

**Data Science**

Specially Appointed Associate Professor Berrar Daniel

Research: Data science, machine learning, bioinformatics

Homepage: http://www.berrar.com

● **Research Purpose and Content**

*Data science* is a relatively young discipline that focuses on intelligent data analysis and knowledge extraction from data. Data science is inherently interdisciplinary, drawing on theories and methods from statistics, mathematics, and computer science, particularly machine learning, artificial intelligence, data warehousing, and high performance computing. The scope of data science is similar to that of the closely related fields of data mining and knowledge discovery from databases. Data science is concerned with the analysis of extremely complex, noisy, and heterogeneous data, particularly big data, which cannot be readily analyzed with standard desktop computers anymore. Indeed, data science as an independent discipline is closely linked to the emergence of big data. My research focuses on the development, evaluation, and application of machine learning and statistical methods for the analysis of such data, specifically in the context of the life sciences.

● **Research Themes**

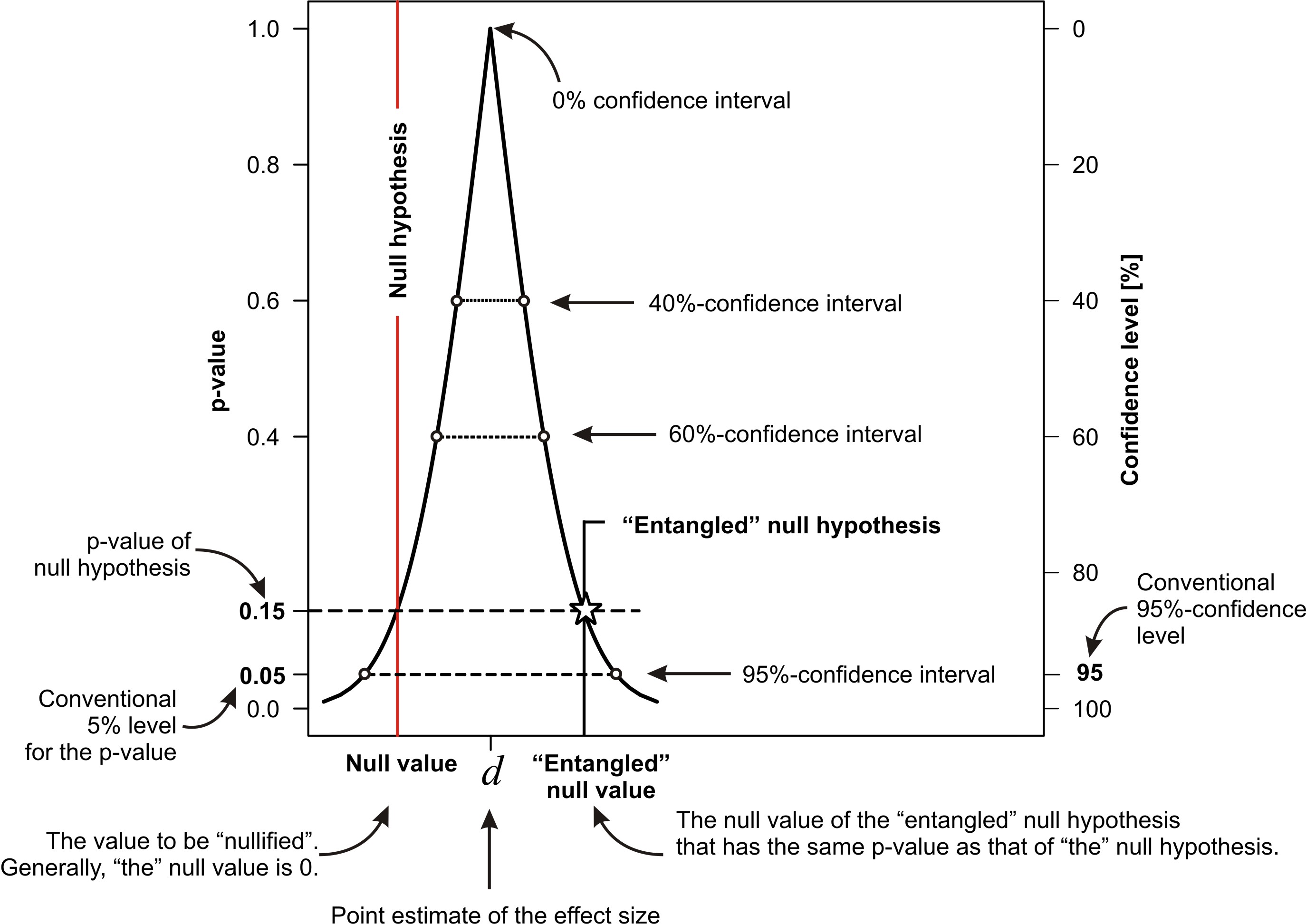
**1. Bioinformatics**

Modern biotechnologies are becoming more and more sophisticated. At the same time, the costs of generating life science data are plummeting, so that even relatively small laboratories can now generate data at a massive scale. Today, the challenge is no longer the generation of such data, but their processing, integration, and intelligent analysis. A key problem is the fact that life science data are often extremely noisy and incomplete. Machine learning and statistical techniques play a pivotal role for the intelligent data analysis in the life sciences [5]. In an ongoing research project with the Parasitology and Epidemiology Group, Aberystwyth University, UK, we are using RNA sequencing and methylated DNA immunoprecipitation to investigate the pathogenic blood fluke *Schistosoma mansoni*. This parasite causes schistosomiasis, a neglected tropical affecting millions of people worldwide. Our research will give new systems-level insights into the complex interplay of the gene regulatory networks and methylation machinery of this parasite and might ultimately lead to the discovery of novel drug targets.

**2. Knowledge extraction from life science data**

　Knowledge extraction refers to the creation of knowledge from structured data (e.g., data organized in databases), unstructured data (e.g., natural language text), and semi-structured data (e.g., unstructured data annotated with meta-data). Knowledge extraction presents considerable challenges in terms of data management, as it is necessary to organize and store data from diverse sources, including pre-processed data and meta-data. In a joint research project with the Children's Memorial Research Center, USA, we used text mining and statistical techniques for knowledge extraction from full-text scientific articles [4]. We we interested in the invasiveness of glioblastoma, which is a type of aggressive brain cancer. To elucidate the effect of sphingosine-1-phosphate (S1P, a signaling sphingolipid) on the malignancy of glioblastoma, we used DNA microarrays to investigate gene expression as a response to S1P administration *in vitro*. We identified a set of candidate genes that were significantly differently expressed as a unique response to S1P. To elucidate the interplay between these genes, we then used text mining and extracted gene-gene interaction patterns from approximately 120,000 full-text scientific articles on cancer research. We developed a data warehouse to integrate the wet lab experimental data with the text mining results. Then, we devised a new algorithm to construct gene-gene interaction networks from the extracted interaction patterns and the gene expression data. Thereby, we could identify a novel pathway linking S1P to the invasive phenotype of glioblastoma.

**3. Machine learning for sports data analysis**

Soccer is a fast-growing multibillion dollar industry that nowadays employs methods from data science to improve the performance of players, equipment, marketing, scouting, etc. In a joint research project with the Sport and Exercise Science Department, University of Evry-Val d’Essonne, France, we created a database of over 200,000 international soccer matches and are currently investigating to which extend the outcome of future soccer matches can be predicted. This project involves data integration and fusion, feature engineering, and predictive modeling.

**4. Performance evaluation of machine learning algorithms**

　Evaluating the performance and robustness of predictive models is a crucial element of machine learning. A major focus of my research is the analysis and development of performance metrics and performance visualization tools, specifically for the statistical comparison of classifiers [1,2].

● **Message to students**

My laboratory is actively collaborating with overseas research institutes. All students with a keen interest in data science (and some programming skills) are welcome to join.

**Figure 3**: Confidence curve [1].

● **Selected publications**

[1] Berrar D. (2016) Confidence curves: an alternative to null hypothesis significance testing for the comparison of classifiers. *Machine Learning* doi:10.1007/s10994-016-5612-6.

[2] Oentaryo R., Lim E.P., Finegold M., Lo D., Zhu F., Phua C., Cheu E.Y., Yap G.E., Sim K., Nguyen M.N., Perera K., Neupane B., Faisal M., Aung Z., Woon W.L., Chen W., Patel D., and Berrar D. (2014) Detecting click fraud in online advertising: a data mining approach. *Journal of Machine Learning Research* 15:99−140.

[3] Berrar D. and Flach P. (2012) Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Briefings in Bioinformatics* 13(1):83−97.

[4] Natarajan J., Berrar D., Dubitzky W., Hack K., Zhang Y., Van Brocklyn J.R., and Bremer E. (2006) Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. *BMC Bioinformatics* 7:373.

[5] Berrar D., Dubitzky W., Granzow M. (eds.) A Practical Approach to Microarray Data Analysis, Springer, Dordrecht/Heidelberg/London, 384 pages, 2002.