Mining the LAMOST Spectral Archive

A-Li Luo*[a], Yan-Xia Zhang[a], Jian-Nan Zhang[b] and Yong-Heng Zhao[a]

[a]National Astronomical Observatories, Chinese Academy of Sciences, 100012, Beijing, P.R.China.

[b] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 100080, Beijing, P.R.China.

ABSTRACT

The Large sky Area Multi-Object fibre Spectroscopic Telescope will yield 10 million spectra of a wide variety of objects including QSOs, galaxies and stars. The data archive of one-dimensional spectra, which will be released gradually during the survey, is expected to exceed 1 terabyte in size. This archive will enable astronomers to explore the data interactively through a friendly user interface. Users will be able to access information related to the original observations as well as spectral parameters computed by means of an automated data-reduction pipeline. Data mining tools will enable detailed clustering, characterization and classification analyses. The LAMOST data archive will be made publicly available in the standard data format for Virtual Observatories and in a form that will be fully compatible with future Grid technologies.

Keywords: LAMOST, data mining, clustering, characterization, classification, VO

1. INTRODUCTION

Astronomers and engineers in China will complete a wide-field multi-fibre spectrographic telescope, the Large sky Area Multi-Object fibre Spectroscopic Telescope (LAMOST) in the coming two years. The telescope will be used to survey 1,000,000 QSOs, 10,000,000 galaxies and 1,000,000 stars. The size of the final data archive of one-dimensional spectra will exceed 1 terabyte. Mining* of the LAMOST spectral archive is expected to be useful for a wide variety of astronomical studies.

The LAMOST spectroscopic survey will target over ten million objects chosen from the SDSS photometric survey, DSS-II and other catalogues such as FIRST and ROSAT. Many targets will be selected on the basis of cross-identifications between more than one of the above catalogues. The spectroscopic survey will utilize 32 multi-fibre medium resolution (R=2000) spectrographs, with a total of 4000 optical fibres. The spectral coverage of each spectrum will be from 3700 to 9000 Å.

The LAMOST data archive will be distributed in two main data sets: a spectroscopic catalogue and a set of individual spectra. The former will contain positions, information related to the observations, other measured parameters such as redshifts (or radial velocities), line intensities (or equivalent widths) and positions of identified emission and absorption lines etc. The latter data set will comprise of one-dimensional spectra for one million quasars, ten million galaxies and one million stars. Catalogue subsets may also be included, such as a narrow-line quasar catalogue. The data sets and their expected sizes are listed in Table 1.

The LAMOST telescope will be situated at the National Astronomical Observatories' (NAOC) Xinglong station. Observational data from the telescope will be shipped on tapes to NAOC headquarters in Beijing, where they will be reduced, analyzed and archived automatically by a pipeline.[2] This pipeline will calibrate, process, parameterize, and classify the data, prior to its publication in a public archive. The archive will provide reference data for stars, galaxies and quasars in FITS format, and will also provide a variety of services including a flexible user interface on the web, which will allow sophisticated queries within the database.

*contact: lal@lamost.bao.ac.cn; phone 86 (0)10-6484-1693; fax 86 (0)10-6487-8240

*The concepts and methodologies of "data mining" and "knowledge discovery" in databases are already described in many papers and books,[1] and are summarized in Section 3 of the present paper.

Table 1. LAMOST data sets and their expected sizes

product	records	size
Spectroscopic catalogue:		
Raw observational data	-	40TB
Redshift catalog	10^7	20GB
Radial velocity catalogue	10^6	2GB
Observation log and file headers	10^6	10GB
Simplified catalogue	4×10^8	80GB
Individual spectra:		
1D spectra	10^7	1TB

The database will support two main kinds of query. Firstly, there will be a simple user-friendly search tool enabling one to retrieve data subsets based on search limits chosen by the user. For example, the user will be able to search for all objects lying within a specified field on the sky, by entering positional limits interactively. The interface will also permit interactive "advance" searches as well as interactive "refined" searches of data subsets.

The second kind of query will give the user the option of supplying his/her search criteria in standard SQL[3](a widely used database language). Users will work using views rather than with heavily-indexed base tables. To speed access, indices are helpful to manage those most frequently accessed attributes, and an SQL query will automatically use those indices covering the most important attributes. Aided by different indices, users will be able to retrieve observational information as well as spectral parameters computed automatically.

2. DATA MINING AND EXAMPLES

In addition to the query options mentioned in Section 2, mining tools are also indispensable in order to extract novel information. The LAMOST software system will contain a spectra-based data miner of knowledge, which incorporates data mining functions such as clustering, characterization, and classification.

Data mining (DM) has many alterative names, such as knowledge discovery in databases, knowledge extraction, data archaeology, data dredging, information harvesting, business intelligence etc. In large scientific databases, it can generally be separated into two subsets: event-based mining and relationship-based mining.[4]

In astronomy, event-based mining includes: (1) the use of existing physical models to locate known phenomena of interest either spatially or temporally within a large database; (2) the use of pattern recognition and the clustering properties of data to discover new astrophysical relationships relating to known phenomena; (3) the use of predictive models for the observational parameters of astrophysical phenomena to predict the presence of previously unseen events within large complex databases; and/or (4) the use of thresholds or trends to identify transient or otherwise unique events, thereby revealing new phenomena.

Relationship-based mining on the other hand refers to searching for associations or correlations among a set of items or objects in a database. For example, clustering techniques can be used to identify events that are co-located within a multi-dimensional parameter space.

In this section, we focus on three concept of data mining: clustering, characterization and classification, and give examples of mining algorithms relevant to solving a variety of astrophysical problems.

2.1. Clustering

Clustering divides a database into different groups. The goal of clustering is to find groups of objects that are very different from one other. Unlike classification (see Subsection 3.3), one does not know a priori either which objects one's clusters will include or by which attributes the data will be clustered. Consequently, someone who

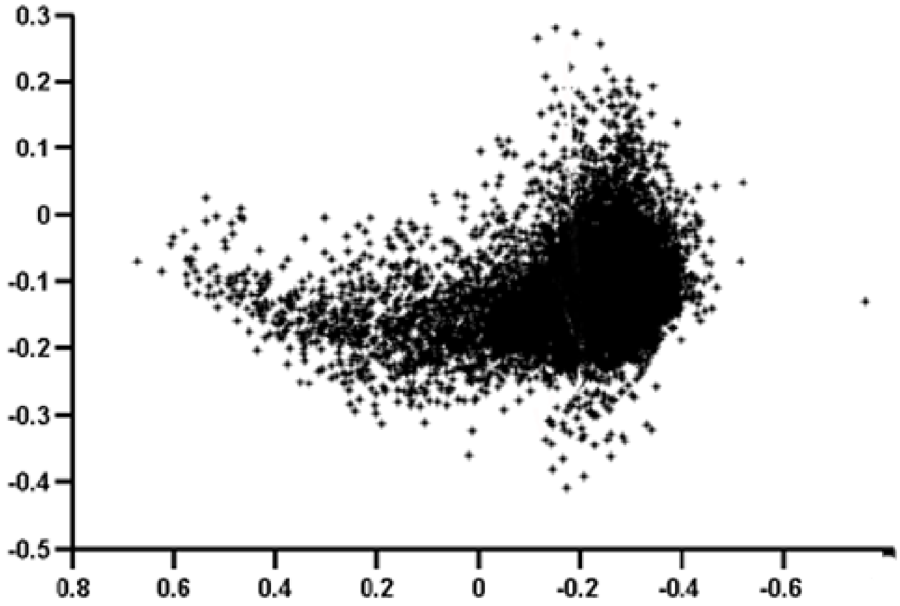

Figure 1. SDSS DR1 QSO spectra for more than 15,000 objects projected onto a 2-d PCA subspace. The x-axis is the first principal component, PC1, while the y-axis is the second principal component, PC2. Each small asterisk in the figure represents a projection of a spectrum. We found that most of the spectra were located within a spherical space. A quick check revealed that most BLQs lie within the spherical space, while most NLQs (which are less numerous) lie outside it. Using a K-mean algorithm, we altered the size of the spherical space in order to achieve an optimal separation between BLQs and NLQs.

is knowledgeable in astronomy is needed in order to interpret the clusters. Often it is necessary to modify the clustering by excluding some of the variables previously employed, because upon examination the user identifies them as irrelevant or not meaningful. After the user finds clusters that segment his/her database meaningfully, these clusters may then be used to classify the new data.

Some of the common algorithms used to perform clustering include[5]:
(1) partitioning-based algorithms, which enumerate various partitions and then score them by some criterion e.g. K-means, K-medoids etc.;
(2) hierarchy-based algorithms, which create a hierarchical decomposition of the set of data (or objects) using some criterion; and
(3) model-based algorithms, in which a model is hypothesized for each of the clusters.

Searching for special objects is one of the tasks of clustering. In Figure 1., we give an example of how a class of special objects known as narrow-line quasars (NLQs) can be identified using a K-mean clustering algorithm. Quasars are active galactic nuclei (AGN) in which two different regions of ionized gas can be distinguished: a broad-line region (BLR) and a narrow-line region (NLR).[6] While NLRs in Seyfert galaxies are already relatively well studied, there are no comparable studies of NLRs in quasars.[7] However, in this example, NLQs can easily be separated from ordinary QSOs in principal component analysis (PCA) space[†]. In most definitions, QSOs are luminous objects, which have broad emission lines superimposed on a non-thermal continuum. The full-width half maxima (FWHM) of their emission lines often exceed 5000 km/s, except that in the cases of NLQs the FWHMs are generally narrower than 1000km/s.

In the LAMOST archive, there will be 10^6 QSO spectra, including large numbers of NLQs amongst them.

[†] Principal Component Analysis (PCA)is widely used in astronomy. The basic goal in PCA is to reduce the dimensions of the multi-parameter space defined by one's data without loss of information.[8] Such a reduction in dimensions has important benefits, especially as projection onto a 2-d or 3-d subspace is often useful for visualizing the data.

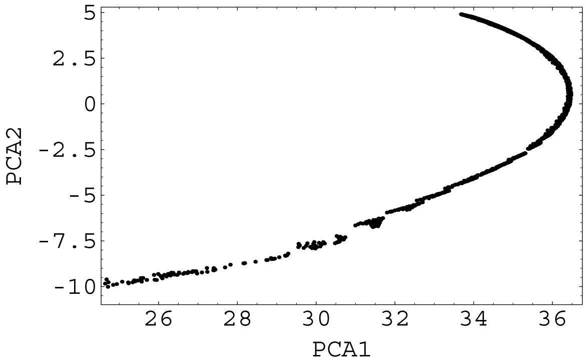

Figure 2. A projection of 1599 stellar spectra onto a 2-d PCA plane.

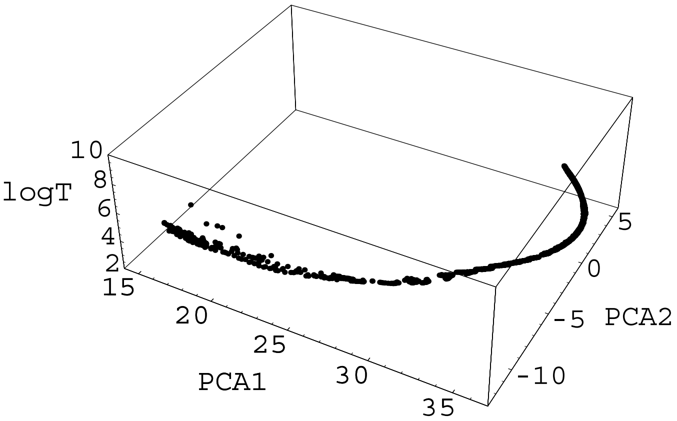

Figure 3. The distribution of (X, $Log_{10}T$) in 3-d space.

However, it will be a simple matter to use an SQL query in order to search for those QSOs with narrow lines, because the width of each identified spectral line will be given in the catalogue. Under the framework of the united AGN model, we will need to compare statistically the spectra of NLQs with those of Seyfert galaxies, and clustering analyses in PCA space is the best tool to do this.

2.2. Characterization

Data characterization is a summarization of general features of objects in target classes, and produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstraction. For example, one may want to characterize the effective temperatures of stars in our archive and obtain the temperature distribution within our Galaxy.

DM methods to estimate stellar parameters are different from traditional methods based on direct measurement. We need not measure each stellar spectrum as we are interested is the temperature distribution. The effective temperature of each star is just one point in such a distribution. Bailer-Jones[9,10] have trained an artificial neural network (ANN) to estimate stellar parameters. Soubiran et al.[11] and Katz et al.[12] on the other hand, have established a template library containing 211 stellar spectra, and used cross-correlation techniques to match their observations with their templates. Here we present a surface-fitting technique to estimate the distribution function of stellar effective temperature.

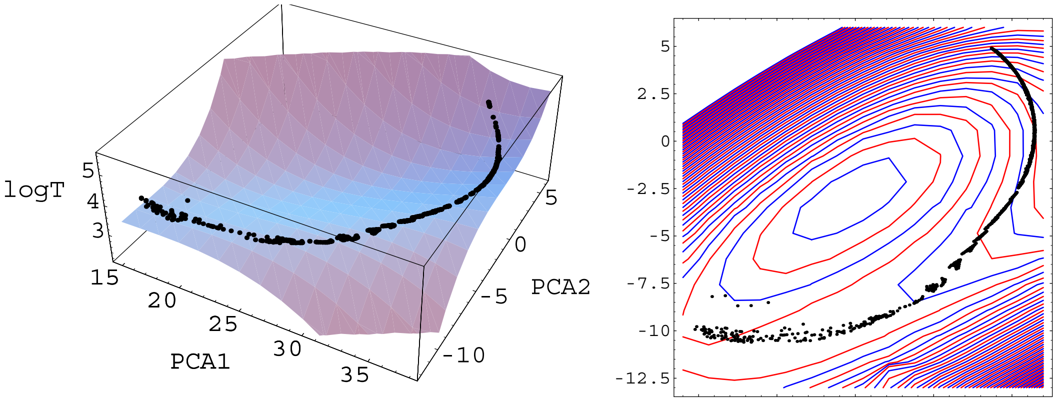

Figure 4. Left: The fitted cubic surface (X , $Log_{10}T$). Right: The isotherm $T = 10^{P(x,y)}$.

The data set we used in this study is a comprehensive library of synthetic stellar spectra from Lejeune et al.,[13] which is based on three original grids of model atmosphere spectra by Kurucz et al.,[14] Fluks et al.,[15] and Bessell et al.[16,].[17] First of all, the spectra in this data set were processed by means of a PCA, yielding Figigure2, in which all 1599 spectra are projected onto a 2-d PCA plane. The data distribution in PCA space is a locus X, and effective temperature T is the function of X: $T = F(X)$. Thus, T is a surface in a 3-d space as shown in Figure 3. By experimentation, we found that the following equations can fit the surface well.

$$T = 10^{P(x,y)} \tag{1}$$

Where P(x,y) is a polynomial of the form:

$$\begin{aligned} P(x,y) = 25.0069 - 1.80461x + 0.0525264x^2 - 0.000450855x^3 + 3.22394y - 0.181638xy \\ +0.00256156x^2y + 0.173964y^2 - 0.00434289xy^2 + 0.00358684y^3. \end{aligned} \tag{2}$$

The surface of effective temperature is shown in Figure 4.(left). Figure 4.(right) gives the isotherm of effective temperature in a PCA plane. When an observational spectrum is projected onto this PCA space, we can judge the effective temperature of the object in question. We are presently working on optimizing characterization algorithms in order to obtain stellar parameters such as T_{eff}, g, and [Fe/H].

2.3. Classification

Classification is also called "predictive data mining", in that the aim is to identify the characteristics of group in advance. This pattern can be used both to understand existing data as well as to predict how new instances will behave.

For the LAMOST data archive, the data analysis pipeline will write the results of its automated spectral classification directly to the spectroscopic catalogue. From the catalogue, users will be able to obtain the classification result e.g. QSO, ordinary galaxy or star of a particular spectral type. For ordinary galaxies, the pipeline will not classify them further, since several very different classification schemes exist. For example, galaxy classifications can depend on strengths of lines, morphology, or some other objective method (e.g. ANN or PCA).

Some groups have classified galaxy spectra according to line strength. For example, Castander et al.[18] classified galaxies from the Coma Cluster into 5 classes depending on the positions of their Balmer breaks as well as their H_α and H_β lines. The classes they defined were absorption-line galaxies, post-starburst galaxies, absorption-line dominated galaxies with emission lines, emission- and absorption-line galaxies, and emission-line dominated galaxies. The MORPHS group (Dressler et al.[19] and Poggianti et al.[20]) classified 10 distant clusters of galaxies into 7 classes depending on [OII] and H_δ. There are other methods based on different lines, such as those of Balogh et al.(1999),[21] Tresse(1999) et al.,[22] and Dessauges-Zavadsky(2000)[23] etc., but space does not permit us to list them all here. The LAMOST archive will include 10^7 galaxies, and there will therefore be plenty of scope for experimenting with different line-based classification methods. We will provide easy access to the catalogue including all identified lines.

Some authors have tried to establish a relationship between morphological type and spectral type, notably Zaritsky et al..[24] It will be difficult for LAMOST to pursue this line of study since the galaxies from LAMOST are very distant and morphological information will not generally be available.

There are many authors trying to classify galaxy spectra based on objective PCA techniques. Castander et al.[18] have used such techniques to classify Coma galaxies into 4 classes using SDSS data; Folkes et al.[25] have classified 2dF spectra into 6 classes; while Bromley et al.[26] have classifed galaxies from the Las Campanas redshift survey into 6 classes. By contrast, Slonim et al.[27] have used the "information bottleneck" technique to classify 2dF galaxies into 5 classes. Note that the SDSS already gives an "e-class" index (which is also from a PCA) for each galaxy, indicating whether the galaxy is likely to be of early or late type. In the LAMOST data set, we will also be able to include an indicator of this type.

More and more authors are realizing that the classification of galaxy spectra is a complex problem, and should be based on evolutionary models. PEGASE (Projet d'Etude des Galaxies par Synthese Evolutive) is such a spectrophotometric evolution model for starbursts and evolved galaxies of the Hubble sequence.[28] The model includes evolutionary tracks, stellar libraries, initial mass functions and star formation rates. When LAMOST galaxy spectra have been obtained, we will be able to classify the data using PEGASE or other evolutionary templates. This should also help to improve the evolutionary models themselves.

3. SUMMARY

An objective of LAMOST DM is to provide software tools that will also be useful for the development of China's Virtual Observatory (VO). All data mining functions will be encapsulated as VO tools, including various mining algorithms. We are designing each function of the software as a form of command line for use in Unix/Linux environments. When international VO standards are decided upon, all of these commands will be easy to encapsulate. All of the software tools will then be available as free packages that can be downloaded and installed on any Unix/Linux system. However, under the framework of Grid architecture, the software tools will not need to be downloaded since they will be be automatically executable on any computer on the grid.

The LAMOST data set, including all its sub-catalogues and FITs files of 1-d spectra, will of course be another important contribution to the VO. Eventually, all of the data will be converted to the chosen standard VO format. Exploring data such as these through VOs is likely to offer many new interesting research directions and will change astronomers research methods. Data mining offers great promise in helping scientists uncover patterns hidden in large data sets. However, data mining tools need to be guided by users who understand the data, and the general nature of the analytical methods involved.

The true relationship between LAMOST and the VO is in using data mining and knowledge discovery to explore the LAMOST data. Building models is only one step in knowledge discovery. It is vital to collect and prepare the data carefully, as well as to check one's models against actual phenomena. The "best" model is often found after building models of several different types, or by trying different technologies or algorithms. Our mission is to choose the right data mining tools with the best basic capabilities, an interface that matches the computer-skill levels of potential users, and features relevant to future directions in astronomical research.

ACKNOWLEDGMENTS

The author would like to thank Prof. C.K. Young (NAOC) for much help in improving the English in this paper.

REFERENCES

1. U. Fayyad, G. Piatesky-Eisenberg, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases: An Overview in Advances in Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park, Calif, 1996.
2. A. L. Luo, "Steps towards a fully automated classification and redshift mesurement pipeline for LAMOST spectra. I. continuum level and wavelength estimation for galaxies," *Chin J. Astron. Astrophys.* **1**, pp. 563–572, 2001.
3. J. Bowman, S. L. Emerson, and M. Darnovsky., *The Practical SQL Handbook: Using SQL Variants, 4th Edition*, Addison Wesley Professional, 2001.
4. J. Han and M. Kamber, *Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems*, Morgan Kaufmann Publishers, 2000.
5. Y. Zhang, Y. Zhao, and C. Cui, "Data Mining and Knowledge Discovery in Database of Astronomy," *Progress in Astronomy* **20**, pp. 312–323, 2002.
6. N. Bennert, H. Falcke, H. Schulz, A. S. Wilson, and B. J. Wills, "Size and Structure of the Narrow-line Region of Quasars," *The Astrophyiscal Journal* **547**, pp. L105–L109, 2002.
7. J. A. Baldwin, R. McMahon, C. Hazard, and R. E. Williams, "QSOs with Narrow Emission Lines," *The Astrophyiscal Journal* **327**, pp. 103–115, 1988.
8. I. Jolliffe, *Principal Component Analysis*, Springer-Verlag, German, 1986.
9. C. Bailer-Jones, "Stellar parameters from very low resolution spectra and medium band filters. $\mathrm{T_{eff}}$, log g and [M/H] using neural networks.," *Astronomy and Astrophysics* **357**, pp. 197–205, 2000.
10. C. A. L. Bailer-Jones, "Determination of stellar parameters with GAIA.," *Astrophysics and Space Science* **280**, pp. 21–29, 2002.
11. R. C. C. Soubiran, D. Katz, "On-line determination of stellar atmospheric parameters $\mathrm{T_{eff}}$, log g, [Fe/H] from ELODIE echelle spectra II. -the library of F5 to K7 stars.," *Astronomy and Astrophysics Supplement* **133**, pp. 221–226, 1998.
12. R. C. D. Katz, C. Soubiran, "On-line determination of stellar atmospheric parameters, $\mathrm{T_{eff}}$, log g, [Fe/H] from ELODIE echelle spectra I. - the Method.," *Astronomy and Astrophysics* **338**, pp. 151–160, 1998.
13. T. Lejeune, F. Cuisinier, and R. Buser, "Standard stellar library for evolutionary synthesis. I. calibration of theoretical spectra," *Astronomy and Astrophysics Supplement series* **125**, pp. 229–246, 1997.
14. R. L. Kurucz, "Model atmospheres for G, F, A, B, and O stars," *Astrophysical Journal Supplement Series* **40**, pp. 1–340, 1979.
15. M. A. Fluks, B. Plez, P. S. The, D. de Winter, B. E. Westerlund, and H. C. Steenman, "On the spectra and photometry of M -giant stars," *Astronomy and Astrophysics Supplement series* **105**, pp. 331–336, 1994.
16. M. S. Bessell, J. M. Brett, P. R. Wood, and M. Scholz, "Colors of extended static model photospheres of M giants," *Astronomy and Astrophysics Supplement series* **77**, pp. 1–30, 1989.
17. M. S. Bessell, J. M. Brett, M. Scholz, and P. R.Wood, "Colors and stratifications of extended static model photospheres of M stars located on the FGB, AGB and supergiant branch," *Astronomy and Astrophysics Supplement series* **89**, pp. 335–336, 1991.
18. F. J. Castander, R. C. Nichol, and A. M. et al., "The First Hour of Extragalactic Data of the Sloan Digital Sky Survey spectroscopic commissioning: The coma cluster," *Astronomical Journal* **121**, pp. 2331–2357, 2001.
19. A. Dressler, I. Smail, and B. M. P. et al., "A Spectroscopic Catalog of 10 Distant Rich Clusters of Galaxies," *The Astrophysical Journal Supplement Series* **122**, pp. 51–80, 1999.
20. B. M. Poggianti, I. Smail, and A. D. et al., "The Star Formation Histories of Galaxies in Distant Clusters," *The Astrophysical Journal* **518**, pp. 576–593, 1999.
21. M. L. Balogh, S. L. Morris, H. K. C. Yee, R. G. Carlberg, and E. Ellingson, "Differential Galaxy Evolution in Cluster and Field Galaxies at z 0.3," *The Astrophysical Journal* **527**, pp. 54–79, 1999.
22. L. Tresse, S. Maddox, J. Loveday, and C. Singleton, "Spectral analysis of the Stromlo-APM survey -I. Spectral properties of galaxies," *The Astrophysical Journal* **310**, pp. 262–280.
23. M. Dessauges-Zavadsky, M. Pindao, A. Maeder, and D. Kunth, "Spectral classification of emission-line galaxies," *Astronomy and Astrophysics* **355**, pp. 89–98.

24. D. Zaritsky, A. I. Zabludoff, and J. A. Willick, “Spectral Classification of Galaxies Along the Hubble Sequence,” *Astronomical Journal* **110**, p. 1620.

25. S. Folkes, S. Ronen, I. Price, and O. L. et al., “The 2dF Galaxy Redshift Survey: spectral types and Luminosity Functions,” *Monthly Notices,The Royal Astronomical Society* **308**, pp. 459–472.

26. B. C. Bromley, W. Press, H. Lin, and R. P. Kirshner, “Spectral Classification and Luminosity Function of Galaxies in the Las Campanas redshift survey,” *The Astrophysical Journal* **505**, pp. 25–36.

27. N. Slonim, R. Somerville, N. Tishby, and O. Lahav, “Objective classification of galaxy spectra using the information bottleneck method,” *Monthly Notices,The Royal Astronomical Society* **323**, pp. 270–284.

28. M. Fioc and B. Rocca-Volmerange, “PEGASE: a UV to NIR spectral evolution model of galaxies. Application to the calibration of bright galaxy counts.,” *Astronomy and Astrophysics* **326**, p. 950.