Big Data Techniques for Applied Geoscience: Compute *and* Communicate

Anya M. Reading*
Sch. of Physical Sci. (Earth Sciences) and CODES Centre of Excellence University of Tasmania
anya.reading@utas.edu.au

Matthew J. Cracknell
Sch. of Physical Sci. (Earth Sciences) and ARC ITR Hub for Transforming the Mining Value Chain, UTAS
m.j.cracknell@utas.edu.au

Stephen Kuhn
Sch. of Physical Sci. (Earth Sciences) and CODES Centre of Excellence University of Tasmania
stephen.kuhn@utas.edu.au

**presenting author*

SUMMARY

Big Data techniques have the potential to be paradigm-changing for applied geoscience if they are used widely. A significant number of such techniques, under the umbrella of Earth informatics, involve Machine Learning applied to high dimensional data to create new forms of value. This contribution presents two case studies of successful Earth informatics computation and the communication of the value of results, which provide insight into the uptake of 'Big Data' in geosciences.

Machine Learning techniques split naturally into either supervised or unsupervised approaches. Supervised algorithms, such as Random Forests™ (RF), support vector machines or neural networks, share the concept of training a classifier using an initial (training) dataset. They are generally applied to predictive tasks, such as our first case study, predicting lithology from remote sensing and airborne geophysical data. Unsupervised algorithms, such as Self-Organising Maps (SOM), allow patterns inherent in the data to emerge without the use of a training dataset. They are generally applied to tasks which seek to explore patterns in data, such as our second case study, which identifies new potentially prospective river catchments. We find that calculating and presenting explicitly the newly extracted value, of the result obtained through computation, is an essential component of the post-compute evaluation.

As strong advocates for the use of a range of Big Data techniques in applied geosciences, we conclude that the benefits to be gained from the way that we 'compute' can be lost if we do not also take considerable care with the ways that we 'communicate'.

Key words: Big Data, Machine Learning, Supervised, Unsupervised, High Dimensional, Communication.

INTRODUCTION

Big Data Approaches and Uptake in Applied Geoscience

Big Data approaches have gained wide popularity in many areas of commerce. The name implies large volumes of data, but the key idea relates to computational analyses "*that we can do with large amounts of data, that we can't do with small amounts, to extract new insights or create new forms of value*" (Mayer-Schönberger and Cukier, 2013). These approaches are based on statistical associations and complement traditional geophysics approaches which are based on deterministic physical relationships. Where available data exist in high dimensions (effectively 4 or more layers), there is great potential to make predictions and to explore for useful patterns between these dimensions using Big Data techniques.

Examples of successful Machine Learning studies in applied geoscience using high-dimensional data inputs include those applied to exploration target generation (Cracknell *et al.*, 2014, Merdith *et al.*, 2015), those applied to multi-disciplinary studies (Cracknell *et al.*, 2015) and those applied to data-rich applications such as drill-hole data (Reading and Gallagher, 2013, Hill *et al.*, 2015). In spite of these successful demonstration studies, and numerous examples from other industries, the uptake of Big Data techniques as routine procedures for mineral resources exploration has been limited.

METHODS AND DATA

Prediction using Random Forests (Case Study 1)

Supervised classification was implemented using the Orange software environment (Demsar *et al.*, 2013) including the graphical programming interface to readily facilitate data input and workflow. RF (Breiman, 2001) is an ensemble supervised classifier that constructs a classification algorithm through multiple randomised decision tree classifiers. Randomness is introduced by randomly subsetting a predefined number of inputs (often $\sqrt{}$ number of variables) to split at each node of a tree and by bagging. Bagging generates training samples for each tree by sampling with replacement a number of samples equal to the number of instances in the training data. The remainder, so called out-of-bag samples, are used for evaluation. The Gini Index is used by RF to determine a best-split threshold at each tree node. The Gini index is defined as $Gini(t) = \sum_{c=1}^{j} g_c(1 - g_c)$, where g_c is the probability of class c at node j, $g_c = \frac{n_c}{n}$, n_c is the number of samples belonging to c, and n is the total number of samples within j. For each candidate split, the threshold t that results in the maximum reduction in class heterogeneity is selected (Breiman *et al.*, 1984). The class membership probability, p_c, that a given sample, i, is one of c in the training data is estimated by dividing the total number of votes for each class by the number of trees, T (Hastie *et al.*, 2009), $p_c = \frac{1}{T}\sum_{c=1}^{T} y_c^i$.

Case Study 1 uses data from an orogenic gold exploration area near St Ives Gold Mine, Eastern Goldfields, WA (Figure 1A). 16 layers of data were used as initial inputs including gravity, magnetic, DTM, Landsat 7 and airborne gamma ray spectrometry data. After the variable ranking process, the potential field, DTM and spectrometry data proved most relevant. RF was trained using 1.4% of available data, 100 samples per lithology class. The trained RF classifier was applied to all samples within the study area and the compute time was 10-15 minutes on a well specified current-model notebook or desktop PC.

Patten detection using a Self-Organising Map (Case Study 2)

Unsupervised classification using a Self-Organising Map (SOM) was implemented using the R statistical programming language (R Core Team, 2015) package kohonen (Wehrens and Buydens, 2007). The Davies-Bouldin Index (Davies and Bouldin, 1979) which provides an estimate of cluster similarity and dispersion, was used to identify an optimal number of merged SOM nodes. Input variables that contributed significantly high or low SOM code-vector values (relative differences) were identified using the ratio between $s_v(i,k)$ and $\frac{1}{c-1}\sum_{j+1} s_v(j,k)$, where $s_v(i,k)$is the mean cluster i code-vectors for input variable k and where $s_v(j,k)$ is the mean cluster j code-vectors for input variable k. This ratio provides an indication of the relative difference in k for cluster i compared to that of all other cluster's mean code-vector value for k (Siponen *et al.*, 2001).

Case Study 2 (Figure 2A) uses data from the National Geochemical Survey of Australia (de Caritat and Cooper, 2011). Multiple input layers of Mobile Metal Ions (MMI^{TM}) assay data derived from top of outlet sediments (coarse fraction) were used. Analysis was carried out on these data for 44 elements (including Au); bulk sediment properties, pH, electrical conductivity and % fractions of clay, silt and sand; 30 geostatistical summaries of geophysics data, and two measures of catchment geometry, a total of 81 layers. The SOM was initiated with 200 randomly positioned seed nodes and run for 10,000 iterations in a compute time of 15-20 minutes.

RESULTS

Case Study 1, Lithology Prediction – Scientific Outcomes

The Machine Learning prediction has resulted in a redefinition of some of the lithology boundaries within the study area (Figure 1). Notable changes include the likely boundary between the granite and the dolerite (C1) being located further to the west; the volcanogenic sediments (C2) apparently spanning a wider area, but terminating at a more southerly point; significant refinements to shape of the komatiite units (C3, C4), and the likely extension of basalt unit (C5) over an area previously mapped as dolerite. These refinements may be in themselves significant to explorers, or they may enable a better result in inversions constrained by, for example, potential field data (Reading *et al.*, 2015). An important step in the evaluation of these refinements is prediction evaluation (Cracknell and Reading, 2014). For this field area, Kuhn *et al.*, (2016) present an investigation that explores the use of information entropy, H, as a proxy for inaccuracy in the prediction result.

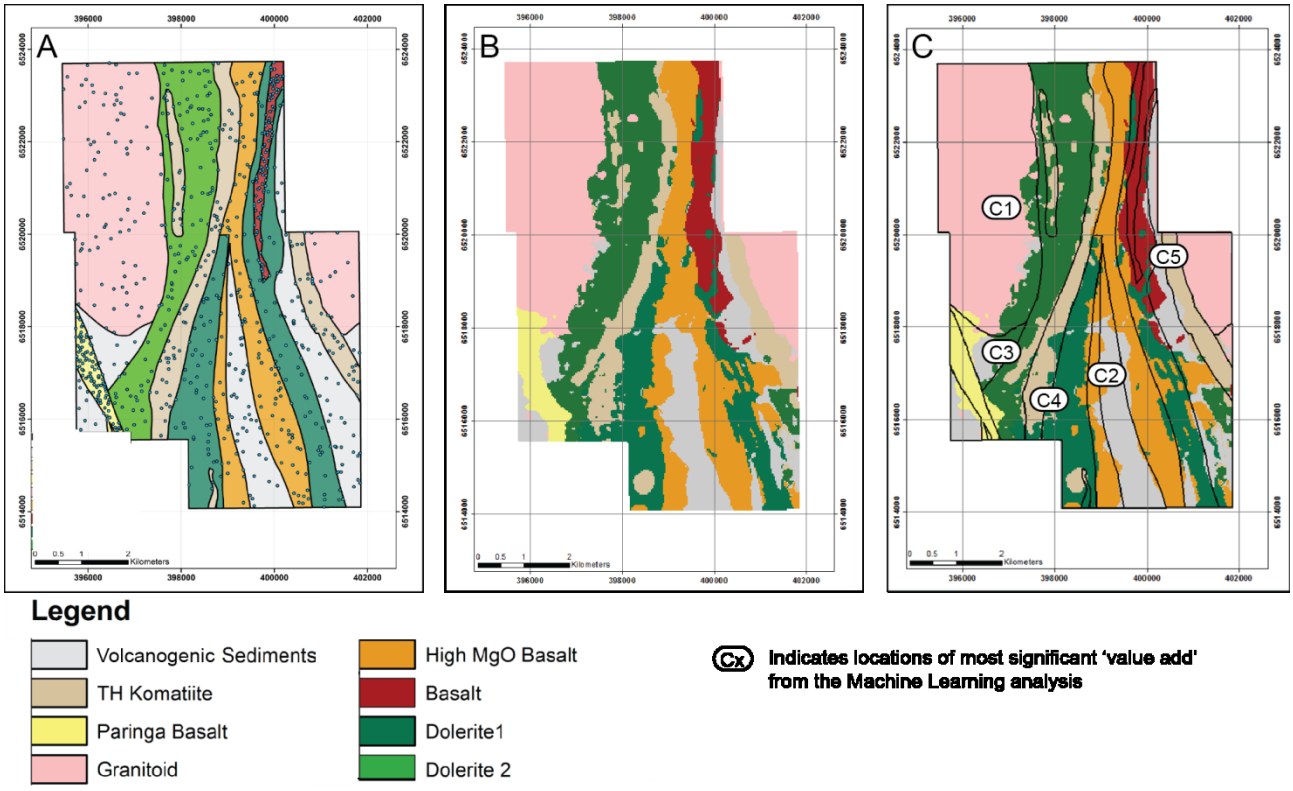

Figure 1: Refinement of a geological map in the area of St Ives Gold Mine, Eastern Goldfields, WA, through Machine Learning prediction using RF (adapted from Kuhn et al, 2015; Kuhn et al., in preparation). A – Original map with locations of training samples. B – Refined map showing the result of the RF classification. This is effectively the 'new' map for subsequent use. C – 'Communication' figure that shows the locations where the Machine Learning has improved the value of the original map through the relocation of lithology contacts.

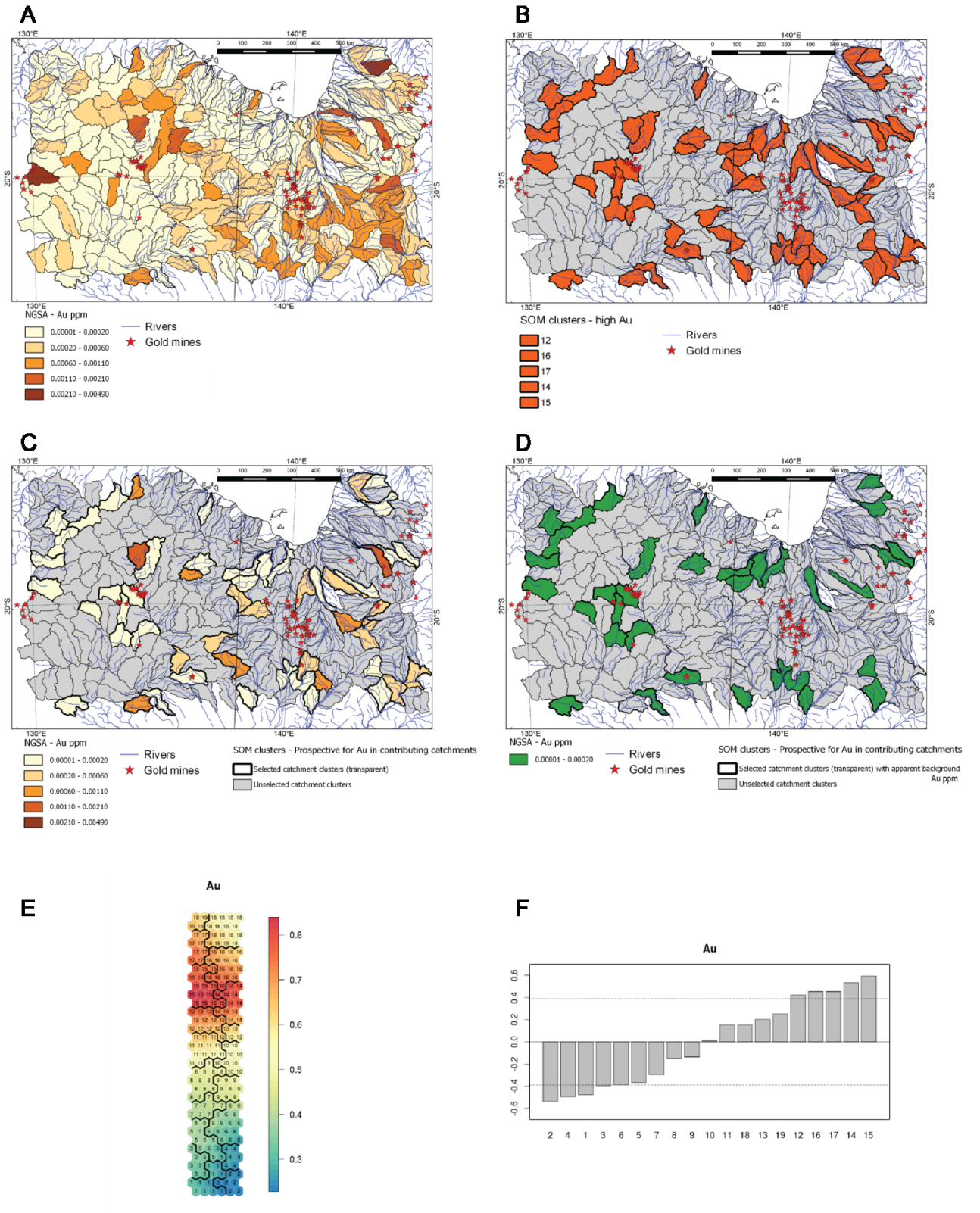

Figure 2: Identification of river catchments potentially prospective for gold through Machine Learning pattern detection using SOM applied to geochemical data in northern central Australia. A – Concentration of gold in river catchments. B – River catchments identified, across many input variables, using SOM as showing a propensity for gold occurrence. C – River catchments shown as concentration of gold overlain with catchments identified in B. This is effectively the new map for subsequent use. D – 'Communication' figure that highlights the catchments identified using SOM that would not have been identified as prospective based solely on the geochemical concentration data. E – Position of identified clusters on the SOM 2D map, warm colours indicate prospective clusters. F – Code-vector ratios, used to define clusters indicating high propensity for gold occurrence.

Case Study 2 – Pattern Identification - Scientific Outcomes

The Machine Learning identified patterns highlight numerous catchments that share groups of features that together indicate elevated prospectivity for Au in contributing catchments (Figure 2B). The relationship of different clusters on the 2D unsupervised SOM that supports this pattern identification is shown (Figure 2E) together with the corresponding code-vector ratios (Figure 2F). A more advanced presentation of input data uses the centred log transform values of the input geochemical data (Cracknell and de Caritat, 2016). We use the geochemical concentration data in this illustration to focus directly on the value added through Machine Learning. Identified catchments are spread through the study area and show considerable variability as to whether they correlate with catchments identifiable through their elevated Au concentration alone (Figure 2C). Thus, we can see the value in using both the 'obvious' indications of prospectivity, and the indications that come through using multiple input layers and a computational approach. Further use of the outputs of this analysis requires bringing together the geochemical data, the insights from the SOM analysis and ancillary information such as river catchment geometry. Collaborating experts can then make best use of all information available towards the common mineral prospecting goal.

DISCUSSION

Communication of Machine Learning Outputs

Machine Learning outputs (e.g. Figure 1B) provide results in a form that often looks very similar to the input information (e.g. Figure 1A). While a domain expert (e.g. exploration geologist) would be well advised to proceed with the updated product (e.g. lithology map, Figure 1B) that contains refinements due to the Machine Learning process, the extent to which the improvements come from this computation is not readily apparent. Routine outputs from Machine Learning do indeed contain "*new insights and new forms of value*" but the newly added value is often not well demonstrated. We suggest that this has contributed to the limited uptake of Big Data techniques in applied geoscience.

One relatively simple additional output that can be presented shows the output information obtained from the RF exercise on high dimensional data set against the lithology map input information, e.g. Figure 1C. In this example, the Machine Learning output is shown as coloured areas, with the contacts as inferred prior to the Machine Learning shown as black lines. This approach is successful for cases of map refinement and may be enhanced with the addition of annotations as in our example. Text annotations become less appropriate for large and complex output explanations but could be replaced with links or augmented reality information in this case.

We illustrate an alternate approach in Figures 2C and 2D making use of transparency / opacity masks to set the new knowledge obtained through the Self-Organising Map exercise on high dimensional data in the context of what was previously understood from the most obvious single input data layer. While all information is contained in Figure 2C, Figure 2D shows explicitly the value added from the Machine Learning. In this case, the value added information is significant. Some catchments with low element concentration in sediments show considerably higher promise from the high dimensional study which brings much more information to the inference process.

Promoting the Uptake of Machine Learning in Applied Geosciences

In the thorough summary of Machine Learning applied to geological mapping by Cracknell *et al.*, 2014, the three parts of the process 1) Data preparation; 2) Machine Learning; 3) Prediction evaluation, were outlined. In this contribution, we have focussed on adding 'communication' figures which show the value added by the Machine Learning. One further, relatively easy, step remains outstanding for Machine Learning to become a natural part of the exploration tool box, and that relates to part 1, that of data preparation. All too often, data are stored in diverse formats and with limited metadata. Increasing use of GIS software and data integration training is improving this situation, but data should be archived in interoperable formats, such that reprocessing is achievable, as routine. In order to promote the uptake of Machine Learning, we therefore make the following recommendations:

- Data should be collected and stored in archives format standards such that each record may be used in automated Machine Learning or other Big Data exercises with minimal reformatting. This reduces the lead-in time for the Machine Learning work and hence lowers the barrier to uptake.
- The outputs of the Machine Learning, and subsequent prediction evaluation, should be displayed in a way that makes clear the improvement (or otherwise) due to the Machine Learning procedures. It is vital that the insights and/or added value gained through the Machine Learning are clear to decision makers, other than computation experts, reviewing the improved product.

CONCLUSIONS

The first case study shows that the Machine Learning algorithm, Random Forests, may be used to predict lithology from geophysical and remote sensing input datasets. The approach improved a lithology map for a region including significant undercover areas. The second case study shows that a Self-Organising Map may be used to identify river catchments of potential new interest to prospectors.

Both case studies result in a geological map or spatial representation that contains added value by virtue of the Big Data process that has been undertaken. This added value translates directly to new insights and the analyst who has carried out the Machine Learning study has a sense of the added benefit. Common to both case studies, these improvements, i.e. the parts of the geographic map where the Machine Learning process has added new knowledge, are not readily evident to the end-user in standard map outputs.

While Machine Learning approaches remain non-routine in applied geoscience, we must take care to present a second set of results that make explicit the benefits to be gained from the 'compute'. These benefits can be lost on the very end-users that stand to gain the most, if we do not also put considerable thought and insight into the ways that we 'communicate'.

ACKNOWLEDGMENTS

This research was conducted in association with the ARC Research Hub for Transforming the Mining Value Chain (project number IH130200004). SK is supported through an Australian Postgraduate Award Scholarship through the University of Tasmania. We would like to thank Gold Fields Ltd for access to data for case 1, and acknowledge data provided by Geoscience Australia through the National Geochemical Survey of Australia for case 2. Map plotting software used for case 1 was Esri ArcGIS, and for case 2 was Open Source QGIS. Our thanks also go to Stephen Hardy, Craig Lindley, Adele Seymon, Robbie Rowe and numerous sceptical colleagues for constructive discussions.

REFERENCES

Breiman, L., 2001. Random forests, *Machine Learning,* 45**,** 5-32.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J., 1984. Classification and Regression Trees, *The Wadsworths & Brooks/Cole Statistics/Probability Series, Pacific Grove, USA*.

Cracknell, M.J. & de Caritat, P., 2016. Catchment-scale gold prospectivity analysis from the National Geochemical Survey of Australia, *26th Goldschmidt Conference, Yokohama, Japan, 26 June - 1 July*.

Cracknell, M.J. & Reading, A.M., 2014. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information, *Computers & Geosciences,* 63**,** 22-33.

Cracknell, M.J., Reading, A.M. & de Caritat, P., 2015. Multiple influences on regolith characteristics from continental-scale geophysical and mineralogical remote sensing data using Self-Organizing Maps, *Remote Sensing of Environment,* 165**,** 86-99.

Cracknell, M.J., Reading, A.M. & McNeill, A.W., 2014. Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer-Mt Charter region, Tasmania, using Random Forests (TM) and Self-Organising Maps, *Australian Journal of Earth Sciences,* 61**,** 287-304.

Davies, D.L. & Bouldin, D.W., 1979. A cluster separation measure, *IEEE Tranactions on Pattern Analysis and Machine Intelligence PAMI-1***,** 224-227.

de Caritat, P. & Cooper, M., 2011. National Geochemical Survey of Australia: The Geochemical Atlas of Australia, *GA Record 2011/20, Geoscience Australia, Canberra, ACT, Australia*.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M. & Zupan, B., 2013. Orange: Data Mining Toolbox in Python, *Journal of Machine Learning Research,* 14**,** 2349-2353.

Hastie, T., Tibshirani, R. & Friedman, J.H., 2009. The elements of statistical learning: data mining, inference and prediction, *2nd Edition, Series in Statistics. Springer, New York, USA*.

Hill, E.J., Robertson, J. & Uvarova, Y., 2015. Multiscale hierarchical domaining and compression of drill hole data, *Computers & Geosciences,* 79**,** 47-57.

Kuhn, S., Cracknell, M.J. & Reading, A.M., 2016. Lithological Mapping via Random Forests: Information Entropy as a Proxy for Inaccuracy, *ASEG Extended Abstracts, 25th International Geophysical Conference and Exhibition, 21-24 August, Adelaide, Australia***,** 1-4.

Mayer-Schönberger, V. & Cukier, K., 2013. Big Data: A Revolution That Will Transform How We Live, Work and Think, *John Murray (Publishers), UK*.

Merdith, A.S., Landgrebe, T.C.W. & Muller, R.D., 2015. Prospectivity of Western Australian iron ore from geophysical data using a reject option classifier, *Ore Geology Reviews,* 71**,** 761-776.

R Core Team, 2015. R: A Language and Environment for Statistical Computing.

Reading, A.M., Cracknell, M.J., Bombardieri, D.J. & Chalke, T., 2015. Combining Machine Learning and Geophysical Inversion for Applied Geophysics, *ASEG Extended Abstracts, 24th International Geophysical Conference and Exhibition, 15-18 February, Perth, Australia***,** 1-4.

Reading, A.M. & Gallagher, K., 2013. Transdimensional change-point modeling as a tool to investigate uncertainty in applied geophysical inference: An example using borehole geophysical logs, *Geophysics,* 78**,** WB89-WB99.

Siponen, M., Vesanto, J., Simula, O. & Vasara, P., 2001. An approach to automated interpretation of SOM, *in: Allinson, N., Yin, H, Allinson, L., Slack, J. (Eds), Advances in Self-Organising Maps, Springer London***,** 89-94.

Wehrens, R. & Buydens, L.M.C., 2007. Self- and super-organizing maps in R: The kohonen package, *Journal of Statistical Software,* 21**,** 1-19.