A Framework for Augmented Analytic Knowledge Discovery

Jon-Paul Boyd
MSc Intelligent Systems, De Montfort University
1st January 2018

1 Summary

Data volumes acquired by enterprise are increasing exponentially. Data complexity is on the rise, with growing numbers of features and a wider variety of data types including unstructured text, images and video. It is now extremely challenging, if not outright impossible, for humans to analyse all possible patterns and take the most significant, unbiased and accurate actions using manual processes. Using todays processes for business problem driven data analytics can take weeks or months to deliver value actions. This research aims to explore, architect, develop and evaluate a framework in which subfields of artificial intelligence including machine learning (ML) and natural language processing (NLP) can be leveraged to augment and support the human with automated knowledge discovery, insight understanding and decision-making support in an integrated, operationalised system.

2 Background

We live in an age of information, with exponential data growth generated by human and machine. According to [1] big data volume sizes "*are reported in multiple terabytes and petabytes*". Untapped data has no value, regardless of quantity. Our ability to consume data with correct analysis transformed into understood, justified and actionable decisions in hours or days is being exceeded. Enterprise needs to accelerate their analytics data processing beyond speeds that humans with their current toolsets are capable of if they are to survive in the fiercely competitive digital economy. [2] agrees, "*the ability to analyze meaningful and relevant data and convert data to information, knowledge, and ultimately action in time to favorably influence an organization is a key competitive differentiator*". The challenge of analytics sustaining business value decision making is compounded in part by the 4 Big Data Vs (volume, velocity, veracity, variety) and also by business users who "*often resort to exploring their own biased hypotheses, miss key findings, and draw incorrect or incomplete conclusions, which adversely affects decisions and outcomes*" [3].

Enterprise can maximise Big Data value by leveraging the right technologies and skillsets. Unfortunately there is a shortage of data science expertise. On average it takes 46 days to fill data scientist job vacancies and the projected 5-year demand growth for this role is 28% [4]. Data science modelling is complex, multi-disciplinary and largely manual, requiring programming skills, understanding maths behind algorithms, statistics experience and business domain knowledge. Gartner tells us data scientists "*are professionals with the capability to derive mathematical models from data to reap clear and hard-hitting business benefits*" [5]. Regarding technologies, [6] informs us that "*to extract information from big data that is of value to an organization, new techniques and advanced tools have to be developed and applied, such as advanced data mining or new artificial intelligence tools*" and "*return on big data is associated with making decisions cheaper, faster and better than before*". Shortage of data scientists can be mitigated in part by automating steps in an analytics workflow, especially complex and time-expensive ones like model generation. A product called Salesforce Einstein, introduced in October 2016, has intelligence capabilities, with "*The idea is to uncover insights, predict outcomes, recommend next-best actions and automate routine, manual tasks that keep people from being more productive*" [7]. However, Salesforce Einstein is Customer Relationship Management (CRM) centric without natural language narration (NLN) to augment outcome understanding, whereas I propose a domain-independent solution including NLN. DataRobot, another automated analytics platform, generates predicative models using ML, evaluates competing models and deploys models so they can be operationalised. DataRobot however does not leverage NLP for translation of the business question or narration of the outcome.

In [8], the authors submit "*A Knowledge Discovery (KD) process is a complex inter-disciplinary task, where different types of techniques coexist and cooperate for the purpose of extracting useful knowledge from large amounts of data. So, it is desirable having a unifying environment*". Rialto provides capability to ingest data from various sources, model data relationships and apply algorithms to mine the data, all via a graphical user interface (GUI) that supports manual modelling of workflows with nodes to represent data and tasks. Their manual approach is orientated to data driven questions with pattern mining, without ML for advanced insight or NLP to augment result evaluation.

My proposal seeks to deliver an integrated, automated knowledge discovery process, with the definition of "model" expanded from just the algorithm to include the complete analytics workflow process, from data ingestion to predicted outcome and deployment. This will realise the vision of [9]: "*opportunity lies in the other phases of the knowledge discovery process where certain tasks could be semi-automated if not completely automated to increase the overall efficiency and effectiveness of the knowledge discovery process*". There is a gap in the body of knowledge for an augmented analytics solution to address the aforementioned challenges, where ML and NLP can support human decision making with: a) asking business questions in natural language, b) automating feature selection from the business-driven question for unbiased search, c) automatic generation, ranking and deployment of predictive models in hours,

and d) automatic narration and intelligent visualisation of findings for aided understanding and transparency.

3 Proposed Work

3.1 Aims and Objectives

The main aim of this research project is to develop an augmented analytics, self-service application for practical use that delivers data science to non-expert users and supports their business-driven decision making through an analytics workflow from data ingestion to insight and understanding with automated predicative analytics.

3.2 Sub Objectives

Note the organisation of sub objectives as hypothesis (Hn) grouped by research question (RQn).

RQ1 Feasibility Can such an augmented analytics application be delivered through to execution?

- H1 Discover the available open source modules that can expedite such an application build
- H2 Determine cloud hosting solutions with data store, memory and processor scaling
- H3 Consider a pilot study

RQ2 Functionality Can fields within A.I. enable automation of processes within the analytics workflow?

- H4 Confirm NLP sensibly transforms natural language questions into automated data science questions
- H5 Verify intelligent algorithms can be used to automate feature selection
- H6 Ascertain if automated feature selection can improve on manual process
- H7 Verify the automatic ability to select the right out-of-box model(s) sensibly
- H8 Confirm the ability to generate the right machine learning algorithms
- H9 Determine if automated model generation or selection leads to better human decision making
- H10 Determine if automated model generation or selection can be validated by data scientists
- H11 Determine if possible to narrate outcomes in natural language with high quality and sense
- H12 Discover if possible to automate intelligent selection of visualisation types
- H13 Verify automated models can be operationally deployed so they can be used

RQ3 User Experience Can human decision making be supported by augmented analytics?

- H14 Verify application enables non-expert users to extract value and make better business decisions
- H15 Ascertain if application reduces the overall manual time spent on knowledge discovery and insight
- H16 Determine if natural language outcome narration gives greater meaning and transparency
- H17 Determine if natural language narration of outcome assists better decision making

RQ4 Technical Performance Can the application perform?

- H18 Compare automated models against known values to check model prediction performance
- H19 Compare predictions from automated models against full datasets to check for overfitting
- H20 Check automated models can scale up, including with parallelisation, when run against full datasets
- H21 Verify automated processes complete in acceptable timeframes

4 Rationale

A number of vendors in the market, including Salesforce and DataRobot, are making headway in addressing the challenge of delivering data science to non-expert end users, with tools supporting automated analytic processes to reduce the time between data ingestion and decision making. There are however current gaps, with predicative analytics approaches data-driven and heavily dependent on traditional data mining techniques to find patterns in data. My opinion is the data-driven approach is the wrong approach, with business-driven models the higher value route to take. Discussing this theme of data-driven vs business-driven with Chris Wright, Global Business Solution Manager for Business Analytics at Nestle, he agrees: "*For me the business driven tends to result in a business decision. Data driven can provide interesting insights but if it doesnt align to a decision-making process it does not result in action. And there are so many possible insights that it is hard to narrow down. So, for me, most useful things start with a hypothesis: the insights and iterations can change dramatically the original hypothesis, but it still starts with a hypothesis. Also in the business world this tends to get better buy in*".

A Harvard Business Review on "Making Advanced Analytics Work For You" [10] supports this position, with "*the desired business impact must drive an integrated approach to data sourcing, model building, and organizational transformation. Thats how you avoid the common trap of starting with the data and simply asking what it can do for you*". They go further with "*One approach that gets inconsistent results, for instance, is simple data mining. Corralling huge data sets allows companies to run dozens of statistical tests to identify submerged patterns, but*

that provides little benefit if managers cant effectively use the correlations to enhance business performance. A pure data-mining approach often leads to an endless search for what the data really say" [10]. Asking open questions like "*How many customers are going to leave me this month?*" or "*How can we adjust pricing to increase sales volume?*" offers new opportunities to leverage natural language processing and query to take the business question and get at the right data without human bias or theory confining the analysis and leading to wrong conclusions.

I will harness NLP and NLN to narrate outcomes, articulating how algorithms arrive at conclusions with a concise story that users can engage with. Not only will this capability humanise automated predications, it will be a mandatory legal requirement. In the EU in 2018, and in the US by 2020, the General Data Protection Regulation (GDPR) will be extended to add "*right to explanation, whereby a user can ask for an explanation of an algorithmic decision that was made about them*" [11], [12]. I believe this demand for transparency is anyway good for business in ensuring they comprehend their own decision-making process, and creates exciting scientific opportunities.

There is a clear need for heavily augmenting business end users with data science to help them make better decisions, but models can only contribute to the business bottom line if lifted out of the training environment and efficiently deployed to production. My research will consider deployment, noting [13] "*analytic models are knowledge intensive products that are not only expensive to build, but also expensive to maintain and deploy rapidly*". The race is on in contributing to the automation of advanced predictive analytics. According to Gartner, "*By 2020, 50% of analytic queries will be generated using search, natural-language processing or voice, or will be auto-generated*" [14], and "*By 2021, the number of users of modern BI and analytics platforms that are differentiated by smart data discovery capabilities will grow at twice the rate of those that are not, and will deliver twice the business value*" [14].

5 Methodology

I will begin with planning ways of working with my supervisor(s) and domain experts: agreed methods on how and when we will communicate to assess progress and discuss open points. A comprehensive literature review will follow, to include the latest augmented analytics current state and prediction reports from research bodies like Gartner, and any filed patents. As the solution looks to automate various steps in the predictive analytics workflow I will review the existing body of knowledge relevant to each component part of the automation pipeline. In taking a business question and translating to a data science question for automation, I can learn from work done in the field of Natural Language Query (NLQ), an example being [15] where "*complex English language sentence is correctly translated into a SQL query, which may include aggregation, nesting, and various types of joins*". Resolving automated feature selection, I will look to existing research that includes "*classification experiments on face datasets and UCI datasets show that DGFS (Decision Graph-based Feature Selection) can reduce the redundancy information contained in feature set effectively, and the selected features have a better ability of discriminant*" [16]. Exploring the selection, generation and tuning of predictive algorithms, I will investigate what native Python, Scikit-Learn, R, Vowpal Wabbit and TensorFlow off-the-shelf open source options are available, but include research covering papers such as [17] that notes "*the demand for machine learning algorithms grows faster than the supply of machine learning experts, there is an increasing need for methods to automatically configure algorithms to the task at hand, even where the task at hand is complex and algorithm performance (such as cross-validation error, or likelihood) expensive to evaluate*". Researching Natural Language Generation (NLG) in order to attach a story to the prediction, I will research the existing body of knowledge, for example papers like [18] which share interesting outcomes including "*We show that the use of Natural Language Generation (NLG) improves decision-making under uncertainty, compared to state-of-the art graphical-based representation methods. We also show that women achieve significantly better results when presented with NLG output (an 87% increase on average compared to graphical presentations*)".

A mixed methods, after-only study design will be employed to evaluate the research. Qualitative in-depth user feedback pre and post experiment from questionnaires containing closed (e.g. "*Are you a data scientist?*" dichotomous variable with yes/no) and open questions ("*What is your impression of the narrated outcome?*") will enrich quantitative empirical measurements collected by the technical framework. The technical solution itself will constitute the main research instrument. Observations will be made on each of the 4 concepts outlined in "Sub Objectives". RQ1 Feasibility will be measured with scales qualitative in nature, as will RQ3 User Experience (except H15 "..overall manual time spent.." measured in time interval scale). RQ2 Functionality is mixed: H11 "..narrate outcomes in natural language.." can be quantitatively measured by ordinal scale (e.g. agree, neutral, disagree) whereas H4 "..NLP sensibly transforms natural language questions.." is qualitative. RQ4 Technical Performance is quantitative. 21 hypotheses in total are being tested, another indicator the study is primarily quantitative in design.

Lets look closer at H7 "Verify the automatic ability to select the right out-of-box model(s) sensibly". I take a training dataset from the well-known UCI Machine Learning Repository (secondary data source) where the actual best predication as determined by an expert data scientist with domain expertise is established. Several algorithms including random forest and regularised generalised linear models can be trained on this dataset. We measure the performance of each model by comparing the predicated outcome vs the actual known outcome and score with ratio interval, the result validated by the expert. This approach can form the basis of our algorithm test framework.

6 Programme of Work

Objectives delivered by the following work packages (WPn). Effort estimates include technical unit testing. An additional 8 weeks will be allocated as "cushion time" to allow for unseen circumstances. 156 weeks total.

WP1 Agree Ways Working (2 weeks) Methods & periodicity with Supervisor(s), Industry & Domain Experts.

WP2 Literature Review (8 weeks) Examine the current body of knowledge as elaborated in "Methodology".

WP3 Feasibility Study (32 weeks) Ambitious research not possible with purely custom development approach/on-premise hosting reliant on open source modules, algorithms and cloud hosting, to be identified, costed and procured. Deliverables include architectural/technical design documents, detailing the generic "engine" and how feature selection, model generation and outcome narration plugged into engine as components and executed in parallel pipelines. Selection of initial dataset from UCI Machine Learning Repository. Build functional, full end-to-end prototype, proving technical build can be successfully operationalised. Evaluate generated models vs known (including by mean f score, continuous ranked, average precision etc). Demonstrate and collect feedback. Hypothesis may be refined.

WP4 Research Instrument Design (8 weeks) Identification of respondents from study population and design data collection, including composing of questionnaires. Identify additional secondary dataset samples, prepare test case documentation. Refine hypothesis and associated measurements used, based on outcome from WP3.

WP5 Engine Core Component (10 weeks) Orchestrate automation of the predictive analytics workflow process as a multi-step pipeline and enable integration of additional components as plugins.

WP6 Data acquisition component (4 weeks) Enable drag/drop of file-based datasets.

WP7 Natural Language Query component (12 weeks) Facilitates end user providing business question in natural language form and translating to data science question, output available to feature selection component.

WP8 Automated Feature Selection Component (8 weeks) Intelligent feature selection for ingested dataset. Includes GUI showing scoring of all dataset dimensions on relevance.

WP9 Model Selection Component (10 weeks) Repository for off-the-shelf models with model scoring against ingested dataset sample. Includes GUI for model score leader board.

WP10 Model Generation Component (14 weeks) Automation of model generation feature for models requiring training and ensemble models where multiple models are synthesized and scored. Includes enhancement of GUI for model score leader board.

WP11 Natural Language Narration Component (13 weeks) Combine dataset & model outcome to narrate story in natural language. Includes GUI enhancements to hide model complexity for simple user interface.

WP12 Intelligent Visualisation Component (4 weeks) Automate selection of best fit visualisation type.

WP13 Model Deployment Component (4 weeks) A feature to operationalise the successful model.

WP14 User Acceptance Testing (8 weeks) End user evaluation of the solution, includes interviews, observations and data collection to measure each hypothesis. Includes end user manual.

WP15 Technical Specification Document (3 weeks) Detailed documenting of the technical implementation.

WP16 Final Research Report (8 weeks) Report for stakeholders, detailing research instrument, research observations, findings and conclusions. Submission for publication to appropriate journals (IEEE, ACM, Others).

7 Professional, Legal and Ethical Issues

Datasets available in the public domain from the UCI Machine Learning Repository will be used to avoid data privacy concerns and not be bound by any dependencies or restrictions if enterprise-supplied data was used. In addition to the literature review existing patent information will need to be searched to ensure the proposal is novel and not breaching any existing patents. There should also be consideration for filing a patent. Data collected from study respondents will be held securely and if shared will be anonymous and by consent. The De Montfort University code of ethics will be followed.

8 Risks

I will require the support of expert data scientists to evaluate and help refine the feature selection, model selection/generation and outcome narration components. One critical goal is to deliver a user experience that fully engages. Only by participating closely with real business end users in a creative design thinking process to shape the solution can there be a successful outcome. The research is at risk if either party is unavailable to support. I hope to leverage support from within my professional network to provide these resources. Possible mitigation can be by crowdsourcing data science and domain expertise via social media and sites like `kaggle.com`. The main financial cost will be cloud hosting of the solution throughout development. This will be funded personally unless "WP3 Feasibility" indicates unreasonable costs, whereby I will approach professional contacts for possible sponsorship. I have mitigated any data privacy concerns by opting to use datasets in the public domain.

9 Research Management Plan

With a professional background in software development and information systems technical architecture, I have experience in using Microsoft Project to document, plan and track technical implementation at granular level that includes task dependency and milestones. Objectives and work package details from this research proposal will be migrated to the technical project plan to form the main planning blocks. Finer technical, testing and administration tasks with resources will be added. The plan will be continuously updated and used to identify risks and act as a source for updating my supervisor during catch-up sessions.

10 Justification of Resources

People The research proposal author alone will be responsible for the study design and execution, including all technical aspects of the research instrument. 1 data scientist expert will be required for evaluation for 3 hours per week. 2 business end users will be required for iterative design and evaluation for 2 hours per week. Regular review sessions (frequency to be agreed) with research supervisor.

Data Datasets freely available from UCI Machine Learning Repository - `https://archive.ics.uci.edu/ml/datasets.html`

Hardware Cloud hosted, normally on paid subscription / hourly rate.

Software Open source freely available, including Python, R, Scikit-Learn, Vowpal Wabbit, Tensorflow

Technical References Many Python/Scikit-Learn/Tensorflow/data science/statistical method books in personal library. Additional free learning materials from `coursera.org`, `udacity.com`, YouTube

References

[1] A. Gandomi and M. HaiderTed, "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management, vol. 35, no. 2, pp. 137-144, 2015.

[2] D. Bumblauskas, H. Nold, P. Bumblauskas and A. Igou, "Big data analytics: transforming data to action", Business Process Management Journal, vol. 23, no. 3, pp. 703-720, 2017.

[3] R. L. Sallam, C. Howson and C. J. Idoine, "Augmented Analytics Is the Future of Data and Analytics", Gartner, 2017.

[4] S. Miller and D. Hughes, "The Quant Crunch - How the demand for data science skills is disrupting the job market", Burning Glass Technologies, 2017.

[5] J. Rivera, "Gartner Says Advanced Analytics Is a Top Business Priority", Gartner, [Online]. Available: `https://www.gartner.com/newsroom/id/2881218`. [Accessed 03 01 2018].

[6] A. Intezari and S. Gressel, "Information and reformation in KM systems: big data and strategic decision-making", Journal of Knowledge Management, vol. 21, no. 1, pp. 71-91, 2017.

[7] D. Henschen, "Inside Salesforce Einstein Artificial Intelligence - A Look at Salesforce Einstein Capabilities, Use Cases", Constellation Research, 2017.

[8] G. Manco, P. Rullo, L. Gallucci and M. Paturzo, "Rialto: A Knowledge Discovery suite for data analysis", Expert Systems With Applications, vol. 59, pp. 145-164, 2016.

[9] S. Sharma, Kweku-Muata, Osei-Bryson and G. M. Kasper, "Evaluation of an integrated Knowledge Discovery and Data Mining process model", Expert Systems with Applications, vol. 39, pp. 11335-11348, 2012.

[10] D. Barton and D. Court, "Making Advanced Analytics Work For You", Harvard Business Review, 2012.

[11] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a right to explanation", in 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, 2016.

[12] "Right to explanation," [Online]. Available: https://en.wikipedia.org/wiki/Right_to_explanation.

[13] Y. Li, M. A. Thomas and K.-M. Osei-Bryson, "A snail shell process model for knowledge discovery via data analytics", Decision Support Systems, vol. 91, pp. 1-12, 2016.

[14] R. L. Sallam, C. Howson, C. J. Idoine, T. W. Oestreich, J. L. Richardson and J. Tapadinhas, "Magic Quadrant for Business Intelligence and Analytics Platforms", Gartner, 2017.

[15] F. Li and H. V. Jagadish, "Understanding Natural Language Queries over Relational Databases", SIGMOD Record, vol. 45, no. 1, pp. 6-13, 2016.

[16] J. He, Y. Bi, L. Ding, Z. Li and S. Wang, "Unsupervised feature selection based on decision graph", Neural Computing and Applications, vol. 28, no. 10, p. 30473059, 2017.

[17] T. Nickson, M. A. Osborne, S. Reece and S. Roberts, "Automated Machine Learning on Big Data using Stochastic Algorithm Tuning", 2014.

[18] D. Gkatzia, O. Lemon and V. Rieser, "Natural Language Generation enhances human decision-making with uncertain information", in 54th annual meeting of the Association for Computational Linguistics, Berlin, 2016.

Appendix A - Project Management - Deliverables with Timings

Task Name	Duration	Start	Finish	Predecessors
Research - A Framework for Augmented Analytic Knowledge Discovery	**770 days**	**03/09/2018**	**13/08/2021**	
WP1 Agree Ways Working	1 wk	03/09/2018	07/09/2018	
WP2 Literature Review	**40 days**	**03/09/2018**	**26/10/2018**	
Academic paper research	3 wks	03/09/2018	21/09/2018	
Business technology report research	3 wks	24/09/2018	12/10/2018	4
Patent research	2 wks	15/10/2018	26/10/2018	5
WP3 Feasibility Study	**160 days**	**29/10/2018**	**07/06/2019**	**3**
Research open source modules	1 wk	29/10/2018	02/11/2018	
Research open source algorithms	1 wk	05/11/2018	09/11/2018	8
Research cloud hosting options incuding cost/terms	1 wk	12/11/2018	16/11/2018	9
Dataset selection	1 wk	19/11/2018	23/11/2018	10
Technical prototyping	21 wks	26/11/2018	19/04/2019	11
Architectural document	2 wks	22/04/2019	03/05/2019	12
Technical design document	2 wks	06/05/2019	17/05/2019	13
Evaluation	2 wks	20/05/2019	31/05/2019	14
Demo and feedback	1 wk	03/06/2019	07/06/2019	15
WP4 Research Instrument Design	**40 days**	**10/06/2019**	**02/08/2019**	**7**
Identification of respondents	3 wks	10/06/2019	28/06/2019	
Design data collection	3 wks	01/07/2019	19/07/2019	18
Dataset selection - additional	1 wk	22/07/2019	26/07/2019	19
Refine hypothesis and associated measurements	1 wk	29/07/2019	02/08/2019	20
WP5 Engine Core Component	**50 days**	**05/08/2019**	**11/10/2019**	**17**
Build and Test	10 wks	05/08/2019	11/10/2019	
WP6 Data acquisition component	**20 days**	**14/10/2019**	**08/11/2019**	**22**
Build and Test	4 wks	14/10/2019	08/11/2019	
WP7 Natural Language Query component	**60 days**	**11/11/2019**	**31/01/2020**	**24**
Build and Test	12 wks	11/11/2019	31/01/2020	
WP8 Automated Feature Selection Component	**40 days**	**03/02/2020**	**27/03/2020**	**26**
Build and Test	8 wks	03/02/2020	27/03/2020	
WP9 Model Selection Component	**50 days**	**30/03/2020**	**05/06/2020**	**28**
Build and Test	10 wks	30/03/2020	05/06/2020	
WP10 Model Generation Component	**70 days**	**08/06/2020**	**11/09/2020**	**30**
Build and Test	14 wks	08/06/2020	11/09/2020	
WP11 Natural Language Narration Component	**65 days**	**14/09/2020**	**11/12/2020**	**32**
Build and Test	13 wks	14/09/2020	11/12/2020	
WP12 Intelligent Visualisation Component	**20 days**	**14/12/2020**	**08/01/2021**	**34**
Build and Test	4 wks	14/12/2020	08/01/2021	
WP13 Model Deployment Component	**20 days**	**11/01/2021**	**05/02/2021**	**36**
Build and Test	4 wks	11/01/2021	05/02/2021	
WP14 User Acceptance Testing	8 wks	08/02/2021	02/04/2021	38
WP15 Technical Specification Document	3 wks	05/04/2021	23/04/2021	40
WP16 Final Research Report	8 wks	26/04/2021	18/06/2021	41
Cussion Time	8 wks	21/06/2021	13/08/2021	42