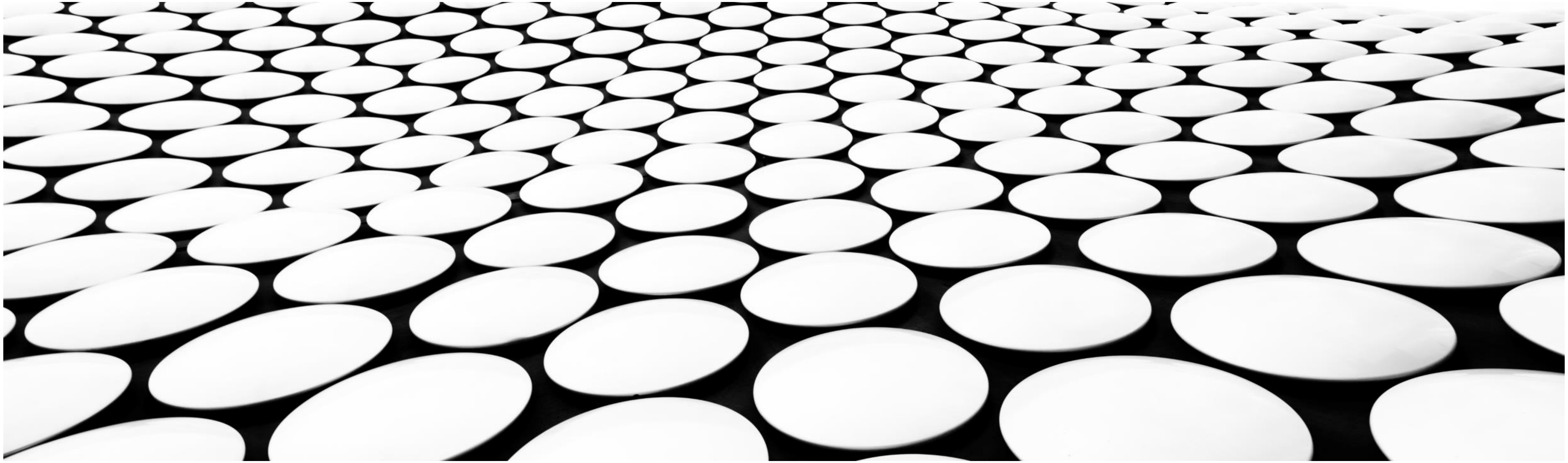
C2T1: Credit One Analysis Plan

Brian Mattis

BAM
Data Science



Problem Statement and Business Goals

Problem Statement

- An increasing number of customers are defaulting on loans

Business Goals

- Develop a better way to predict the creditworthiness of new and existing customers
  - Evaluate whether new customers should be approved
  - Compute an appropriate credit maximum based on customer information

Data Science Framework

- BADIR Data Science Framework methodology
  - Business Question
  - Analysis Plan
  - Data Collection
  - Insights
  - Recommendation
- BADIR chosen as it encourages initial thought of the most important question – what the analysis strives to answer
- BADIR places focus on the Insights phase, where both rear-looking analysis and predictive models work together iteratively to find answers to the business questions

Data Sources

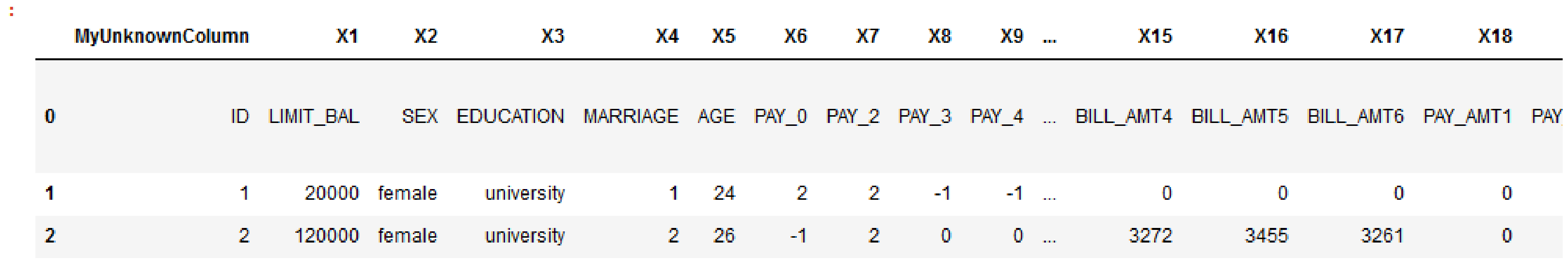
- Data source resides on a remote MySQL database
  - Data will be imported into Jupyter Notebook and converted into a Pandas dataframe for further manipulation
- Data contains customer demographics, historical data, and account status
  - Demographic: education, marriage status, age, and gender
  - Historical Data: Prior billing amounts, prior payment amounts
  - Account Status: Credit limit, Monthly repayment status, client’s default status

Data Management

- The data will be imported into Jupyter Notebook, with copies of the data exported to .csv and saved locally in case the connection to the remote MySQL server becomes unavailable.
- Analysis techniques will be done locally for optimum performance

Known Issues

- Incoming data set has the intended column names buried within the data



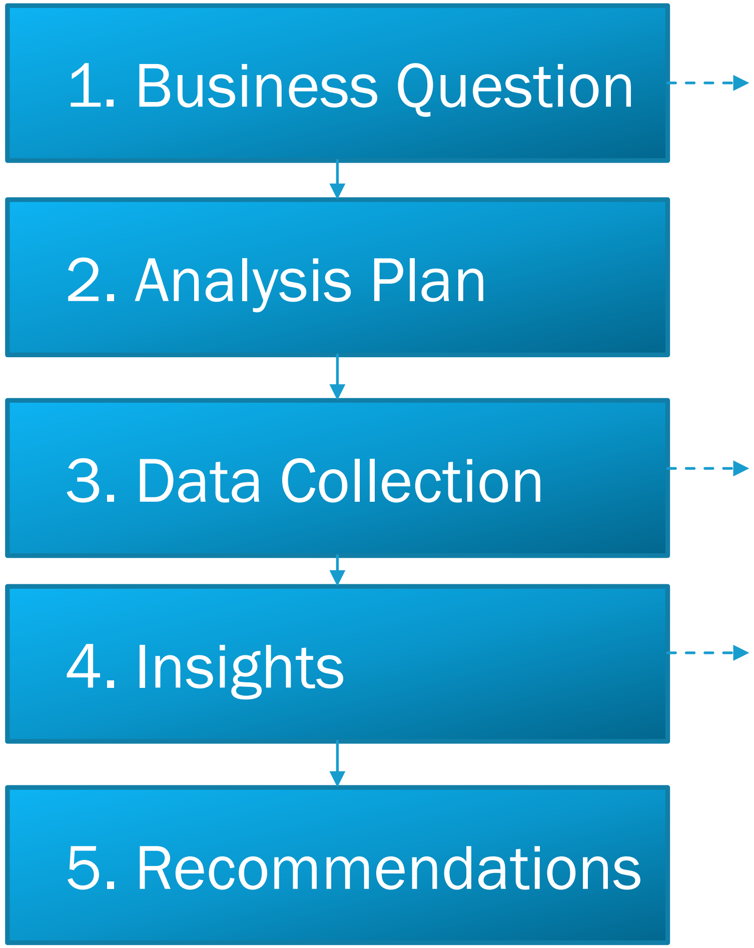
	MyUnknownColumn	X1	X2	X3	X4	X5	X6	X7	X8	X9	...	X15	X16	X17	X18	
0	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY
1	1	20000	female	university	1	24	2	2	-1	-1	...	0	0	0	0	
2	2	120000	female	university	2	26	-1	2	0	0	...	3272	3455	3261	0	

Imported column names

Intended column names

- The data must be thoroughly cleaned before in-depth analysis can start:
  - Add descriptive column names
  - Remove character values from numeric fields
  - Remove duplicate entries
  - Convert columns to the proper numerical data types

BADIR Flowchart



1. Business Question

2. Analysis Plan

3. Data Collection

4. Insights

5. Recommendations

Pitfall: The criteria for placing account in default are not defined
Solution: Base analysis and predictions on existing data only

Pitfall: Non-numeric and null values in data disrupt analysis
Solution: Clean data thoroughly

Pitfall: Original hypotheses from analysis plan don’t yield results
Solution: Investigate new hypotheses based on correlation plots, decision trees, and other data grouping

Initial Insights

- The provided bill, payment, and monthly status data is not sufficient to calculate whether an account becomes labeled as *defaulted*.
  - Other prior payment and usage history data that is not available must drive this decision
- More than 20% of accounts are in default (22%)
- Significantly more female customers than male customers
  - 18100 vs 11900