Mining Big Data in Statistical systems of Monetary Financial Institutions

Afshin Ashofteh

www.linkedin.com/in/statas

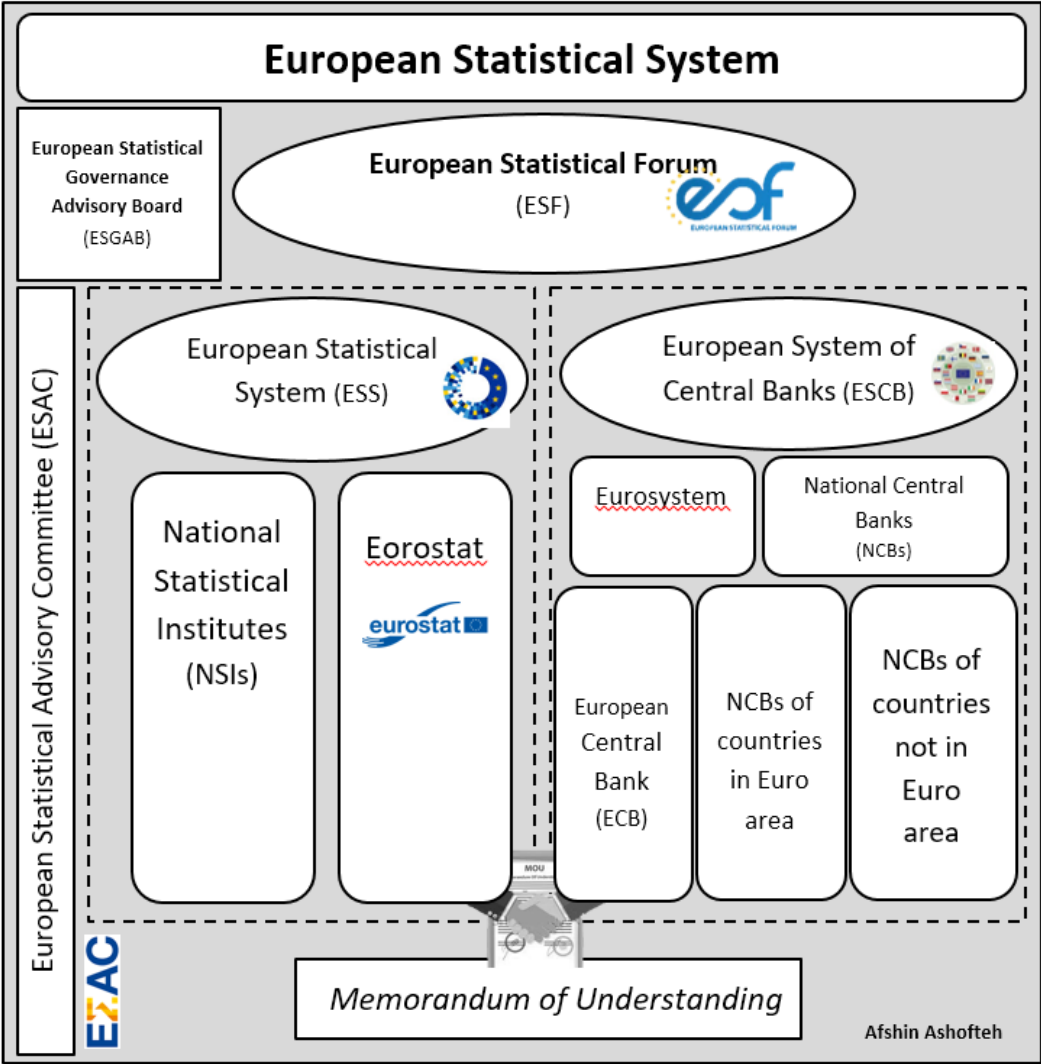# Section 1

Information Management and Statistics in Banking

Integrated Information Management in MFI's

Level 1	Board	Decision
Level 2	Specialised Committee of Information and Technology Management — Strategic Management and Control	Strategic Coordination
Level 3	Information Management: Statistics Department — Departments — IT Management: IT Department	Operational Coordination
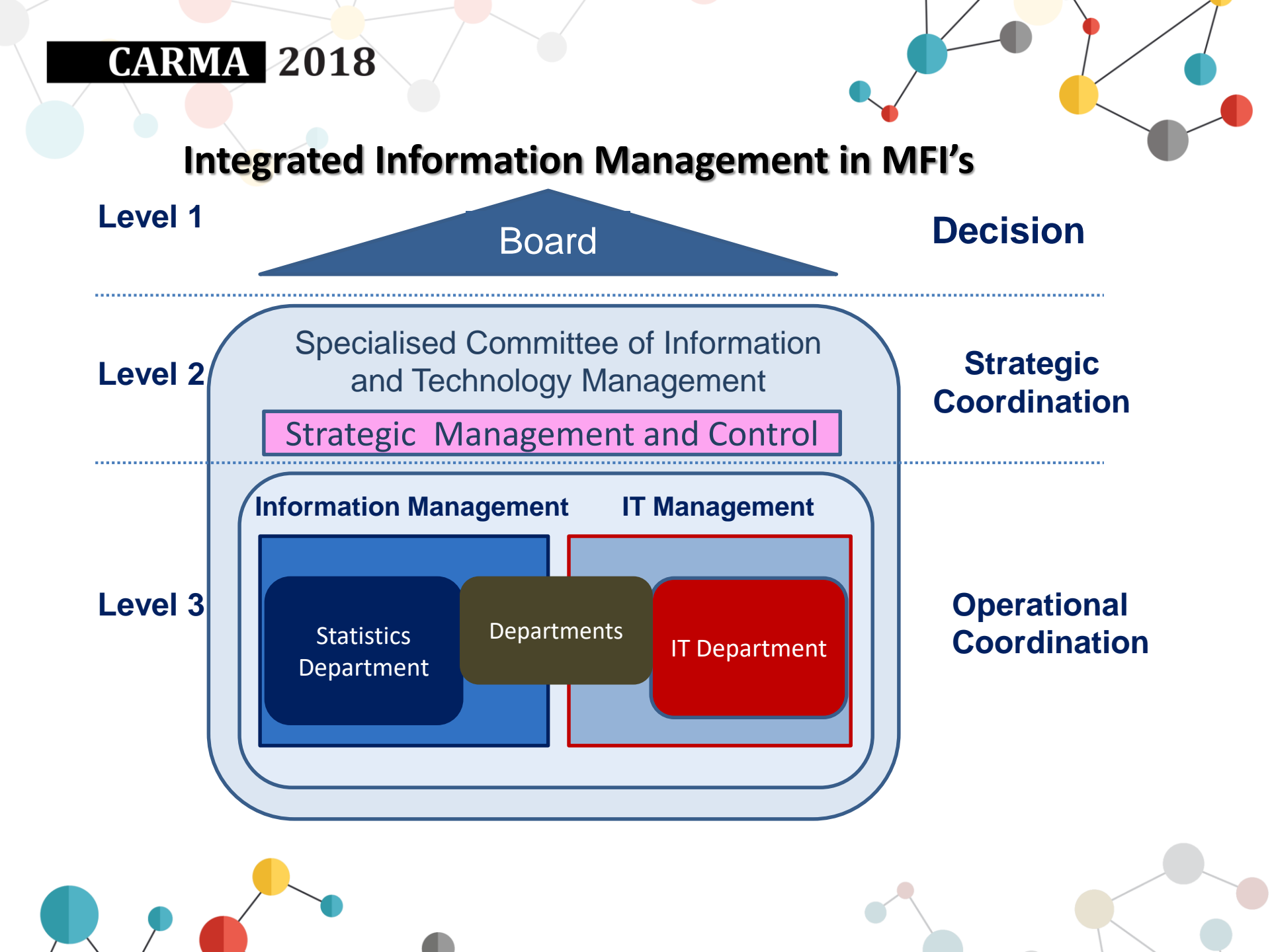
Important points:

1. Statistics and IT departments are able to make big improvements and even a big bang.

Section 2

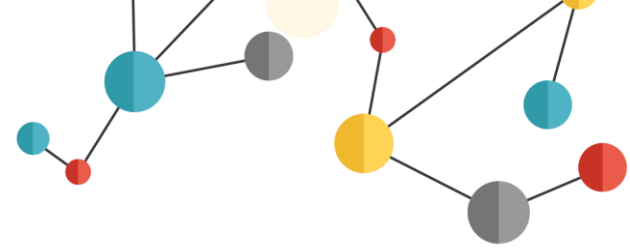Mining Big Data: in statistical systems of the monetary financial institutions (MFIs)

Past – Present - Future

An Insight into Banking System's Big Data

INTRODUCTION

OLAP

Multi-dimensional data tables

BUSINESS INTELLIGENCE

Data driven decisions and reporting tools

ANALYTICS

Statistical and mathematical analysis

BIG DATA

Large, semi-structured data

COGNOS

Business Objects

Ssas.

hadoop

1997

2000

2010

PRESENT

Volume	Velocity	Variety	Veracity*
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

http://www.rosebt.com/blog/data-veracity

- **Variability**: data constantly changing

CARMA 2018

% of respondents

Main problems with Big Data
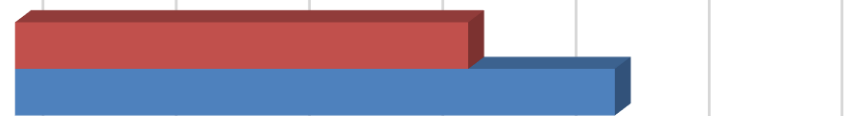
Lack of awareness of big data...
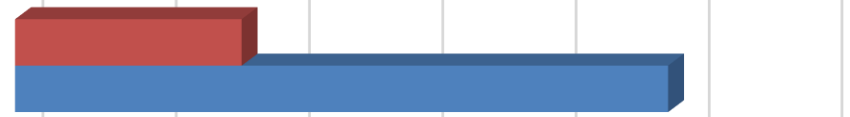
Expertise

Technical challenges of data...

Budgetin/setting priorities

Data Protection & Security

■ Capgemini   ■ Irving Fisher Committee
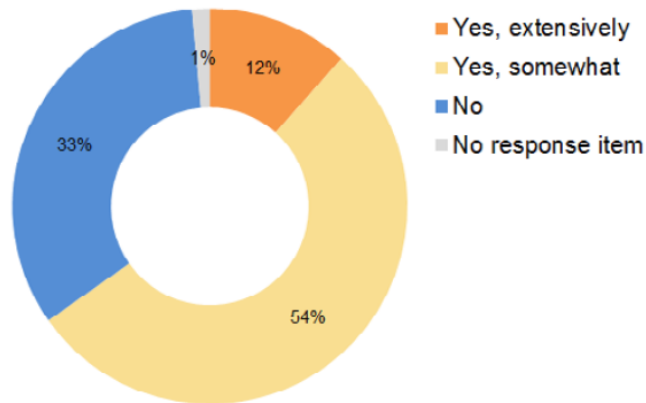
0   10   20   30   40   50   60

Irving Fisher Committee on Central Bank Statistics Report on Central banks' use of and interest in "big data"

(October 2015 – 69 Central Banks were participated)

Europe, IFC members, Turkey and Banco Central del Paraguay

Is the topic of big data being formally discussed within your central bank?   Graph 2

- Yes, extensively
- Yes, somewhat
- No
- No response item

Answers	Percentage	Count
Yes, extensively	12%	8
Yes, somewhat	54%	37
No	33%	23
No response	1%	1
Respondents		69

Irving Fisher Committee on Central Bank Statistics Report on Central banks' use of and interest in "big data"

(October 2015 – 69 Central Banks were participated)

Europe, IFC members, Turkey and Banco Central del Paraguay

How do you rate the interest of your central bank, as expressed at the senior policy level, in the topic of big data?

Graph 3

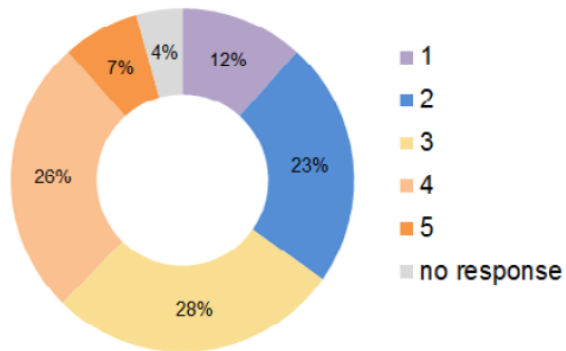|	Percentage	Count
1 (Very low)	12%	8
2	23%	16
3	28%	19
4	26%	18
5 (Very high)	7%	5
No response	4%	3
Respondents		69

**Irving Fisher Committee on Central Bank Statistics
Report on Central banks' use of and interest in "big data"
(October 2015 – 69 Central Banks were participated)**

- **Conclusion 1:** There is strong interest in big data in the central banking community, in particular at senior policy level.

- **Conclusion 2:** Central banks actual involvement in the use of big data is currently limited.

- **Conclusion 3:** Big data can be useful for conducting central bank policies.

- **Conclusion 4:** Big data are perceived as a potentially effective tool in supporting macroeconomic and financial stability analyses.

Irving Fisher Committee on Central Bank Statistics
Report on Central banks' use of and interest in "big data"
(October 2015 – 69 Central Banks were participated)

- **Conclusion 5:** Big data may also create new information/research needs.

- **Conclusion 6:** International cooperation can add value.

- **Conclusion 7:** Exploring big data is a complex, multifaceted task.

- **Conclusion 8:** Regular production of big data-based information will take time, especially because of resource issues.

Big Data does not replace banks' current analytical infrastructure but simply extends its scope

It has now become conceivable to conduct analyses based on the whole spectrum of data available, not just a limited sample.

Important points:

1. Statistics and IT departments mutually are able to make big improvement and even a big bang.

2. new algorithms and technologies are helping banks continually, they are also interested, but they need a platform to have personal data of clients and at the same time take care of TRUST and SATISFACTION of them.

PAST…

Extracting information from official information of each clients.

HDFC BANK
We understand your world

HDFC bank is the 5th largest bank in India by assets with an estimated value of £39 billion. HDFC uses their knowledge of the channel journey to personalize the customers experience on their channel of choice. For example, an ATM would recognise the customer's preferred language from previous interaction on the organization's website. Coupled with other preferences measured across a wide range of channels has decreased time at ATMs by 40% and helped reduce operational costs.

URL referrals

COOKIES

Customer segmentation by decision tree

OCBC Bank

OCBC bank is headquartered in Singapore and has assets estimated to be over £59 billion. OCBC identified a significant market segment moving to online channels and the expectations of these channels are growing with rapid pace of technology. One of OCBC's core principles is to deliver a high quality service through the channel of choice, by adapting to what the costumer expects.

中国银行
BANK OF CHINA

Bank of China has created a new online banking platform to examine data from various channels accessed by its global customer base. The online platform integrates customer-facing systems such as branch, phone, mobile and web services for Bank of China's 100 million customers. This enables Bank of China to provide the right content at the right channel.

URL metrics

HSBC ◆◆

HSBC found the primary barriers for internet banking were:
- Customer habit
- Security concerns
- And a lack of confidence.

They now have an active migration strategy to address these concerns.
Part of the HSBC migration strategy is to enable customers to undertake increasingly more complicated banking activates via internet.

Banque de France

Data science and big data experimentations on Banknotes Processing Machines

- Data science and Big Data
- Advanced statistics and Machine Learning
- Start-up collaboration

Banknotes

Context

Banque de France carry out a broad range of cash-related activities :

- **Issuing** new banknotes and coins
- **Maintaining the quality** of banknote

Banque de France ensures banknotes **recirculation** by using **120 processing machines** to check over **7 billions** banknotes **per year**

120 banknotes processing machines

Huge, detailed and available data

- **400 information** generated from embedded censors and scanners
- **100 Mb logs data** generated on each banknote for every scanning process (**3 Tb total / year**)

Big Data and Data Science allow new opportunities in data analytics

Banque de France

Predictive maintenance on banknotes processing machines : How to predict next failure time ?

- Data science
- Machine learning : Random forest algorithm
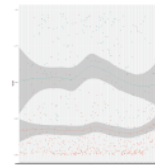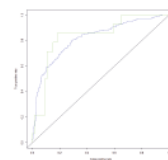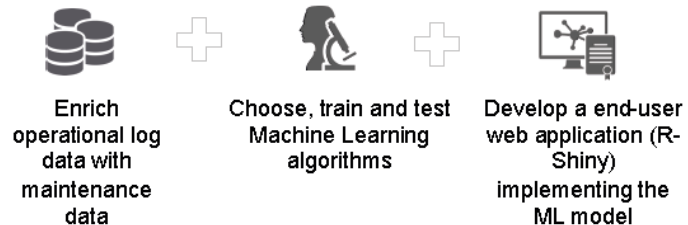
Banknotes

Context & Objectives

➢ Machines failures or malfunctions are too costly
➢ **Find a Machine Learning algorithm able to predict if a processing machine will fail within next days**

Project key points

– Iterative and agile project management
– Realisation time : 8 months with 1 data scientist
– PoC (Proof of Concept) completed. PoV (Proof of Value) on progress

Main benefits

✓ Reduce operational cost : Transform corrective maintenance into preventive maintenance
✓ Increase production by reducing fixing time
✓ Higher Quality of Service

Enrich operational log data with maintenance data

Choose, train and test Machine Learning algorithms

Develop a end-user web application (R-Shiny) implementing the ML model

Machine Learning algorithm able to predict 70% of failures occuring next day

NOW …

Extracting information from registered data by each clients.

Banque de France

CashCycle Project :
Big data advanced analysis of banknotes circulation
from issue to destruction

- Big Data
- Advanced analytics

Banknotes

Context & Objectives

- ➢ Cash cycle knowledge and banknotes security are core concerns of National Central Banks
- ➢ **Find efficient and scalable analytics to fully process big data volumes and to improve operational requirements**

Project key points

- − Only 20 machines are used (serial number scanning capability)
- − Realisation time : 10 months. 1 data scientist, 1 data engineer and 1 business expert
- − 1 billion banknotes processed during the PoC

Main benefits

- ✓ Performance and activity monitoring
- ✓ Better estimation of banknotes needs
- ✓ Analysis of banknotes degradation over the time
- ✓ Security signs improvement
- ✓ Analysis of rejected and unfit banknotes reasons

Data pipeline :
Collection,
storage and
processing

Set of advanced
analytics for cash
cycle analysis

Data visualization
and discovery tools
for business expert

logs

Operational data

Transactional data
(CRM, maintenance)

Hadoop Cluster

Spark

hadoop

Big data Infrastructure

tableau

Business Expert
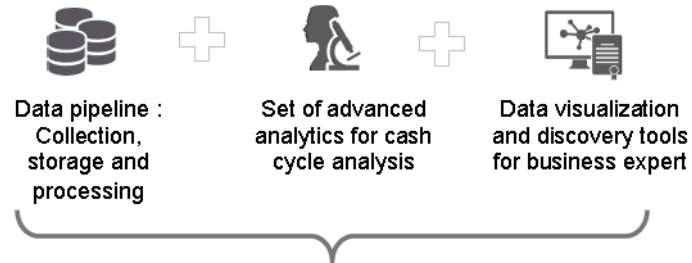
Big data processing pipeline

Banque de France

Detecting improper business practices in insurance and banking with social media listening

Prudential Supervising

- Natural Language Processing (NLP)
- Social listening
- Text mining

Context & Objectives

➢ ACPR (Banque de France) is responsible of prudential supervising of French banks and insurance companies
➢ **Use social listening to early detect contract violations and poor business practices**

Project key points

– Platform R/Python machine learning models
– Pretty good failure prediction ratio (about 70%)
– Big data processing workflow (statistics, KPIs..)

Main benefits

✓ Early detection of contract violation
✓ enhance business and legal intelligence
✓ Take advantage of open and social data to supplement internal data

Collect data from social media, forums and blogs

Use text mining algorithms to make sense of the social web

Develop a web application to monitor and analyse the information
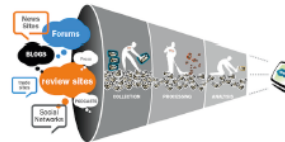
Important points:

1. Statistics and IT departments mutually are able to make big improvement and even a big bang.

2. new algorithms and technologies are helping banks continually, they are also interested, but they need a platform to have personal data of clients with taking care of TRUST and SATISFACTION of them.

3. Till now, registered data and official information of clients were used for different purposes.

Now let's talk about the Big Bang!

The Big Bang is Future!

The big bang is applying each client's desires, decisions, risk appetite, wishes, profile, characteristics and personality as a data source into decision making and continuous improvement process of banks in benefit of both sides. Banks and Customers!
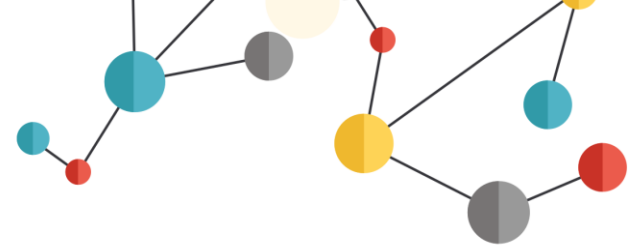
FUTURE …

Extracting information from behaviors and desires of each clients.
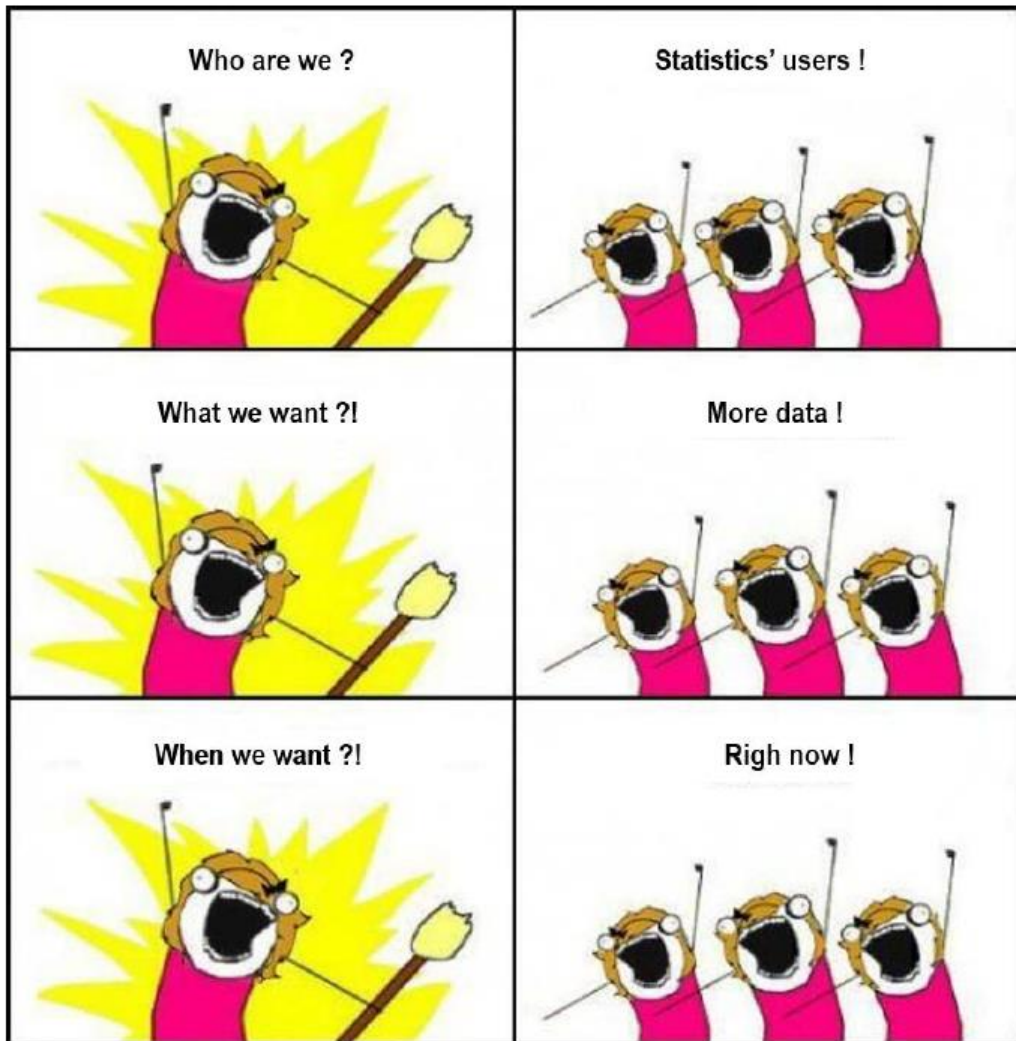
We want to offer a solution to:

- Maximize Capital Adequacy Ratio of Bank to have more free capital.
- Minimize the Risks.
- Maximize the Benefits of Clients.
- Maximize the Satisfaction of Clients.
- Minimize the requisite Client's Trust to the bank's activities.

ALL TOGETHER!

IMPOSSIBLE

MoneyMakers(M&M)

Try making **Money** by **Investing** and enjoy it like playing a **GAME**!

Imagine M&M mobile app based on Imbalanced Big Data mining and AI that recommend the investment opportunities with different risk levels to each of clients, help them to play a real investment game & give them the tips on their own risk appetite one by one!

Business field: Banking, Finance & Insurance.

Technology area: Big Data & Machine Learning

Business model: B2B

Main Problem of Customers

Investment is **Risky** and **Unpleasant** activity

- Lack of blindness trust
- Lack of Knowledge
- Lack of Experience
- Lack of Skills
- Lack of Customization

MoneyMaker(M&M)

A Mobile Application Platform: Imbalanced BigData mining + Machine Learning + Text Mining + Structural equation modeling (SEM) + Enterprise risk management (ERM)

M&M Advantage 1

Adding pleasure into INVESTING activities like playing a GAME

M&M Advantage 2

A platform for the customers to play in Stretch zone
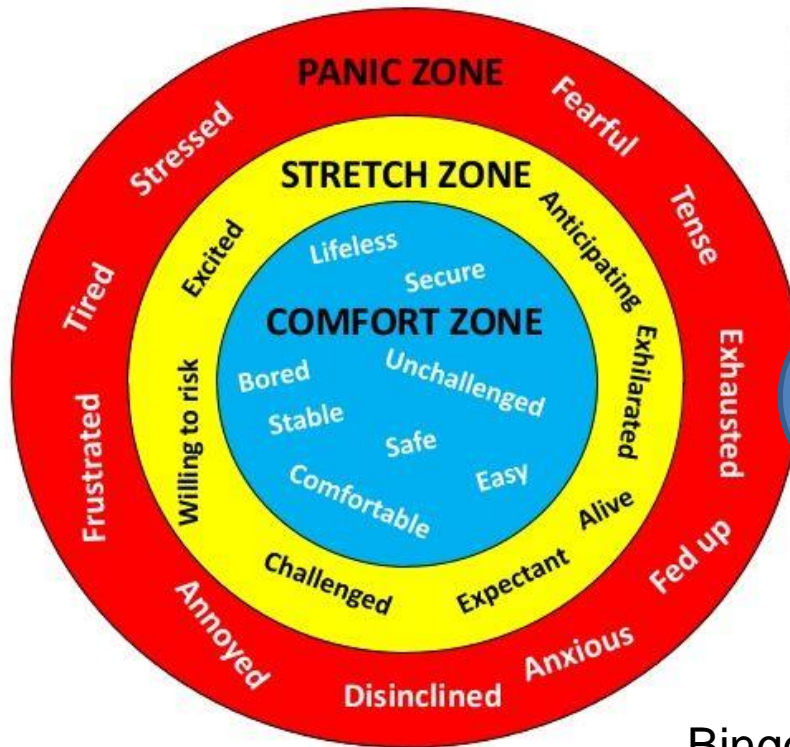
Customers will **ENJOY** investment and **TAKE the RISK** but in an almost **SAFE zone** out of panic zone

Bingo!

Customers requirements to play in Stretch zone

What else can go wrong and how are risks interconnected?

What are all the risks to our business strategy and operations?

What are we doing about the risks?

How much risk are we willing to take?

How well do we manage the risks?

How good are we at overseeing risk taking?

How do we determine the size and scope of the risks and report the results?

How do we ensure we have the right information to manage risk?

Stress Testing

Coverage

Response

Risk Appetite

Control Environment

Culture

Governance & Policies

Measurement, Evaluation and Communication

Risk Data & Infrastructure

- Circular depiction is highly intentional
- Components are meant to be dynamic (reviewed back/forth in any sequence)
- Having the right culture is key

CARMA 2018

The M&M digital platform Model

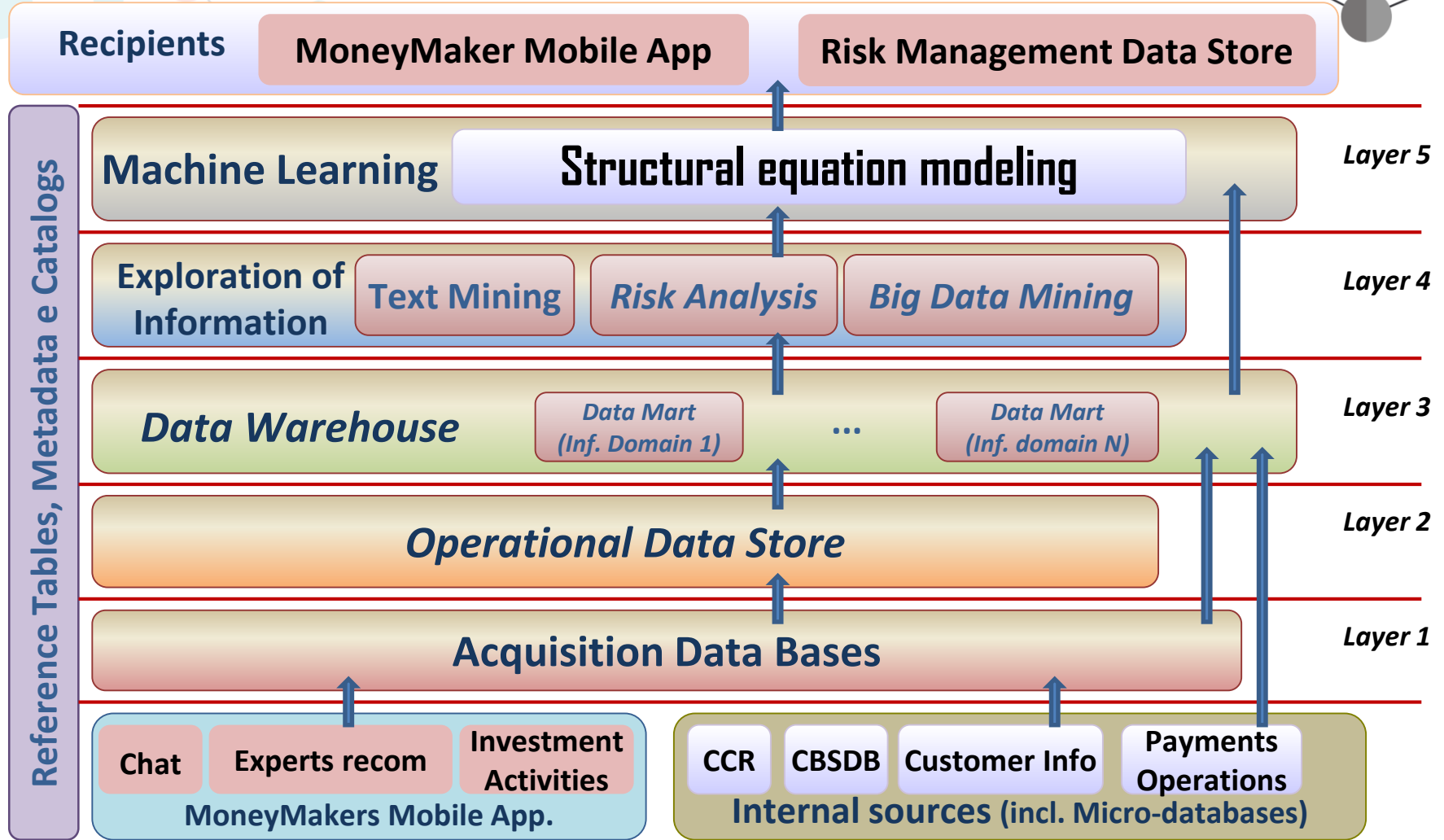**Reference Tables, Metadata e Catalogs**

Recipients	MoneyMaker Mobile App	Risk Management Data Store

Machine Learning — Structural equation modeling — *Layer 5*

Exploration of Information — Text Mining | *Risk Analysis* | *Big Data Mining* — *Layer 4*

Data Warehouse — Data Mart (Inf. Domain 1) ... Data Mart (Inf. domain N) — *Layer 3*

Operational Data Store — *Layer 2*

Acquisition Data Bases — *Layer 1*

Chat	Experts recom	Investment Activities

MoneyMakers Mobile App.

CCR	CBSDB	Customer Info	Payments Operations

Internal sources (incl. Micro-databases)

Requirements

Imbalanced Big Data

distress situations for banks and financial systems are relatively infrequent events.
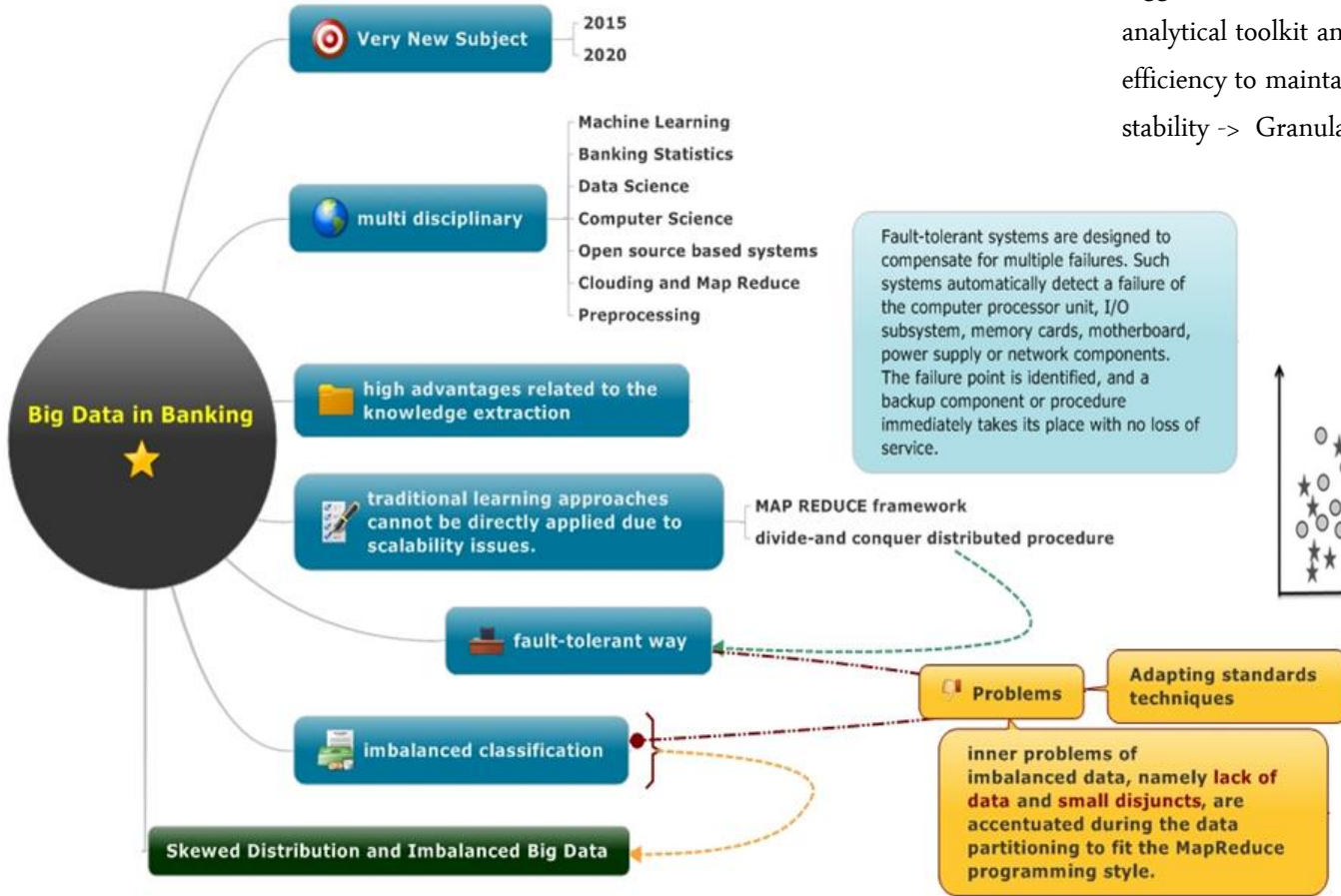
Robust Machine Learning

regulatory changes in incentives are likely to change the environment that generates the data.
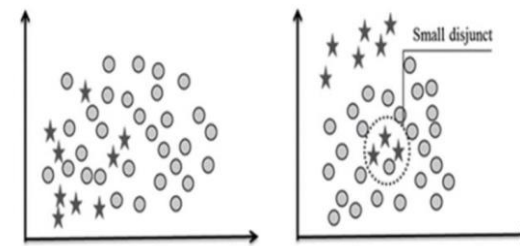
Structural Equation Modeling

To alleviate the risks, banks must build an effective early warning system to protect their assets.

Requirements

bigger and better data might enhance the Bank's analytical toolkit and improve its operational efficiency to maintain monetary and financial stability -> Granular data & Quality of data
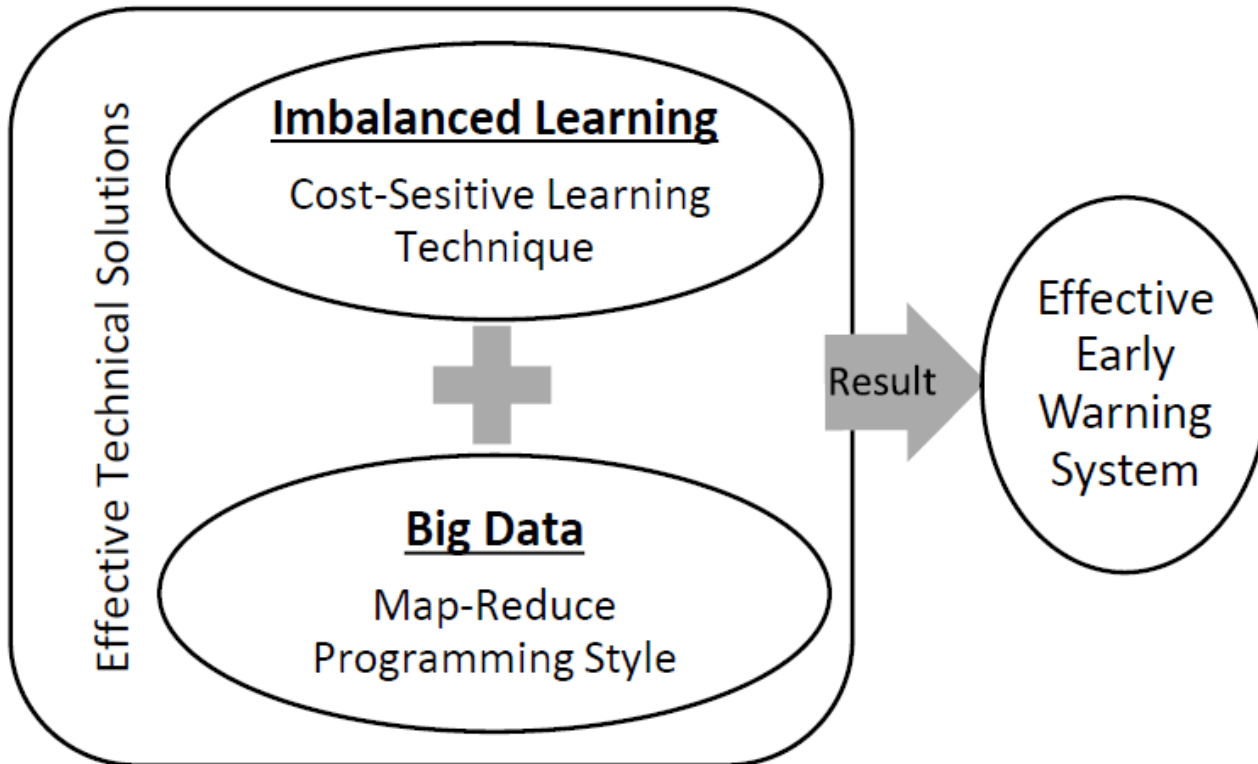
Big Data in Banking ⭐

- Very New Subject
  - 2015
  - 2020

- multi disciplinary
  - Machine Learning
  - Banking Statistics
  - Data Science
  - Computer Science
  - Open source based systems
  - Clouding and Map Reduce
  - Preprocessing

- high advantages related to the knowledge extraction

- traditional learning approaches cannot be directly applied due to scalability issues.
  - MAP REDUCE framework
  - divide-and conquer distributed procedure

- fault-tolerant way

- imbalanced classification

- Skewed Distribution and Imbalanced Big Data

Fault-tolerant systems are designed to compensate for multiple failures. Such systems automatically detect a failure of the computer processor unit, I/O subsystem, memory cards, motherboard, power supply or network components. The failure point is identified, and a backup component or procedure immediately takes its place with no loss of service.

Small disjunct

Problems

Adapting standards techniques

inner problems of imbalanced data, namely lack of data and small disjuncts, are accentuated during the data partitioning to fit the MapReduce programming style.

Requirements

The <u>MapReduce</u> execution environment is the most common framework used in the scenario of Big Data.
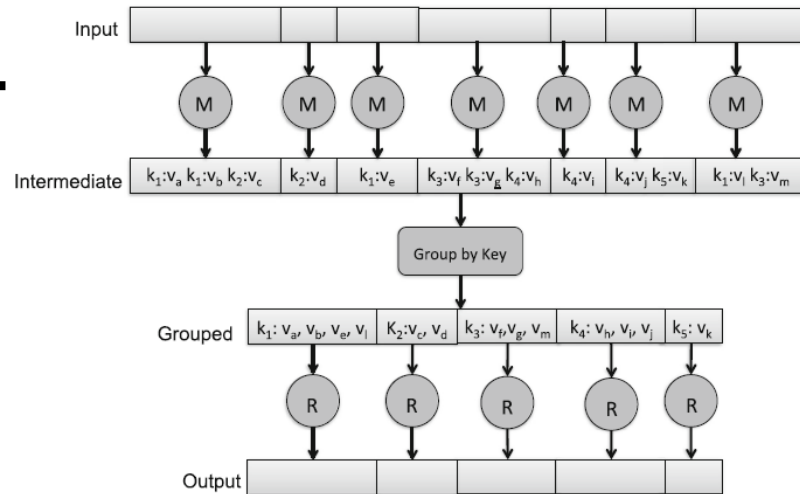
Apache Spark is clearly emerging as a more commonly embraced platform for implementing **Machine Learning** solutions that scale with **Big Data**.
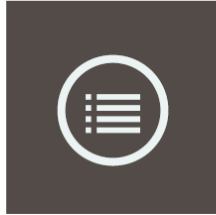
Imbalanced Learning

Cost-Sesitive Learning Technique

Effective Technical Solutions

Big Data

Map-Reduce Programming Style

Result

Effective Early Warning System
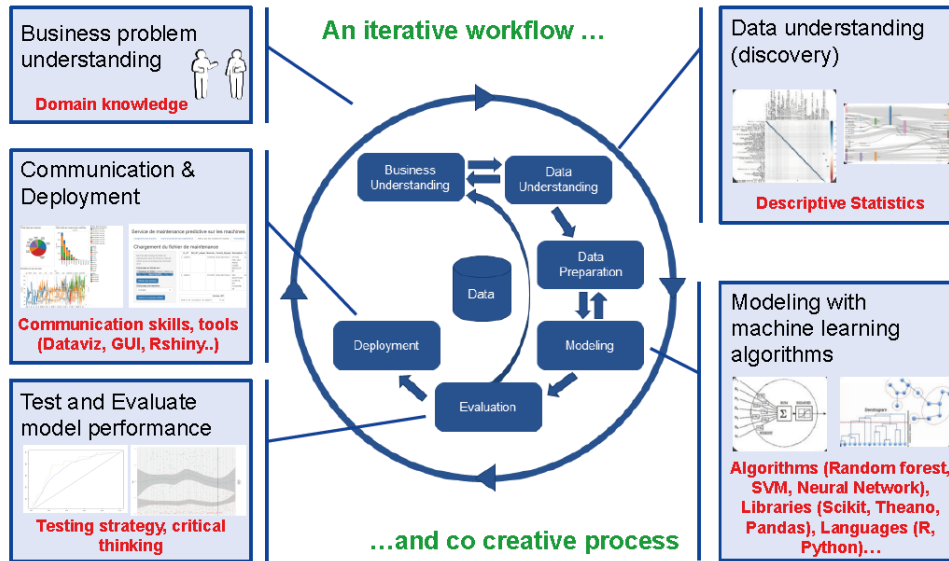
Requirements

structural model

| δ₁ | → | Indicator x₁ |

δ_1 → Indicator x_1

δ_2 → Indicator x_2

δ_3 → Indicator x_3

δ_4 → Indicator x_4

δ_5 → Indicator x_5

δ_6 → Indicator x_6

Credit Risk – Market Risk – Operational

Corporate & Regulatory Governance

ζ

Customer Stretch Zone

Risk Appetite ← ε_1

Total Asset ← ε_2

Characteristics ← ε_3

measurement model of the latent exogenous variables
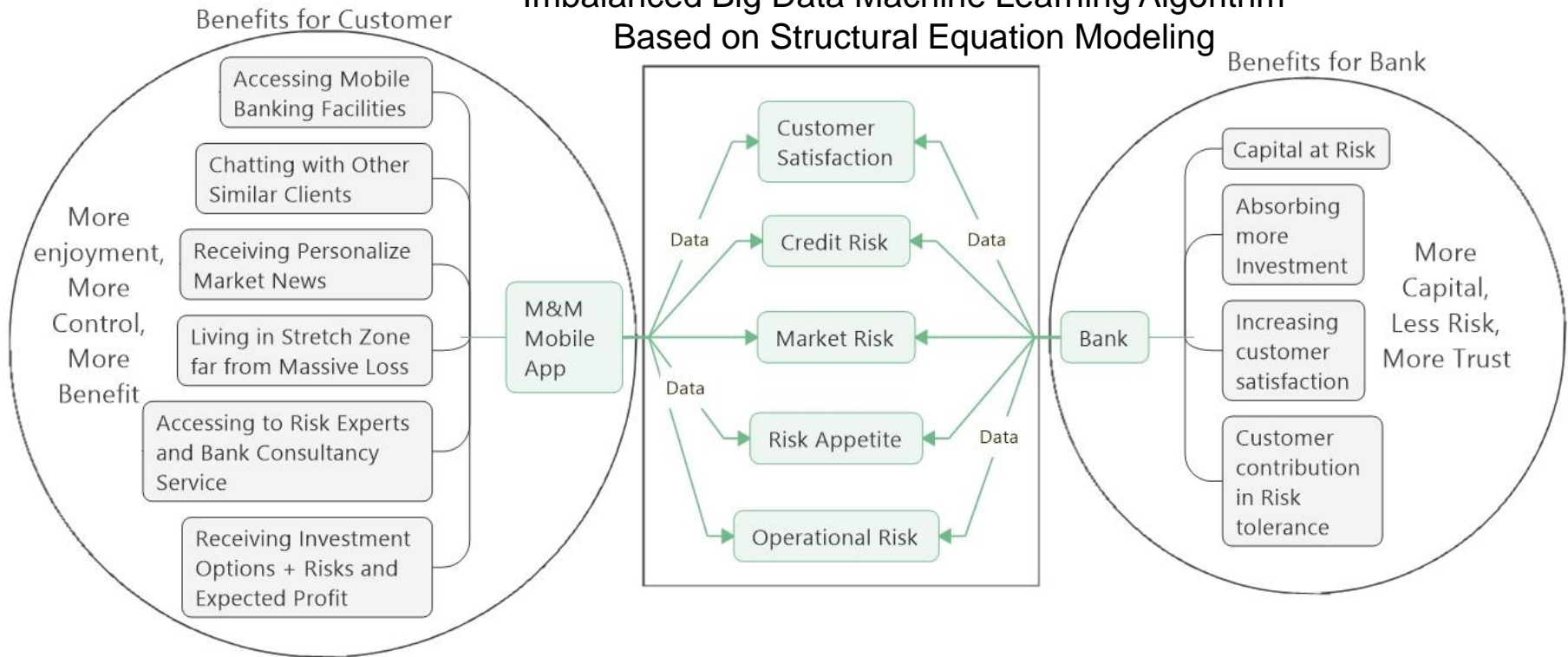
measurement model of the latent endogenous variables
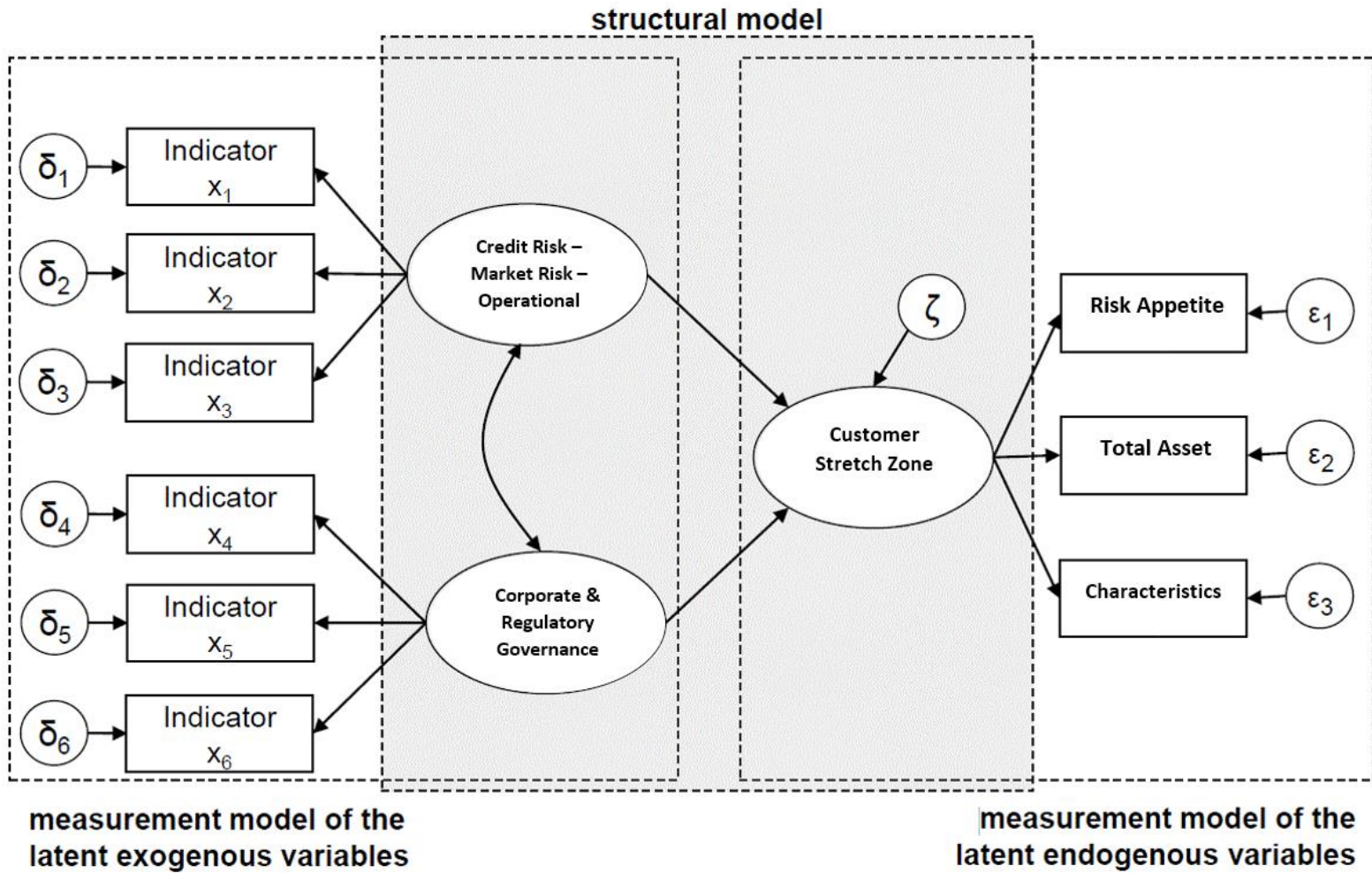
Structural Equation Modeling
CBSEM or PLS

M&M STARTUP

Imbalanced Big Data Machine Learning Algorithm Based on Structural Equation Modeling

Right Information at the Right Time for the Right Client

Thank You!

www.linkedin.com/in/statas