IDENTIFYING COMPLEX SOURCES IN LARGE ASTRONOMICAL DATA USING A COARSE-GRAINED COMPLEXITY MEASURE

GARY SEGAL,1, 2 DAVID PARKINSON,1, 3 RAY P NORRIS,4, 2 AND JESSE SWAN5, 2

1 School of Mathematics and Physics, University of Queensland, St Lucia, Brisbane, QLD 4072, Australia
2 CSIRO Astronomy and Space Science, PO Box 76, Epping, 1710, NSW, Australia
3 Korea Astronomy and Space Science Institute, Daejeon 34055, Korea
4 Western Sydney University, Locked Bag 1797, Penrith South, 1797, NSW, Australia
5 School of Natural Sciences, University of Tasmania, Private Bag 37, Hobart 7001, Australia

ABSTRACT

The volume of data that will be produced by the next generation of astrophysical instruments represents a significant opportunity for making unplanned and unexpected discoveries. Conversely, finding unexpected objects or phenomena within such large volumes of data presents a challenge that may best be solved using computational and statistical approaches. We present the application of a coarse-grained complexity measure for identifying interesting observations in large datasets. This measure, which has been termed apparent complexity, has been shown to model human intuition and perceptions of complexity. Apparent complexity provides a computationally efficient alternative to supervised learning and traditional outlier detection methods for identifying the most interesting observations in very large datasets. Unlike supervised learning approaches it does not learn features associated with known interesting observations positioning the approach as a candidate for identifying unknown unknowns. Furthermore, the approach can be implemented at worst case linear time complexity, providing an advantage when processing very large datasets. We show using data from the Australia Telescope Large Area Survey (ATLAS) that the approach can be used to distinguish between images of galaxies which have been classified as having simple and complex morphologies. We also show that the approach generalises well when applied to new data after being calibrated on a smaller dataset.

Keywords: methods: statistical, techniques: image processing, radio continuum: galaxies

gp.segal@gmail.com
davidparkinson@kasi.re.kr
Ray.Norris@csiro.au
1. INTRODUCTION

The Universe is very large, and we have only just begun to scratch the surface in terms of identifying the different objects and events that it contains. Each generation of instrumentation and infrastructure has expanded our knowledge and sample size significantly, often resulting in unexpected scientific results. For example, of the 10 greatest discoveries by the Hubble Space Telescope, only one was listed in its key science goals (Norris 2017a). The next generation of astrophysical instruments will be collecting petabytes of data per day. This represents a significant opportunity for making unplanned discoveries due to the large volume of observational data that will be made available, but is beyond the limit for the amount of information that can be examined directly by the human astronomical community on any reasonable timescale.

One good example of this expansion in surveying and collection capability is in the area of extragalactic radio astrophysics. We currently know of about 2.5 million extragalactic radio sources, but future surveys will increase this number by several orders of magnitude. The Australian Square Kilometre Array Pathfinder (ASKAP) and the Evolutionary Map of the Universe (EMU) survey is predicted to increase this number to about 70 million (Norris 2017b) in the continuum. The Square Kilometre Array (SKA) may increase this number into the billions (Jarvis et al. 2015).

Progress that has been made with supervised machine learning approaches such as Convolution Neural Networks have demonstrated the potential effectiveness of these approaches for identifying and classifying observations in astronomical surveys based on their features (Aniyan & Thorat 2017; Karpenka et al. 2013; Kim & Bailer-Jones 2016; Kessler et al. 2010; Dieleman et al. 2015; Huertas-Company et al. 2015; Charnock & Moss 2017). Supervised learning approaches however may not be appropriate for discovering new types of interesting objects and phenomena. Supervised approaches may become highly effective at identifying observations that have been previously considered interesting, and overlook new observations whose features have little in common with past interesting observations. Such an approach can be said to suffer from an expectation bias discussed by Norris (2017a) and Robinson (1987). Approaches that do not rely on learning specific features representing past interesting or non-interesting observations, such as the application of unsupervised learning methods, may therefore be considered better candidates for detecting the unexpected.

A good example of an unsupervised learning approach for finding outliers inside the ensemble was the use of random forests by Baron & Poznanski (2017). Here they have utilised the correlation structure in feature space to identify interesting observations using learned features based on the correlation structure of galaxy spectra. They used random forests to identify interesting features in galaxy spectra, by learning the difference between real and synthetic observations, where the correlation structure was removed from the feature space in the synthetic case. Outliers were then identified based on a measure of similarity between objects, where the authors counted how often every pair of real objects were classified as real in the same leaf of a given tree. This approach demonstrates the use of learned interesting features, but learning not based on previous observations, to perform subsequent outlier detection. A potential drawback of this approach, as with other unsupervised outlier detection methods, is the computational burden of having to compute pairwise comparisons.

This paper presents the application of a coarse-grained complexity measure for identifying the most interesting observations in large scientific datasets. We present the specific ap-
plication of an automated approach, that we consider a novel addition to existing machine intelligence methods, for identifying radio galaxies with complex morphologies in large astronomical surveys.

Aaronson et al. (2014) have demonstrated the effectiveness of a coarse-grained complexity measure, termed apparent complexity, at capturing human intuition and perceptions of complexity. This measure provides a quantitative description of a notion of complexity informally proposed by Gell-Mann (1994) as a phenomena that first increases and then decreases with the rising entropy of a closed system. A potential drawback of apparent complexity as a formal description of complexity is that it relies on assumptions regarding human perceptions. Conversely, this very connection to human perceptions suggests that this measure should be effective at identifying complex observations that are likely to be of interest to a human observer. This paper shows that by learning an appropriate smoothing function, the apparent complexity can be used to partition a sample based on the interestingness of observations.

Unlike supervised learning approaches it does not learn features associated with known interesting observations making the approach a candidate for identifying the unexpected in very large datasets. Furthermore, the use of apparent complexity is fast and computationally efficient, as the method only requires applying a smoothing function and compression algorithm, both of which can be implemented at worst case linear time complexity:

$$T(n) = O(n)$$ (1)

Implementations of constant time $O(1)$ and linear time $O(n)$ median filters are detailed in Perreault & Hebert (2007). Being able to leverage fast and efficient approaches that generalise well is likely to be desirable when identifying interesting and unexpected observations in very large scientific datasets such as those that will be produced by ASKAP and the SKA. Using data from the Australia Telescope Large Area Survey (ATLAS, Norris et al. (2006)) we show that apparent complexity can be used to distinguish between images of galaxies which have been classified as having simple and complex morphologies. We also show that the approach generalises well when applied to new data.

The paper is structured as follows: section 2 frames the theory in terms of Kolmogorov and algorithmic complexity and discusses the theoretical merits of apparent complexity as an attractive candidate for identifying interesting astronomical observations in large datasets. Section 3 outlines empirical methods and results that show apparent complexity can be used to distinguish between images of galaxies with simple and complex morphologies. It will also show that a smoothing function calibrated on a small labelled sample with few interesting observations is able to generalise well when applied to a much larger sample containing a larger collection of complex morphologies. Finally in section 4 we summarise our conclusions.

2. THEORY

The notion of interesting appears somewhat subjective, since what is interesting to some may not be interesting to others. To quantify the “interestingness” of an observation requires defining the context in which the observation is made.

In this paper we measure interestingness based on the complexity $C$ of some observation $x$, subject to a function $f$ that extracts only the non-incidental information from measurements. As such, the measured complexity will depend on the function $f$, that can be calibrated to align with the interests and perceptions of the scientific observer, and so should be considered as $C(f(x))$.

In section 2.1 we introduce the concept of apparent complexity as defined by Aaronson et al.
(2014), which motivates this work and achieves our goals as a definition of interestingness. In section 2.2 theory is presented on the description and behaviour of comprehensible and random information used later to inform experimental methods.

2.1. Apparent Complexity

Apparent complexity has been defined by Aaronson et al. (2014) as the entropy $H$ of an object $x$ after applying a smoothing function $f$, $H(f(x))$. The Shannon entropy of a probability distribution $P$ can be defined as the expected number of random bits that are required to produce a sample from that distribution:

$$H(P) = - \sum_{x \in X} P(x) \log P(x). \quad (2)$$

By Shannon’s Noiseless Coding Theorem the minimum average description length $L$ of a sample is close to the Shannon entropy:

$$H(P) \leq L \leq H(P) + 1. \quad (3)$$

The Kolmogorov complexity $K(f(x))$ can be used as a proxy for the entropy of the smoothed function $H(f(x))$, as proposed by Aaronson et al. (2014). The seeming analogy between the concept of entropy and program size has been previously recognised (Chaitin 1975). The Kolmogorov complexity, or prefix complexity, of $x$ is the length of the shortest binary program $l(p)$, for the reference universal prefix Turing machine $U$, that outputs $x$; it is denoted as $K(x)$:

$$K(x) = \min_p \{l(p) : U(p) = x\}. \quad (4)$$

A thorough treatment is provided by Li & Vitanyi (2008). The Kolmogorov complexity has the advantage of being well-defined for a particular description of a system such as an image of a galaxy. This is not the case for the Shannon entropy which is defined in terms of the possible states of the system. While the Kolmogorov complexity is uncomputable, its upper bound can be reasonably approximated by the compressed file size $C(f(x))$ using a standard compression program (Aaronson et al. 2014), such as gzip.

The issue with using the approximated Kolmogorov complexity directly as measure of complexity is that it is maximized by random information. Intuitively a complexity measure should provide low values for random data that does not contain structure that is of interest to the observer (Zenil et al. 2012). Aaronson et al. (2014) have shown that the apparent complexity measure is able to achieve this by applying a smoothing function $f$ to the input $x$.

While the Kolmogorov complexity of a random sequence is large, the apparent complexity of the same sequence becomes small with smoothing, as fluctuations are removed where the average or median information content becomes homogeneous at the coarse-grained resolution. Accordingly, we define the apparent complexity as the compressed description of regularities and structure after discarding all that is incidental. The apparent complexity will be small for both simple and random sequences.

The apparent complexity measure does not rely on existing data (i.e. training sets) to learn the features of interesting observations. The approach is invariant under rigid motions and makes only explicit assumptions regarding the choice of coarse-graining level and the scale of the image. Previous data is therefore used only to calibrate the coarse-graining level.

Apparent complexity runs into obstacles as a well-defined measure of complexity. Firstly, the uncomputability of the Kolmogorov complexity prohibits the concept from being defined in terms of an optimal compression. It has been proven by Chaitin et al. (1995) that there can be no procedure for finding all theorems that would allow for further compression. Furthermore the problem of distinguishing between meaningful
structure and incidental information, especially in finite data, may fail to be well-defined. Different smoothing functions and different coarse-graining levels will retain different distinct regularities in the data.

These theoretical challenges in objectively defining the apparent complexity can be circumvented when the approach is applied to the segmentation of observations by complexity. Here the apparent complexity can be calibrated to coincide with notions of complexity adopted by the observer.

2.2. Comprehensibility

The objective of applying the smoothing function \( f \) when deriving \( C(f(x)) \) is to remove incidental or random information that is incomprehensible to the observer even though it may have a physical basis. Comprehensibility here is defined with respect to the observer of information, in this case scientists with specific interests. Comprehensible information has a structure within feature space, which in the case of images refers to the spatial distribution of bits of information across available channels.

Li & Vitanyi (2008) define the algorithmic entropy of a system as a combination of the entropy constrained by macroscopic parameters, \( H_x \) and the prefix complexity \( K(x) \):

\[
S_A = K(x) + H_x,
\]

where \( H_x \) describes the uncertainty of the microstates of the system and \( K(x) \) is the description of the system resulting from measurement.

In the case where the system under observation is comprehensible, and hence appears not entirely random, the observation of the system reduces the algorithmic entropy of the system as measurements are taken to inform \( K(x) \). This compression occurs through the macroscopic description of the system. The coarser macroscopic description, accessible to an observer, reduces uncertainty regarding the systems microstates. As measurements are taken, the increase in \( K(x) \) will be slower than the reduction in \( H_x \). This effectively makes scientific description a form of data compression when the system under observation is at least partially non-random.

In the other case, when the system appears random, the description of the system is merely a description of the uncertainty of its microstates, and measurements simply replace \( H_x \) with \( K(x) \). Figure 1 represents the difference between these two scenarios.

**Figure 1.** In the above illustration \( K(x) \) is the prefix complexity of \( x \), \( H_x \) represents ignorance about the micro state of the system constrained by the macroscopic parameters \( x \) and the sum of both terms is defined as the algorithmic entropy. The top row is a reproduction from Li & Vitanyi (2008) depicting the change in algorithmic entropy with measurement for a system in a random micro state (left) and nonrandom micro state (right). The bottom row depicts the application of progressively coarser smoothing functions in terms of the measured complexity, illustrating the equivalency of this process to the erasure of measurements, for a system in a random micro state (left) and nonrandom micro state (right).
In the terms of algorithmic entropy, the application of coarser smoothing functions effectively remove measurements describing the system as illustrated in figure 1. A system that is more comprehensible will accordingly provide greater compression in its description, and so $K(x)$ will reduce more gradually as a smoothing function is applied due to the preservation of information as part of its macroscopic structure. Conversely, a system that contains more random information will provide worse compression in its description and will show a rapid initial reduction in information as a smoothing function is applied due to the detail contained at finer measurement resolutions and the lack of macroscopic structure.

3. ANALYSIS OF RADIO CONTINUUM DATA

Segmentation based on apparent complexity can be used to identify complex images or complex regions within an image. We demonstrate the application this approach using radio continuum images from the Australia Telescope Large Area Survey (ATLAS) survey, to distinguish between simple and complex radio sources. Here we define “simple” sources as single unresolved components, and “complex” sources as anything else, including bent-tail galaxies and extended radio sources (e.g. Fanaroff-Riley I, Fanaroff-Riley II) containing bright radio components in combination with diffuse plume-like jets. Figure 2 provides examples of complex radio sources and figure 3 provides examples of simple unresolved sources.

It would be expected that two simple radio sources (unresolved sources), representing spatially separated galaxies that are randomly associated, will contain a significant amount of shared information in their morphologies, due to the similarity of the basic components. Conversely a true complex source is likely to contain components of a differing nature, such as radio lobes of differing luminosity and jets and plumes with luminosity gradients. By the unique decompression property distinct components will require additional bits of information in the combined compressed description (Cilibrasi & Vitanyi 2005). Accordingly the apparent complexity of a complex radio source should be larger than the apparent complexity of two simple sources in proximity.

The measure of apparent complexity can be used as a proxy for how interesting a radio image of a galaxy is, on the basis that radio sources with a larger apparent complexity contain a larger number of distinct and meaningful components that are likely to be of interest to an observer. We first demonstrate this intuition by distinguishing between artificially generated doubles and true complex sources, before applying the approach to segment the ATLAS data.

3.1. Method

We approximate the apparent complexity measure by applying a median filter $f$ with a window size of 10 pixels to an image $x$, and then calculating the gzip (Levine 2012) file size $C(f(x))$ as an upper bound on the Kolmogorov complexity $K(f(x))$. We apply this approach
to the 256 by 256 pixel radio continuum images as follows:

1. Load a centred image as a 256 by 256 matrix of 8 bit channel pixel intensity values
2. Crop from the centre of the image to create a 64 by 64 matrix
3. Filter the matrix using a percentile based threshold (P90) for pixel intensity values
4. Apply a median filter using a learned window size \( h = 10 \) to produce a smoothed 64 x 64 matrix (to remove random information and retain structural information)
5. Compress the smoothed array using gzip
6. Measure the compressed image size to estimate an upper bound of the Kolmogorov complexity

We adopt a median filter with window size \( h \), based on structuring element \( \Pi^h_k \) centred at location \( k \), where the the coarse-scale version of an element \( x_i \) at location \( k \) is defined:

\[
x^h_{i=k} = \text{median}\{x_i : x_i \in \Pi^h_k\}.
\]

The choice of window size is the only free parameter in this method, and is learned from some small training set, as described in section 3.4.

3.2. Validation using synthetic data

Synthetic radio images were constructed by placing simple radio sources in close proximity, comparable to the distances between components associated with real complex sources with two components (i.e. real doubles). In accordance with the unique decompression property the distinct components of true complex sources should produce larger apparent complexity values. The apparent complexity was calculated for images containing only simple radio sources, synthetic sources combining two simple sources and real complex radio sources as shown in figure 4. The apparent complexity values distinguish between the true complex sources and synthetic sources consisting of two simple sources in close proximity, with true complex sources producing larger complexity values.

![Image](image_url)

**Figure 4.** The apparent complexity (in bytes) calculated for images containing only simple radio sources, synthetic sources combining two simple sources and real complex radio sources. These results show that the apparent complexity values can be used to distinguish between the true complex sources and simple sources in close proximity.

3.3. Survey sample

ATLAS data consists of deep radio continuum imaging of the the Chandra Deep Field South
Table 1. ATLAS DR1 and DR3 samples (CDFS) and the European Large Area ISO Survey (ELAIS). The data is described in Norris et al. (2006) & Middelberg et al. (2008) (DR1), and Franzen et al. (2015) (DR3).

Table 1 shows the number of distinct identified radio sources from within each field and data release. The table also provides a breakdown between sources that have been classified by human inspection from Norris et al. (2006) and Norris et al (direct communication) as having simple and complex morphology.

The sources were provided as 256x256 pixel Portable Network Graphics (PNG) files. The images provided were pre-processed as detailed in Norris et al. (2006); Franzen et al. (2015).

Labels were provided for ATLAS data release 1 (DR1) files identifying which sources had been classified as simple and complex. The data was then used to learn the smoothing function window size using images of galaxies that were manually labeled as having complex and simple morphologies. The window size of the smoothing function was chosen as to maximise the difference between the average apparent complexity of observations labelled complex and simple in the ATLAS DR1 sample.

As shown in table 1, the ATLAS DR3 data provides a much larger sample containing more complex sources. The analysis for data release 3 (ATLAS DR3), was conducted ‘blind’, where labels were not provided with the source files. The success of the approach could therefore be judged independently using the DR3 data.

3.4. Smoothing function calibration

The appropriate window size for a smoothing function can be learned from a training set by maximising the difference between the apparent complexity of observations expertly labelled as complex and simple. As suggested by Aaronson et al. (2014) there may also be natural choices for selecting the smoothing function suggested by our physical ability to actually observe systems and our knowledge of the systems properties.

Applying an appropriate smoothing function to the radio images appears to remove random information and retain information contained within a macroscopic structure comprised of distinct regions with different mean or median pixel values. The isolation of structure in a complex source is shown through the progressive application of coarser median filters in figure 5. In accordance with the theory introduced in section 2.2, images containing more random information show an initial rapid reduction in apparent complexity. In this sense, the rate of change of the apparent complexity with the increasing window size of the smoothing function, appears to act as a proxy for the comprehensibility of the information content.

Figure 5. ATLAS DR1 sources with progressive smoothing function window sizes and complexity scores in bytes
Figure 6 compares the average apparent complexity of complex and simple ATLAS DR1 images across changes in the smoothing function window size. As the smoothing function window size increases, the apparent complexity of the simple and complex sources decreases, since information is being removed, but importantly they decrease at different rates. The rate of change in these curves resemble respectively the rate of change of $K(x)$ for random and structured (comprehensible) microstates as depicted in figure 1. The near exponential shape of the curve representing simple sources suggests that the measured complexity of these images consists of random information at the baseline pre-processing level. Conversely, the changing apparent complexity of the images labelled as complex, that reverts to a closer to linear rate of decrease, suggests a greater content of coarser, more comprehensible, information. Where the apparent complexity curves of both complex and simple sources become flatter and converge, the smoothing function window size is large enough to remove both random and comprehensible content.

![Figure 6. A comparison of the average apparent complexity between simple and complex ATLAS DR1 sources across changes in the size of the smoothing function window.](image)

This analysis shows that there is an appropriate smoothing function window size that provides a clear separation, on average, between images that have been expertly classified as complex and simple sources in the ATLAS DR1 data. The separation between the average apparent complexity values of simple and complex sources becomes clear between a window size of 5 to 25 pixels, with the largest separation achieved toward the centre of this range. We fixed our choice to a diameter of 10 pixels to use when applying this function to the ATLAS DR3 data. The separation disappears at larger coarse-graining levels where the smoothing filter blacks out the image, removing all content. At null and low smoothing levels the images represent the baseline pre-processing level of the ATLAS images.

It is interesting to note that at the baseline pre-processing level, the average apparent complexity of images representing simple sources exceeds the average apparent complexity of images representing complex sources. The rapid reduction in complexity of these images with increased smoothing suggest that this higher initial apparent complexity is likely to be attributable to a higher content of random information. This is in contrast to the truly complex images that retain a greater percentage of the measured complexity at equivalent smoothing levels.

3.5. Classification Results

We computed the apparent complexity of the ATLAS DR3 images, using the method described above. The resulting distribution of values is shown in figure 7, showing that the apparent complexity values for expertly classified simple and complex sources slightly overlap but with complex sources concentrated in the heavier right tails of the distributions.

Without reference to the expertly classified labels for the images, we adopted a boundary of the approximate 90th percentile of the apparent complexity distribution (of the combined sample) segmenting the heavy tails that
we interpreted as being influenced by a second overlapping distributions assumed to represent complex sources. In this way classification was achieved by using an apparent complexity threshold of 300 bytes, equivalent to selecting approximately the top 10% of complexity values.

Binary classification results in comparison to the true nature (expert classification) using this scheme are shown in table 2. These results show that the apparent complexity measure, at the selected partition boundary, can be used to correctly identify 86% (i.e. a recall of 0.86) of the interesting observations from the combined DR3 samples with a 91% reduction in the non-interesting data volume following classification.

We also partitioned complexity scores by applying k-means followed by spectral clustering. We used a nearest neighbours affinity matrix to perform spectral clustering. Both methods employed a Euclidean distance metric. Binary classification results using this approach are shown in table 3. Results based on this approach correctly identify 96% (i.e. a recall of 0.96) of the interesting observations from the combined DR3 samples with a 82% reduction in the non-interesting data volume following classification.

Figure 7. Distributions of the apparent complexity of the total sample (black solid line), subdivided into simple (blue) and complex (red) radio sources (human classified) in the CDFS (left) and ELAIS (right) fields from ATLAS DR3. The black dashed vertical line gives the 300 byte boundary we assume in order to partition the two populations by apparent complexity.

|        | CDFS (n=2861) | ELAIS (n=1964) |
|--------|---------------|----------------|
|        | Complex obs   | Simple obs     |
| Prediction: Complex | 81 | 244 |
| Prediction: Simple | 16 | 2520 |
|        | Complex obs   | Simple obs     |
| Prediction: Complex | 65 | 167 |
| Prediction: Simple | 7 | 1725 |

Table 2. Confusion matrix (P90 complexity cut)

|        | CDFS (n=2861) | ELAIS (n=1964) |
|--------|---------------|----------------|
|        | Complex obs   | Simple obs     |
| Prediction: Complex | 94 | 558 |
| Prediction: Simple | 3 | 2206 |
|        | Complex obs   | Simple obs     |
| Prediction: Complex | 69 | 463 |
| Prediction: Simple | 3 | 1429 |

Table 3. Confusion matrix (Complexity partition based on K-means and Spectral clustering )

Type II errors, representing the incorrect classification of complex sources as simple, may be due to the removal of meaningful information by the smoothing function or potentially by the sparse representation of complex features, discernible to a human observer, but having little impact on the information content of the image. Classification errors may also be attributable to the allocation of the partition boundary.

Type I errors, representing the incorrect identification of simple sources as complex, may be due to the presence of non-random information deemed by a human observer to be incidental and not contributing to the complexity of the source itself. An example could be a telescope imaging artifact containing structure, such as a point spread function originating from a brighter source. Examples of telescope imaging artifacts include the diagonal lines shown at baseline in figure 5.
Alternatively, type I errors may be explained by the retention of random information not removed by the smoothing function. Figure 6 shows that there is a large amount of random information in the simple sources at baseline, and this suggests there is a risk that in some images random information will take the form of incidental structure that may not be removed as smoothing increases. Where random information is retained after smoothing the classification is likely to be improved by incorporating thresholds in both the apparent complexity and the signal-to-noise ratio (SNR) as random information is likely to be distributed more uniformly across the available channel intensity values. The SNR can be calculated as the reciprocal of the coefficient of variation for the channel intensity values.

Figure 8 shows the ATLAS DR3 samples partitioned using both the apparent complexity and the SNR. In this figure the true complex radio sources are clustered at larger apparent complexity and SNR values, towards to the top right edge of the scatter plot. This figure shows the effectiveness of the approach at segmenting simple and complex sources. Based on these distributions, approximately 10% of observations can be classified as interesting, by selecting sources from within the identified regions. Binary classification using this scheme is presented in table 4.

To allocate a partition boundary using both complexity and SNR values we fitted a two component Gaussian mixture model (GMM) using the Expectation Maximisation algorithm. Binary classification results using this approach are shown in table 5.

In assessing classification results, the most important factor is the ability of the classifier to reduce the type II error rate, thereby identifying as many of the interesting observations as possible, while providing a significant reduction in the volume of non-interesting data flagged for further investigation, equivalent to minimising the type I error rate. Given a significant reduction in the total data volume and a very low type II error rate, the contamination of the segregated sample as measured by precision was not deemed to be of primary concern. For this reason the Informedness measure, as described in Appendix A, was chosen to assess performance. The Informedness measure incorporates both Type I errors (False Positives) and Type II errors (False Negatives) and describes the improved performance of the measured classifier with respect to chance, costing true positives and false positives in a way analogous to how a bookmaker fairly prices the odds (Powers 2011). A detailed description of these metrics are provided in Appendix A.

Based on the results shown in table 5 the CDFS sample provides a recall of 0.86 and informedness of 0.81 while the ELAIS sample provides a recall of 0.90 and informedness of 0.84. By reducing the likelihood thresholds used for

![Figure 8. Scatter plot demonstrating the effectiveness of apparent complexity and SNR to partition simple (blue) and complex (red) radio sources in the CDFS (filled circle) and ELAIS (empty circle) fields from ATLAS DR3.](image-url)
classification the recall can be further improved at the expense of the false positive rate.

Based on the results shown in table 4 the CDFS sample provides a recall of 0.96 and informedness of 0.90 while the ELAIS sample provides a recall of 0.89 and informedness of 0.83. These results show that the apparent complexity measure when combined with the SNR, and the selected partition boundary, segments 93% of the interesting observations across both DR3 samples and achieves a 94% reduction in the non-interesting data volume carried forward for further investigation. These results also show that the smoothing function window size learned from the smaller DR1 sample is able to generalise well when applied to the larger DR3 samples containing a larger collection of complex morphologies.

To demonstrate the importance of applying an appropriate smoothing function, an experiment was run to partition the data without smoothing. The images were classified by selecting approximately the top 10% of observations based on apparent complexity score as was also done using the smoothed images. The drastic reduction in performance shown by comparing the results in table 2 and 6, demonstrates the importance of the smoothing function when partitioning interesting observations. These results suggest that the learned smoothing function is successfully able to isolate comprehensible content associated with the meaningful structural information used by astronomers to manually classify the radio sources.

4. CONCLUSIONS

Using ATLAS data we have shown that apparent complexity combined with the signal to noise ratio is able to partition images of galaxies into those with simple and complex morphologies. Using the distribution of these values with reference to expert classification we are able to partition complex sources with a recall of 0.93 and informedness of 0.87 across both CDFS and ELAIS samples. Partitioning the data by fitting Gaussian Mixture Models using the Expectation Maximisation algorithm we are able to implement an automated process to classify complex sources with a recall of 0.90 and informedness of 0.82 across both CDFS and ELAIS samples. Results also show that over 96% of expertly classified complex sources were contained

|                | Complex obs | Simple obs |
|----------------|-------------|------------|
| CDFS (n=2861)  | 93          | 163        |
| Prediction: Complex |            |            |
| Prediction: Simple    |            |            |
| ELAIS (n=1964) | 64          | 118        |
| Prediction: Complex |            |            |
| Prediction: Simple    |            |            |

Table 4. Confusion matrix (complexity & SNR boundary)

|                | Complex obs | Simple obs |
|----------------|-------------|------------|
| CDFS (n=2861)  | 83          | 140        |
| Prediction: Complex |            |            |
| Prediction: Simple    |            |            |
| ELAIS (n=1964) | 65          | 126        |
| Prediction: Complex |            |            |
| Prediction: Simple    |            |            |

Table 5. Confusion matrix (complexity & SNR GMM partition)

|                | Complex obs | Simple obs |
|----------------|-------------|------------|
| CDFS (n=2861)  | 4           | 321        |
| Prediction: Complex |            |            |
| Prediction: Simple    |            |            |
| ELAIS (n=1964) | 2           | 230        |
| Prediction: Complex |            |            |
| Prediction: Simple    |            |            |

Table 6. Confusion matrix (P90 complexity cut without smoothing)
within the largest 20% of complexity values. We have shown that the apparent complexity with a smoothing function window size learned from the smaller ATLAS DR1 sample (n=720) is able to generalise well when applied to the much larger ATLAS DR3 sample (n=4825) containing a larger collection of complex morphologies. Unlike supervised learning approaches, apparent complexity does not learn features associated with known interesting observations. The approach can be implemented at worst case linear time complexity, making it a candidate for assisting with the identification of new and interesting observations in the very large datasets that will be produced by the next generation of astrophysical instruments. A potential application of the coarse-grain complexity measure is to identify complex sources for further analysis using more computationally expensive or targeted approaches.

While a large number of the most interesting observations are likely to be more complex than less interesting observations, there may also be interesting outliers in scientific datasets that are relatively simple. Further work to understand the intersection of the sets of outlying and complex observations will clarify the potential limitations and best applications of a complexity based approach.

The use of apparent complexity to segment images and identify interesting observations should generalise well across new and varied samples including different types of observational data beyond radio images. Further testing with large radio samples and new types of data will be needed to test this hypothesis.

ACKNOWLEDGEMENTS
The authors would like to thank Geoffrey McLachlan, Lawrence Park, Evan Crawford and Michael Drinkwater for some helpful discussions and early contributions.

APPENDIX

A. EVALUATION MEASURES
For a binary classification problem a 2x2 contingency table can be constructed to represent counts of False Positives (FP), True Positives (TP), False Negatives (FN) and True Negatives (TN) as depicted by table 7. Such a table is referred to as a Confusion Matrix and depicts both the counts of Type I errors (False Positives) and Type II errors (False Negatives).

| Prediction + | Class + | Class - | Total |
|--------------|---------|---------|-------|
| TP           |         |         |       |
| FP (Type I error) |         |         | Predicted Positives (PP) |
| FN (Type II error) |         |         | Predicted Negatives (PN) |
| Total        | Real Positives (RP) | Real Negatives (RN) |       |

Table 7. Confusion Matrix for binary classification problem

Quantities can be derived using the information contained within a binary Confusion Matrix to measure the performance of a classifier. Two such quantities are precision

\[
\text{Precision} = \frac{TP}{TP + FP},
\]

and recall

\[
\text{Recall} = \frac{TP}{TP + FN}.
\]
Precision determines the number of correct positive classifications as a fraction of positive classifications, while recall determines the number of correct positive classifications as a fraction of the total number of actual positives (so the fraction of positive objects that have been missed would be $1 - \text{recall}$, in the binary classification case).

An alternative framework for measuring performance involves the use of Receiver Operating Characteristic (ROC) curves. The use of ROC curves to construct a comparative framework has been adopted in the machine learning literature (Fünkranz & Flach 2005). These approaches account for chance level performance and can also be used to account for the cost weightings of negative and positive cases. ROC analysis examines the false positive rate ($FP/RN$) versus the true positive rate ($TP/RP$) and presents equivalent information to ratios calculated in the vertical direction of the Confusion Matrix presented in table 7.

The maximum distance of the receiver operating characteristic (ROC) curve from the 45 degree chance line is known as the Youden’s J statistic (Youden 1950) or as the Informedness measure (Powers 2011). The Informedness measure is equivalent to the subtraction of the false positive rate ($FPR$) from the true positive rate ($TPR$) as follows:

$$\text{Informedness} = \frac{TP}{TP + FN} - \frac{FP}{TN + FP} = TPR - FPR$$

(A3)

This measure is also equivalent to a chance adjusted version of Recall:

$$\text{Informedness} = \text{recall} - FPR$$

(A4)

Powers (2011) shows that Informedness is an unbiased estimator of above chance performance. The measure incorporates both Type I errors (False Positives) and Type II errors (False Negatives) and describes the improved performance of the measured classifier with respect to chance, costing true positives and false positives in a way analogous to how a bookmaker fairly prices the odds (Powers 2011). For this reason the measure is also referred to as Bookmaker Informedness. The Informedness measure is defined on a (-1,1) interval and gives equal weighting to the true positive and false positive rate.

Informedness appears appropriate for evaluating the effectiveness of alternative approaches at detecting and classifying interesting observations in large astronomical data. The Informedness measure relates to the following objectives of classification:

1. **Maximise true positive rate** (i.e. minimise the type II error rate) - providing assurance that actual interesting observations are available for analysis.

2. **Minimise false positive rate** (i.e. minimise the type I error rate) – to minimise the data burden and place minimal unnecessary burden on data transmission and storage infrastructure.

Removing false positives reduces storage and data handling requirements and the associated costs. Retaining a smaller subset of observations that are likely to be interesting allows these to be directed to low latency storage options where they can be easily retrieved.

Due to the likely small number of actual interesting observations compared to normal observations the metric is likely to be more sensitive to small changes in the true positive count and less sensitive to small changes in the false positive count.
REFERENCES

Aaronson, S., Carroll, S., & Ouellette, L. 2014, http://arxiv.org/abs/1405.6903v1
Aniyan, A. K., & Thorat, K. 2017, ApJS, 230, 20, doi: 10.3847/1538-4365/aa7333
Baron, D., & Poznanski, D. 2017, MNRAS, 465, 4530, doi: 10.1093/mnras/stw3021
Chaitin, G. 1975, J. Assoc. Comput. Mach. 22, 329-340
Chaitin, G., Arslanov, A., & Calude, C. 1995, CDMTCS Research Report Series CDMTCS-008
Charnock, T., & Moss, A. 2017, The Astrophysical Journal Letters, 837, L28, doi: 10.3847/2041-8213/aa603d
Cilibrasi, R., & Vitanyi, P. 2005, IEEE Trans. on Information Theory, 51(4)
Dieleman, S., Willett, K., & Dambre, J. 2015, MNRAS, 450, 1441–1459, doi: 10.1093/mnras/stv632
Franzen, T. M. O., Banfield, J. K., Hales, C. A., et al. 2015, Monthly Notices of the Royal Astronomical Society, 453, 4020, doi: 10.1093/mnras/stv1866
Fürnkranz, J., & Flach, P. 2005, Mach Learn, 58, 39–77
Gell-Mann, M. 1994, The Quark and the Jaguar: Adventures in the Simple and the Complex.
Huertas-Company, M., Gravet, R., Cabrera-Vives, G., et al. 2015, The Astrophysical Journal Supplement Series, 221, doi: 10.3847/2041-8213/aa603d
Jarvis, M., Bacon, D., Blake, C., et al. 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 18, https://arxiv.org/abs/1501.03825
Karpenka, N. V., Feroz, F., & Hobson, M. P. 2013, MNRAS, 429, 1278, doi: 10.1093/mnras/sts412
Kessler, R., Bassett, B., Belov, P., et al. 2010, PASP, 122, 1415, doi: 10.1086/657607
Kim, D.-W., & Baier-Jones, C. A. L. 2016, A&A, 587, A18, doi: 10.1051/0004-6361/201527188
Levine, J. 2012, The ‘application/zlib’ and ‘application/gzip’ Media Types, Tech. Rep. RFC 6713, doi: 10.17487/RFC6713
Li, M., & Vitanyi, P. 2008, An Introduction to Kolmogorov Complexity and Its Applications
Middelberg, E., Norris, R. P., Cornwell, T. J., et al. 2008, AJ, 135, 1276, doi: 10.1088/0004-6256/135/4/1276
Norris, R. P. 2017a, PASA, 34, e007, doi: 10.1017/pasa.2016.63
—. 2017b, Nature Astronomy, 1, 671, doi: 10.1038/s41550-017-0233-y
Norris, R. P., Afonso, J., Appleton, P. N., et al. 2006, AJ, 132, 2409, doi: 10.1086/508275
Perreault, S., & Hebert, P. 2007, IEEE Transactions on Image Processing, 16, 2389
Powers, D. M. 2011, Journal International Journal of Machine Learning Technology, 2, 37
Robinson, B. J. 1987, PASA, 7, 220, doi: 10.1017/S1323358000022256
Youden, W. J. 1950, Cancer, 3, 32, doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
Zenil, H., Delahaye, J.-P., & Gaucherel, C. 2012, Complexity, 17, 26, doi: 10.1002/cplx.20388