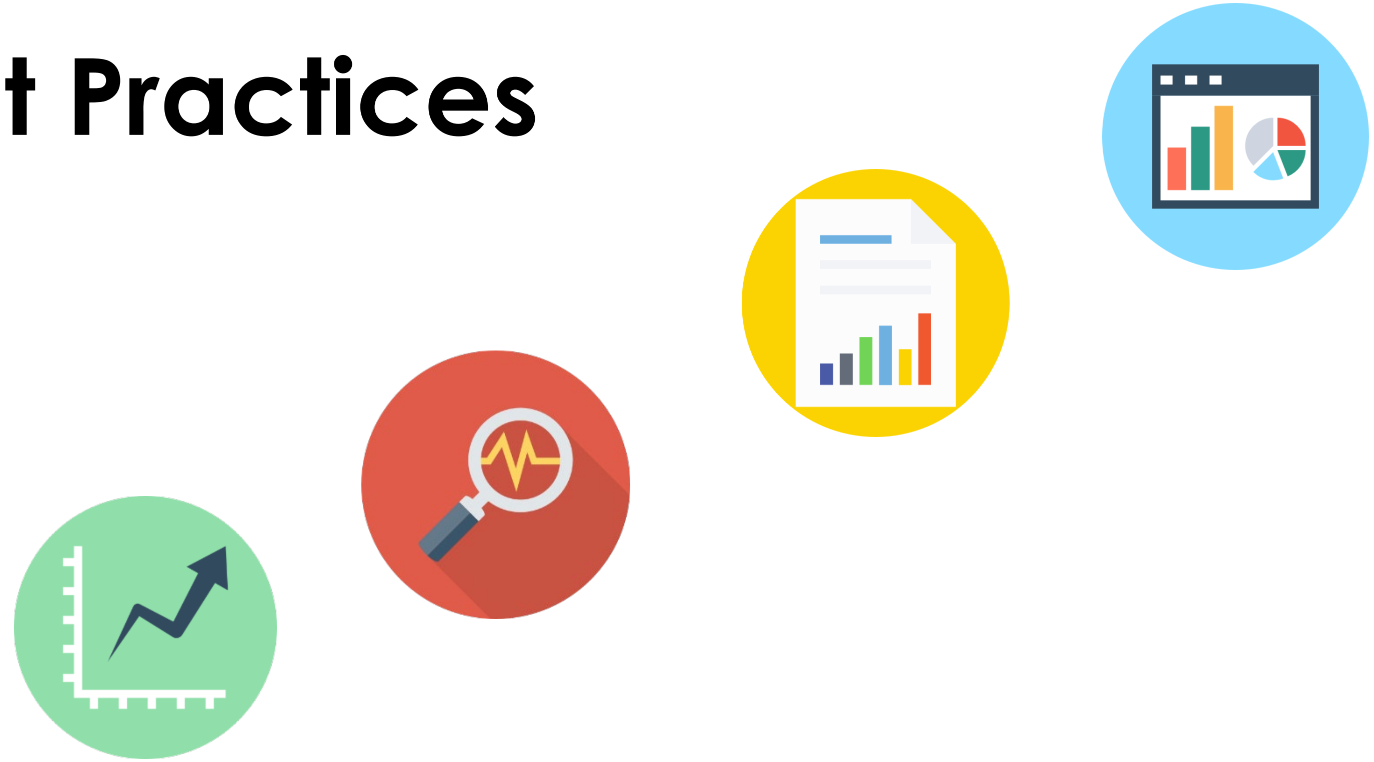
Data Analytics Best Practices

2.1



Agenda

Day 1

- Analytics basics
  - Analytics Processes
  - Data Requirements
  - Types of datasets in the industry
  - Data issues
  - Data Cleaning
  - Data Integration
- Workshop on arranging data elements
  - Functions
  - Data Formats
  - Date-time in R

Day 2

- Analytics best practices
  - Data Transformation
  - Exploratory Visualisation
  - Feature Engineering
  - Decision Engineering
  - Model Deployment
  - Model Maintenance
  - ROI Models
- Workshop on Analytics best practices
  - Intro to Data Cleaning
  - Data Preparation

Day 3

- Workshop on Data exploration
  - Visual Data Exploration
  - Non-Visual Data Exploration
- Data Warehousing Basics
  - Data Warehousing Introduction
  - Data Modelling Essentials

Data Transformation

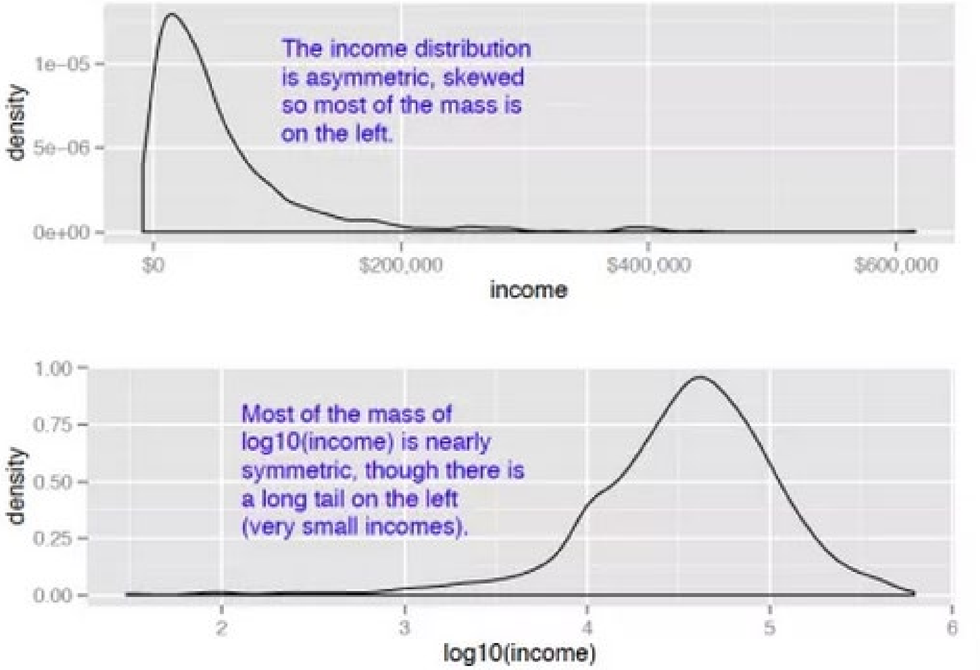
Data Analytics Best Practices

Data transformation

- Data might need to be transformed for the following reasons
  - Rolling up to the appropriate granularity
  - Creating a single row per customer
  - Merging data from different sources
  - Creating aggregations or summaries
  - Removing skews
  - Bringing multiple variables to the same scale
  - Creating new features (feature engineering)
  - Capping to remove extreme values
  - Creating data appropriate for the downstream technique
- Data visualisation usually helps with suggesting data transformations
- Best Practice benchmarks would create 3x derived variables from raw data

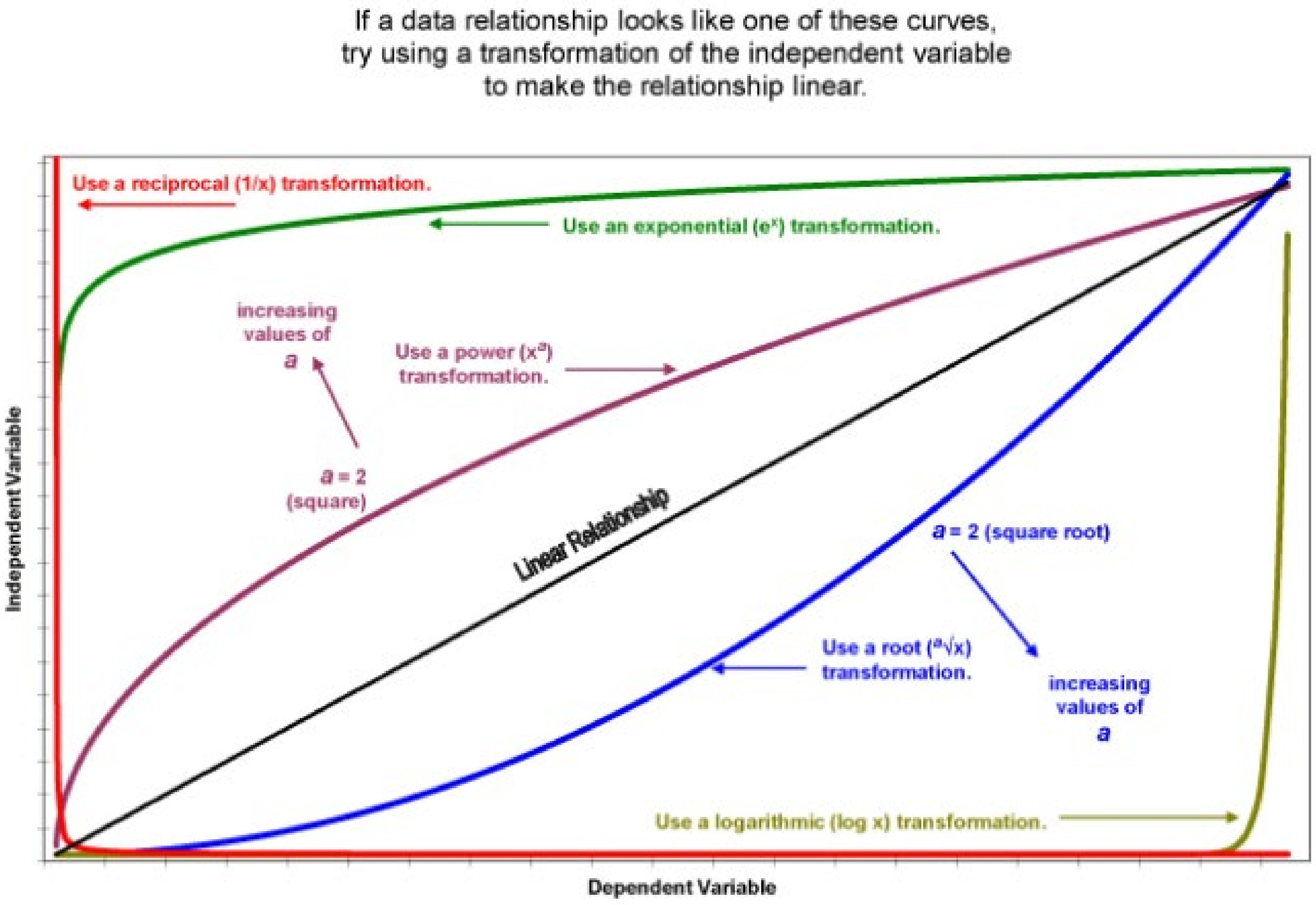
Log Transformations

- Log Transformation
  - Makes a skewed attribute more symmetric
  - Reduces the magnitudes
  - Common bases 10, 2, *e* *(which base to use is often not important)*



- Incomes, customer value, account or purchase sizes—are commonly encountered sources of skewed distributions in data science applications.
- Often they are log-normally distributed: the log of the data is normally distributed

Log Transformations



Data Normalization

- Reduces outlier distortion and enhances linear predictability
- Ensure all variables have approximately the same scale
  - E.g. variable *Age* vs *Income*: a distance of 10 "years" may be more significant than a distance of $1000, yet $1000 swamps 10 when they are added in calculating distance
- Normally re-center and rescale the data to be around zero, in the range from 0 to 1, etc.
- Common Methods…..

$$v' = \frac{v - min_A}{max_A - min_A}$$

Min-max scaling

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

Z-score scaling

$$v' = \frac{v}{10^j}$$

Where *j* is the smallest integer such that Max(|*v'*|)<1

Decimal scaling

Handling Categorical Data

- Many modeling methods require numerical inputs
  - One major exception is decision tree methods
- How to convert categories into numbers without introducing an unintended ordering?
- E.g. Which of these is the best mapping?

- Small ->1
- Medium -> 2
- Large -> 3

- Small ->3
- Medium -> 2
- Large -> 1

- Small ->2
- Medium -> 3
- Large -> 1

- What about this?

- Yishun->1
- Clementi -> 2
- Tuas-> 3
- Queensway -> 4

Handling Categorical Data

- How to handle…
  - Marital status = single, married, divorced, widowed?
- Could convert to…
  - Marital status = 0,1,2,3 where
    0 = single, 1=married, 2=divorced, 3=widowed
- Better to create four new T/F variables
  - Single = 0,1
  - Married = 0,1
  - Divorced = 0,1
  - Widowed = 0,1

- Caution:
  - For visualization and decision tree models, it's best to leave as one field called "marital status" with values = single, married, divorced, widowed

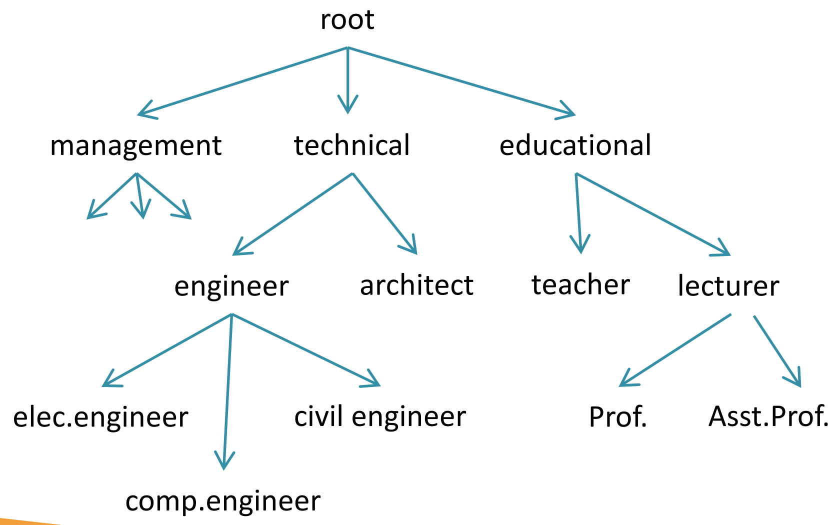
Handling Categorical Data

- **If** there is no obvious ordering within the categories then converting to a series of binary (1 => true and 0 => false) inputs is preferable
- This is often also called "one-hot" encoding or "dummy" variable encoding
- Example

Obs.	Colour	Colour_Red	Colour_Green	Colour_Blue
1	Green	0	1	0
2	Blue	0	0	1
3	Blue	0	0	1
4	Red	1	0	0
5	Green	0	1	0
6	Red	1	0	0

Handling Categorical Data

- Simplify categorical variables that have too many categories before doing binarisation
- Simple grouping may help
  - E.g. transform states into groups: western, eastern etc.
- If a concept hierarchy exists then categories can be merged by climbing the hierarchy
- Example



Gender	Profession	Bought PEP
M	teacher	Y
M	professor	Y
F	Asst. professor	Y
M	Civil engineer	N
F	Comp.engineer	N
F	Elec. engineer	N
M	architect	N

Gender	Profession	Bought PEP
M	educational	Y
M	educational	Y
F	educational	Y
M	technical	N
F	technical	N
F	technical	N
M	technical	N

Exploratory Visualisation

Data Analytics Best Practices

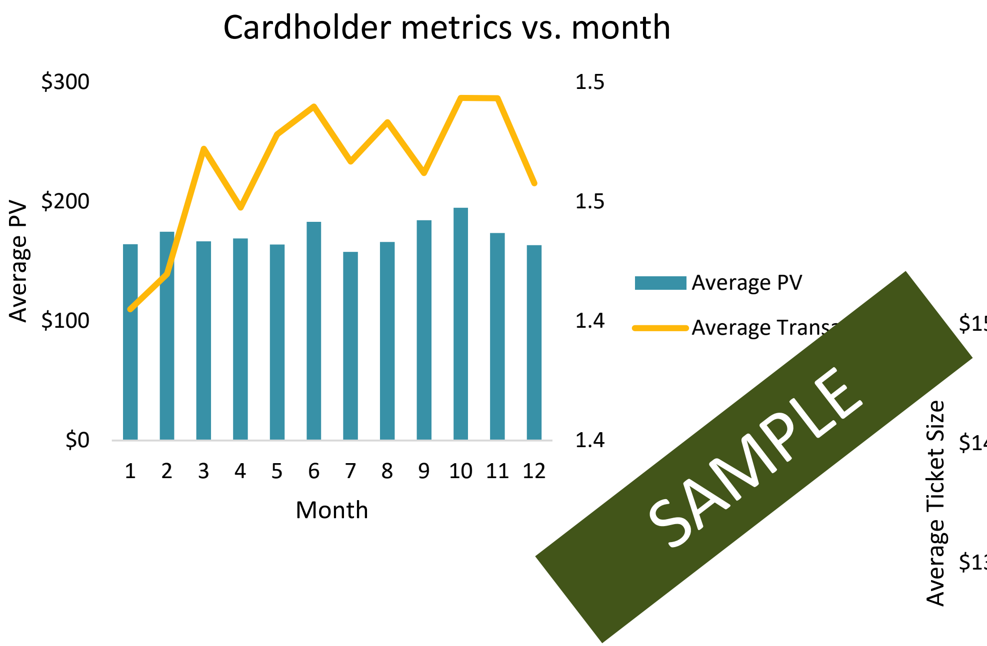
Visualization Phase Deliverables

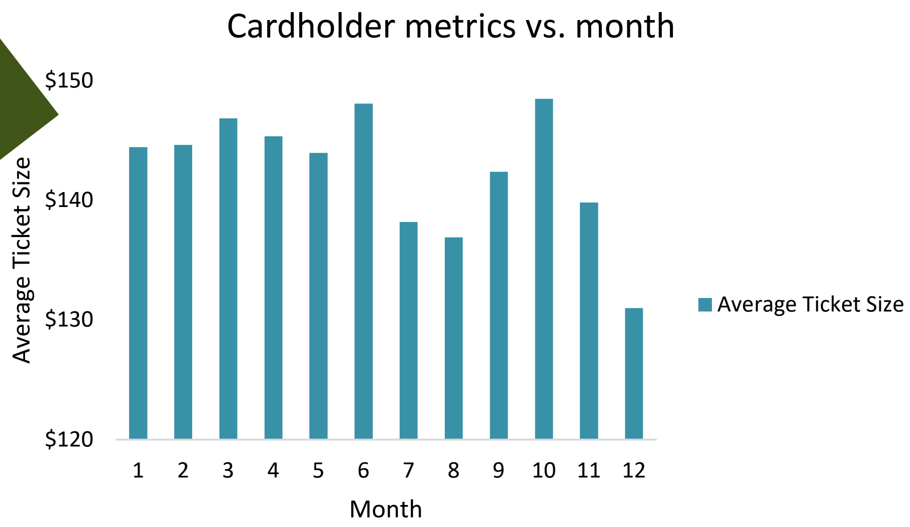
- Data Visualisation Presentation
  - Hypotheses improvements
  - Validation by Business
  - Often leads to interesting side projects
  - Often leads to new variable creation (more data preparation)

Summary Trends

- High level summaries of data trends are useful to start the visualisation with
- Usually time wise trends provide basic patterns to study
- To drill deeper into time across multiple categories, heat maps could be used

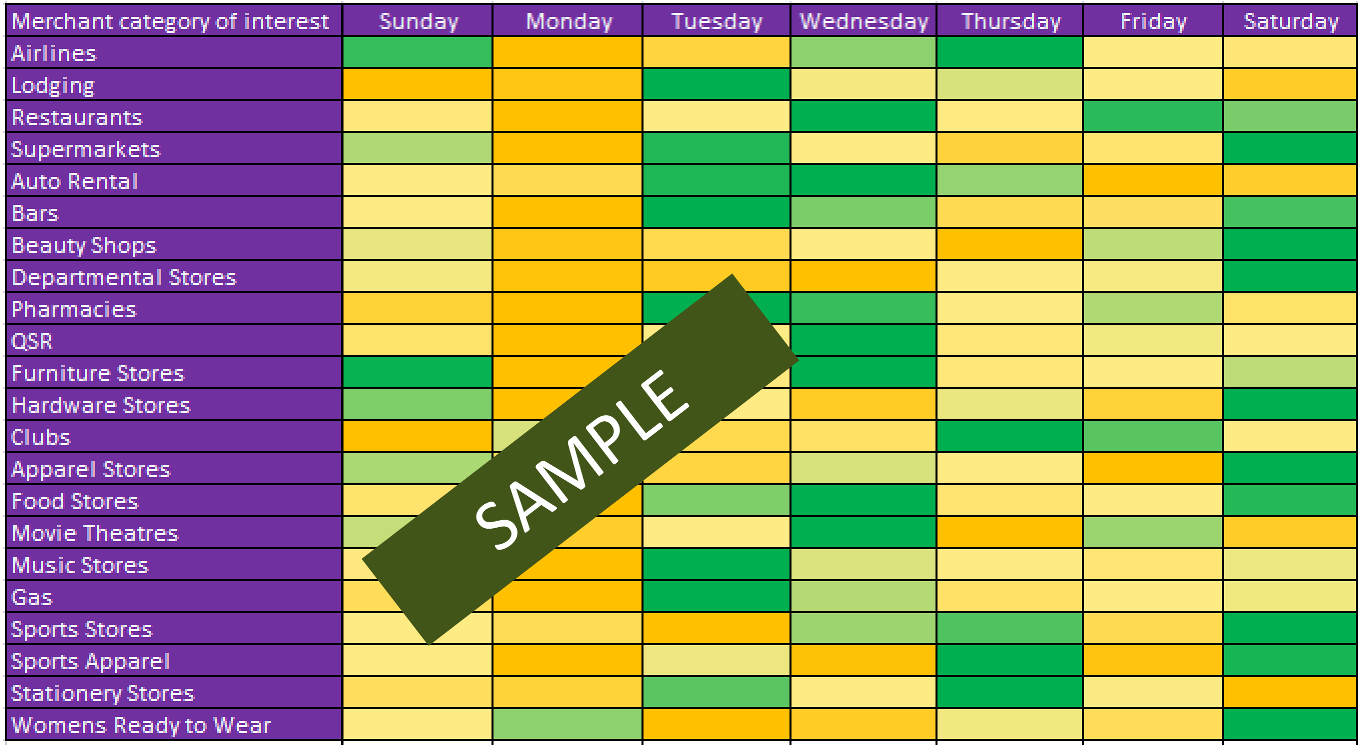
October has the highest PV, no. of transactions and ticket size





Average PV: (Total Purchase Volume for Ownership Category)/(Number of Unique Cardholders)
Average Transactions: (Total number of transactions for Ownership Category)/(Number of Unique Cardholders)
Average Ticket size: Average Spend per transaction for that Category

Wednesday has maximum transactions, followed by Saturday



Average Transactions: (Total number of transactions for Ownership Category)/(Number of Unique Cardholders)

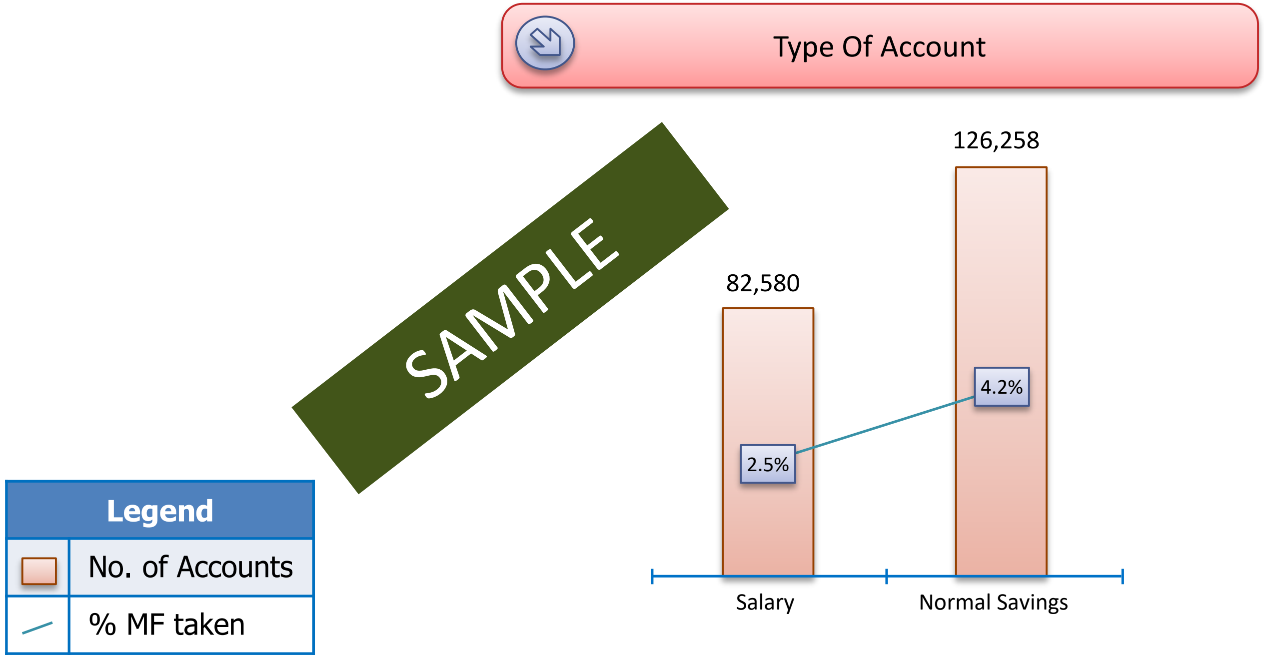
Univariates with Target variable

- The next set of visualisations should be the univariate graphs with target variables
- Each graph should show the population distribution and the target variable distribution
- A disproportionate reading on a category compared to the population is a sign of an interesting variable
- Often examining this set of visualisations leads to ideas about creating newer derived variables
- Once newer variables have been created, they need to be visualised as well

Univariate Visualisation example

Effect of type of account

Salaried customers re conservative in investment patterns, compared to Saving customers

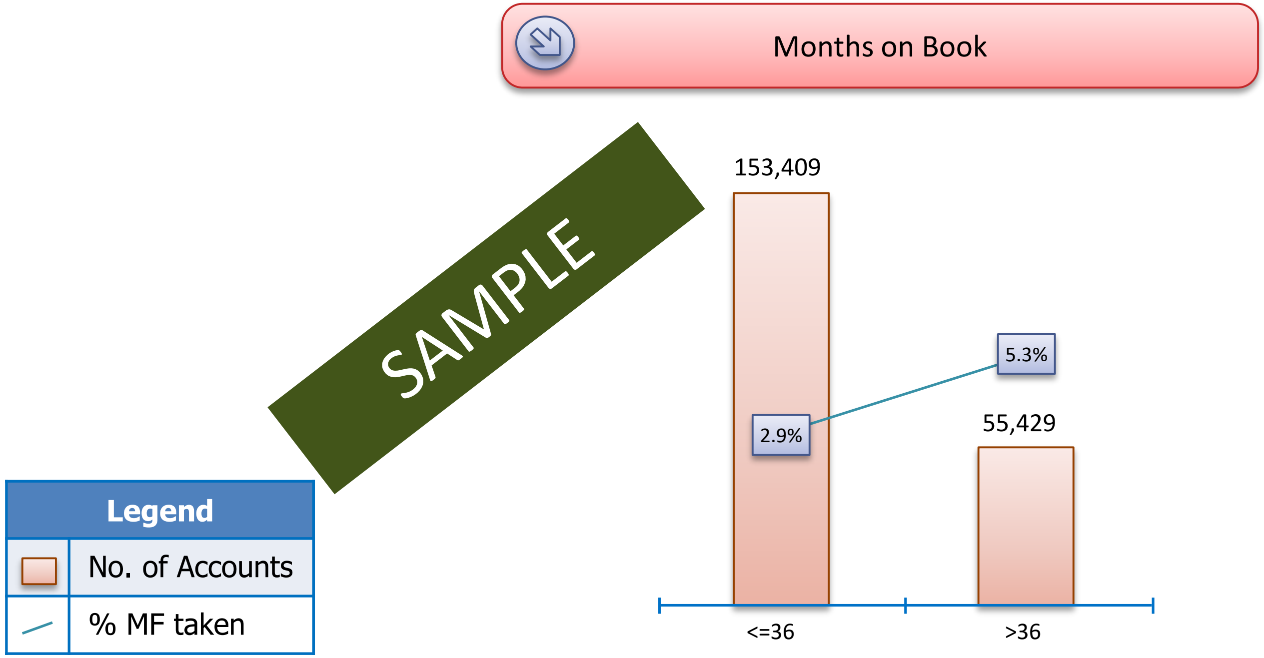


Based on an Indian Bank

Univariate Visualisation example

Effect of of time on books of the bank

Longer the period since the customer is on book, greater is the chance of investing in Mutual Funds

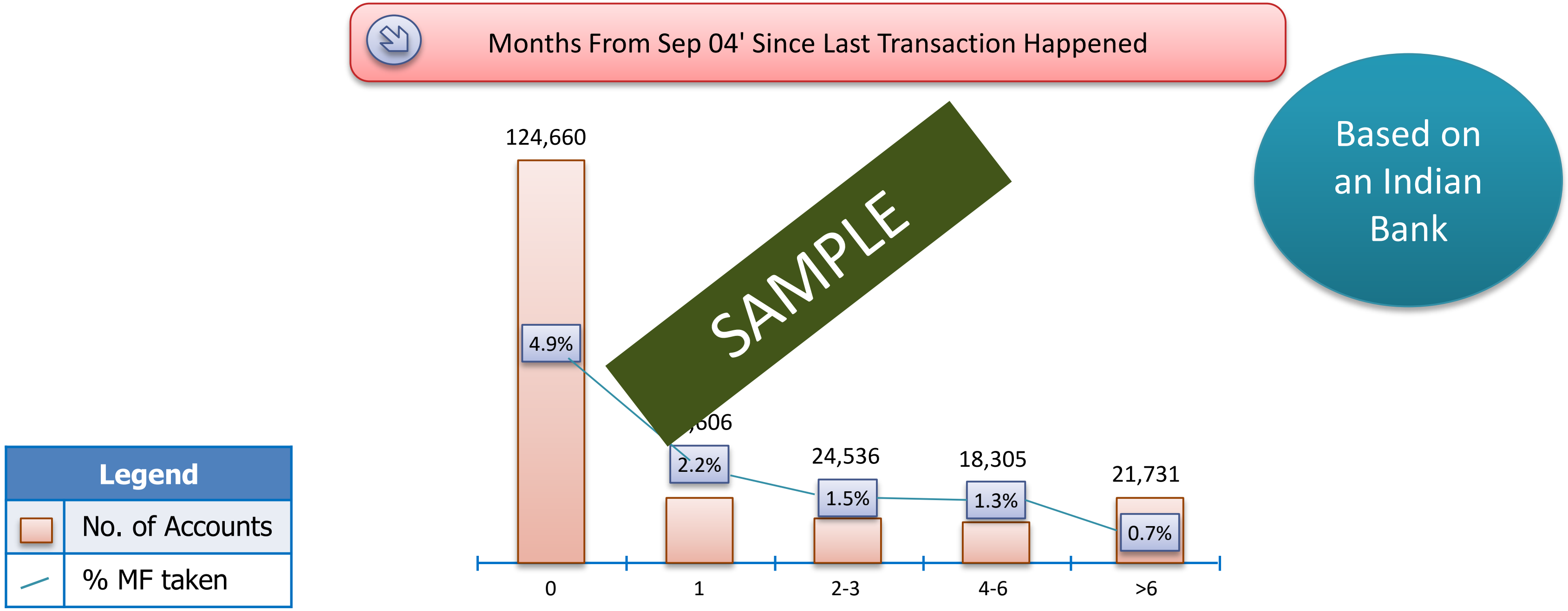


Based on an Indian Bank

Univariate Visualisation example

Effect of latency between transactions

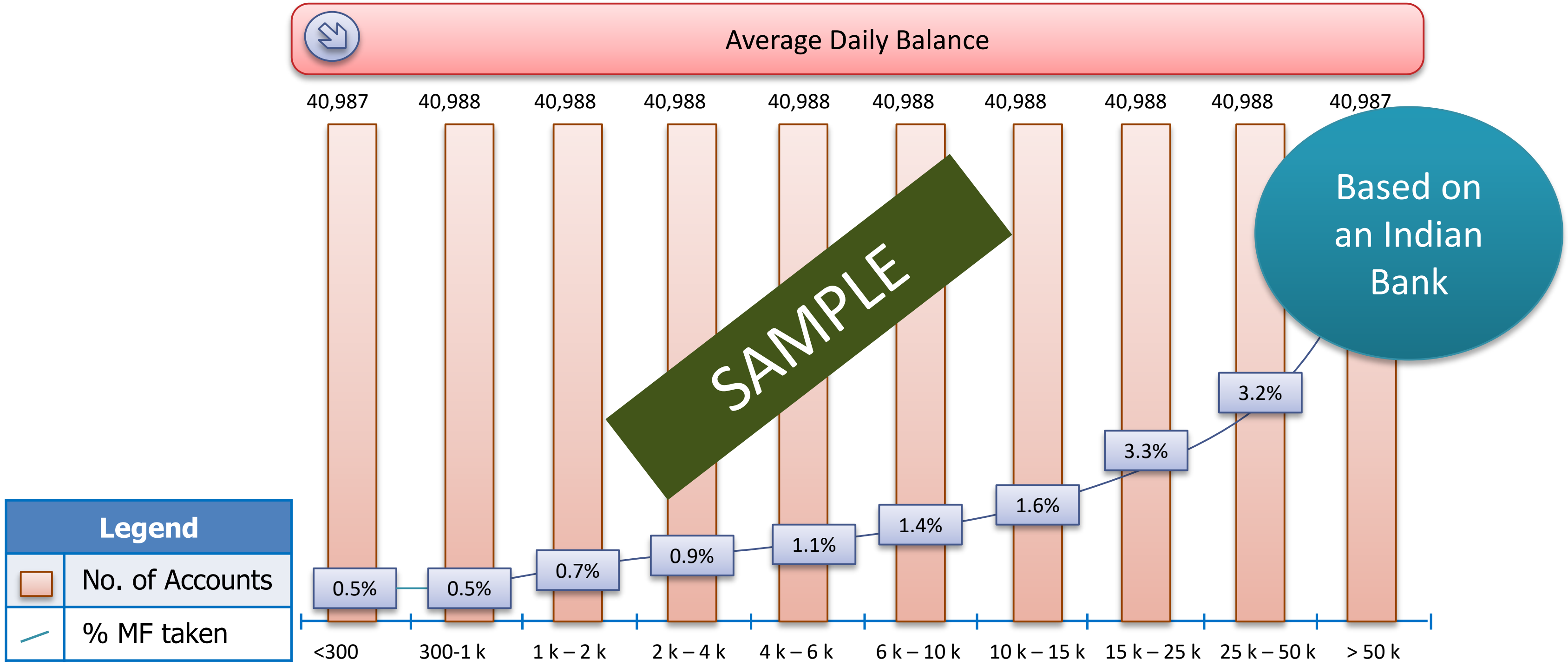
Longer the period between two transactions, lesser is the likelihood of taking up Mutual Fund as it indicates lower level of involvement



Univariate Visualisation example

Effect of increasing balance

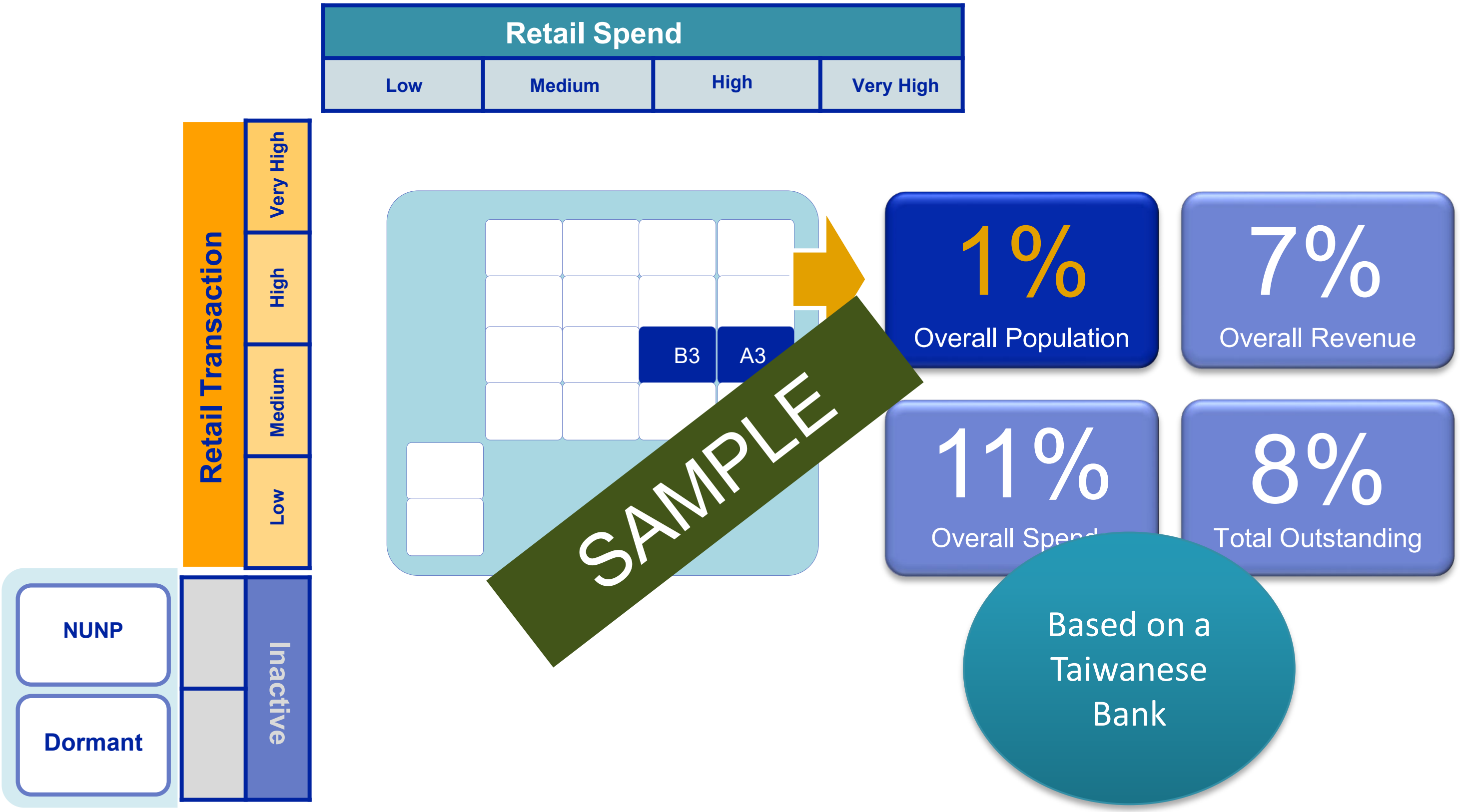
With increasing balance, there is more disposable funds and hence greater propensity to invest in Mutual Fund



Bivariate Cross tabs

- Cross tabs across two variables shows multiple trends that can reveal interesting insights
- Profiling the segments that emerge from such cross tabs could provide interesting insights
- Often just these segment analysis could provide enough insights for business teams to take action upon

Sample High Spenders Segment



Example Cross tabs

		Dimension 2		
		L	M	H
Dimension 1	1	L1 0% 0%	M1 0.26% 0.83%	H1 0.10% 0.19%
	2	L2 2.68% 7.99%	7.49% 14.41%	H2 2.86% 1.87%
	3	L3 56.82% 63.75%	M3 4.3% 2.96%	H3 25.50% 8.00%

TXN AMT %
TXN FREQ %

SAMPLE

- L3 contributes the most to both transaction amount and transaction frequency
- H3 contributes majorly to transaction amount following L3
- Other segments do not contribute significantly

NUS National University of Singapore
ISS INSTITUTE OF SYSTEMS SCIENCE

Feature Engineering

Data Analytics Best Practices

Some Feature categories to consider

- Aggregations
- Benchmarks vs. Aggregations
- Slices of time
- Other dimension slices
- Momentum/Rate of Change
- Industry classifications
- Cross tabs of two variables

Feature Construction

- Decomposing compound features into simpler components, e.g.…

ID	Product Holdings	Purchased Service
1.	ProdA + ProdC	Y
2.	ProdB + ProdC	N
3.	ProdA + ProdD	N
4.	ProdB + ProdD	Y
…		

ProdA	ProdB	ProdC	ProdD	Svc
1. 1	0	1	0	Y
2. 0	1	1	0	N
3. 1	0	0	1	N
4. 0	1	0	1	Y

Feature Construction

- Deriving a value that is more useful / making something more explicit
- E.g.

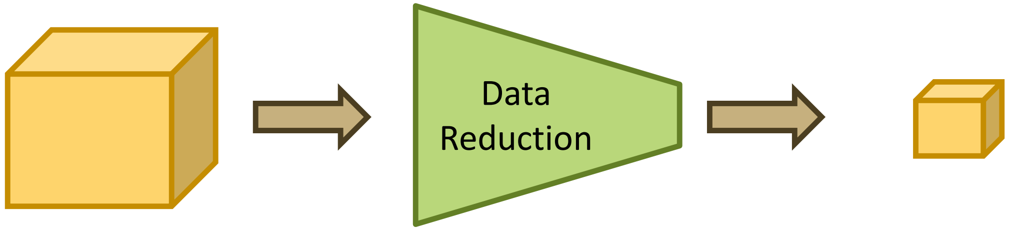
ID	Cost per unit	Units purchased
1.	10	10
2.	15	5
3.	8	8
4.	10	5

⇩

ID	Cost per unit	Units purchased	Total $ Revenue
1.	10	10	100
2.	15	5	75
3.	8	8	64
4.	10	5	50

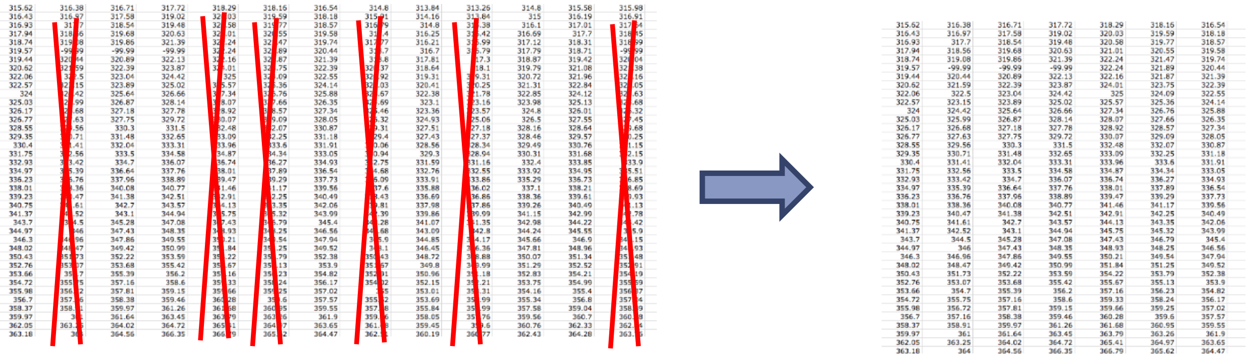
Data Reduction

- Complex data analytics may take a very long time to run on the complete data set
- Data Reduction
  - Obtain a reduced representation of the data set that is much smaller in volume yet produces the same (or almost the same) analytical results
- Data Reduction Strategies
  - Dimensionality reduction—reduce the number of attributes
  - Numerosity reduction – reduce by finding alternate, smaller data representations
    - Parametric methods: - fit data into models, store model parameters, discard the data
    - Non-parametric methods - histograms, clustering, sampling

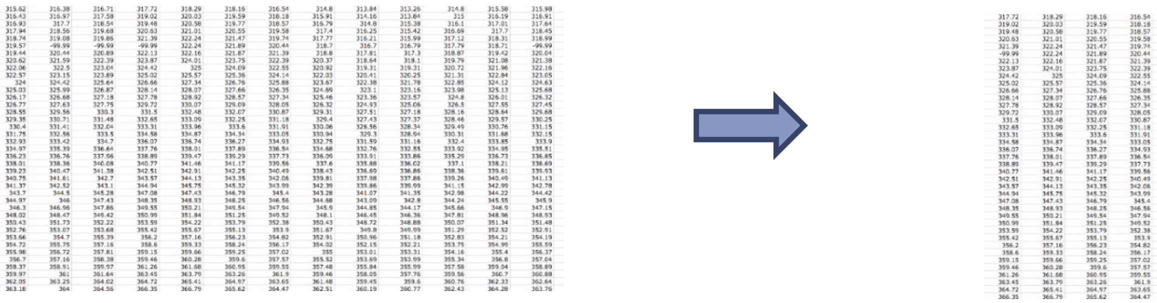


Dimensionality Reduction

- **Feature Selection** (attribute subset selection)
  - Selecting the most relevant attributes



- **Feature Extraction**
  - Combining attributes into a new reduced set of features



Original Data Reduced Data

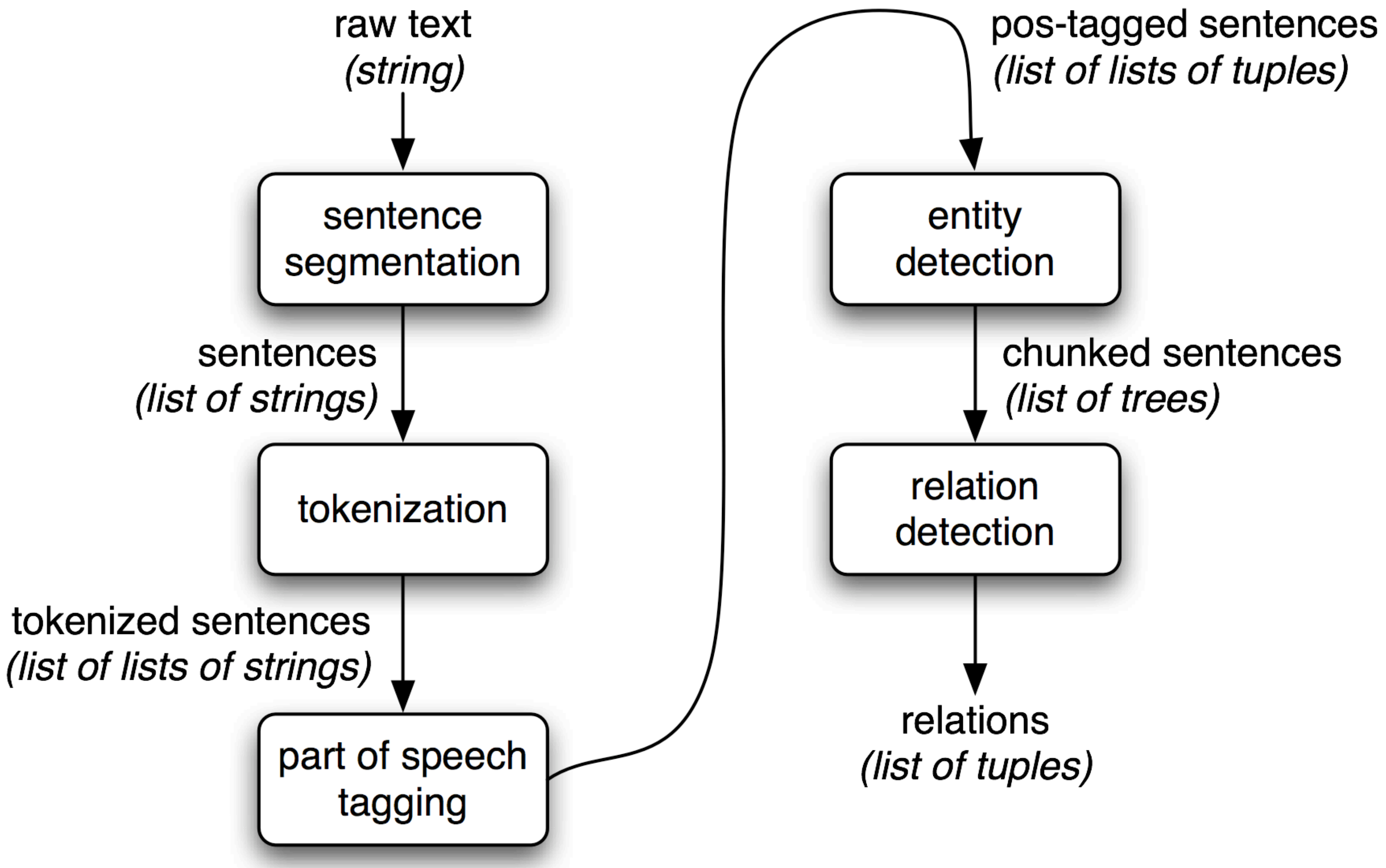
Feature Extraction

- Also attribute reduction process by combining the original attributes
- Leading to a much smaller and richer set of attributes
- Methods exist which work well for linear between-variable relationships
  - Principle component analysis
  - Factor analysis

Unstructured data processing

- Speech data processing
  - Speech data is considered as unstructured data, Step 1 in analysis is to convert it to structured data
  - Speech data could be converted to text and text could be further processed
  - Alternatively speech data could be mined for affect (tone of voice)
- Text processing
  - Before text data can be mined for insights, a lot of pre processing needs to be in place
  - Steps in pre processing include tokenisation, POS tagging, Named Entity Recognition
  - Insights could be mined using rule based approaches or machine learning techniques
  - Applications include polarity analysis, concept extraction and sentiment mining
- Example: customer service conversation transcripts

A typical text processing pipeline



Logical Data Models

- Definition
  - A repository of data dictionaries, raw fields, transformed variables, metadata and transformation scripts
- Benefits
  - Enables maintenance and easy deployability of multiple analytics data processing workflows
  - Enables best practices to be transferred across multiple teams
- Maintenance
  - Every time new processing pipelines are created, the data model gets appended

Feature expansions

- Statistical features
  - Averages
  - Variances
  - Ranges
  - Distances
- Combination features
  - Cross tabs
  - Categorisations
  - Multiple Hierarchies
- Benchmark features
  - Deviation from average
  - Direction of deviation
  - Count of deviations

Decision Engineering

Data Analytics Best Practices

Decision Engineering

- Decision Engineering is the process of converting the analytics insights into decisions
- Decision Engineering could be used to implement policies
- Example: Collection strategy: From a behavioural score and amount at risk matrix, arrive at collection policy
  - Type 1 to Type 4 treatment could vary in decreasing intensity of follow up.

2 - Variable Heat Map

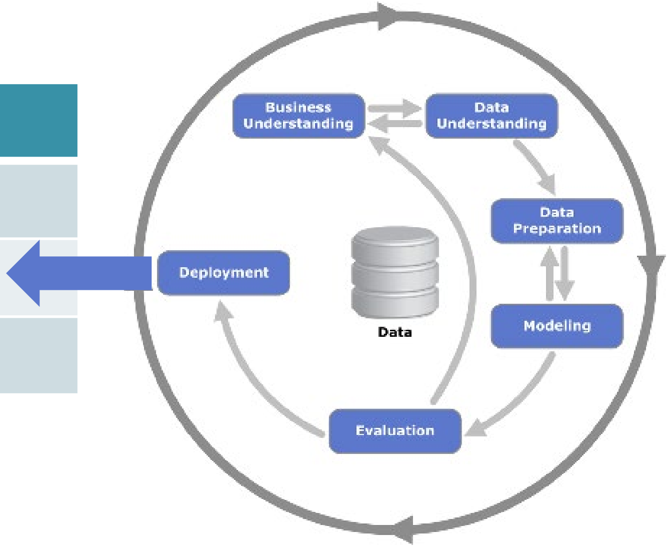
Embed into processes

Application Scorecard

Application Score	Decision Matrix	Process
Below 350	Automatic Rejection	Straight through processing
350 – 650	Manual Review	Review by Credit Department
Above 650	Automatic Acceptance	Straight through processing

Anatomy of a Decision Making Unit

Component	Function
Sensors	Data Collection and Processing
Brain	Model Scoring
Actuators	Decision Execution Processing



Model Deployment

Data Analytics Best Practices

What if we could predict cash outs?

- ✓ Cash-outs down by 80 %
- ✓ 30,000 hours of customer wait time eliminated
- ✓ Trips required to reload network down by 20 %
- ✓ Leftover cash returned to the bank decreased by 40 %

- ✓ 1,100 ATMs with optimized operations
- ✓ 4 million customers spared inconvenience

Deployment needs infrastructure

Modify Process

Collect Route specific Information

Collect Data across ATMs

Process Collected Data

Test Results

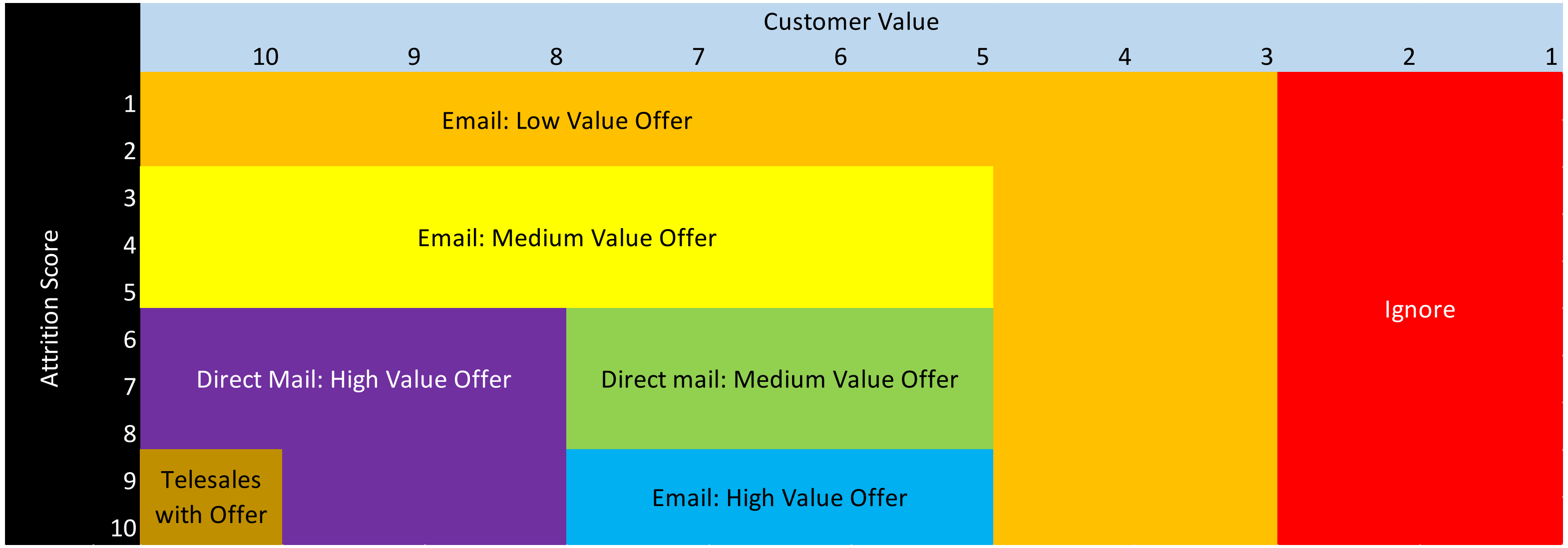
Optimise Schedule

Predict Cash-out per ATM

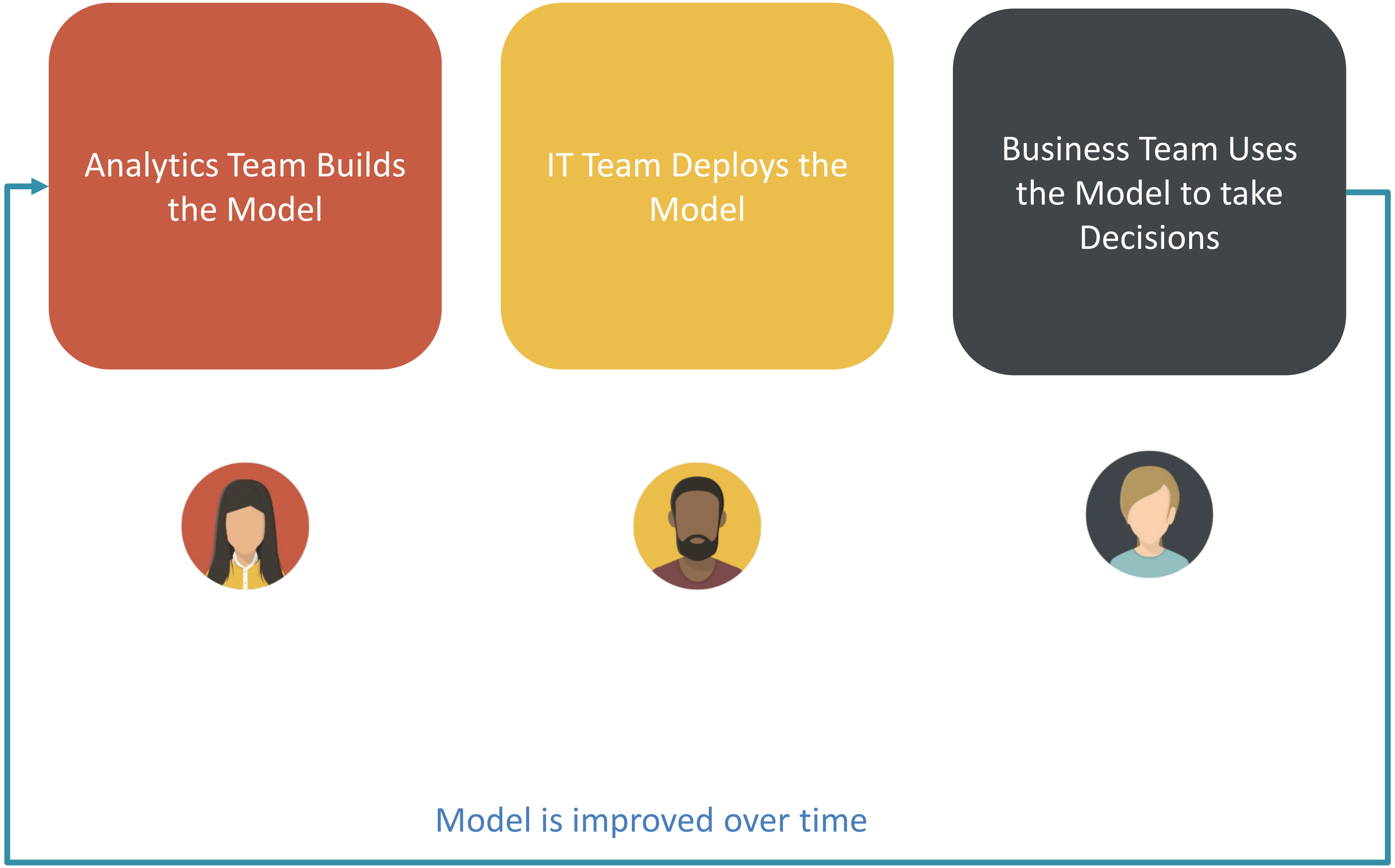
Predict Traffic per ATM

Deployment strategy could involve multiple models

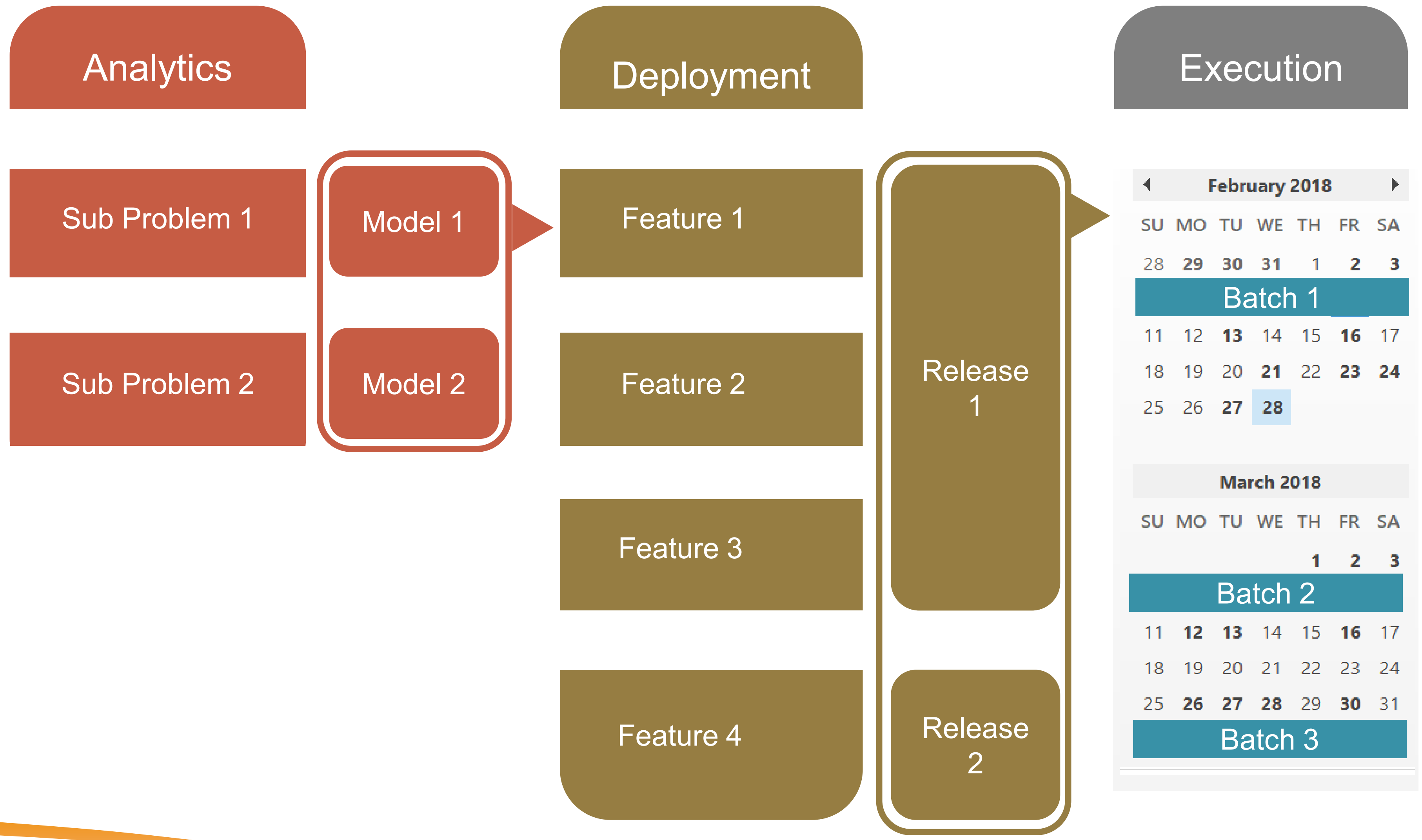
- Implementation of analytics based treatment should be based on sound business logic
- Here is an example treatment strategy for attrition prevention



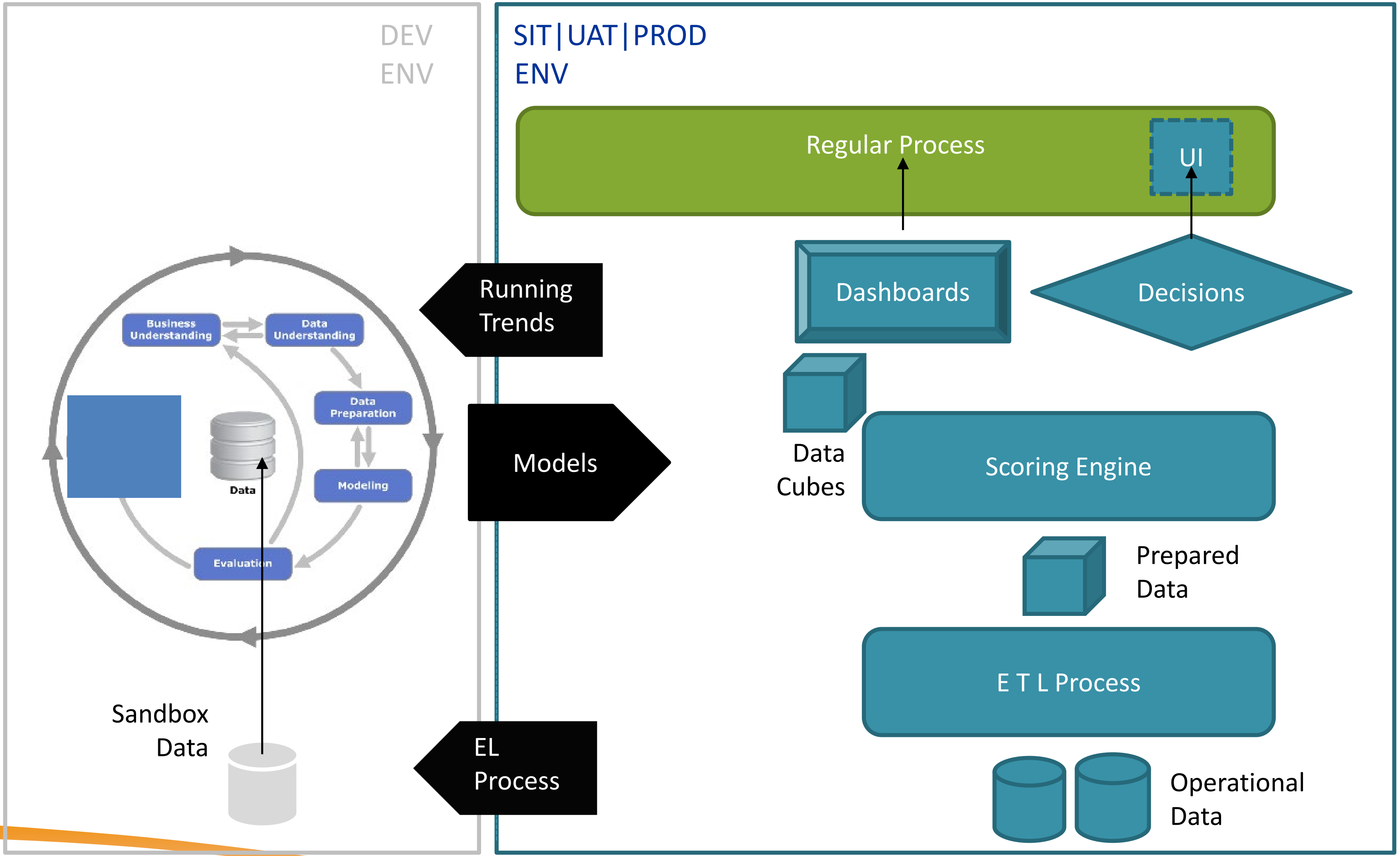
The Entire Handover



End to End cycles



Implementation Lifecycle



Tools required at various stages

- Sandbox Data, Prepared Data, Operational Data
  - Databases: Oracle, Hadoop
- ETL Process, EL Process
  - Data Wrangling systems: SQL, Informatica
- Data Preparation
  - Data Preparation workbenches: R, SAS
- Modelling
  - Modelling Workbenches: R, SAS
- Models
  - Model Formats: XML, JSON
- Scoring Engine
  - Rule Engines: SQL Scripts, Drools
- Dashboards, Running Trends
  - Reporting Engines: Tableau, Qlikview
- Data cubes
  - OLAP Databases: Tableau, Oracle
- UI (Optional)
  - User Interface: Web/Mobile Apps

Model Maintenance

Data Analytics Best Practices

Population stability Index

- If the underlying population has changed, we might need to recalibrate the model
- A threshold on the population stability index is tracked to determine this
- In this example, the PSI crossed the review threshold, the deviation was from the web channel
- Look into the details of why this happened before taking further action

Channel	# records last 3 months	# records last month	% prev 3 months	% recent month	change	ratio	WoE	PSI portion
Social	6,000	2,200	7.6%	13.0%	5.4%	1.71	0.534	0.029
Web	25,000	1,600	31.7%	9.5%	-22.3%	0.30	-1.211	0.270
Email	3,000	900	3.8%	5.3%	1.5%	1.40	0.333	0.005
Print	3,500	977	4.4%	5.8%	1.3%	1.30	0.261	0.003
InStore	41,252	11,250	52.4%	66.5%	14.1%	1.27	0.238	0.034
Total	**78,752**	**16,927**	**100%**	**100%**				**0.341**

PSI	Strategy
< 0.1	No change
0.1 to 0.25	Closely Monitor
> 0.25	Review

Population stability Index for a credit score

- PSI should be calculated across score bands too

Credit Score	# records last 3 months	# records last month	% prev 3 months	% recent month	change	ratio	WoE	PSI portion
<500	6,234	2,200	15.1%	24.1%	9.0%	1.59	0.465	0.042
500-599	7,780	1,600	18.9%	17.5%	-1.4%	0.93	-0.075	0.001
600-699	5,700	900	13.8%	9.9%	-4.0%	0.71	-0.339	0.013
700-799	6,320	977	15.3%	10.7%	-4.6%	0.70	-0.360	0.017
800-850	8,700	1,325	21.1%	14.5%	-6.6%	0.69	-0.375	0.025
>850	6,500	2,134	15.8%	23.4%	7.6%	1.48	0.393	0.030
Total	**41,234**	**9,136**	**1**	**1**				**0.127**

PSI	Strategy
< 0.1	No change
0.1 to 0.25	Closely Monitor
> 0.25	Review

Characteristic Analysis Report

- Stability index

$$\sum(\%Actual - \%Expected) \times \ln\left(\frac{\%Actual}{\%Expected}\right)$$

- Characteristic reports

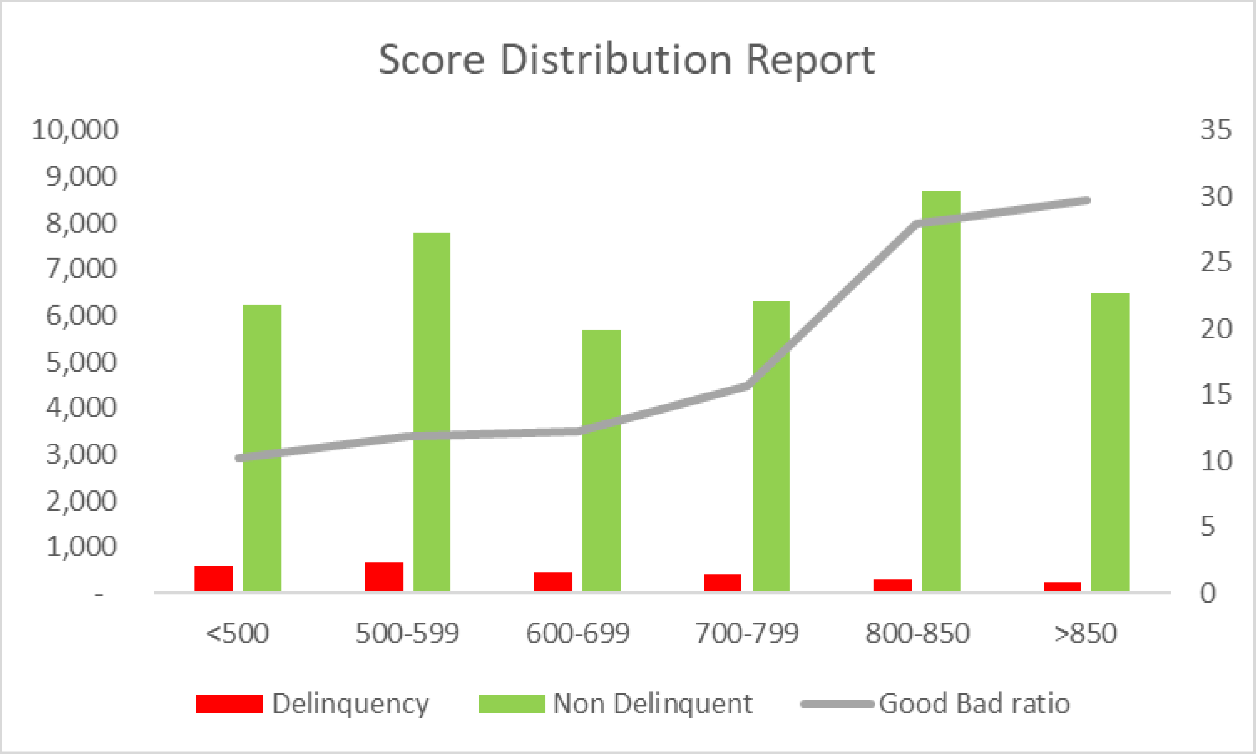
Age	Expected	Actual	Points	Index
18–24	12%	21%	10	0.9
25–29	19%	25%	15	0.9
30–37	32%	28%	25	−1
38–45	12%	6%	28	−1.68
46+	25%	20%	35	−1.75
				−2.63
Time at Res				
0–6	18%	29%	12	1.32
7–18	32%	32%	25	0
19–36	26%	22%	28	−1.12
37+	24%	17%	40	−2.8
				−2.6
Region				
Major Urban	55%	58%	20	0.6
Minor Urban	26%	24%	25	−0.5
Rural	19%	18%	15	−0.15
				−0.05
Inq 6 mth				
0	63%	34%	40	−11.6
1–3	19%	31%	30	3.6
4–5	10%	16%	15	0.9
6+	8%	19%	10	1.1

# Delq	Expected	Actual	Points	Index
0	80%	65%	45	−6.75
1–2	12%	21%	20	1.8
3–5	5%	8%	12	0.36
6+	3%	6%	5	0.15
				−4.44
Utilization at Bureau				
0	12%	8%	15	−0.6
1–9	10%	19%	40	3.6
10–25	14%	20%	30	1.8
26–50	22%	25%	25	0.75
50–69	11%	6%	20	−1
70–85	13%	9%	15	−0.6
86–99	14%	8%	10	−0.6
100+	4%	5%	5	0.05
				3.4

Source: https://www.slideshare.net/VijayDesai9/transactionbasedanalytics2010

Scorecard performance report

- The scorecards are analysed against actual performance to see if they are working well



Maintenance options

- Score shelf life varies across different scores
  - Fraud scores have a lower shelf life as fraudsters change techniques
  - Credit scores are relatively stable
  - Response models are likely to be less stable as market conditions change often
- Score deterioration options
  - Recalibration
    - Inexpensive
    - Remap old score to new score
  - Retraining
    - More expensive
    - Keep same variables, change weights
  - Rebuilding
    - Most expensive
    - Rebuild the model from scratch

ROI Models

Data Analytics Best Practices

Take costs into account

Predicted

Model A	No Infection	Infection
Actual No Infection	630	50
Infection	170	150

Predicted

Model B	No Infection	Infection
Actual No Infection	480	200
Infection	70	250

22% misclassificaion

27% misclassificaion

Predicted

Actual	No Infection	Infection
No Infection	$0	$2,000
Infection	$10,000	$0

$1.8 Million

$1.1 Million

Take costs into account

Model A	Transactions not flagged	Transactions Flagged suspicious
Normal cases	600,000	80,000
Actual Laundering cases	170	150

Model B	Transactions not flagged	Transactions Flagged suspicious
Normal cases	480,000	200,000
Actual Laundering cases	30	210

12% misclassification

29% misclassification

Cost of misclassification	Transactions not flagged	Transactions Flagged suspicious
Normal cases	-	$ 1,000
Actual Laundering cases	$ 1,000,000	-

$250 Million

$230 Million