

# **Detecting Money Laundering Actions Using Data Mining and Expert Systems**

**Ekrem Duman**

**Dogus University**

**Industrial Engineering Department**

**Istanbul, TURKEY**

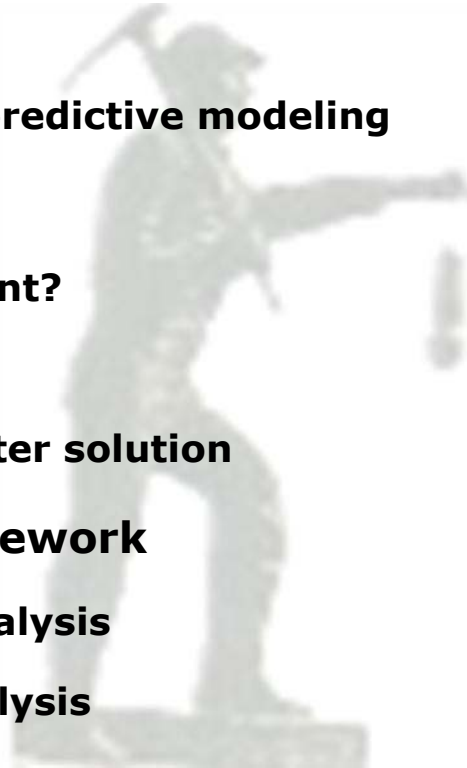
**Joint work with Ayse Buyukkaya**



# Outline

---

- ✓ **Our experience in Banking**
  - ✓ **Introduction**
  - ✓ **Success Factors in predictive modeling**
- ✓ **Motivation for AML**
  - ✓ **Why AML is important?**
  - ✓ **Existing solutions**
  - ✓ **Motivation for a better solution**
- ✓ **Suggested AML Framework**
  - ✓ **Exploratory data analysis**
  - ✓ **Inferential data analysis**
  - ✓ **Expert system**



# Introduction

---

## Basic Objectives in Using DM:

### 1. Descriptive

- ✓ Clustering / Segmentation
- ✓ Basket (association) analysis
- ✓ Sequence (pattern) analysis

### 2. Predictive

- ✓ Classification
- ✓ Time series analysis - regression

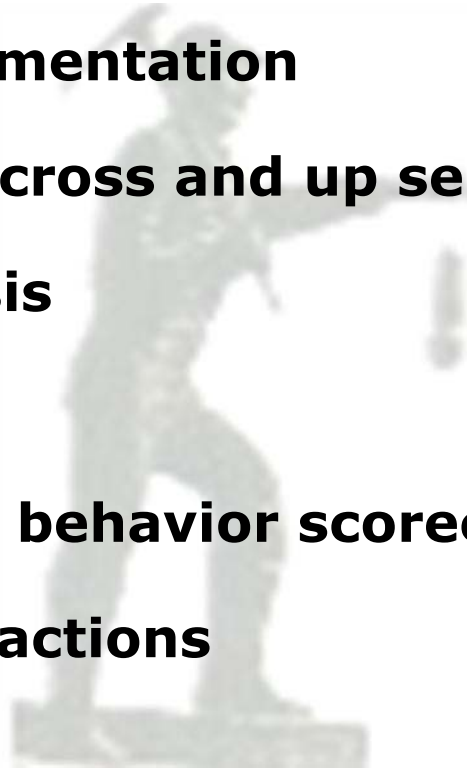


# Introduction

---

## Some of the DM projects we made:

- ✓ **Clustering / Segmentation**
- ✓ **Product specific cross and up sell models**
- ✓ **Sequence analysis**
- ✓ **Customer churn**
- ✓ **CC, Loan and OD behavior scorecards**
- ✓ **Suspicious transactions**



# Success Factors

---

**Some technical factors that can affect the success of a project:**

- ✓ **The adequacy of data**
- ✓ **Selecting input variables**
- ✓ **The way of using input variables**
- ✓ **Forming the training set**
- ✓ **Determining model application period**
- ✓ **Algorithm selection**
- ✓ **Determining model assessment criteria**
- ✓ **Feeding the campaign results back**

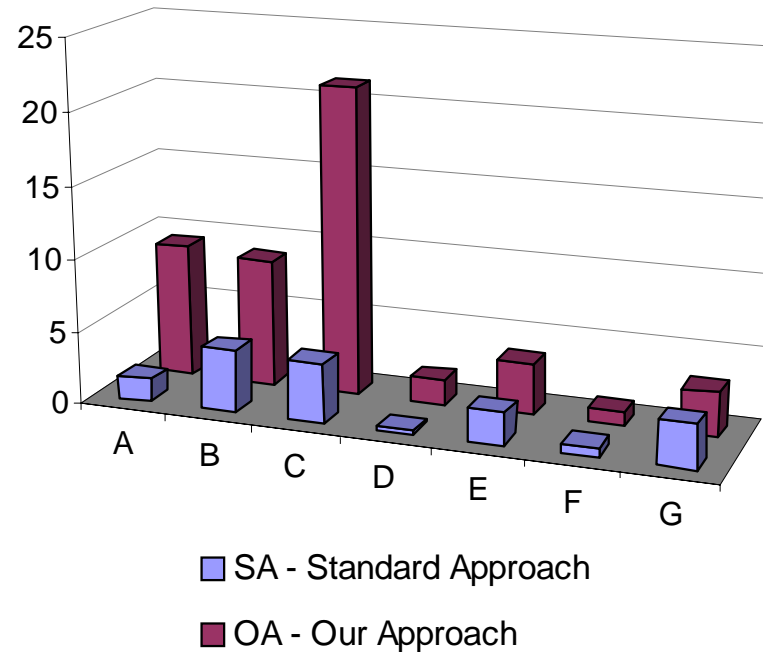
# Success Factors

Everything is the same but two factors are handled differently;

	80% SA %*	80% OA %*	80% Ratio
PROD A	1.6	9.2	5.7
PROD B	4.3	8.8	2.1
PROD C	4.1	21.3	5.2
PROD D	0.3	1.7	5.1
PROD E	2.3	3.6	1.5
PROD F	0.7	1.0	1.4
PROD G	3.0	3.1	1.0
<b>Average</b>	2.3	6.9	<b>3.2</b>

<b>Avg Target</b>	10.578	10.039	
-------------------	--------	--------	--

\* All models are built by CART



\*The percentage of actual buyers in the model target list within the following month

**Need for developing specific models for each product!**

# Success Factors

---

**Some technical factors that can affect the success of a project:**

- ✓ **The adequacy of data**
- ✓ **Selecting input variables**
- ✓ **The way of using input variables**
- ✓ **Forming the training set**
- ✓ **Determining model application period**
- ✓ **Algorithm selection**
- ✓ **Determining model assessment criteria**
- ✓ **Feeding the campaign results back**

# Success Factors

---

## The adequacy of data

- ✓ **Do we have the necessary and sufficient variables in data mart?**
- ✓ **Are the data complete?**
  - ✓ **How should we handle the missing data?**
- ✓ **What is the data quality?**



# Success Factors

---

## Selecting input variables

- ✓ **Few variables** → interpretable, fast, less accurate (?) models
- ✓ **Moderate number of variables** → ?
- ✓ **Many variables** → non-interpretable, slow, highly accurate (?) models

**How can we determine the right variables?**

**Our Experience:**

**NN** ⇒ few variables

**DT** ⇒ many variables

# Success Factors

---

## The way of using input variables

- ✓ In original form?
- ✓ Transformed?
  - ✓ Log
  - ✓ Categorization
  - ✓ Other

**Our Experience:**

Clustering ⇒ Categorization  
Cross Sell ⇒ Original

Average balance in current account

Range(\$)	Value
0-1	0
1-100	1
100-1000	2
1000-10000	3
10000+	4

# Success Factors

---

## Forming the training set

- ✓ **Who should take place in the training set?**
  - ✓ **Whole customer base?**
  - ✓ **Some clusters only?**
    - ✓ **How to cluster customers?**
- ✓ **What should be the ratio of positive to negative samples in the training set?**

**Our Experience:**

**Cross Sell**     $\Rightarrow$     **1 - 1**

**Scorecard**     $\Rightarrow$     **1 - k (k > 1)**

# Success Factors

---

## Determining model application period

- ✓ **What is the campaign period?**
- ✓ **What is to be done for campaign timing to meet the needs of business?**
- ✓ **In which period the customer behavior will be analyzed?**

**Our Experience:**

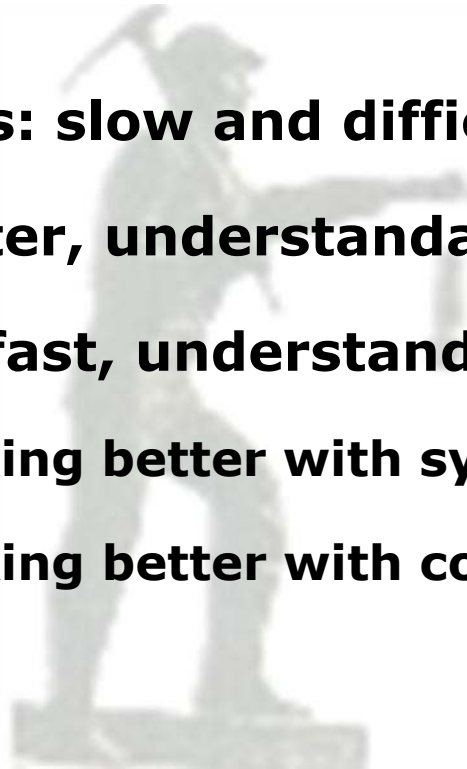
<b>Asset</b>	⇒	<b>1 month</b>
<b>Liability</b>	⇒	<b>3 months</b>

# Success Factors

---

## Algorithm selection

- ✓ **Neural networks: slow and difficult to understand**
- ✓ **Regression: faster, understandable**
- ✓ **Decision trees: fast, understandable**
  - ✓ **The ones working better with symbolic variables**
  - ✓ **The ones working better with continuous variables**



# Success Factors

---

## Determining model assessment criteria

- ✓ ~~Accuracy (confusion) matrix – ROC~~
- ✓ Hit rate
- ✓ Capture rate
- ✓ LIFT

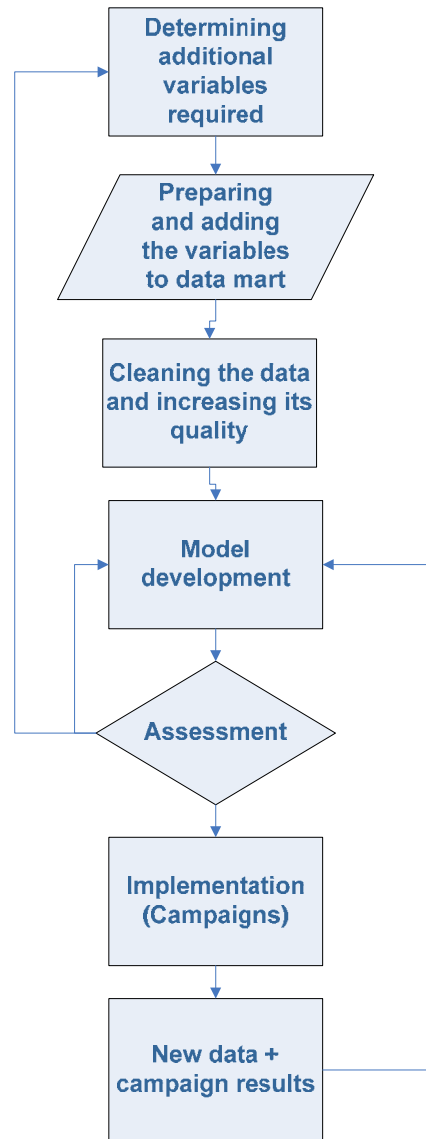


# Success Factors

Algorithm	Conf.	Hit %	Capture %	LIFT	Accuracy (0, 1)	
CART 7	70	16.2	41.8	8.3	54	79
C50	70	15.8	48.5	8.0	74	74
ECHAID7	70	19.1	37.9	9.7	67	68
QUEST	73	16.6	29.7	8.5	83	46
C50	80	13.3	5.8	6.8	74	74
ECHAID7	80	21.9	15.9	11.2	67	68
<i>Avg (70%)</i>	<i>71</i>	<i>16.9</i>	<i>39.5</i>	<i>8.6</i>	<i>70</i>	<i>67</i>
<i>Avg (80%)</i>	<i>80</i>	<i>17.6</i>	<i>10.8</i>	<i>9.0</i>	<i>71</i>	<i>71</i>

# Success Factors (CROSS SELL)

## Feeding the campaign results back



# Success Factors

After some improvement studies:

	Old			New				
	80%	80%	80%	80%	90%		80%	90%
	SA %	OA %	Ratio	OA %	OA %		Ratio	Ratio
PROD A	1.6	9.2	5.7	9.3	40.8		5.8	25.5
PROD A-12	1.0			6.8	34.6		6.8	34.6
PROD B	4.3	8.8	2.1	25.7	46.5		6.0	10.8
PROD B-6	1.1			10.0	27.8		9.1	25.3
PROD C	4.1	21.3	5.2	28.5			7.0	
PROD D	0.3	1.7	5.1	0.8	1.6		3.1	6.2
PROD E	2.3	3.6	1.5	2.6	3.6		1.2	1.6
PROD F	0.7	1.0	1.4	2.2	5.5		3.2	8.1
PROD G	3.0	3.1	1.0	8.2	17.3		2.7	5.8
<b>Average</b>	2.0	6.9	<b>3.2</b>	<b>10.5</b>	<b>22.2</b>		<b>5.0</b>	<b>14.7</b>

# Some Campaigns We Made

---

## First Term Deposit (TD) Model and Campaign:

- ✓ The customers who did not have a TD are targeted
- ✓ Sales/reached = 26% (lift effect = 45)
- ✓ Avg account size = 2.5 times the general avg

## Second Term Deposit Model and Campaign:

- ✓ The customers who did not have a TD are targeted
- ✓ Sales/reached = 40% (lift effect = 98)
- ✓ Avg account size = 2.5 times the general avg

# Some Campaigns We Made

---

## Bill Payment Order Campaign:

- ✓ The customers who did not have a payment order before are targeted
- ✓ Sales/reached = 6% (lift effect = 30)
- ✓ One more bill payment order, differentiated by the sales to lower segments

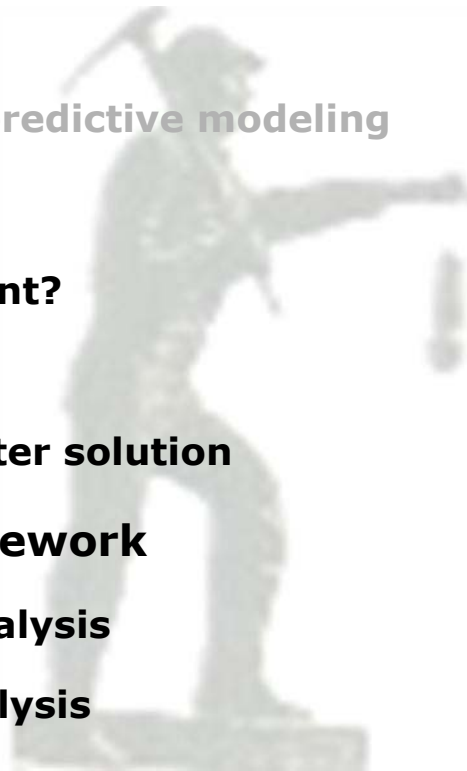
## Overdraft Campaign:

- ✓ Sales/reached = 38% (lift effect = 55)
- ✓ The actual users in the first three months are 12% more than the general average

# Outline

---

- ✓ **Our experience in Banking**
  - ✓ Introduction
  - ✓ Success Factors in predictive modeling
- ✓ **Motivation for AML**
  - ✓ Why AML is important?
  - ✓ Existing solutions
  - ✓ Motivation for a better solution
- ✓ **Suggested AML Framework**
  - ✓ Exploratory data analysis
  - ✓ Inferential data analysis
  - ✓ Expert system



# Motivation for AML

---

- ✓ **Terrorism is the main threat to everybody.**
- ✓ **Before 9/11 it was related to underdeveloped countries.**
- ✓ **After 9/11 it was understood that it threatens everybody**
- ✓ **One of the ways of struggling terrorism is to block their financial activities.**
- ✓ **Terrorism finance is a type of black money and AML (anti money laundering) techniques can be used to combat it.**

***(e.g. Kurdish terrorist group PKK and its revenues from illegal drugs)***

# Motivation for AML

---

- ✓ **There are many commercial packages available for AML**
- ✓ **They perform standard checks**
  - ✓ **Is the account holder on OFAC list?**
- ✓ **They are mostly rule based.**
  - ✓ **Look at the (past) transactions**
  - ✓ **Identify irregularities by some predefined rules**



***"Which customers made EFT more than 50 times last month?"***

# Motivation for AML

---

## Basic deficiency of AML commercial packages:

- ✓ **Too few “AND” rules**
  - ✓ list is too big to inspect
- ✓ **Too many “AND” rules**
  - ✓ an actually fraudulent transaction/person could be missed



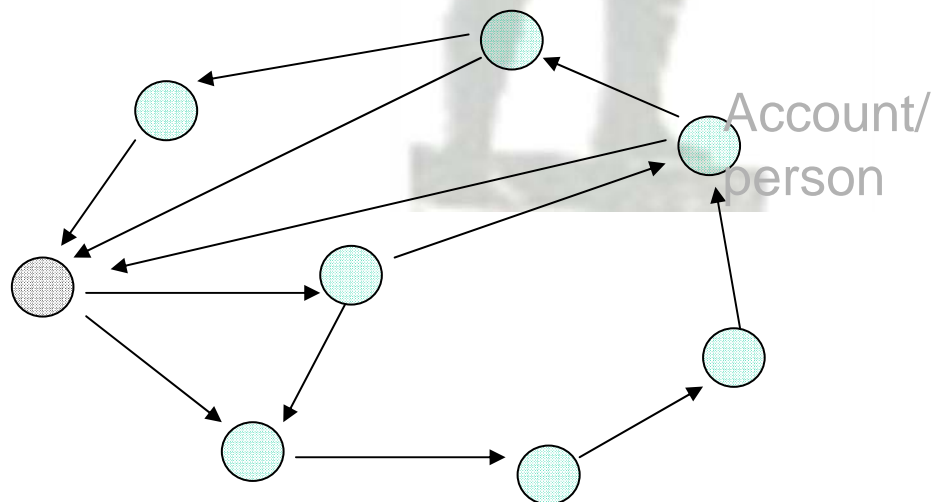


# Motivation for AML

---

## Money laundering actions can be inspected at two levels:

- ✓ **Individual account/person level**
  - ✓ **look at the (past) transactions of an account/person and identify irregularities**
- ✓ **Network Level**
  - ✓ **look at all accounts/customers and identify suspicious loops**



# Suggested Framework

---

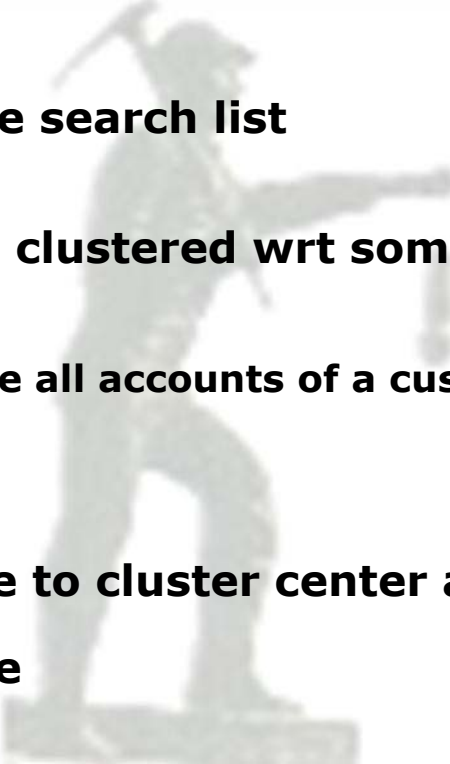
## A three phase AML solution framework:

- ✓ **exploratory data analysis**
    - ✓ **descriptive data mining (DM) to determine unusualities**
  - ✓ **inferential data analysis**
    - ✓ **predictive DM to determine cases that need to be inspected**
  - ✓ **expert system**
    - ✓ **coding the inspection process**
- 

# Suggested Framework

---

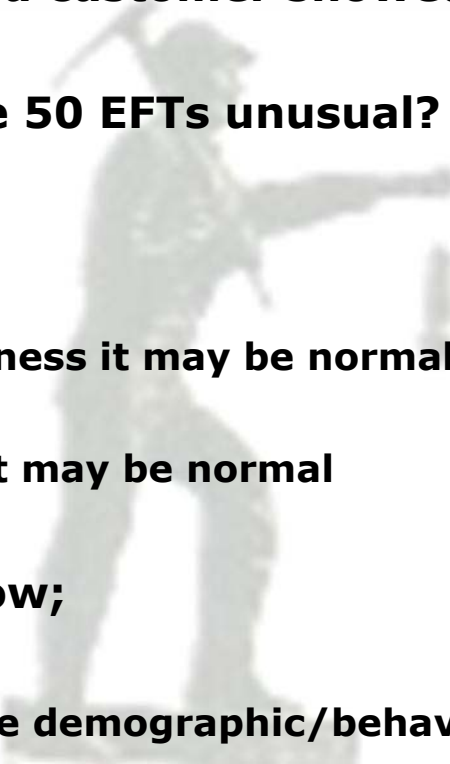
## **I - Exploratory data analysis phase:**

- ✓ **used to narrow down the search list**
  - ✓ **customers/accounts are clustered wrt some behavioral variables**
    - ✓ **we prefer to consolidate all accounts of a customer and work on customer list**
  - ✓ **Customers who are close to cluster center are taken as “normal” and not inspected anymore**
- 

# Suggested Framework

---

## **I - Exploratory data analysis phase:**

- ✓ **How can we decide that a customer showed unusual behavior?**
  - ✓ **Is a customer who made 50 EFTs unusual?**
    - ✓ **It could be...**
    - ✓ **If he owns a small business it may be normal**
    - ✓ **If he always does this it may be normal**
  - ✓ **To judge we need to know;**
    - ✓ **the average value in the demographic/behavioral segment of the customer**
    - ✓ **the routine (historic) behavior of the customer**
- 

# Suggested Framework

---

## I - Exploratory data analysis phase:

- ✓ The clustering to be made should take both aspects into account
- ✓ First aspect is handled automatically
- ✓ For the second aspect, we calculated the deviation values and used them in clustering

**Deviation = (value in last month – avg of the last six months) /  
std. dev. in the last six months**

# Suggested Framework

---

## **I - Exploratory data analysis phase:**

- ✓ **Variables should be selected in accordance with the purpose:**

BD_CA_NUM_CRD_TRX
BD_CA_NUM_CRD_TRX_DEV
BD_CA_NUM_ACCT_TL_OP
BD_CA_NUM_ACCT_TL_OP_DEV
BD_DBC_NUM_TRX
BD_DBC_NUM_TRX_DEV
BD_INT_LOGON_NUM
BD_INT_NUM_TRX
EFT_IN_TRX_NUM
EFT_OUT_TRX_NUM

# Suggested Framework

---

## I - Exploratory data analysis phase:

- ✓ For a clustering with  $k$  variables and  $n$  clusters;
- ✓ Cluster centers are determined
  - ✓ This is a point in  $k$  dimensional space which take averages of all customers in that cluster
- ✓ All customers in the cluster have some deviation from the center

**Variable Deviation <sub>$k$</sub>  = dev of the customer's variable  $k$  value  
from cluster center value**

**Customer Deviation = total of the variable deviations**

# Suggested Framework

---

## I - Exploratory data analysis phase:

**Customer anomaly index = (customer deviation) / (average customer deviation in the cluster)**

- ✓ **Customers are sorted with a non-increasing value of anomaly index values;**
- ✓ **The lower part of this list is the normal or usual part**
- ✓ **The upper part can be inspected more**

# Suggested Framework

---

## I - Exploratory data analysis phase:

An. Ind.	CI	Primary Var.	VCM	Secondary Var.	VCM
5.000	3	BD INT LOGON NUM	0.618	BD CA NUM ACCT TL	0.216
5.000	1	BD CA NUM DEB TRX DEV	0.962	BD CC NUM ACCT	0.026
4.924	3	BD CHEQUE NUM TRX DEV	0.341	BD BRANCH NUM TRX DEV	0.29
4.589	1	EFT IN TRX NUM DEV	0.979	EFT OUT TRX NUM	0.008
4.285	3	BD CHEQUE NUM TRX	0.293	BD CA NUM CRD TRX DEV	0.259
4.123	3	BD ACTIVE CARD NUM	0.936	BD CC NUM TRX	0.017
4.000	1	BD CA NUM ACCT CL DEV	0.925	BD CC NUM ACCT	0.052
3.923	3	BD CA NUM ACCT FX	0.273	BD CA NUM CRD TRX	0.169
3.758	1	BD CA NUM ACCT CL DEV	0.93	BD CC NUM ACCT	0.058
3.501	2	BD CA NUM ACCT CL DEV	0.625	BD CA NUM CRD TRX	0.1
3.284	7	BD CA MAX BAL RATE	0.551	BD CHEQUE NUM TRX	0.105
3.001	1	BD CA NUM ACCT CL DEV	0.962	BD INT LOGON NUM	0.026
2.992	4	BD CURRENCY PURCHASE NUM TRX	0.212	BD CA NUM ACCT	0.201
2.841	3	BD INT LOGON NUM DEV	0.772	BD DBC NUM ACCT	0.075
2.552	3	BD BILL NUMBER	0.763	BD CA NUM DEB TRX	0.074
2.528	4	BD CA NUM ACCT	0.198	BD CA NUM ACCT TL	0.182
2.428	3	BD BRANCH NUM TRX	0.932	BD ACTIVE CARD NUM	0.025

**Variable contribution measure (VCM) = variable deviation / customer deviation**

# Suggested Framework

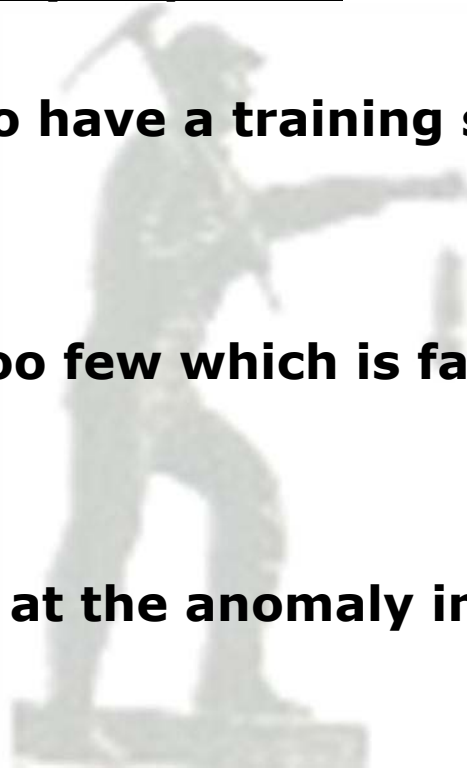
---

## II - Inferential data analysis phase:

**It would be very nice to have a training set to learn who are laundering money.**

**The known cases are too few which is far beyond being sufficient.**

**Let the inspectors look at the anomaly indexes and decide which ones to inspect;**



# Suggested Framework

---

## II - Inferential data analysis phase:

**Anomaly List:** The list produced by anomaly indexes

**Inspect List:** The ones that the inspectors see the need to inspect

**Anomaly List**

**Training Set**

**Inspect List**



# Suggested Framework

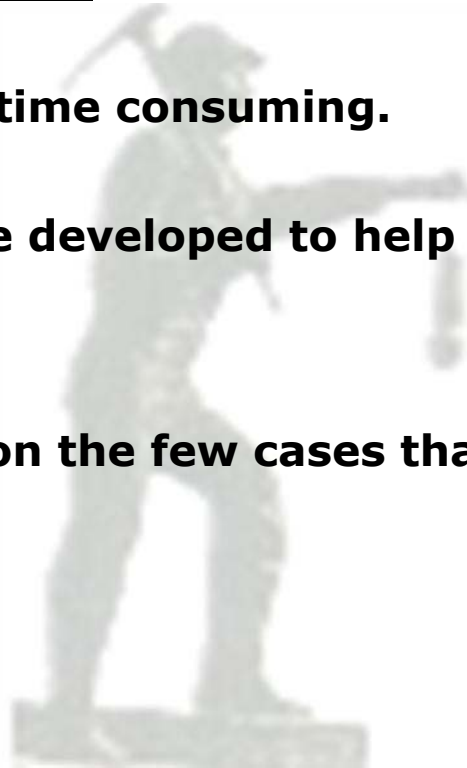
---

## III - Expert system phase:

**Inspection process is too time consuming.**

**An expert system could be developed to help the inspection process.**

**The inspectors can focus on the few cases that the expert system finds suspicious.**



# Summary and Conclusions

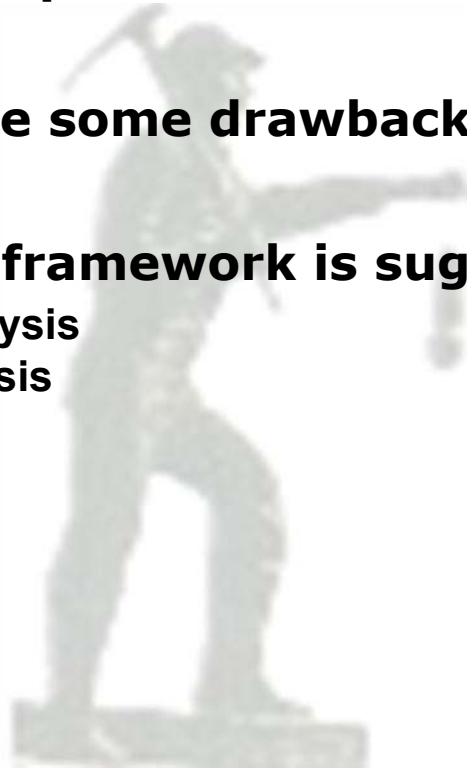
---

**Combating ML is very important.**

**Available solutions have some drawbacks.**

**A three phase solution framework is suggested.**

- exploratory data analysis
  - inferential data analysis
  - expert system

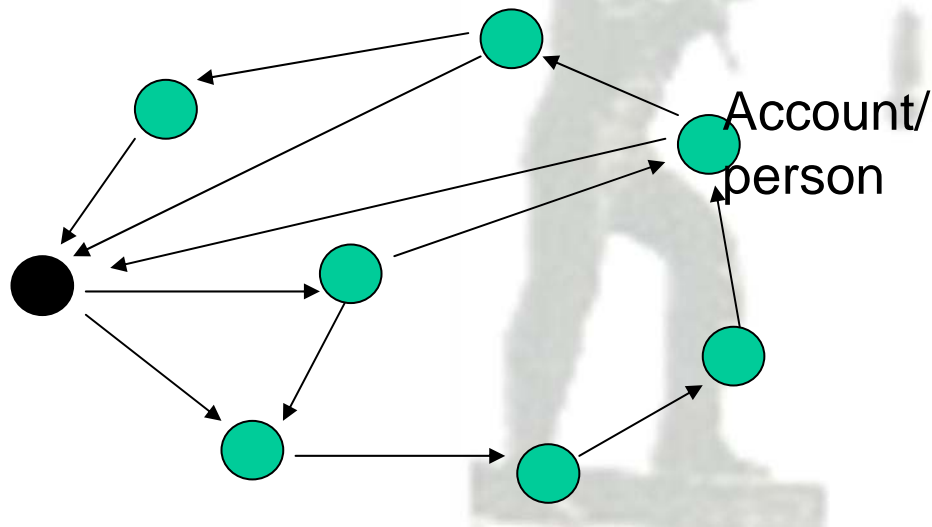


# Future Work

---

**To mature and implement our solution framework.**

**To find solutions for the network level problems.**



**We are open to collaboration...**

# Any Questions?

---



**Thank you for listening to us.**