RESEARCH SCRIPT IJRCS - International Journal of Research in Computer Science ISSN: 2349-3828

Data Science: Reducing Environmental Complexities

Trivenu Vembuluru
[1](Department of Computer Science, Vemu Institute Of Technology, Chitoor, Andhra Pradesh)

Abstract— *Data Science has emerged as an ambitious new scientific field, related debates and discussions have sought to address why science in general needs data science and what even makes data science a science. However, few such discussions concern the intrinsic complexities and intelligence in data science. As data science focuses on a systematic understanding of complex data and business related problems. The core objective of data science is exploration of the complexities, among these complexities Environmental complexities is an important factor. By using some algorithms this complexity can be reduced.*

Keywords— Data science, Environmental complexities, algorithms

1. INTRODUCTION

The concept of data science was originally proposed within the statistics and mathematics community, where it essentially concerned data analysis. Data science today goes beyond specific areas like data mining and machine learning or whether it is the next generation of statistics.
Definition: Data science is a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, including statistics, informatics, computing, communication, management, and sociology, to study data following "data science thinking"
Consider this discipline-based data science formula
Data science = {statistics ∩ informatics ∩ computing ∩ communication ∩sociology ∩ management | data ∩domain ∩ thinking }
Where "|" means "conditional on."

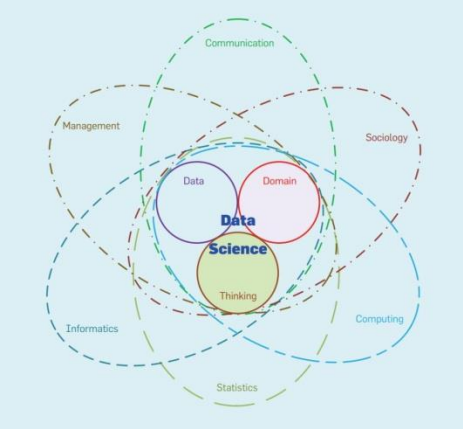

Figure-1: Transdisciplinary Data science

Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from data in various forms,either structured or unstructured, similar to data mining.

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics,statistics, information science, and computer science, in particular from the subdomains ofmachine learning, classification, cluster analysis, data mining, databases, andvisualization.

2. KEY INSIGHTS

Data science problems require systematic thinking, methodologies, and approaches to help spur development of machine intelligence. The conceptual landscape of data science assists data scientists trying to understand, represent, and synthesize the complexities and intelligence in related problems. Data scientists aim to invent data and intelligence-driven machines to represent, learn, simulate, reinforce, and transfer human-like intuition, imagination, curiosity, and creative thinking through human-data interaction and cooperation.

Data objects as independent and identically distributed variables (IID)?"; "What problems not solved well previously are becoming even more complex, as when quantifying complex behavioural data?''; and "What could I not do better before, as in deep analytics and learning?" As data science focuses on a systematic understanding of complex data and related business problems,5,6 I take the view here that data science problems are complex systems3,19 and data science aims to translate data into insight and intelligence for decision making. Accordingly, I focus on the complexities and intelligence hidden in complex data science problems, along with the research issues and methodologies needed to develop data science from a complex-system perspective.

3. CONCEPTUAL LANDSCAPE

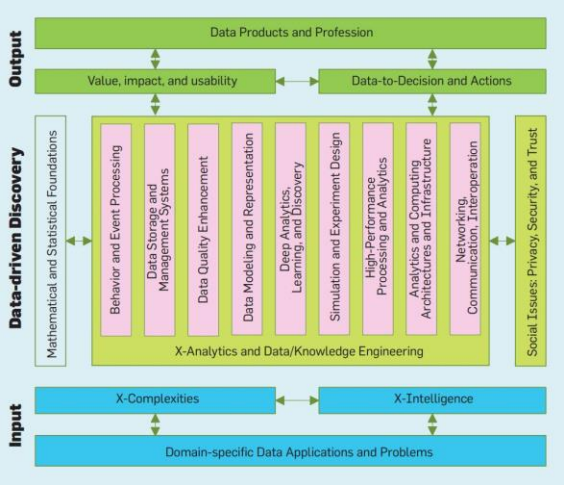

Figure-2:Landscape of Data Science

Data science landscape. The X-complexity and X-intelligence in complex data science systems and widening gap between world invisibility and capability/capacity immaturity yield new research challenges that motivate development of data science as a discipline. Figure 4 outlines the conceptual landscape of data science and its major research challenges by taking an interdisciplinary, complex-system-based, hierarchical view.

The data science landscape consists of three layers: "data input," including domain-specific data applications and systems, and X-complexity and X-intelligence in data and business problems; "data-driven discovery" consisting of discovery tasks and challenges; and "data output" consisting of results and outcomes.

Research challenges and opportunities emerge in all three in terms of five areas not otherwise managed well through non-data-science methodologies, theories, or systems:
Data/business understanding. The aim is for data scientists, as well as data users, to identify, specify, represent, and quantify the X-complexities and X-intelligence that cannot be managed well through existing theories and techniques but nevertheless are embedded in domain-specific data and business problems.

Mathematical and statistical foundation. The aim is to enable data scientists to disclose, describe, represent, and capture complexities and intelligence for deriving actionable insight. Data/knowledge engineering and X-analytics. The aim is to develop domain-specific analytic theories, tools, and systems not available in the relevant body of knowledge to represent, discover, implement, and manage the data, knowledge, and intelligence and support the corresponding data and analytics engineering.

Data quality and social issues. The aim here is to identify, specify, and respect social issues in domain-specific data, business-understanding, and data science processes, including use, privacy, security, and trust, and make possible social issues-based data science tasks not previously handled well.

Data value, impact, utility. The aim is to identify, specify, quantify, and evaluate the value, impact, and utility of domain-specific data that cannot be addressed through existing measurement theories and systems.

Data-to-decision and action-taking challenges. The aim is to develop decision-support theories and systems to enable data-driven decisions and insight-to-decision transformation, incorporating prescriptive actions and strategies that cannot be managed through existing technologies and systems.

X-complexity and X-intelligence pose additional challenges to simulation and experimental design, including how to simulate the complexities, intelligence, working mechanisms, processes, and dynamics in data and corresponding business systems and how to design experiments to explore the effect of business managers' datadriven decisions. Big-data analytics requires high-performance processing and analytics that support large-scale, real-time, online, high-frequency, Internet-based, cross-organizational data processing and analytics while balancing local and global resource objectives. Such an effort may require new distributed, parallel, high-performance infrastructure, batch, array, memory, disk, and cloud-based processing and storage, data-structure and-management systems, and data to-knowledge management.

4. X-COMPLEXITIES IN DATA SCIENCE

Here, complexity refers to sophisticated characteristics in data science systems. I treat data science problems as complex systems involving comprehensive system complexities, or X-complexities, in terms of data (characteristics), behavior, domain, social factors, environment (context), learning (process and system), and deliverables.

Data complexity is reflected in terms of sophisticated data circumstances and characteristics, including large scale, high dimensionality, extreme imbalance, online and real-time interaction and processing, cross-media applications, mixed sources, strong dynamics, high frequency, uncertainty, noise mixed with data, unclear structures, unclear hierarchy, heterogeneous or unclear distribution, strong sparsity, and unclear availability of specific sometimes critical data. An important issue for data scientists involves the complex relations hidden in data that are critical to understanding the hidden forces in data. Complex relations could consist of comprehensive couplings that may not be describable through existing association, correlation, dependence, and causality theories and systems. Such couplings may include explicit and

implicit, structural and nonstructural, semantic and syntactic, hierarchical and vertical, local and global, traditional and nontraditional relations, and evolution and effect.

Data complexities inspire new perspectives that could not have been done or done better before. For example, traditional large surveys of sensor data, including statisticians' questions and survey participants, have been shown to be less effective, as seen in related complications (such as wrongly targeted participants, low overall response rate, and questions unanswered).

Behaviour complexity refers to the challenges involved in understanding what actually takes place in business activities by connecting to the semantics and processes and behavioural subjects and objects in the physical world often ignored or simplified in the data world generated by physical-activityto-data conversion in data-acquisition and -management systems.

Social complexity is embedded in business activity and its related data and is a key part of data and business understanding. It may be embodied in such aspects of business problems as social networking, community emergence, social dynamics, impact evolution, social conventions, social contexts, social cognition, social intelligence, social media, group formation and evolution, group interaction and collaboration, economic and cultural factors, social norms, emotion, sentiment and opinion influence processes, and social issues, including security, privacy, trust, risk, and accountability in social contexts.

5. Environmental Complexities

Environment complexity is another important factor in understanding complex data and business problems, as reflected in environmental (contextual) factors, contexts of problems and data, context dynamics, adaptive engagement of contexts, complex contextual interactions between the business environment and data systems, significant changes in business environment and their effect on data systems, and variations and uncertainty in interactions between business data and thebusiness environment. Such aspects of the system environment have concerned open complex systems20 but not yet data science. If ignored, a model suitable for one domain might produce misleading outcomes in another, as is often seen in recommender systems.

Other requirements for managing and exploiting data include appropriate design of experiments and mechanisms. Inappropriate learning could result in misleading or harmful outcomes, as in a classifier that works for balanced data but could mistakenly classify biased and sparse cases in anomaly detection.

The complexity of a deliverable data product, or "deliverable complexity" becomes an obstruction when actionable insight is the focus of a data science application. Such complexity necessitates identification and evaluation of the outcomes that satisfy technical significance and have high business value from both an objective and a subjective perspective. The related challenges for data scientists also involve designing the appropriate evaluation, presentation, visualization, refinement, and prescription of learning outcomes and deliverables to satisfy diverse business needs, stakeholders, and decision support. In general, data deliverables to business users must beeasy to understand and interpretable by nonprofessionals, revealing insights that directly inform and enable decision making and possibly having a transformative effect on business processes and problem solving.

6. Algorithms for Environmental Complexities

Following are the Eight Data Science algorithms that reduce the Environmental Complexities thatconcentrates on various entities of the external and internal environment.

7. Ordinary Least Squares Algorithm:

To have a best fit regression line : Used for forecasting and quantify marginal effect

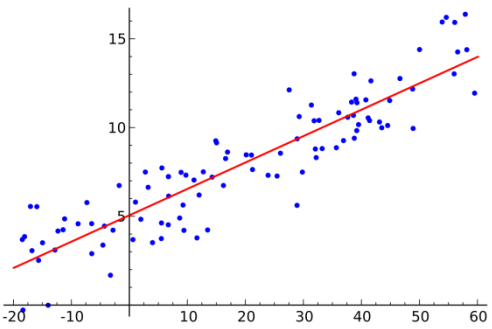

Figure-3: Ordinary Least Squares

Pros:

Move count is decreased by about 4 moves compared to normally doing the last F2L pair, then OLL.

It requires less look ahead if implemented into solves, compared to doing the last F2L pair and OLL. So, although it only saves 4 moves, decreased look ahead can help reduce your solve times.

Increased chance of a last layer skip.

Cons

There are a total of at least 864 algorithms, including mirrors.

Because of the first point, this means that if the solver were to learn full VLS and HLS, it would take over a year to learn if 2 algorithms were learned per day

Decision Tree Algorithm:

Binary classifier to choose two from two gives decision that uses features

Useful for classification and segmentation

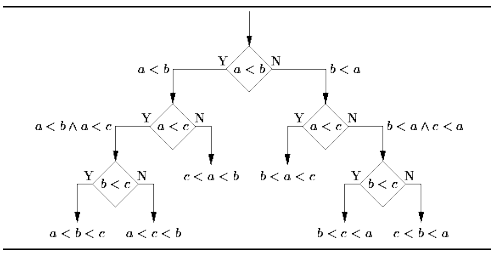

Figure-4: Decision Tree

Entropy:
Entropy, also called as Shannon Entropy is denoted by H(S) for a finite set S, is the measure of the amount of uncertainty or randomness in data.

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

Information Gain:
Information gain is also called as Kullback-Leibler divergence denoted by IG(S,A) for a set S is the effective change in entropy after deciding on a particular attribute A. It measures the relative change in entropy with respect to the independent variables.

$$IG(S, A) = H(S) - H(S, A)$$

Alternatively,

$$IG(S, A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

Naive Bayesian Classification Algorithm:
Uses Bayesian theorem for data classification
Uses strong independent assumption between features
Used for text mining, Email spam classification-popular in text classification

Using Bayes theorem, the conditional probability can be decomposed as:
P(Cₖ|X)=p(Cₖ)p(x|Cₖ)/p(X)

P(*c*|*x*) is the posterior probability of *class*
P(*c*) is the prior probability of *class*.
P(*x*|*c*) is the likelihood which is the probability of *predictor* given *class*.
P(*x*) is the prior probability of *predictor*.

Logistic Regression Algorithm:

Used for modelling binomial target variable
Can be extended to multinomial logistic regression
Simple and Linear in nature
Popular in credit scoring, marketing campaign analysis

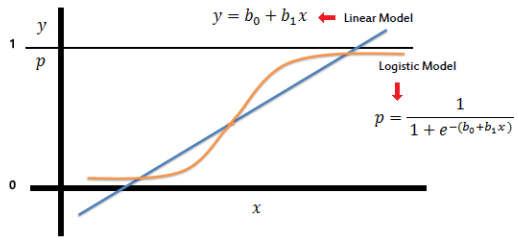

Figure-5: Logistic Regression
In logistic regression, the dependent variable is binary or dichotomous (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, e

$$logit(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

Support Vector Machine Algorithm:

Uses hyperplanes to separate binary classes
Popular classification algorithm that takes in to non linearity in data
Highly scalable

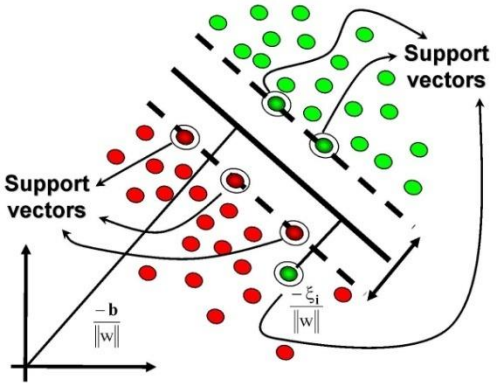

Figure-6: Support Vector Machine
When SVM creating a hyperplane basically there are four things to considers.
x, are the data points
y, are the labels of the classes
w, is the weight vector
b, is the bias

$$w^T x + b = 0$$

Clustering Algorithms:
Group Objects that are similar
Does classification where there is no target variable
Used for dimension reduction and segmentation

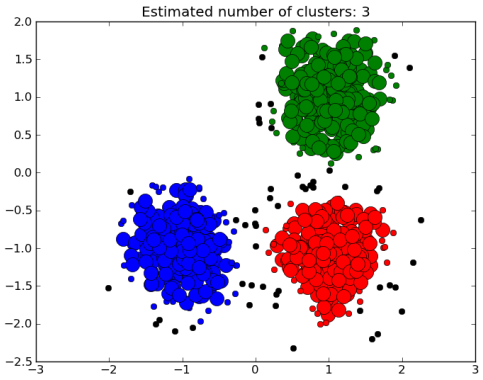

Figure-7: Clustering

Script output for above Clusters:
Estimated number of clusters: 3
Homogeneity: 0.942
Completeness: 0.81
V-measure: 0.874
Adjusted Rand Index: 0.900
Adjusted Mutual Information: 0.815
Silhouette Coefficient: 0.577

Principal Component Analysis Algorithm:

Orthogonal transformation of correlated variables to uncorrelated principal components
Used for dimension reduction
Used along with regression known PC regression

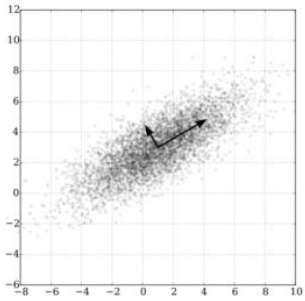

Figure-8: Principal Component Analysis

PCA of a multivariate Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the (0.866, 0.5) direction and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue, and shifted so their tails are at the mean.

Linear Discriminant Analysis Algorithm:

Used for classification of classes: two or more by taking linear combination of features
Compete with LR algorithm
Used in marketing science, pattern recognition in medical data etc...

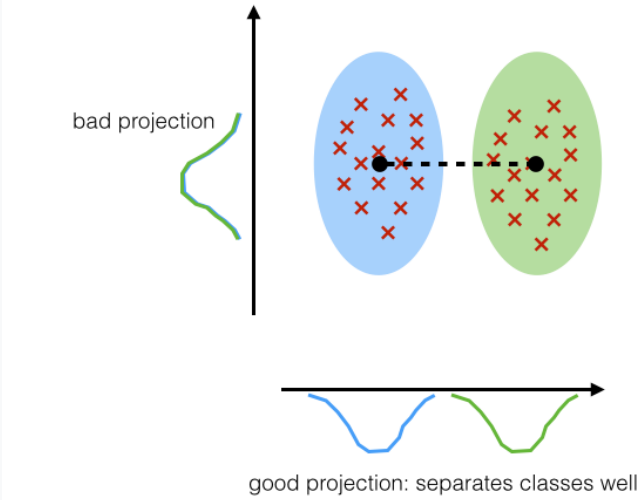

Figure-9: Linear Discriminant Analysis
– In order to find a good projection vector, we need to define a measure of separation
– The mean vector of each class in x-space and y-space is $\mu i = 1/N_i \sum_{x\in\omega i} x$ and $\mu_i = 1/N_i \sum_{y\in\omega i} y = 1/N_i \sum_{x\in\omega i} w^T x = w^T \mu_i$

7. Conclusion

Increasingly manufacturers implement lean practices to improve operational performance. In addition, manufacturers operate in ever more complex and volatile environments. This research investigates the effects of environmental complexity and dynamism on lean operations and lean purchasing practices. It empirically examines these relationships using archival and survey data from 126 manufacturers. The results show that environmental complexity positively moderates the effects of lean operations and lean purchasing on performance. However, environmental dynamism reduces the benefits of lean operations on performance, but enhances the benefits of lean purchasing on performance. Robustness tests further confirm the contingent effects of complexity and dynamism on lean operations and lean purchasing. This research offers a more nuanced understanding of the effect of external environmental context on lean practices, and suggests that practitioners should carefully consider the external environment when implementing different types of lean practices.

References

[1] Jeff Leek (2013-12-12). "The key word in "Data Science" is not Data, it is Science". Simply Statistics.

[2] The Journal of Data Science. (2003, January). Contents of Volume 1, Issue 1, January 2003. Retrieved from http://www.jds-online.com/v1-1

[3] Barlow, Mike (2013). The Culture of Big Data. O'Reilly Media, Inc.

[4] Donoho, David (September 2015). "50 Years of Data Science" (PDF). Based on a talk at Tukey Centennial workshop, Princeton NJ Sept 18 2015.

[5] Data Science Journal. (2012, April). Available Volumes. Retrieved from Japan Science and Technology Information Aggregator, Electronic: http://www.jstage.jst.go.jp/browse/dsj/_vols

[6] K. Karimi and H.J. Hamilton (2011), "Generation and Interpretation of Temporal Decision Rules", International Journal of Computer Information Systems and Industrial Management Applications, Volume

[7] Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4

[8] Brenner, N., Bialek, W., & de Ruyter van Steveninck, R.R. (2000).

[9] Andrecut, M. (2009). "Parallel GPU Implementation of Iterative PCA Algorithms". Journal of Computational Biology. 16 (11): 1593–1599. doi:10.1089/cmb.2008.0221. PMID 19772385

[10] McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience. ISBN 0-471-69115-1. MR 1190469

[11] Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering" (2001) Journal of Machine Learning Research, 2: 125–137.

[12] R. Ng and J. Han. "Efficient and effective clustering method for spatial data mining". In: Proceedings of the 20th VLDB Conference, pages 144-155, Santiago, Chile, 1994.

[13] Rennie, J.; Shih, L.; Teevan, J.; Karger, D. (2003). Tackling the poor assumptions of Naive Bayes classifiers

[14] Everitt, Brian (1998). The Cambridge Dictionary of Statistics. Cambridge, UK New York: Cambridge University Press. ISBN 0521593468.