SC Student Projects
2015-2016

Strategic Consulting - Student Data Science Projects

INVESTORS IN PEOPLE | Gold

INVESTORS IN PEOPLE | Champion

Contents

Amey SC Student Projects 2015-2016

The following pages provide a list of projects that could be carried out by MSc/MEng/BSc students during a period of 3 to 6 months (depending on course).

The projects intentionally provide a high level and open ended description that will allow flexibility of implementation. The goal is to be able to accommodate students from different disciplines and skill levels that will be able to deliver an agile software infrastructure (suitable for the current and future type of work delivered by Amey Strategic Consulting).

Notice to Students and Program Coordinators

All the projects described here require a high level of competence and/or interest in computing (programming and software engineering) and mathematics (statistical/machine learning).

The projects will involve a significant amount of programming (approx. 80%). The specific languages employed will strongly depend on the nature of the service/micro-service. For example:

- Machine/statistical learning services could be developed in languages such as R, Python or Java/Scala;
- Web Applications could be developed in Java/C#, Python or Ruby and will involve understating and development in other languages like HTML, CSS and JavaScript (the team is currently making heavy use of Libraries and Frameworks like Twitter Bootstrap, D3, Leaflet, AngularJS, among many others);
- Relational Data Base querying will be performed with SQL (e.g. Stored Procedures);
- DevOps and Infrastructure will take place within Amazon Web Services and could be scripted with Cloud Formation (services that could be used include EC2, EC2-Containers, S3, RDS, VPC, Elastic BeanStalk, among others);
- Version control is performed with Git while continuous integration could be done with tools such as Jenkins.

As it is well known, development of micro-services or service oriented architectures typically starts with simple monolithic applications in order to understand the problem at hand (identification of ideal component boundaries). In this regard, R represents one of the best options available (easy scripting language, with huge library to perform any type of analytics that also provides a simple way to deploy interactive GUI - http://shiny.rstudio.com/).

In terms of mathematics, the projects will focus on the implementation of statistical/machine learning techniques and linear algebra. A wide range of techniques/models could be employed depending on the data and problem under consideration, for example: Supervised vs Unsupervised, Classification vs Regression and Parametric (e.g. Polynomial Regressions, Logistic Regression) vs Nonparametric (e.g. Artificial Neural Networks, Support Vector Machines). Student will be particularly exposed to supervised models and will need to understand the importance of Cross Validation and the Bias-Variance Trade off when determining the best suitable model for an application.

Student Placement Modalities at Amey Strategic Consulting.

	Internships	**Final Year Industrial Projects (MSc/BSc/etc)**
Short Description	This placement gives the student complete experience on the day to day workings of Strategic Consulting. It involves working with different projects and clients.	This placement exposes students to state of the art research in Infrastructure Asset Management by focusing on delivering end to end data products.
Time Frames	6 months to 1 year (Full Time – 40 hours/week)	3 to 6 months (Full Time – 40 hours/week)
Start Dates	Any	Any
Space at Strategic Consulting Head Office (3rd Floor Chancery Exchange | 10 Furnival Street | London | EC4A 1AB)	Full time (5 days a week)	1 day a week. The rest of the time students should work from university/home. Students can stay in touch with Supervisors via Email at all times.
Amey Logging Account / Email / PC (Laptop)	Yes	No (Students need to own their own laptop). Amey will provide cloud resources – Amazon Web Services.
Nationality Constraints	Must demonstrate a right to work in the UK.	None.
Academic Quality Thesis	**No. As students will be performing a multitude of tasks, it becomes too difficult to assemble a coherent Thesis.**	**Yes. The work will be supervised by staff at PhD level and extensive academic research experience.**
IP / Non-disclosure Agreements Required (each case will be examined individually)	Yes	Yes (may not use anything other than for dissertation purposes).
Recruitment Process	CV, Test, Interview (SC London office)	CV, Test, Interview (SC London office or Skype Video)
Payment /Employment Contract	£20K p.a. (negotiable) / Yes	None / None
Allocation of SC resource time	A line manager will be assigned as per the standard line management setup – line managers have 80% utilisation targets (including leave).	Informal management / mentoring by the supervisor – supervisor has a 80% utilisation target (including leave). Supervision time expectations should be agreed at the project start.

Project Name: Analysis of Smart Meter Data

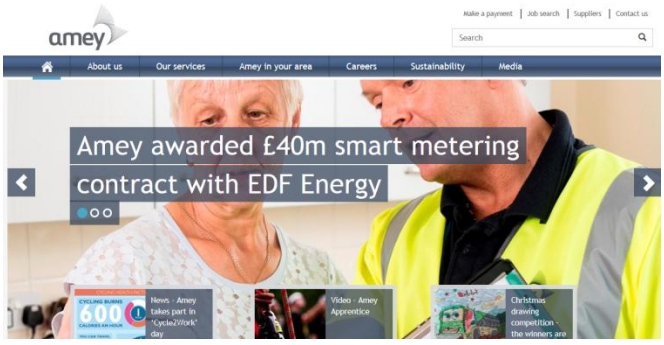
Project Code: AmeySC-1
Required Core Skills: Programming Skills (R, Python, Java/C#, Scala, etc) and Mathematics (Machine Learning, Linear Algebra, etc).

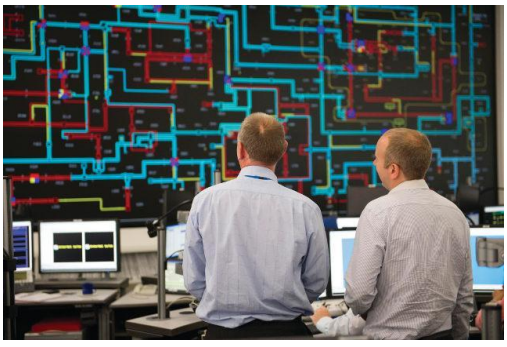
Project Background [1]

- The prevention of customer churn is a major issue for many utility companies, especially in deregulated markets. Usage and customer profile can allow suppliers to create tailored packages/special-deals to suit individual customers.
- Big-data analytics can help to improve the utilities own operations, particularly when it comes to forecasting demand in order to optimise supply, predicting likely outages, and identifying leaks and/or fraud.
- Market analyst GTM Research predicts global utility company expenditure on data analytics will grow from $700m in 2012 to $3.8bn in 2020, with gas, electricity, and water suppliers in all regions of the world increasing their investment.
- Sample projects based on open data sources and apps [2]:
    - http://data.gov.uk/dataset/energy-consumption-for-selected-bristol-buildings-from-smart-meters-by-half-hour
    - http://data.gov.uk/dataset/dcms-energy-consumption
    - http://data.gov.uk/dataset/decc-live-energy-data
    - http://data.gov.uk/apps/energy-use-stats-by-location
    - http://data.gov.uk/apps/the-interactive-uk-energy-consumption-guide

Project Description

UK National Grid's infrastructure is adapting to new regulatory requirements outlined by the Revenue = Incentives + Innovation + Outputs (RIIO) framework which governs the revenues that the UK's 14 electricity Distribution Network Operators are allowed to collect during April 2015 to March 2023 [3]. The present project will focus on identifying efficiencies by being able to match supply more closely to demand at regional and customer levels. Important predictors that will be included are weather [4] and demographics. The project will require the development of a Service Oriented Architecture (in particular Microservices) that will include: Real Time Web based GUI and GIS, Data Warehousing (SQL/NoSQL) and Data Analytics (Machine Learning).





References

1. http://eandt.theiet.org/magazine/2014/01/data-on-demand.cfm
2. http://data.gov.uk/
3. https://www.ofgem.gov.uk/network-regulation-%E2%80%93-riio-model
4. http://www-03.ibm.com/press/us/en/pressrelease/41310.wss

Project Name: Analysis of Road Traffic Information

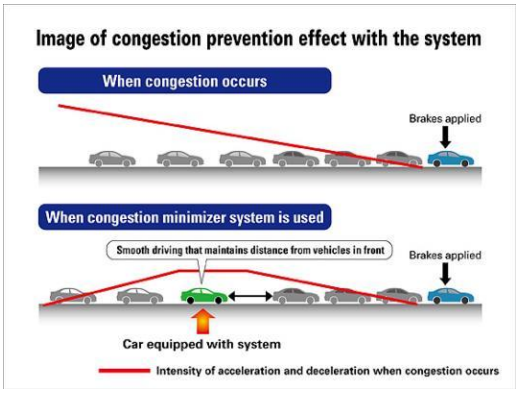
Project Code: AmeySC-2
Required Core Skills: Programming Skills (R, Python, Java/C#, Scala, etc) and Mathematics (Machine Learning, Linear Algebra, etc).

Project Background

- Increasing motorist population means that governments around the world face difficulties in addressing issues related to road traffic, such as reducing congestions or improving road safety. Fortunately, the increasing amount of data from traffic information systems (e.g. GPS, cameras, sensors) provides the mechanism to solve many of the current and future challenges.
- Open data UK provides live traffic information data showing traffic information on the strategic road network in England, maintained by the Highways Agency [1].
- Small sample projects that work with traffic/congestion/roads/cars:
    - http://www.inrix.com/
    - http://www.smh.com.au/technology/technology-news/ibm-smartphone-app-predicts-traffic-jams-20110413-1deiy.html
    - http://www.wsj.com/articles/SB10001424052702303444204577460552615646874
    - http://www.t-systems.co.uk/abouttsystems/big-data-in-traffic-52-million-cars-and-zero-congestion/1021692
    - http://www.roadtraffic-technology.com/projects/hong-kong/
    - http://data.gov.uk/apps/uk-road-accident-map
    - http://data.gov.uk/apps/drive-time-maps
    - http://data.gov.uk/apps/uk-motorcycle-accident-hotspots
    - http://data.gov.uk/apps/geenfile-no-traffic-jam
    - http://data.gov.uk/apps/traffic-london-uk

Project Description

Analysis of live traffic information has the potential to address issues related to environment (e.g. emissions reductions), economic productivity (e.g. congestion reduction), safety (e.g. traffic accidents reduction), among others. The present project will aim to provide management, monitoring, analysis and control of traffic related activities. It is expected that the system will be able to identify efficiency gains in critical areas such as congestion reduction and safety. The project will require the development of a Service Oriented Architecture (in particular Microservices) that will include: Real Time Web based GUI and GIS, Data Warehousing (SQL/NoSQL) and Data Analytics (Machine Learning).



References

1. http://data.gov.uk/dataset/live-traffic-information-from-the-highways-agency-road-network

Project Name: Smart Public Transport System

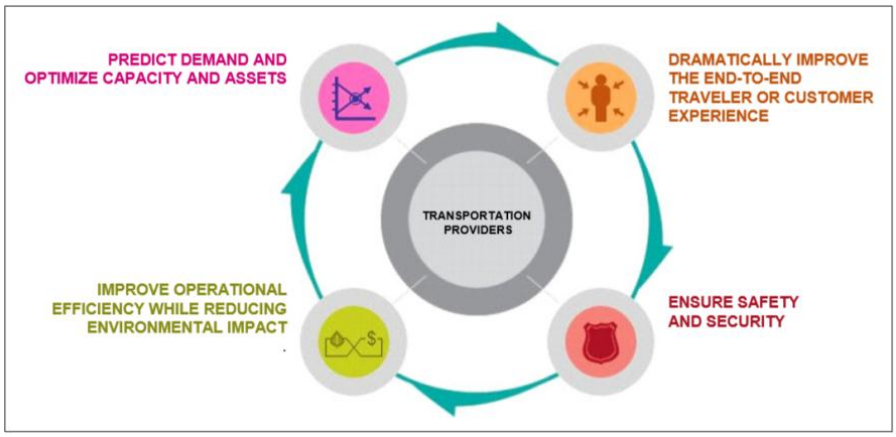
Project Code: AmeySC-3
Required Core Skills: Programming Skills (R, Python, Java/C#, Scala, etc) and Mathematics (Machine Learning, Linear Algebra, etc).

Project Background

- In all societies, the transportation system is vital because it connects people, goods, and services. In recent decades, significant increases in urbanization have placed a burden on most cities transport systems around the world. The typical approach to solve transport problems include: increase infrastructure capacity (build more roads, rails, etc) or increase number of vehicles (more buses, trains, etc). These approaches have reached their limit in many cities and for this reason it is time for a new approach to these challenges. The future goal should focus in optimising the use of existing capabilities in order to provide safer, cleaner and more efficient travel.
- Small sample projects that work with open data for public transport:
    - http://data.gov.uk/apps/mapumental
    - http://data.gov.uk/apps/locrating-commuting-map
    - http://data.gov.uk/apps/allschedulesschedules-stops-and-routes
    - http://data.gov.uk/apps/visualising-public-transport-networks
    - http://data.gov.uk/apps/uk-bus-times-live-bus-scout
    - http://data.gov.uk/apps/ldn-busdar
    - http://data.gov.uk/apps/uk-bus-checker
    - http://data.gov.uk/apps/bustimesorg
    - http://data.gov.uk/apps/london-traffic
    - http://data.gov.uk/apps/london-bus-live
    - http://data.gov.uk/apps/london-live-bus-map
    - http://data.gov.uk/apps/railgb
    - http://data.gov.uk/apps/london-next-bus-arrivalsinseconds
    - http://data.gov.uk/apps/bustop-london
    - http://data.gov.uk/apps/trainsim

Project Description

When it comes to travel, time and condition of commuting is more important than distance. For this reason the goal of the present project is to provide a holistic view of an entire public transport network of a city like London in order to deliver the following objectives: predict demand and optimise capacity, assets and infrastructure; improve end-to-end experience for users; increase operation efficiency while reducing environmental impact and ensuring safety and security. The project will require the development of a Service Oriented Architecture (in particular Microservices) that will include: Real Time Web based GUI and GIS, Data Warehousing (SQL/NoSQL) and Data Analytics (Machine Learning).



Project Name: Prediction of Blockage in Sewer Systems.

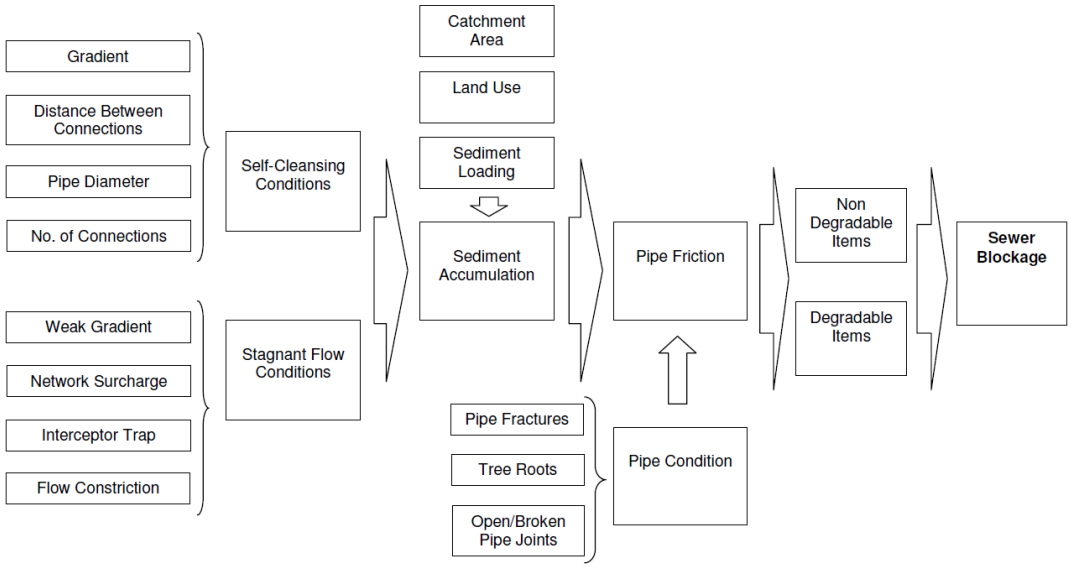
Project Code: AmeySC-4
Required Core Skills: Programming Skills (R, Python, Java/C#, Scala, etc) and Mathematics (Machine Learning, Linear Algebra, etc).

Project Background (Commercial Relevance) [1]

- Sewer blockages are responsible for the majority of sewer flooding incidents. They cause the discharge of raw sewage effluent into homes and into natural watercourses and are immensely expensive to the water industry. Moreover, aging infrastructure and increase loads due to ever increasing population are contributing to an increase of reported blockages.
- In its conclusion the Government stated that a TOTEX approach similar to the one proposed by Ofgem in the energy sector has "considerable attractions". TOTEX, according to Ofwat, would impact on three areas of the regulatory framework: Incentive effects – there would likely be some rebalancing between CAPEX and OPEX costs and how Regulatory Capital Value (the value of a water company's assets) is calculated; Cost assessment – how Ofwat assesses the performance and efficiency of a process or asset; Cost recovery – how the companies recover costs and earn a return.

Project Description

Despite advances in monitoring technology, blockage prediction still remains a challenge for the water industry due to cost constrains. Unlike the energy industry (in which smart meters are becoming the norm), the water industry still has thousands of users without any type of metering devise. The objective of the present project is to provide Water companies with the possibility to capitalise on the public's addiction to social media as well as many other sources of open and/or live data that can be used to improve their current model (based purely on internal data). One of the hypothesis that will be considered is that if blockage is caused primarily by specific sediments such as "fat" and "wet wipes", will a machine learning model that incorporates different types of demographics (e.g. male vs female, single vs married, restaurant vs residential homes) with other open source data such as weather forecast provide better predictive capabilities than existing technologies? These and many other variables (see graph and references) will be considered in the analysis.



References

1. L. Berardi, O.Giustolisi, D.A. Savic, Z. Kapelan, An effective multi-objective approach to prioritisation of sewer pipe inspection, 11th Int. Conf. on Urban Drainage, Edinburgh, 2008.
2. D. A. Savic, O. Giustolisi and D. Laucelli, Asset deterioration analysis using multi-utility data and multi-objective data mining, Journal of Hydroinformatics | 11.3–4 | 2009.
3. R. Harvey, E. McBean, Predictive and Spatial Analytics for Planning Inspections of Sewer Infrastructure, Int. J. of Environmental Protection Apr. 2014, Vol. 4 Iss. 4, PP. 48-57.

Project Name: DS2.0 - Enhancing Data Science with ABMS

Project Code: AmeySC-5
Required Core Skills: Programming Skills (R, Python, Java/C#, Scala, etc) and Mathematics (Machine Learning, Linear Algebra, etc).

Project Background (Commercial Relevance)

- Data science has been referred as the sexiest job of the century and it is among the fastest growing professions due to its potential application (e.g. cancer research, marketing, e-commerce, finance, etc). Unfortunately, its full potential is usually hindered by:
    - Data quality: It is commonly said that 50% to 90% [1] of a data scientist time is spent on data preparation (cleaning, formatting, etc). This is highly dependent on applications, for example, sectors which are highly digitised (e.g. web based) and use automatically generated data could spend less than 50% while sectors that are not digitised (paper based) and use human input could exceed 90%.
    - Lack of proper Investment: Before any company decides to invest in data, it typically faces questions such as: which data should we study/analyse? How much it will cost? And how much will we get back? None of these are easily answered by data science alone.
- Agent Based Modelling and Simulation (ABMS) has been praised as the most exciting and practical development in business simulation since the invention of relation databases [2]. The technique has wide spread use in social sciences, and the understanding of business complexity [2]. Unfortunately the technique has some weaknesses; with Verification & Validation the most important aspect currently limiting its wide spread adoption.

Project Description

The long term goal of the present project is to find optimal or synergetic coupling strategies of Data Science and ABMS (combining strengths to reduce their individual weaknesses). As demonstrated by recent developments in academia and industry (see for example [3,4,5]), the combined use of these techniques can provide the answers that future business leaders are looking for. The specific objective of this project is to develop an ABMS model that is able provide answers related to Return on Investment (RoI) for data science initiatives (see sample figure on the right [2]). Key questions that the model should aim to answer are: what is the RoI of analysing existing data sets? What is the RoI to start collecting/recording new data sets? What is the RoI of improving metrics of our current data quality dimensions (in particular completeness)?. The model will be focused on applications related to Infrastructure Asset Management and will be verified and validated following experiences from large UK infrastructure owners (e.g. London Underground, Network Rail, Heathrow).

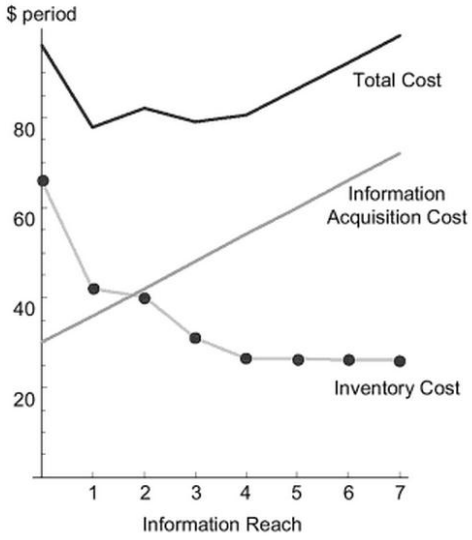


FIGURE 5.21 Results of information simulation experiments.

References:

1. http://insidebigdata.com/2015/07/03/nearly-a-third-of-bi-professionals-spend-50-90-of-time-cleaning-raw-data-for-analytics/
2. M.J. North and C.M. Macal, "Managing Business Complexity", Oxford University Press (2007).
3. http://www.concentricabm.com/simulation/
4. O. Baqueiro, Y.J. Wang, P. McBurney, and F. Coenen, "Integrating Data Mining and Agent Based Modelling and Simulation", P. Perner (Ed.): ICDM 2009, LNAI 5633, pp. 220–231 (2009).
5. M. Wang, X. Hu, Simulation Modelling Practice and Theory, v56, pp. 36–54 (2015).


INVESTORS
IN PEOPLE
Gold


INVESTORS
IN PEOPLE
Champion


Part of Ferrovial
ferrovial