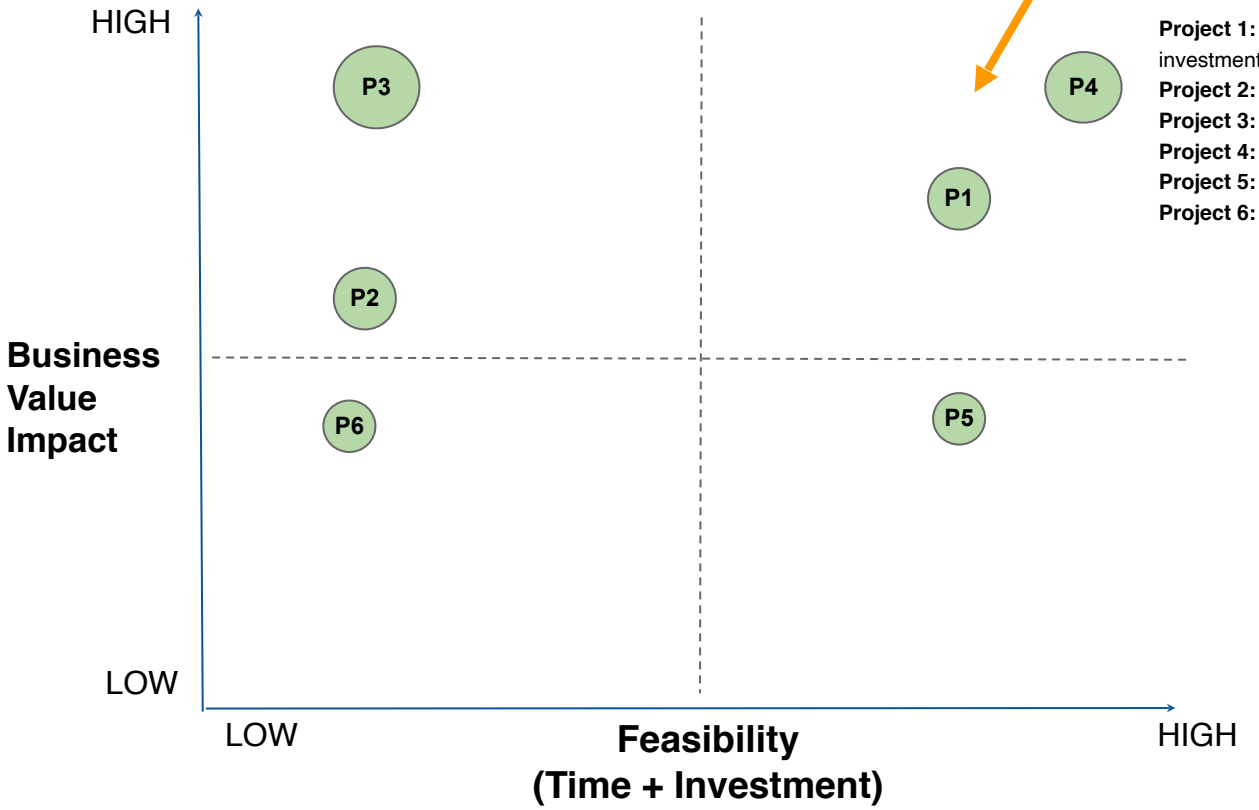
100 Day Data Science Plan: Building a data strategy

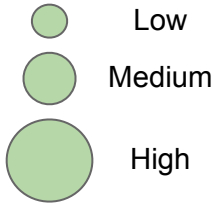
Udacity Capstone Project Presentation - Andra Tolbus

Step 2, Part 2: Complete the "Data Science Opportunity Matrix" below by modeling each of the six projects in terms of feasibility (time & investment), business value impact, and likelihood of value capture



Project 1: Marketing Mix Modelling for a core brand with large investments
Project 2: Predicting customer churn
Project 3: Short term forecasting automation
Project 4: Predicting launch curves for future products
Project 5: Supplier Quality Optimization
Project 6: Localisation of global ML model for vendor selection

Likelihood of Value Capture



AA Consumer Goods Corporation

Establishing the Data Science practice in a local operating company

Name : Andra Tolbus
Position : Advanced Analytics Lead

Date: March 27, 2021

Executive Summary

Purpose of 100-day plan

- Establish the foundations of a Data Science practice in the local OpCo
- Showing immediate benefits of data science in improving business operations

Approach

- Run data science MVPs to gain momentum by leveraging existing corporate data and analytics infrastructure as much as possible with a newly established local data science team

Results

- Direct improvements in business operations and foundational infrastructure and process to further develop the area

Scope of Work for First 100 Days

- Develop a clear understanding of the corporate data and analytics infrastructure and ensure the right level of access for the local teams
- Develop data intake processes together with IT which can support external data integration
- Build a team of data science specialists that can support in executing the planned initiatives
- Prioritise work on high impact projects with high likelihood of productionalisation while working on moving the internal mindset towards “ML models” as a service as the ideal future state
- Initiate engagement and change management activities to support the new activities
- Initiate data literacy training programs

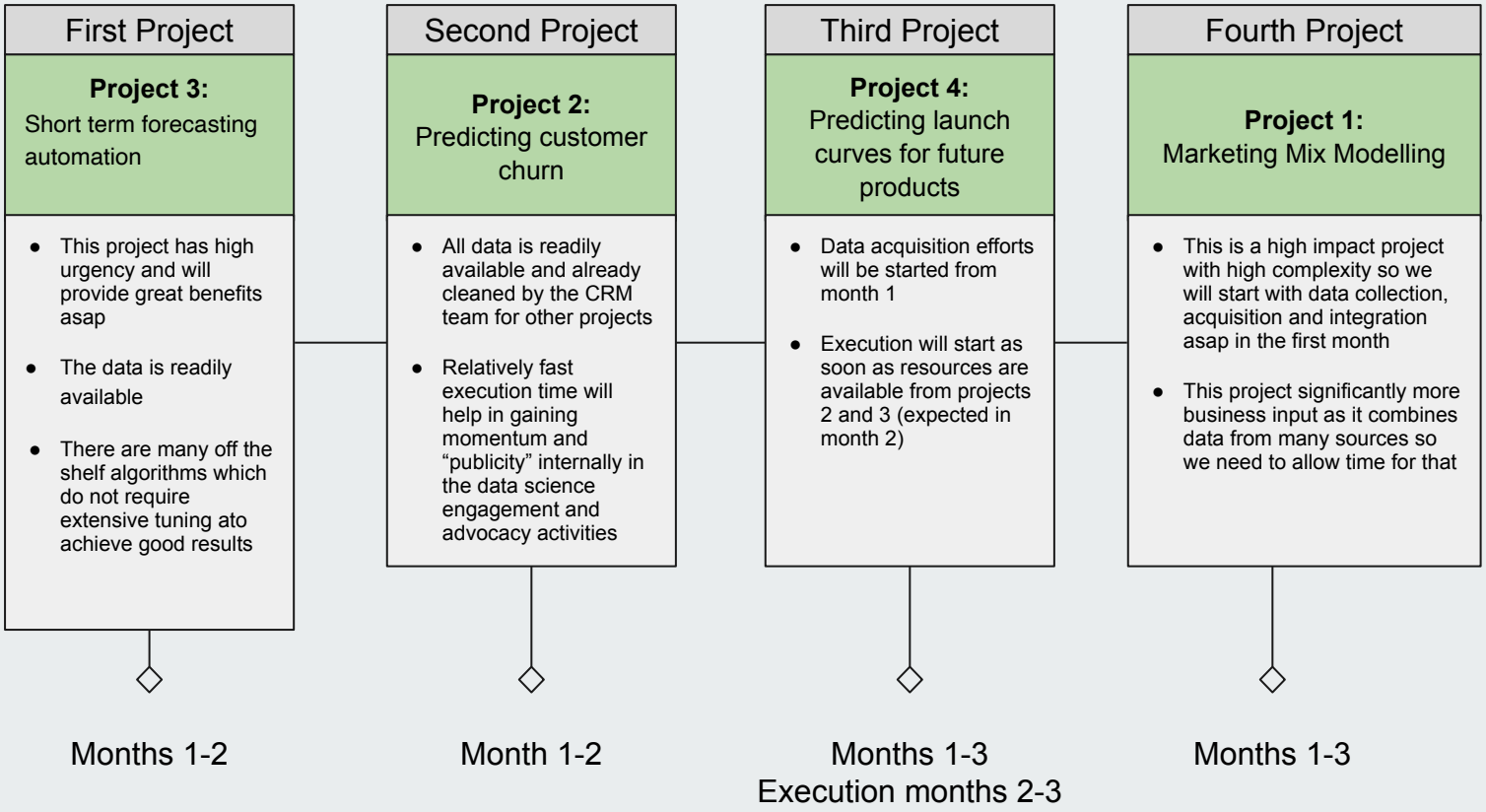
Candidate Data Science Projects

	Functional Area	Project Description
Project 1: Marketing Mix Modelling for a core brand with large investments	Marketing	Identifying the main marketing channels which are diving sales for one of the biggest brands with the ultimate goal to increase profits and ROI by optimizing the marketing investments.
Project 2: Predicting customer churn	Marketing	Identifying webshop customers who are most likely to churn as part of the overall company's retention strategy.
Project 3: Short term forecasting automation	Finance	Automate short term revenue forecasting for the finance teams who are currently spending hours doing this job manually in Excel spreadsheets.
Project 4: Predicting launch curves for future products	Finance	Predict the first couple of years of sales for new products based on similar product launches in the past.
Project 5: Supplier Quality Optimization	Supply Chain	Using data on suppliers, past quality assurance programs and shipments we can predict future non adherence to quality standards.
Project 6: Localisation of global ML model for vendor selection	Procurement	Adapting a global ML model for vendor selection to the local market.

Step 2, Part 3: Complete the “Data Science Road Map” below with the first four data science projects chosen for implementation.

Order	Project	Order Justification
1	**Project 3:** Short term forecasting automation	There is an urgent need as the financial teams workload is very high and the company has experienced a decrease in the continuous forecast accuracy. One of the company’s corporate KPIs is to increase revenue forecasting accuracy across the planning cycles. This project also has a low cost and low implementation effort with a high likelihood for success which will help in getting traction for the other upcoming data science projects.
2	**Project 2:** Predicting customer churn	The strategic goal during the pandemic is to grow the online business. In order to do so, we need to better understand our direct customers and ensure that we have a retention strategy
3	**Project 4:** Predicting launch curves for future products	This project has a high complexity, but it is linked to overall corporate strategic goals and can have a large long impact on business growth.
4	**Project 1:** Marketing Mix Modelling for a core brand with large investments	This project has the highest complexity in terms of the data needed (availability and quality). A successful completion would have a large impact as it would ensure better budget allocation, increase profit and growth.

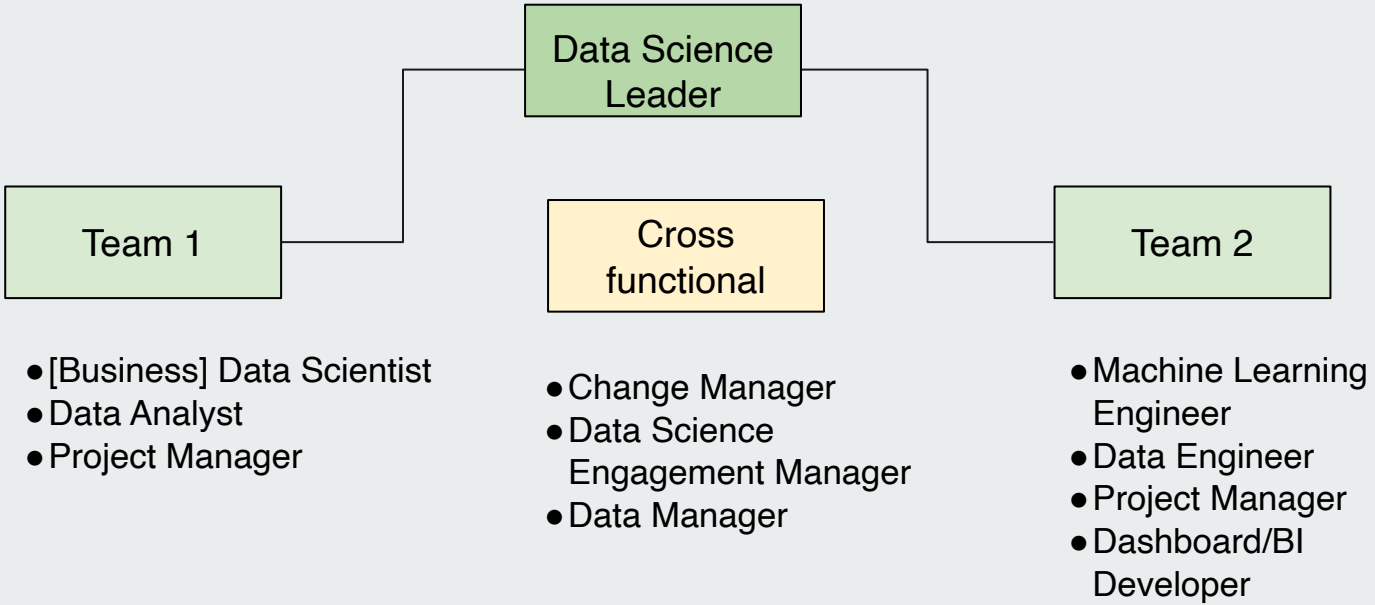
Step 2, Part 3: Complete the “Data Science Road Map” below with the first four data science projects chosen for implementation.



Our Highest-Priority Data Science Projects

Order		Direct Alignment with Strategic Goals?	Cost	Complexity of Implementation	Certainty of Value Capture	Magnitude of Benefit
		1=Low; 5=High	1=High; 5=Low	1=High; 5=Low	1=Low; 5=High	1=Small; 5=Large
First	Project 3: Short term forecasting automation	4	5	4	5	4
Second	Project 2: Predicting customer churn	4	5	4	5	5

Initial Structure of the Data Science Team



I have identified six strategies for promoting a data-driven culture in our business

Strategies for promoting a data-driven culture

Strategy 1: Executive Data Science Education - training leadership role in what data science is the benefits of incorporating data driven approaches in business practice.

Strategy 2: Engagement programs through sharing sessions where Data Science POC results are shared across the organisation

Strategy 3: Data literacy training across all roles

Strategy 4: Use external consultants for unique skills/ first time projects only and ensure that they work on the infrastructure together with the internal teams rather then delivering the results in ppt decks.

Strategy 5: Create concrete learning programs for traditional analysts who would like to transition into a data science career. They will be great assets for these initiatives in the long term.

Strategy 6: Encourage a try & fail fast culture for data science initiatives in the initial phases to stimulate the interest across teams for trying out new things and incorporating learnings into the next project.

Technical Infrastructure Needed to Support the Data Science Organization

Data Requirements	What data should be included in the Data Strategy?	• The financial and CRM data is readily available via the centralised data lake. • The next steps for the OpCo would be to integrate their marketing data from various sources and external partners as well as develop processes for external data acquisition that can help answering the specific questions
Data Governance	Data Availability	• If the data is highly sensitive (such as financials) , it will be made available for the project team with the appropriate clearance only. • If the data is not sensitive, it will be made available for access to the appropriate data dictionaries for all team members . • Regardless of access level, the set up needs to be very easy to understand so I would make sure that the data is clearly labelled and documented.
	Usability	• Providing use cases from the central teams where they have worked on similar data • Providing external use cases either published or from startup partnerships where they have managed to get insights from similar data.
	Integrity	• All data coming from the internal systems will be subjected to a thorough process of cleaning, documentation and validation which will follow an automated process and periodic quality and security reviews. • Security certification will be required to operate with the internal data • The external data required for a POC will not be uploaded to our environment before it has been approved by IT, legal (GDPR and privacy compliance) and security.
	Security	• Data can only be accessed in the company's secure environment (AWS via virtual machines or via the data stack already enable on the cloud infrastructure) • Each user will receive the right level of access depending on their role and scope of project .

Technical Infrastructure Needed to Support the Data Science Organization (continued)

Technology	Data Architecture Components	• Centralised Data Lake for all internal sources (Amazon Redshift) • Amazon S3 buckets for one-off /POC projects where users can upload their data directly • AutoML tool (DataRobot) • Amazon AWS EC2 for advanced ML development with Python • Data pipeline (Apache Airflow : open source) • Tableau for data visualisation • RStudio server for R code development
Skills and Capacity	Data literacy skills and organizational capacity	• While the individual data science specialists are skilled within their own areas, they would require additional training on the bigger picture and how the business operates. This will help that the specialists do not work in silos and solutions which are far fetched from how the business operates. • SImilarly, other teams outside data sciences will require some high level training in data science across the ranks. This will make sure that everyone speaks the same language. • Developing internal learning paths for data upskilling across roles. These would be internal “data champions” that will support in different capacities our data initiatives.
Support for Machine Learning	Machine learning architecture	• DataRobot (autoML) will be connected directly to the Datalake and it will help with data access , preparation and basic ML for the easy use cases. The users can also upload their data in S3 and access it directly via this infrastructure. The output from the ML models can be pushed into S3 and visualised in Tableau. • DataRobot provides Python code integration so we can use the above setup for end-to-end ML in most cases • If we need to implement something that cannot be easily done with the above, we can do so using Python , Apache Airflow and Docker from the same data source. • This would be a more rare use case but suitable in the initial phase (up to 2 years)