# DATA MONETIZATION CASE STUDY

**Note for every step documented –**
**1 – Given data in RED.**
**2 – Technique in BLUE.**
**3 – Output of technique in PURPLE.**

**DATA SET #1 (purchased from an alumni magazine publisher)**

1. First name
2. Last name
3. College or university attended
4. Year of graduation
5. Major or majors
6. Marital status
7. Number of children
8. Current city
9. Email domain
10. Financial net worth
11. Binary variables (one for each interest in the publisher's long list of various sports, activities, hobbies, games) showing whether each one was or wasn't listed by each person

**DATA SET #2 (purchased from a credit bureau)**

1. First name
2. Middle name
3. Last name
4. Marital status
5. Sex
6. Year of birth
7. Current city
8. Whether they ever owned real estate
9. Email domain
10. List of monthly payment status over the last five years for credit cards, mortgages, rent, utility bills, etc. – for each month and each payment:
    a. What type of payment it was – for credit cards, it would say "Visa", "American express", etc., not just "credit card"
    b. How much was owed
    c. How much was paid
    d. Whether the person was considered to be in default

**DATA SET #3 (collected by the company using web site tracking code)**

1. Title
2. First name
3. Middle initial
4. Last name

5. Credit card type
6. Credit card number
7. List of products purchased in the past, with date of purchase and ship-to address
8. Which web pages the person looked at
9. How long the person spent on each page
10. What the person clicked on each page
11. Estimate of how long the user's eyes spent on each page viewed (for customers where the software was able to take over the device's camera)

**PROBLEM GOAL –**

> ***To identify ways in which monetary value can be generated from this data, and evaluating appropriate analytical methods to use in generating that value –***
>
> ***1. using one dataset at a time***
> ***2. using multiple datasets at a time***

**1. WITH DATASET 1**

**Monetization Use-Case –** The first dataset may be used to build a psychology and return-on-time-investment based lead generation system and introductory-message-formatting system for prospective clients that may be assisted with financial advisory aimed at gaining their trust on decision making so they may be able to maximize the returns on their savings and investments and make their life quality better.

Step 1 – Personality, Economic and Lifestyle Modeling

Given: Data from hobbies, financial worth, university, marital status, and number of children, current city, and year of graduation.
Do: Build three models –
1. one for hobbies – K-means clustering
2. one for university, current city, and year of graduation – K-means clustering
3. one for marital status, number of children, financial worth and current city – K-means clustering
To: Build clusters for tailored plans in approaching these leads –
1. in terms of lifestyle and recreation
2. in terms of influence from old friends and circles
3. in terms of family economy and expense structures

Step 2 – For each of the models above – outreach formats

Given: Clusters from step 1 and the individual categories for hobbies, financial worth, university, marital status, and number of children, current city, and year of graduation AND name and email data.

Do: Build a sentence formatting function using NLTK for natural language generation. Design the introduction by speaking about what they'd like and use the second half to speak about the tailored descriptions for life-plans that your company can give them, based on clusters from step 1.
To: Convey what difference you can make, why it improves financial and lifestyle quality, and how you plan on helping.

**2. WITH DATASET 2**

**Monetization Use-Case –** The use-case here seems quite suitable for determining whether a customer should be granted a loan for a specific amount or not.

Step 1 – Extreme points removal

Given: Data subset – real estate owner ship and year of birth.
Do: Outlier detection.
To: Clean data of outliers.

Step 2 – Payment behavior analysis

Given: (Amount paid – amount owed) difference in time series format for each customer.
Do: Exponential smoothing – first order to obtain predicted value of difference for +T time.
To: Assess tendency of customer to default loans or delay loan payment and set score.

Step 3 – Loan decision system

Given: Score from loan defaulting and payment delay, real estate ownership, age, and marital status.
Do: SVM classification with kernel adjusted to be biased in favor of increasing false positives – customers who weren't given the loan but are deserving of it.
To: Determine which customer may be granted a loan and playing it safe instead of taking high-risk and incurring defaulted loan losses.

**3. WITH DATASET 3**

**Monetization Use-Case –** This dataset would support a system that matches customers to what they'd like to buy first and what they might like next, based on their activity and their purchase tendency and capability.

Step 1 – Missing data

Given: Which web pages the person looked at, how long the person spent on each page, what the person clicked on each page, estimate of how long the user's eyes spent on each page viewed (for customers where the software was able to take over the device's camera).
Do: Build a graph structure for the path followed and simulate paths that weren't taken based on relative paths' probabilities, and theory-driven research.

To: Create a probable exploration map for customers and their interests

Step 2 – Exploitation 1

Given: Credit card and financial information for discount avails.
Do: CART algorithm to classify the discount group into which the customer is likely to fall.
To: Present best offers available on the site.

Step 3 – Exploitation 2

Given: Past purchases information per customer.
Do: Run combinations of features – utility, price, reviews, aesthetics, brand, clicks and views count – and build multiple Bayesian models based for conditional probability that a certain product would be interesting. Cut-off the conditional probability for each model at experimented thresholds - using an artificial neural network giving weights to different Bayesian models - and derive the selected products from each Bayesian model.
To: Construct a set of products that are suitable for a customer based on his/her overall across all time past choices.

Step 4 – Exploitation 3

Given: Past purchases across all customers.
Do: Map each product to the next product bought by the customer within the same click-trail. Run a pruned Apriori algorithm to assess confidence scores of each pair. And take all pairs above a certain threshold.
Apriori algorithm: https://www.iasj.net/iasj/download/277bbbb8979bf3d3
To: Find out likeliness of buying a product based on previous product within short time across overall chain buying patterns followed by customer base.

Step 5 – Leveraging experiment and exploitation

Given: Decisions from experiment and exploitation models.
Do: First find intersections between all three exploitation models. Present top intersections from the three exploitation models. Use bandit scores for reinforcement learning to determine when to deviate and show them new products based on lower probability exploitation recommendations, AND products from simulated click-trails from the experiment model in step 1.
To: To present novelty, accuracy, and serendipity in recommendations.

## 4. ACROSS DATASETS

**Monetization Use-Case –** The idea I had for this one isn't something just associated with one goal but rather multiple socio-economic use-cases. It is aimed at understanding which of these factors may be related in a significant way to another and leveraging it any field where only a subset of these factors is available, and the rest of it is needed, but would have to be extrapolated.

Step 1 – Apart from name, email, and sex information, all other data is taken, repeats are removed, missing values are regressed and imputed, and all features are combined into possible pairs based on an SME domain score.

Step 2 – Correlations are taken for each X → Y mapping. And all those above 0.7 and below -0.7 are taken to step 3.

Step 3 – Naïve bayes conditional probability is taken for all selected pairs for Y given X. And Those pairs passing a threshold of 0.7 are verified for causal inference.

Step 4 – Case based simulation is conducted for these pairs of features while varying several domain-related parameters – first in a constrained manner, and then in a macroscope.

Step 5 – For each case, the relation between the predictors is studied, and categorized based type of feature that was changed and difference in relation.

Step 6 – The final report presents all surviving pairs that maintained constant relations during maximum cases, and the deviations under which the other lesser surviving pairs varied with the deviation type given as a reason.

Step 7 – This isn't completely cause-effect driven but it gives a good sense of causation relationships which can be used to deduce unseen case results and missing data.