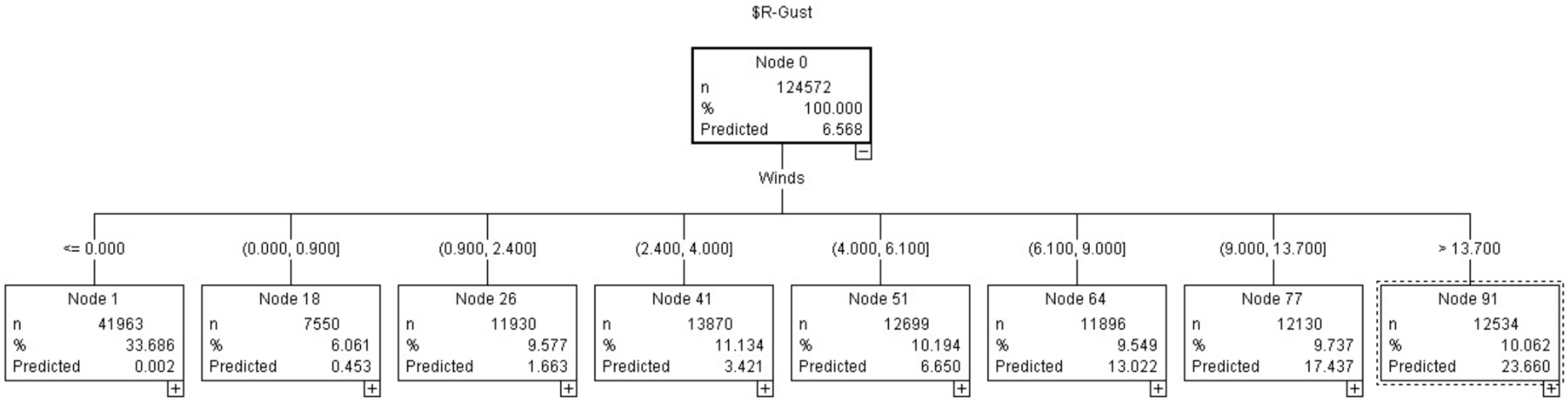
ANOVA(a)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4898160057.877	7	699737151.125	23396356.073	.000(b)
	Residual	4239482.448	141751	29.908		
	Total	4902399540.325	141758			
a. Dependent Variable: Gust						
b. Predictors: (Constant), Winds, Pressure_Rel, Out_Hum, Wind_Dir, Windc, Out_Temp, Dewp						

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.671	.127		20.956	.000
	Dewp	8.27E-002	.003	.005	26.786	.000
	Out_Hum	-7.68E-002	.001	-.007	-75.270	.000
	Out_Temp	1.06E-002	.002	.001	5.748	.000
	Pressure_Rel	4.18E-003	.000	.006	41.164	.000
	Wind_Dir	-3.59E-003	.000	-.003	-31.307	.000
	Windc	-8.35E-004	.001	.000	-.734	.463
	Winds	.999	.000	.999	9914.912	.000
a. Dependent Variable: Gust						

CHAID


$R-Gust
Node 0
n 124572
% 100.000
Predicted 6.568
Winds
<= 0.000
(0.000, 0.900]
(0.900, 2.400]
(2.400, 4.000]
(4.000, 6.100]
(6.100, 9.000]
(9.000, 13.700]
> 13.700
Node 1
n 41963
% 33.686
Predicted 0.002
Node 18
n 7550
% 6.061
Predicted 0.453
Node 26
n 11930
% 9.577
Predicted 1.663
Node 41
n 13870
% 11.134
Predicted 3.421
Node 51
n 12699
% 10.194
Predicted 6.650
Node 64
n 11896
% 9.549
Predicted 13.022
Node 77
n 12130
% 9.737
Predicted 17.437
Node 91
n 12534
% 10.062
Predicted 23.660

PCA

Factor Analysis

Communalities

	Initial	Extraction
Pressure_Rel	1.000	.830
Out_Temp	1.000	.893
Out_Hum	1.000	.998
Dewp	1.000	.939
Windc	1.000	.991
Winds	1.000	.958
Wind_Dir	1.000	.999

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.889	41.274	41.274	2.889	41.274	41.274
2	1.295	18.494	59.768	1.295	18.494	59.768
3	1.087	15.527	75.295	1.087	15.527	75.295
4	.920	13.140	88.435	.920	13.140	88.435
5	.418	5.967	94.402	.418	5.967	94.402
6	.298	4.254	98.656			
7	9.41E-002	1.344	100.000			

Extraction Method: Principal Component Analysis.

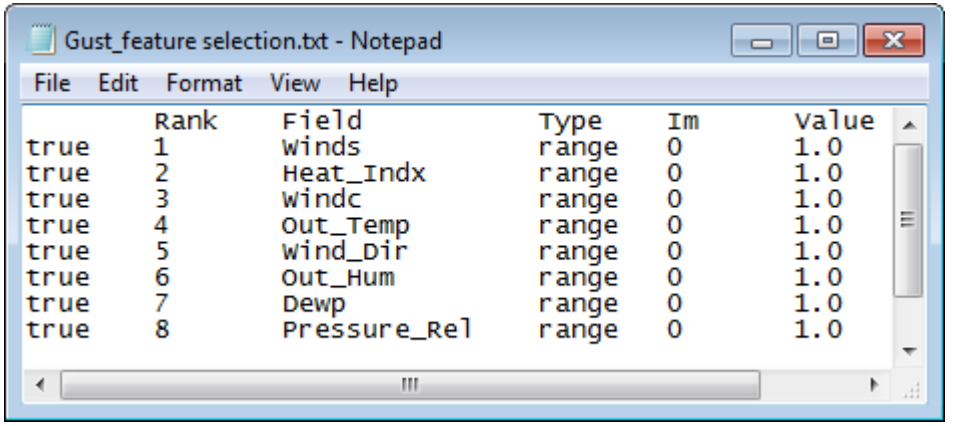
Component Matrix(a)

	Component				
	1	2	3	4	5
Pressure_Rel	-.821	.389	4.26E-002	-2.47E-002	5.26E-002
Out_Temp	.900	-3.19E-002	9.75E-003	3.07E-002	.283
Out_Hum	-7.15E-002	.287	-.744	.590	-9.50E-002
Dewp	.909	-.247	-.144	.171	3.37E-002
Windc	.663	.469	.317	1.15E-002	-.480
Winds	.300	.878	7.86E-002	-6.47E-002	.293
Wind_Dir	-.204	-8.87E-002	.636	.732	9.20E-002

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

Conclusions



	Rank	Field	Type	Im	Value
true	1	Winds	range	0	1.0
true	2	Heat_Indx	range	0	1.0
true	3	Windc	range	0	1.0
true	4	Out_Temp	range	0	1.0
true	5	Wind_Dir	range	0	1.0
true	6	Out_Hum	range	0	1.0
true	7	Dewp	range	0	1.0
true	8	Pressure_Rel	range	0	1.0

- Different primary predictors
  - C5.0=>pressure relative
  - C&RT => wind speed
  - CHAID => wind speed
  - Regression test model => wine speed,

    pressure relative, outdoor humidity, wind direction, wind chill, outdoor temperature, dew point
  - PCA=> pressure relative
    - outdoor temperature, outdoor humidity, dew point, wind chill, w speed, w direction
- Future work
  - Deploy online
  - Test other location data

PIXEL CLUSTERING IN SPATIAL DATA MINING

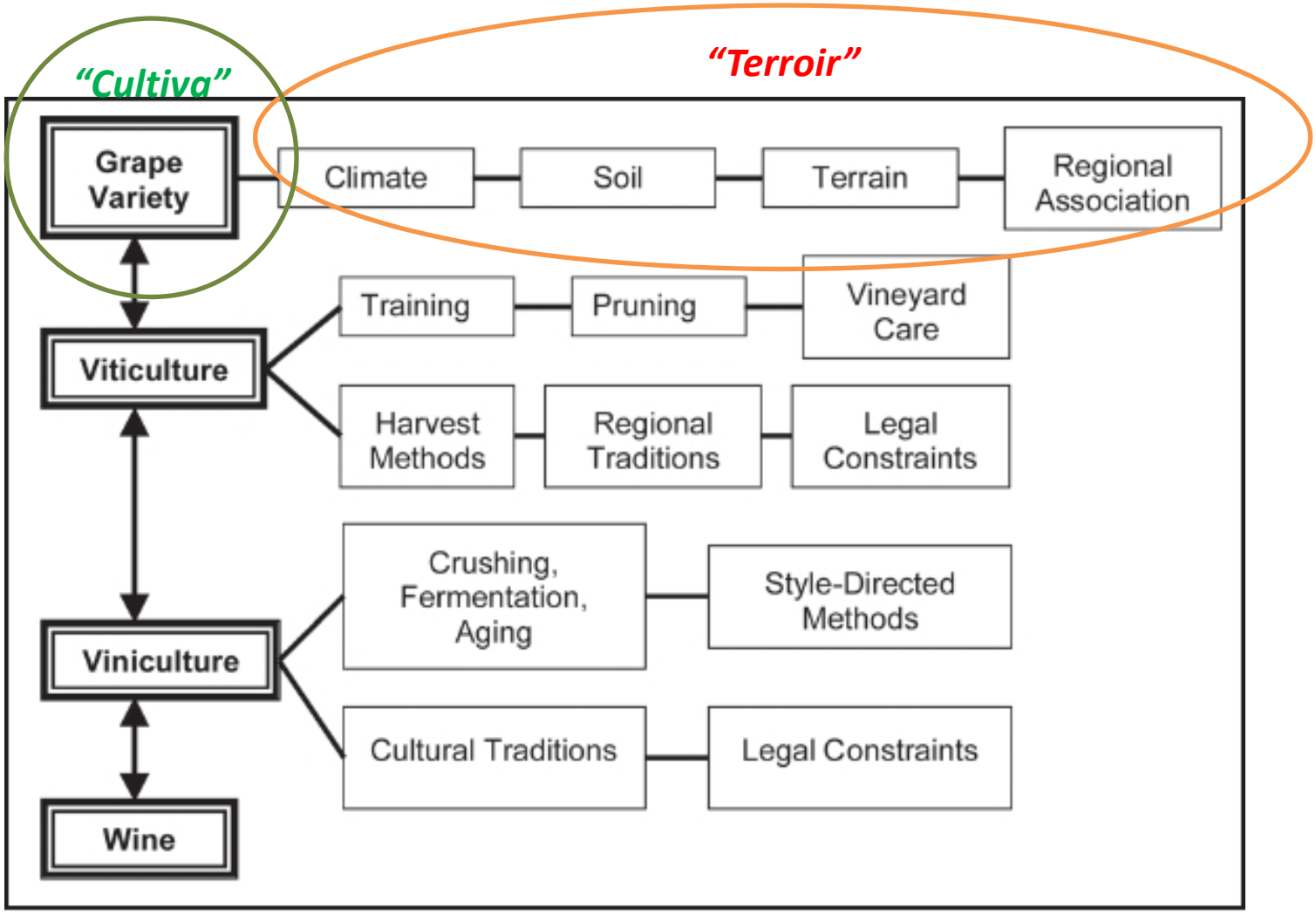


Figure 1 Factors encountered in the grape growing to wine production continuum. Note that viniculture is the study or science of making wines.

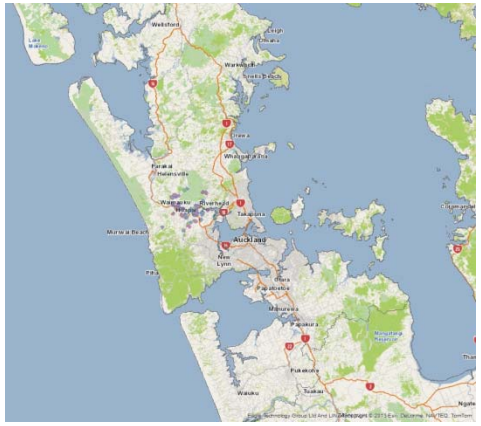
JONES, Gregory V.; SNEAD, Nicholas; NELSON, Peder. Geology and Wine 8. Modeling Viticultural Landscapes: A GIS Analysis of the Terroir Potential in the Umpqua Valley of Oregon. **Geoscience Canada**, [S.l.], dec. 2004. ISSN 1911-4850. Available at: <http://journals.hil.unb.ca/index.php/GC/article/view/2779/3266>. Date accessed: 08 Oct. 2013. doi:10.12789/gs.v31i4.2779.

viticulture zoning

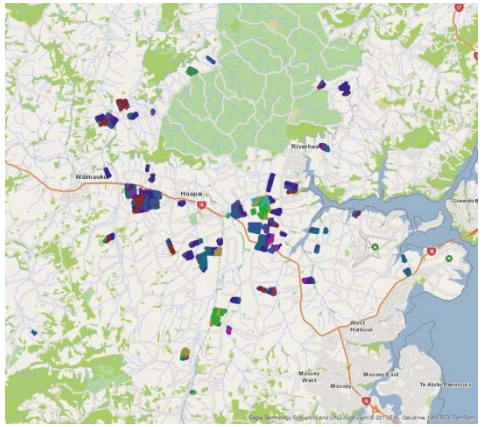
- Requires extensive knowledge on *"Terroir"* properties
- Makes it difficult for zoning "*new terroirs*" / new world wine regions, such as New Zealand, Chilli, South Africa
- New approach is presented→ SOM clustering & TDIDT (Top-Down Induction of Decision Trees) using Kumeu wine region, New Zealand
  www.cs.utsa.edu/~bylander/cs6243/decision-tree.pdf

background

New Zealand



Kumeu



Kumeu vineyards

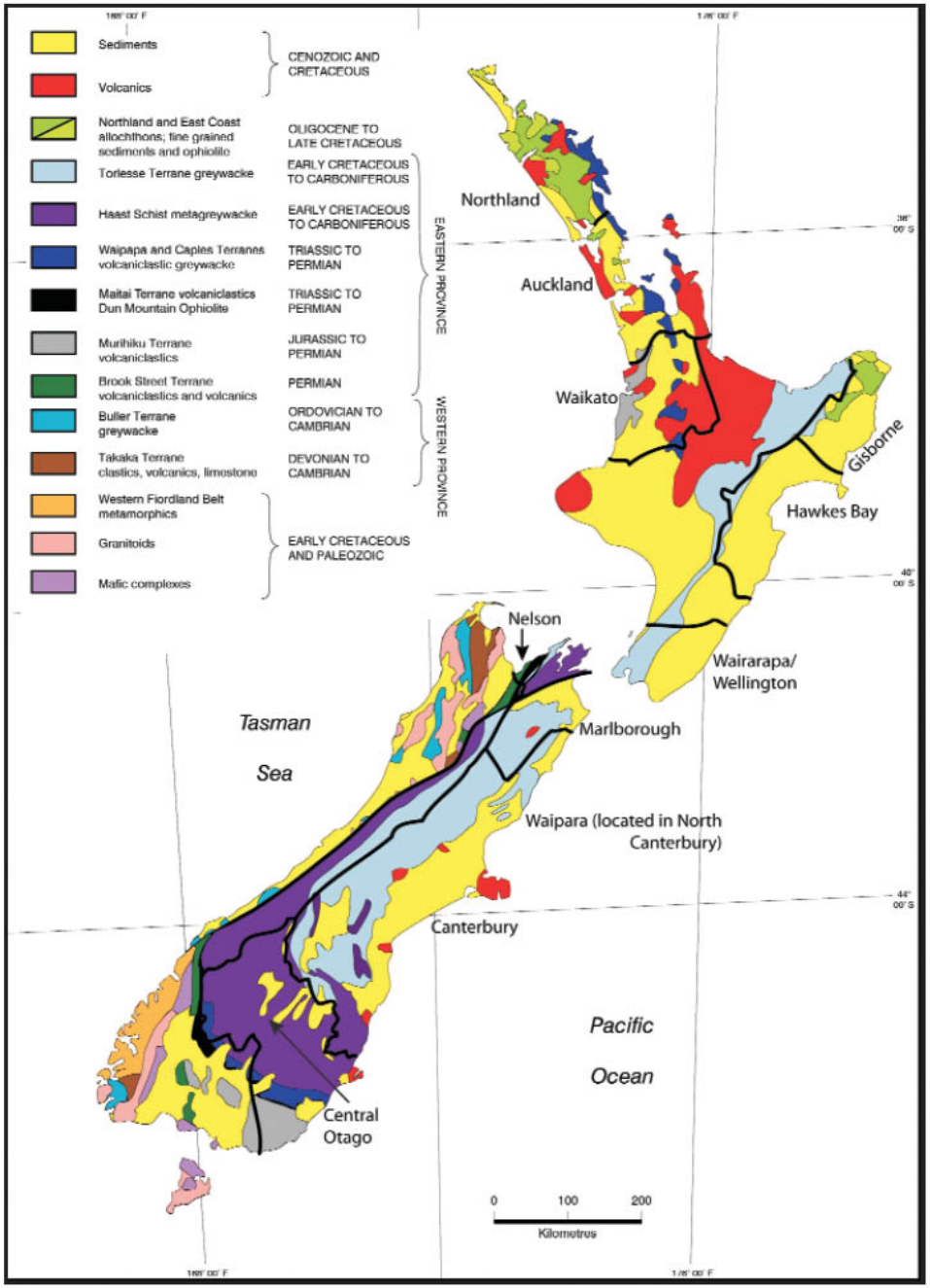


Figure 1. Map of New Zealand geology showing locations and boundaries of major winegrowing regions (geological data of New Zealand was sourced from GNS Science).

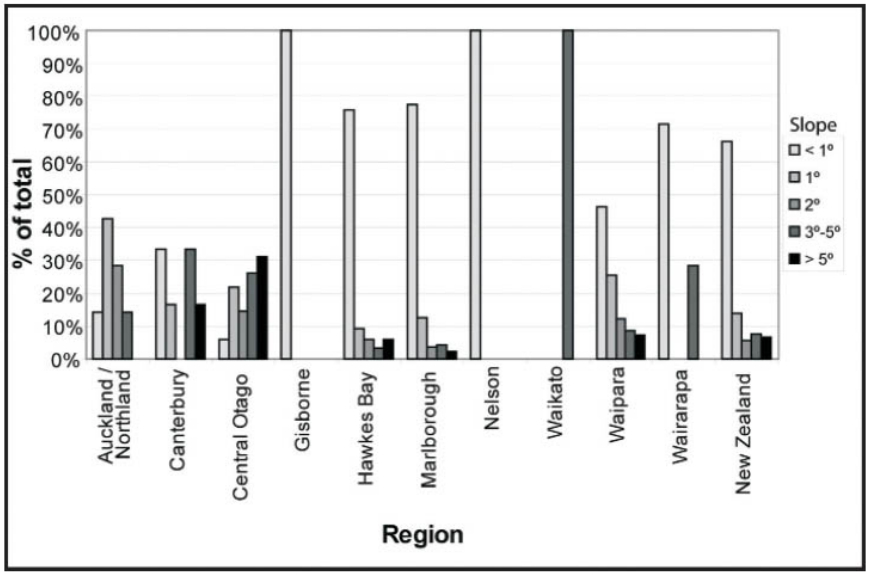


Figure 4. Distribution of vineyard slopes in the major winegrowing regions of New Zealand (Leathwick et al. 2002; this study).

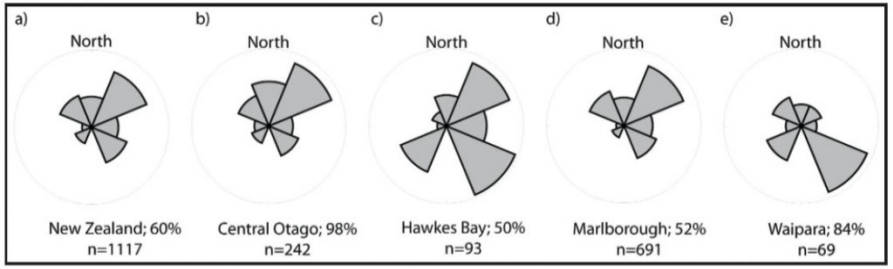
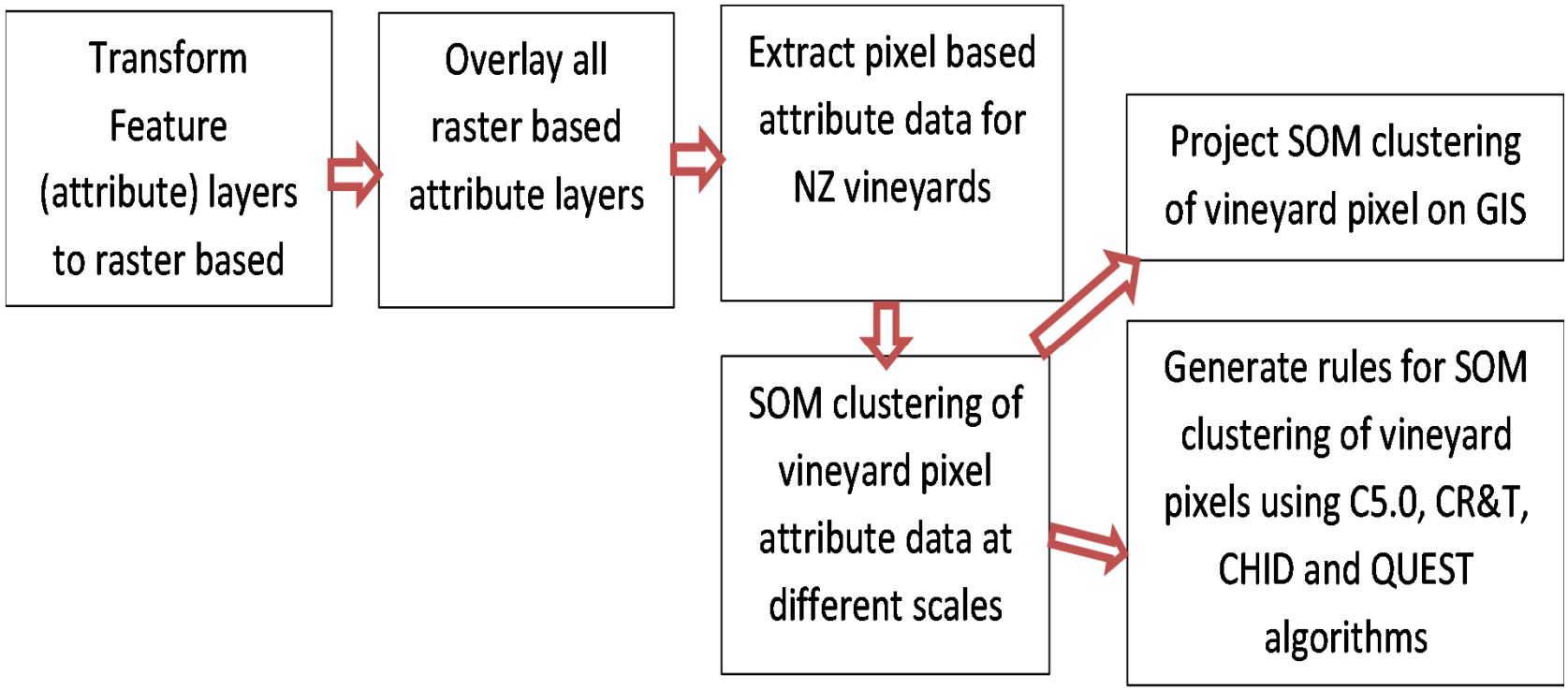


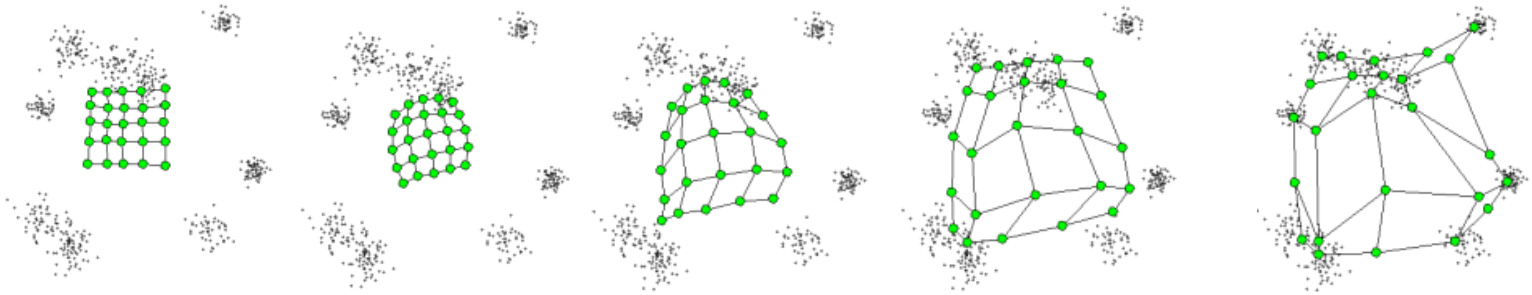
Figure 5. Aspects of vineyards in New Zealand (a) and in selected regions (b-e). The percentage value refers to the percentage of vineyards that have an aspect value, and are therefore represented in the figure. The remaining vineyards are flat and have no aspect value (Leathwick et al. 2002; this study).

IMRE, Stephen P.; MAUK, Jeffrey L.. Geology and Wine 12. New Zealand Terroir. **Geoscience Canada**, [S.l.], dec. 2009. ISSN 1911-4850. Available at: <http://journals.hil.unb.ca/index.php/GC/article/view/12670/13563>. Date accessed: 08 Oct. 2013. doi:10.12789/gs.v36i4.12670.

methodology



SOM clustering



Each neuron *i* of the SOM = a weight model vector, mi = [*mi1,……, min*]T,
neighbourhood *Ni* of the neuron *I*

distances (such as the similarities) between the vector x and all codebook vectors are computed.

The best matching unit (BMU) c, map unit

x: ||x - mc || = min {||x - mi||}

the weight vectors are updated

The BMU and its topological neighbours are moved closer to the input vector in the input space.

The update rule for the weight vector of unit i is:

mi*(t)* + □*(t)* [x*(t)* - mi(t)], *i* □ *Nc (t)*

mi (*t*+1) = mi(*t*), *i* □ *Nc (t)*

where *t* denotes time. *Nc(t)* is the non-increasing neighbourhood function around the winner unit c and 0 < $\alpha(t)$ < 1 is a learning coefficient, a decreasing function of time

variables used

Climate variables

1. Mean annual temperature: strongly influences plant productivity.
2. Mean minimum winter Temperature: influences plant survival.
3. Mean annual solar radiation: determines potential productivity.
4. Monthly water balance ratio: indicates average site “wetness”.
5. Annual water deficit: gives an indication of soil dryness, it is calculated using mean of daily temperature, daily solar radiation and rainfall (Leathwick, Morgan, Wilson, Rutledge, McLeod, & Johnston, 2002)

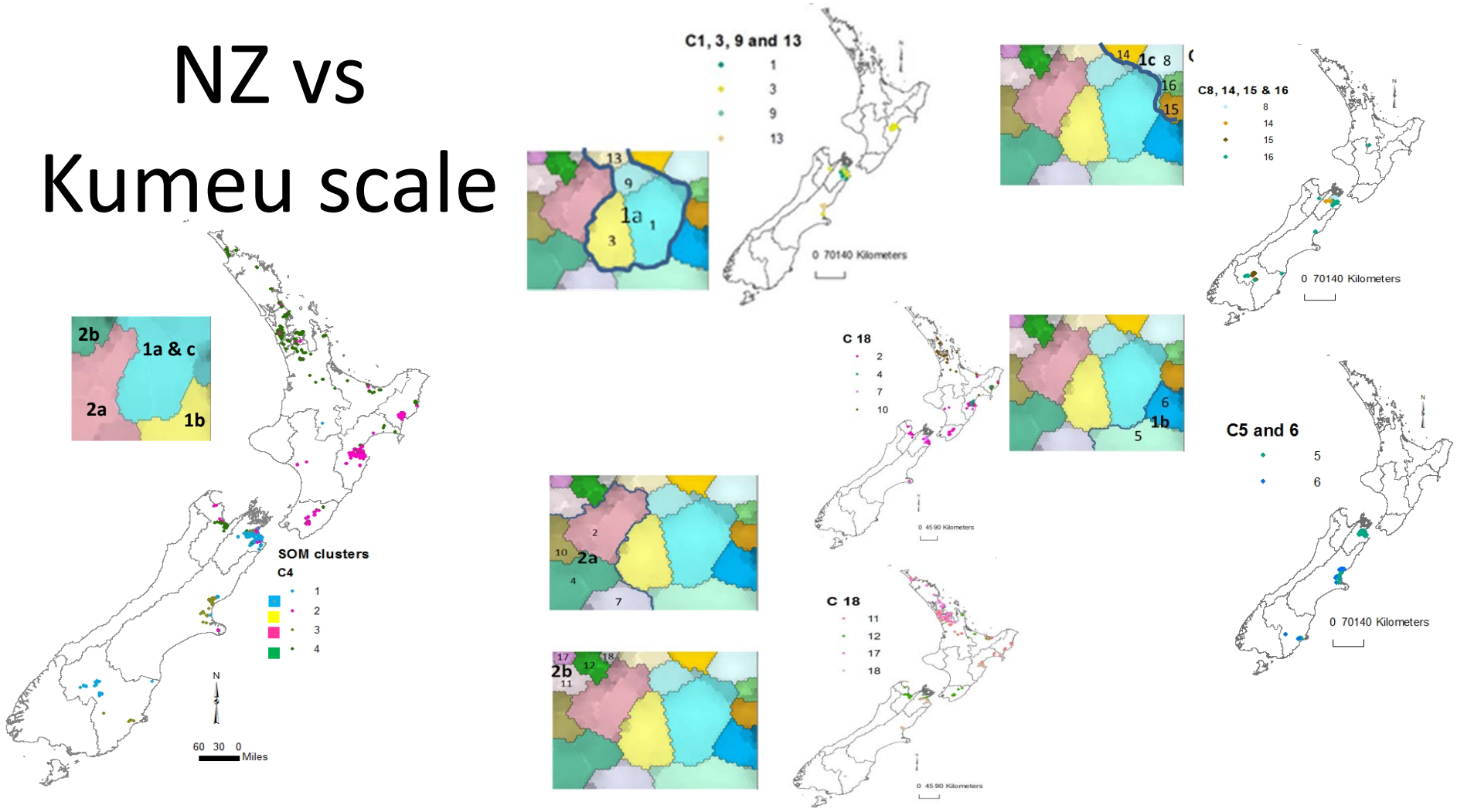
Land form variables

1. Elevation
2. Slope: Major driver of drainage, soil rejuvenation and microclimate
3. Aspect : the direction in which a slope faces
4. Hill shade

Soil variables

1. Drainage: influences the oxygen availability in upper soil layers.
2. Acid soluble phosphorous: indicates a key soil nutrient
3. Exchange calcium: both a nutrient and a determinant of soil weathering.
4. Induration (hardness): determines soil resistance to weathering.
5. Age: separates recent, fertile soils from older less fertile soils.
6. Chemical limitation of plant growth: indicates the presence of salinity of ultramafic substances.

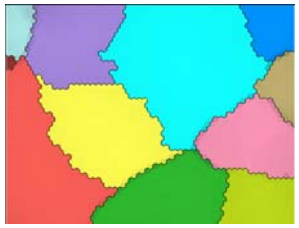
NZ vs Kumeu scale

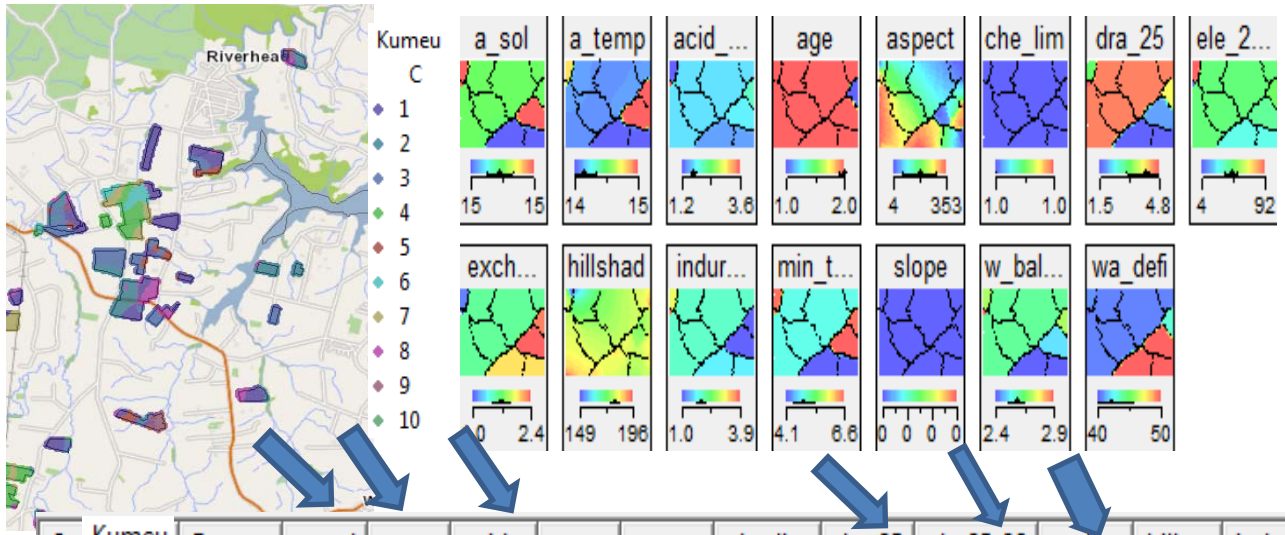


Cluster	pixel count	Ele vation	Ave Temp	A mint Temp	A sol Radiation	Indu ratin	Exch Cation	Aci sol P	Che limit	Age	Slope	Draina ge	Wat Ba R	Water deficit
1a&c	177191	128.59	12.04	1.57	14.92	3.11	1.97	3.79	1.00	1.87	0.06	4.34	1.62	219.95
1b	93607	62.37	11.62	1.09	14.07	3.31	2.01	3.86	1.00	1.16	0.03	4.88	1.70	208.26
2a	127694	36.85	13.35	3.20	14.72	1.23	2.21	2.46	1.07	1.37	0.04	3.28	1.76	179.55
2b	**39396**	**93.84**	**13.74**	**4.59**	**14.89**	**2.28**	**1.42**	**1.62**	**0.94**	**1.71**	**0.06**	**3.74**	**2.67**	**54.10**
Total	437888													

Figure 1b: SOM cluster profiles, Wat Ba R: monthly water balance ratio

SOM clustering

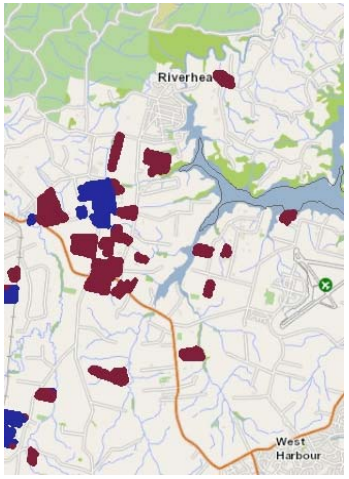


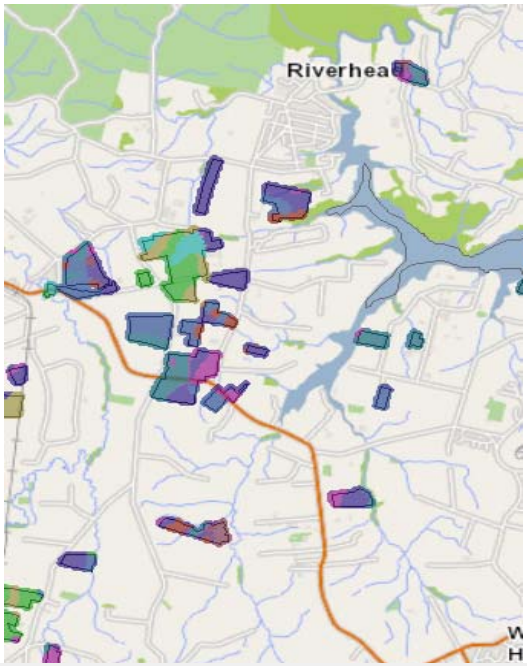


Kumeu C	Frequ...	a_sol	a_t...	acid...	age	aspect	che_lim	dra_25	ele_25_26	exch...	hills...	indur...	m..._...	sl...	w_b...	wa_...
	0.00%															
1	29.05%	14.90	14.10	1.700	2.000	81.2	1.000	4.600	45.00	1.600	177.7	2.000	4.800	0	2.600	40.16
2	20.88%	14.90	14.10	1.702	2.000	318.9	1.000	4.600	44.95	1.599	184.4	2.002	4.802	0	2.600	40.16
3	12.99%	14.90	14.10	1.700	2.000	218.7	1.000	4.600	45.00	1.600	179.5	2.000	4.800	0	2.600	40.16
4	7.89%	14.80	14.00	1.500	2.000	309.3	1.000	1.500	28.00	2.100	182.3	1.700	4.100	0	2.400	50.42
5	10.32%	14.90	14.10	1.700	2.000	134.8	1.000	4.600	45.01	1.600	169.4	2.000	4.800	0	2.600	40.16
6	2.76%	15.00	15.00	1.800	2.000	187.0	1.000	1.900	40.00	2.400	181.2	1.000	6.600	0	2.500	50.28
7	5.28%	14.80	14.00	1.500	2.000	76.4	1.000	1.500	28.00	2.100	179.3	1.700	4.100	0	2.400	50.42
8	7.90%	14.90	14.10	1.700	2.000	18.7	1.000	4.600	45.00	1.600	182.6	2.000	4.800	0	2.600	40.16
9	1.01%	14.90	15.00	2.000	1.001	109.0	1.000	3.899	47.99	2.300	178.9	1.000	6.499	0	2.700	42.96
10	1.91%	15.00	14.80	1.202	2.000	153.6	1.000	3.103	91.87	1.101	176.6	3.098	6.497	0	2.700	40.44

- Annual solar radiation, annual average and minimum temperatures, acid soluble phosphorous, drainage, elevation, cation exchange, induration, monthly water balance and annual water deficit show similarity.
- Aspect and hill shade show variability that can be used for zoning purposes.
- Age (soil) has one cluster that is 1 year (new fertile) and rest of the clusters are 2 years old (less fertile).

NZ vs Kumeu scale





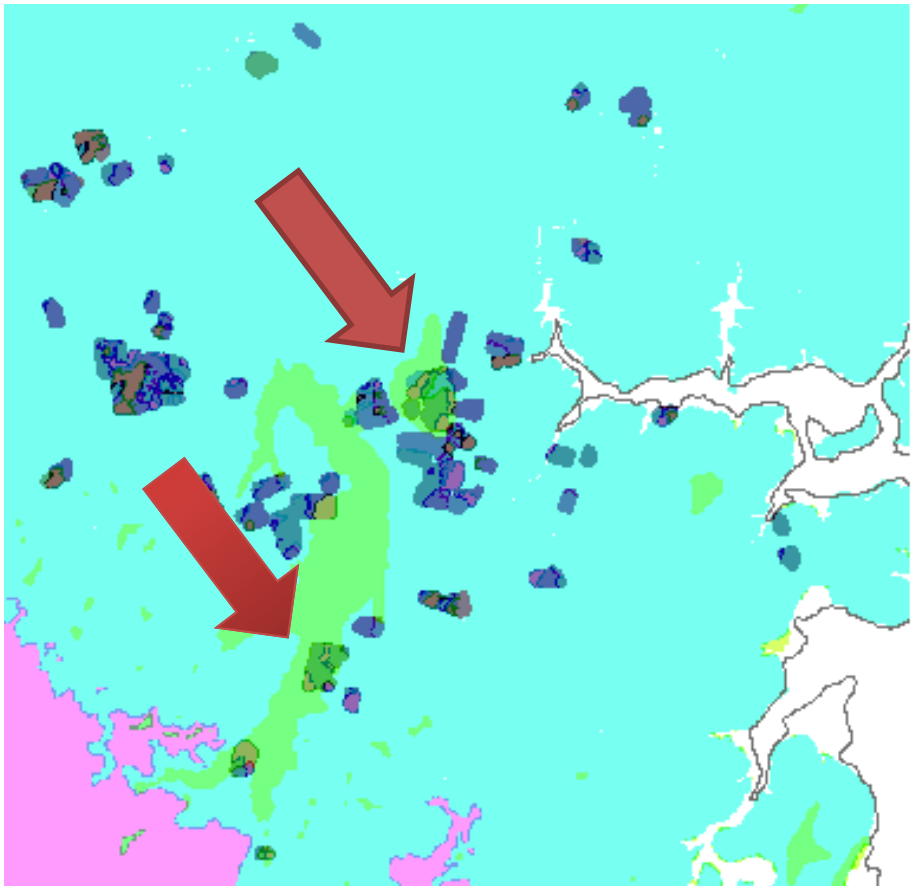
Kumeu C	Frequ...	a_sol	a_t...	acid...	age	aspect	che_lim	dra_25	ele_25_26	exch...	hills...	indur...	min_...	sl...	w_b...	wa_...
	0.00%															
1	29.05%	14.90	14.10	1.700	2.000	81.2	1.000	4.600	45.00	1.600	177.7	2.000	4.800	0	2.600	40.16
2	20.88%	14.90	14.10	1.702	2.000	318.9	1.000	4.600	44.95	1.599	184.4	2.002	4.802	0	2.600	40.16
3	12.99%	14.90	14.10	1.700	2.000	218.7	1.000	4.600	45.00	1.600	179.5	2.000	4.800	0	2.600	40.16
4	7.89%	14.80	14.00	1.500	2.000	309.3	1.000	1.500	28.00	2.100	182.3	1.700	4.100	0	2.400	50.42
5	10.32%	14.90	14.10	1.700	2.000	134.8	1.000	4.600	45.01	1.600	169.4	2.000	4.800	0	2.600	40.16
6	2.76%	15.00	15.00	1.800	2.000	187.0	1.000	1.900	40.00	2.400	181.2	1.000	6.600	0	2.500	50.28
7	5.28%	14.80	14.00	1.500	2.000	76.4	1.000	1.500	28.00	2.100	179.3	1.700	4.100	0	2.400	50.42
8	7.90%	14.90	14.10	1.700	2.000	18.7	1.000	4.600	45.00	1.600	182.6	2.000	4.800	0	2.600	40.16
9	1.01%	14.90	15.00	2.000	1.001	109.0	1.000	3.899	47.99	2.300	178.9	1.000	6.499	0	2.700	42.96
10	1.91%	15.00	14.80	1.202	2.000	153.6	1.000	3.103	91.87	1.101	176.6	3.098	6.497	0	2.700	40.44

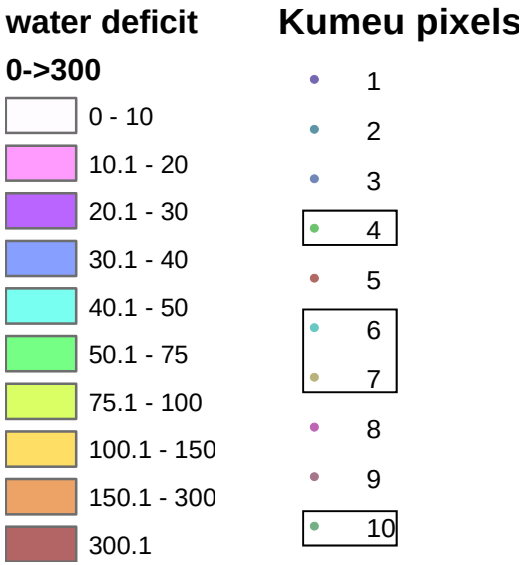
CRT & C5 rules

CRT tree rules created with Kumeu pixels alone show water deficit > 40.3.and aspect (asp>/< 151.99) as major discerning factor followed by hill shade/elevation and min temp.
A temp has been used in 2 rules

Rule no	Instance; confidence	Rule asp; aspect, hs; hill shade, wd; water deficit, ele; elevation 25m resolution	SOM cluster
1	46; 1.0	if wd<= 40.16 and asp <= 106.56 and > 29.15 and hs <= 173 and > 172	one
2	309; 1.0	if wd<= 40.16 and asp <= 136.38 and > 29.15 and hs <= 175 and > 173	one
3	5; 1.0	if wd<= 40.16 and asp <= 145.11 and > 136.38 and hs <= 174 and > 173	one
4	1,916; 1.0	if wd<= 40.16 and asp <= 151.34 and > 29.15 and hs <= 180 and > 175	one
5	4; 1.0	if wd<= 40.16 and asp <= 156.37 and > 151.34 and hs <= 176 and > 175	one
6	2; 1.0	if wd<= 40.16 and asp <= 154.45 and > 151.34 and ele_25 in [45] and hs <= 177 and > 176	one
1	2; 1.0	if wd<= 40.16 and asp <= 277.27 and > 151.34 and ele_25 in [0] and hs <= 182 and > 176	two
2	7; 1.0	if wd<= 40.16 and asp <= 277.27 and > 264.29 and ele_25 in [45] and hs <= 182 and > 181	two
3	126; 1.0	if wd<= 40.16 and asp <= 277.27 and > 176 and hs > 182	two
4	63; 0.984	if wd<= 40.16 and asp <= 284.39 and > 277.27 and hs > 180	two
5	1,425; 1.0	if wd<= 40.16 and asp > 284.39	two
1	13; 0.923	if wd<= 40.16 and asp <= 154.45 and > 151.34 and ele_25 in [45] and hs <= 181 and > 177	three
2	958; 1.0	if wd<= 40.16 and asp <= 277.27 and > 154.45 and ele_25 in [45] and hs <= 181 and > 176	three
3	39; 1.0	if wd<= 40.16 and asp <= 264.29 and > 151.34 and ele_25 in [45] and hs <= 182 and > 181	three
4	8; 1.0	if wd<= 40.16 and asp <= 284.39 and > 277.27 and hs <= 180	three
1	620; 1.0	if wd> 40.16 and asp > 190.35 and min_temp <= 4.8	four
1	484; 1.0	if wd<= 40.16 and asp <= 151.34 and hs <= 172	five
2	22; 1.0	if wd<= 40.16 and asp <= 136.38 and > 106.56 and hs <= 173 and > 172	five
3	11; 1.0	if wd<= 40.16 and asp <= 151.34 and > 136.38 and hs <= 173 and > 172	five
4	9; 1.0	if wd<= 40.16 and asp <= 151.34 and > 145.11 and hs <= 174 > 173	five
5	10; 1.0	if wd<= 40.16 and asp <= 151.34 and > 136.38 and hs <= 175 and > 174	five
6	22; 1.0	if wd<= 40.16 and asp <= 156.37 and > 151.34 and hs <= 175	five
7	253; 1.0	if wd<= 40.16 and asp <= 277.27 and > 156.37 and hs <= 176	five
1	217; 1.0	if wd> 40.16 and min_temp > 6.5	six
1	415; 1.0	if wd> 40.16 and asp <= 190.35 and min_temp <= 4.8	seven
1	52; 1.0	if wd<= 40.16 and asp <= 29.15 and hs <= 180 and hs > 172	eight
2	591; 1.0	if wd<= 40.16 and asp <= 151.34 and hs > 180	eight
1	79; 1.0	if wd> 40.16 and a_temp > 14.8 and min_temp <= 6.5 and > 4.8	nine
1	150; 1.0	if wd> 40.16 and a_temp <= 14.8 and min_temp <= 6.5 and > 4.8	ten

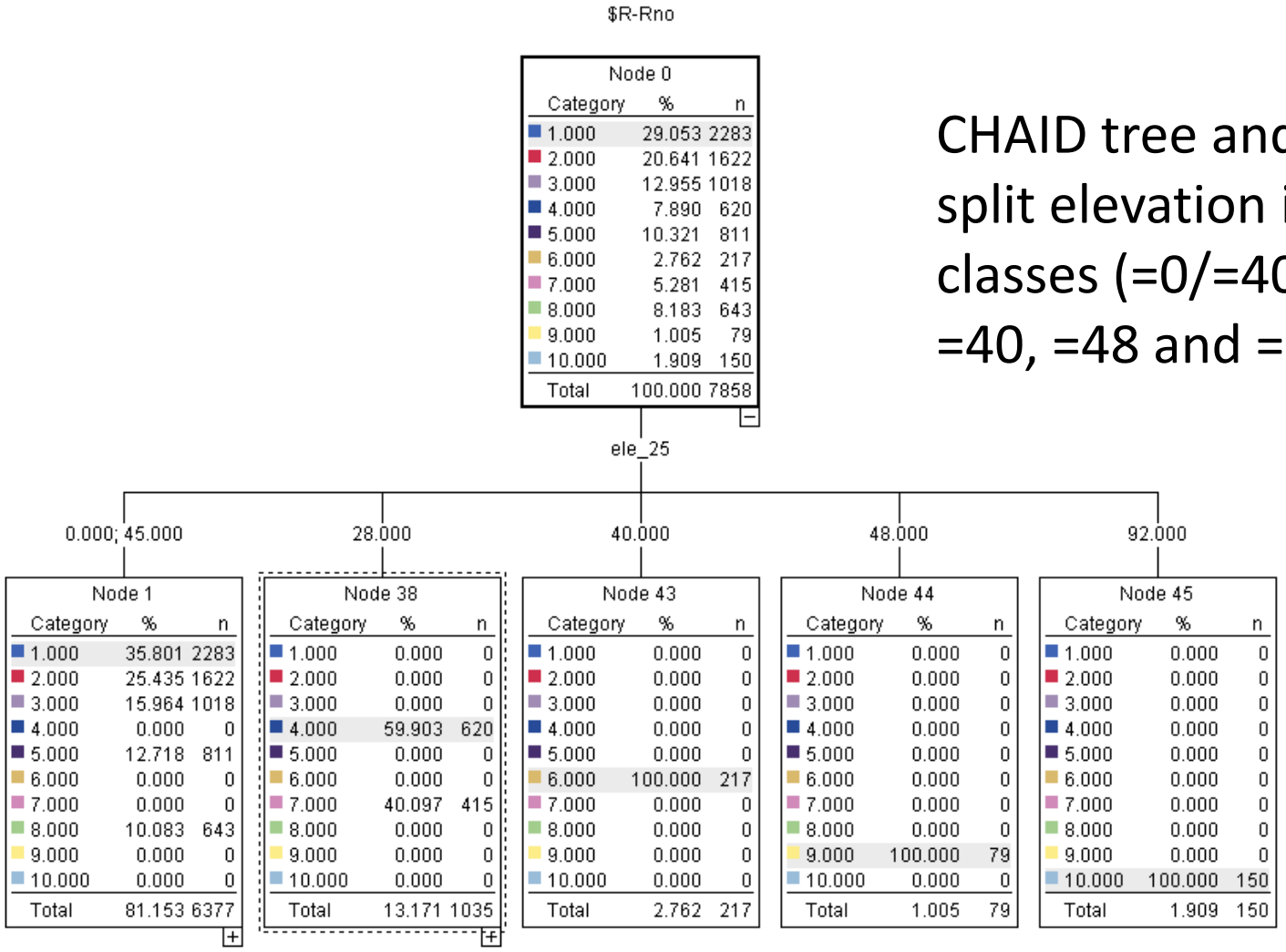
water deficit in Kumeu





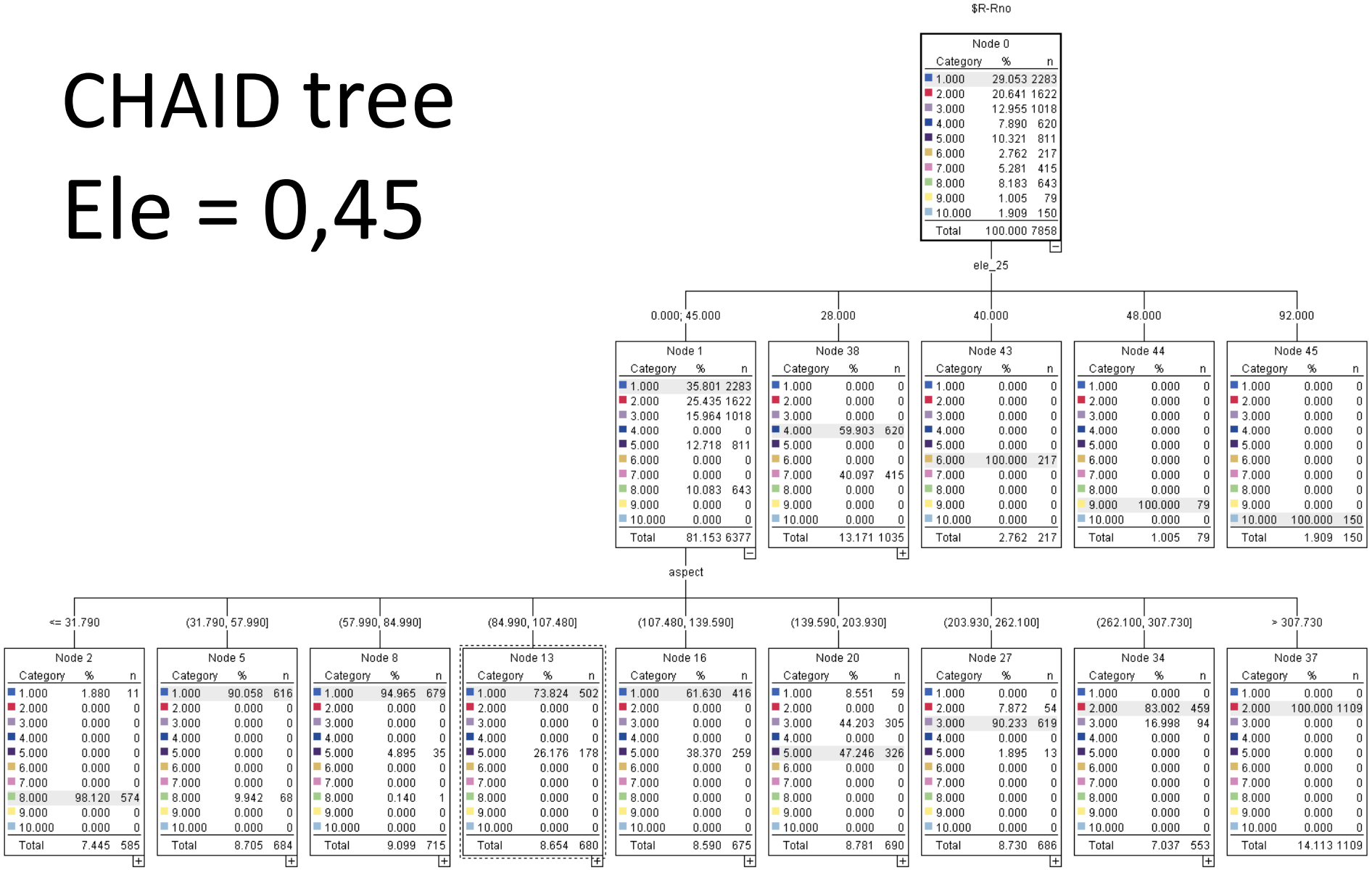
water deficit > 40.16 or <= 40.16

CHAID tree



CHAID tree and rules split elevation into 5 classes (=0/=40, =28, =40, =48 and =92

CHAID tree
Ele = 0,45



CHAID rules

ele = 0 or ele = 45 [Mode: one] (6,377)
- asp <= 31.7900 [Mode: eight] (585)
  - hs <= 180 [Mode: eight] => eight (63; 0.825)
  - hs > 180 [Mode: eight] => eight (522; 1.0)
- asp > 31.7900 and asp <= 57.9900 [Mode: one] (684)
  - hs <= 180 [Mode: one] => one (616; 1.0)
  - hs > 180 [Mode: eight] => eight (68; 1.0)
- asp > 57.9900 and asp <= 84.9900 [Mode: one] (715)
  - hs <= 180 [Mode: one] (714)
    - hs <= 173 [Mode: five] => five (50; 0.7)
    - hs > 173 [Mode: one] => one (664; 1.0)
  - hs > 180 [Mode: eight] => eight (1; 1.0)
- asp > 84.9900 and asp <= 107.4800 [Mode: one] (680)
  - hs <= 173 [Mode: five] => five (209; 0.852)
  - hs > 173 [Mode: one] => one (471; 1.0)
- asp > 107.4800 and asp <= 139.5900 [Mode: one] (675)
  - hs <= 173 [Mode: five] => five (255; 1.0)
  - hs > 173 and hs <= 176 [Mode: one] => one (177; 0.977)
  - hs > 176 [Mode: one] => one (243; 1.0)
- asp > 139.5900 and asp <= 203.9300 [Mode: five] (690)
  - hs <= 173 [Mode: five] => five (198; 1.0)
  - hs > 173 and hs <= 176 [Mode: five] => five (150; 0.853)
  - hs > 176 and hs <= 178 [Mode: three] (237)
    - sp <= 0.06 [Mode: three] => three (156; 0.987)
    - sp > 0.06 [Mode: three] => three (81; 0.815)
  - hs > 178 [Mode: three] => three (105; 0.81)
- asp > 203.9300 and asp <= 262.1000 [Mode: three] (686)
  - hs <= 182 [Mode: three] (633)
    - *a_temp <= 14.1 [Mode: three] (632)*
      - *hs <= 176 [Mode: five] => five (13; 1.0)*
      - *hs > 176 [Mode: three] => three (619; 1.0)*
    - *a_temp > 14.1 [Mode: two] => two (1; 1.0)*
  - hs > 182 [Mode: two] => two (53; 1.0)
- asp > 262.1000 and asp <= 307.7300 [Mode: two] (553)
  - hs <= 181 [Mode: three] => three (154; 0.604)
  - hs > 181 [Mode: two] => two (399; 0.997)
- asp > 307.7300 [Mode: two] => two (1,109; 1.0)

ele = 28 [Mode: four] (1,035)
- asp <= 203.9300 [Mode: seven] (418)
  - asp <= 139.5900 [Mode: seven] => seven (353; 1.0)
  - asp > 139.5900 [Mode: seven] => seven (65; 0.954)
- asp > 203.9300 [Mode: four] => four (617; 1.0)

ele = 40 [Mode: six] => six (217; 1.0)

ele = 48 [Mode: nine] => nine (79; 1.0)

ele = 92 [Mode: ten] => ten (150; 1.0)

Aspect and hillshade also used in the rules

for clusters three, five and two annual average temperature is used (in *italics*).

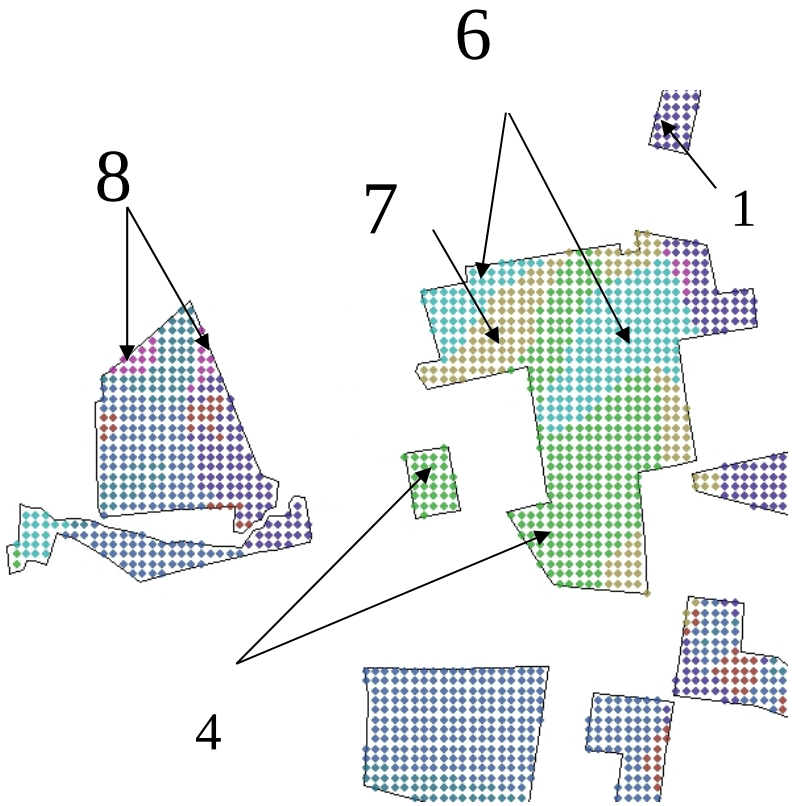
SOM clusters six, nine and ten are defined purely on elevation with 217, 79 with 150 instances respectively all at 100% confidence.

Clusters seven and four vary in elevation and aspect

Quest tree rules

- ele_25 = 0 or ele_25 = 45 or ele_25 = 92 [Mode: one]
  - ele_25 = 92 [Mode: ten] ⇨ ten
  - ele_25 = 0 or ele_25 = 45 [Mode: one]
    - aspect <= 177.7973 [Mode: one]
      - hillshad <= 172.589 [Mode: five] ⇨ five
      - hillshad > 172.589 [Mode: one]
        - aspect <= 143.0628 [Mode: one] ⇨ one
        - aspect > 143.0628 [Mode: three] ⇨ three
    - aspect > 177.7973 [Mode: two]
      - aspect <= 267.2804 [Mode: three] ⇨ three
      - aspect > 267.2804 [Mode: two] ⇨ two
- ele_25 = 28 or ele_25 = 40 or ele_25 = 48 [Mode: four]
  - ele_25 = 28 [Mode: four]
    - aspect <= 216.3665 [Mode: seven] ⇨ seven
    - aspect > 216.3665 [Mode: four] ⇨ four
  - ele_25 = 40 or ele_25 = 48 [Mode: six]
    - min_temp <= 6.5000 [Mode: nine] ⇨ nine
    - min_temp > 6.5000 [Mode: six] ⇨ six

1
2
3
4
5
6
7
8
9
10



Regression

Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed	Method
1	wa_defi, age, hillshad, slope, aspect, min_temp, acid_s_p, induration(b)	.	Enter
a. Dependent Variable: Rno			
b. Tolerance = .000 limits reached.			

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.638(a)	.407	.407	1.90878
a. Predictors: (Constant), wa_defi, age, hillshad, slope, aspect, min_temp, acid_s_p, induration				

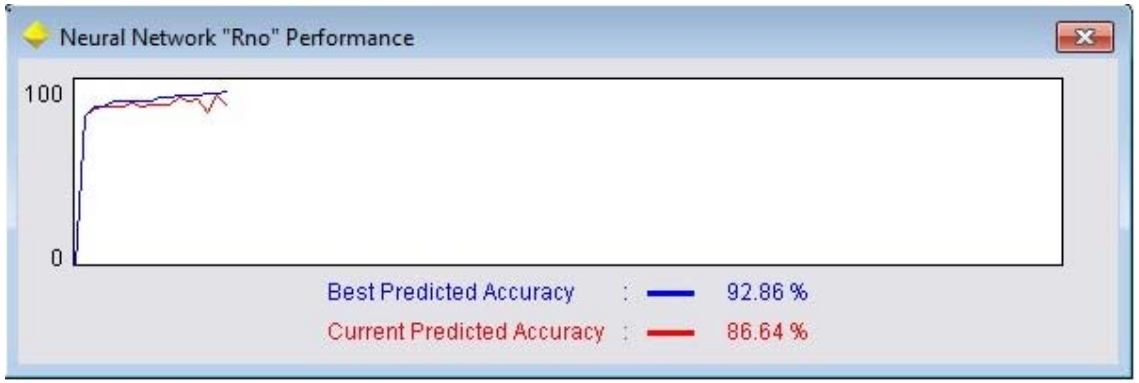
ANOVA(a)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	19656.497	8	2457.062	674.379	.000(b)
	Residual	28597.388	7849	3.643		
	Total	48253.884	7857			
a. Dependent Variable: Rno						
b. Predictors: (Constant), wa_defi, age, hillshad, slope, aspect, min_temp, acid_s_p, induration						

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-6.480	2.014		-3.218	.001
	aspect	-5.89E-003	.000	-.270	-28.249	.000
	min_temp	1.504	.050	.297	30.117	.000
	hillshad	8.71E-003	.005	.018	1.932	.053
	induration	2.539	.217	.272	11.713	.000
	acid_s_p	-3.624	.463	-.157	-7.826	.000
	age	-5.875	.255	-.237	-23.083	.000
	slope	-2.852	.425	-.061	-6.703	.000
	wa_defi	.360	.016	.544	22.101	.000
a. Dependent Variable: Rno						

ANN test results



Analysis

Estimated accuracy: 92.861
Input Layer: 20 neurons
Hidden Layer 1: 3 neurons
Output Layer: 10 neurons
Relative Importance of Inputs

aspect	0.335183
hillshad	0.30243
ele_25	0.242831
age	0.145853
dra_25	0.101321
wa_defi	0.0725418
induration	0.0692616
a_temp	0.0647437
min_temp	0.0521352
a_sol	0.0475153
w_bal_ra	0.0293795
exch_cal	0.0234538
acid_s_p	0.0108849
slope	0.00307043

Fields

Target

Rno

Build Settings

Use partitioned data: false
Method: Quick
Stop on: Default
Set random seed: false
Prevent overtraining: true
Sample %: 50.0
Optimize: Memory

Training Summary

Model type: Neural net
Stream: Stream1
User: sshanmug
Date built: 25/06/13 23:24
Application: Clementine 10.1
Elapsed time for model build: 0 hours, 0 mins, 3 secs

conclusions

- Traditional approaches to zoning require extensive knowledge→ makes zoning new areas/ "*terroirs*" impossible
- It s possible to identify main/major contributory attributes using even low res thematic maps
- SOM and TDIDT approach gives a means overcome this
- For Kumeu
  - Water deficit, elevation (along with aspect and hill shade)
  - annual minimum and average temperatures (same as New Zealand regional/ macro-scale GDD)
- Regression test results: water deficit, age, hill shade, slope, aspect, min temp, acid sol phosphorous, induration as predictors with .407 adjusted R^2