

Stories that Data Tells: A Practitioner's Perspective

Shonali Krishnaswamy



AI Driven Analytics

AI

Artificial Intelligence

Unsupervised Machine Learning -
Novelty and Drift

DL

Deep/Multi-Layered Learning

Award winning proprietary multi-layered ML
methodology

AA

Advanced Analytics

AI plus DL operating on multimodal data
- Structured and Unstructured

AIDA Team's Past Track Record



AIA taps on A*Star's data analytics to study insurance consumers' needs

By Lee Meixian | leemx@sph.com.sg | @LeeMeixianBT

MORE

'Engaging industry' a key priority in boosting S'pore as an R&D hub: NRF

AIA Group to up stake in Indian life insurance JV with Tata

SMRT partners A*Star to develop solutions to improve transport reliability



AIA Group on Tuesday said it has signed a multi-year joint collaboration agreement with A*Star's Institute for Infocomm Research (I2R), Singapore's largest information and communications tech research institute. ST PHOTO

MORE FROM THE BUSINESS TIMES



Expect currency wars in 2016, DBS chief warns



Temasek fund invests undisclosed sum in homegrown



Hot stocks: Keppel and Sembcorp Marine fall more



DBS mentioned as possible buyer in 3 potential sales



COMPUTERWORLD SINGAPORE

Standard Chartered partners A*Star's I2R to leverage data to gain business insights | Zafirah Salim | Jan. 25, 2016



NEC Analytics Joint Lab

top news

thesundaytimes April 14, 2013

S'pore team tops in predicting flight timings

It beats 170 teams with solution that could help airlines save millions of dollars

Grace Chng Senior Correspondent

Travelers often have to put up with long flight delays and even cancellations. Flights can leave late but arrive early. It is a puzzle that has frustrated airlines and travelers alike.

American conglomerate General Electric decided to find a solution. It launched the GE Flight Quest, with a prize of US\$100,000 (S\$24,000) to anyone who could develop a solution that lets airlines better predict flight arrival times.



(From left) Adviser Li Xiao Li and team members Mr. Conort, Dr. Cao, Dr. Phua and Dr. Yap. The fifth member, who is not in the picture, is Dr. Chua.



Dr. Krishna Kumar showing a smartphone app with the Liveness application, which tells users how many people are at a particular spot and how much activity going on there. PHOTO: DAVID VINCENT FOR THE STRAITS TIMES

App tells you if place is hot spot or dead town



Software identifies crooked sellers faster

Built by Visa and A*Star, i

Fighting crime through data mining

Institute for Infocomm Research works with leading credit card company to detect fraud using advanced data analytics technology. OO GIN LEE reports

A Singapore research institute is helping a credit card company weed out fraudulent online merchants. The three-year collaboration started in October last year between the Institute for Infocomm Research (I2R) and the local office of the credit card company. The company cannot be named because of a commercial confidentiality agreement.

Data analytics scientists at the institute studied data from millions of credit card purchases involving thousands of online merchants and devised a scientific methodology to distinguish bona fide online vendors from Internet fraudsters. The company already knew which merchants were crooks but did not tell the institute. The customers and their purchases were not identified. Dr. Shonali Krishnaswamy (left), acting head of the institute's Data Analytics Department, said the scientists called



Researchers at I2R are able to take raw data and apply sophisticated algorithms to identify information red within

I2R has announced its partnership with SingTel to set up a joint laboratory to develop advanced data analytics for innovative, personalised and relevant information/services to mobile users.



plied only basic transaction and this is not a new data science. It can be used to generate an analysis of the data in a way that can be used to identify patterns and trends. The data is then used to create a model that can be used to predict future behavior. The model is then used to create a recommendation system that can be used to suggest products and services to users. The recommendation system is then used to create a personalized experience for each user. The personalized experience is then used to create a more engaging and interactive user interface. The user interface is then used to create a more efficient and effective user experience. The user experience is then used to create a more successful and profitable business.

Industry speak for software - over the next three months, the team will be contributing to the software. Earlier this month, the team also came up with a detection competition at the Singapore Management University here. The first detection in Mobile Ad competition was part of the Conference of Machine Learning. The teams were given anonymous online user data. Their task was to identify the users who were most likely to be fraudulent. The team that identified the most fraudulent users won the competition. The team that identified the most fraudulent users won the competition. The team that identified the most fraudulent users won the competition.

Her team has done research that can even detect a user's gender, just by looking at such raw data.

TAIEX 104.03

CHANNEL NEWSASIA

Strategic Partnerships

Working with Industry to Create Innovation

MIT Technology Review

Forbes

GE Report

GigaOm

InformationWeek

THE BUSINESS VALUE OF TECHNOLOGY

thesundaytimes

THE STRAITS TIMES asia report

InformationWeek THE BUSINESS VALUE OF TECHNOLOGY

GigaOm

GE Report

Forbes

MIT Technology Review

kaggle

Multi-Award Winning Machine Learning Team

AIDA Won the Monetary Authority of Singapore Hackcelerator
Inaugural Singapore FinTech Festival 2016

<http://www.fintechfestival.sg/hackcelerator/>
www.aidatech.io

Kaggle
Grandmaster

Almost 😊 All PhD
Team

100 Problem Statements, 650 Submissions, 20 Finalists and 3 Winners



2011

**1st Place in ALL 4
Categories:**

EU OPPORTUNITY Mobile
Activity Recognition
Challenge

2015

1st Place: IJCAI Repeat Buyer
Prediction in E-Commerce – **Beat
754 Data Scientists**

1st Place: ACM KDD Cup Predictive
Analytics – **Beat 821 intl' teams**

2015

1st Place: Springleaf Sponsored
Kaggle Marketing Response
Competition – **Beat 2,225
international teams**

IES Prestigious Engineering Award

2016

IES Prestigious
Engineering Award
ASEAN Outstanding
Engineering
Achievement



Key Take Away #1: Its ...the FEATURES...!

AI Driven Analytics

E-Commerce Repeat Buyer Prediction

International Joint Conference in Artificial Intelligence 2015
1st Place out of 754 international teams

- 260,000+ “loyal” customer + merchant pairs provided as training data
- Predict the repeat buying probability of other 261,000 user and merchant pairs

user	item	category	merchant	brand	time	event
048300	436606	0204	0731	5417	1110	add2cart
328862	844400	1271	2882	2661	0829	click
328862	81766	0614	4605	7622	0709	buy
328862	524981	0664	2382	1272	0602	click
328862	440930	1271	2882	2661	0829	click
244974	218624	0451	1503		1109	add2cart
328862	575153	1271	2882	2661	0829	favourite
...

Question:

User “048300” will be a loyal customer of merchant “0731” in the future?

Solution:



65.00 %

AIDA Scientist

Part of Winning Team

70.50 %

Direct Email Marketing Campaign

Kaggle - Springleaf Marketing Response Competition 2015

1ST Place out of 2,225 international teams

- Customer data: Anonymized, messy (strings and numbers) and high dimensional
- Column names range from VAR001 to VAR1934
- Column value can be a number or date or hash string – **No meaningful explanations**

[illegible]

Question:

Will user “X” respond to a direct email campaign?

Solution:

AIDA

Team Member Winning Solution –

80.39%

Machine Learning Process

Feature Engineering

Feature Selection

Feature Generation

Key Features for Predicting are Selected

Machine Learning

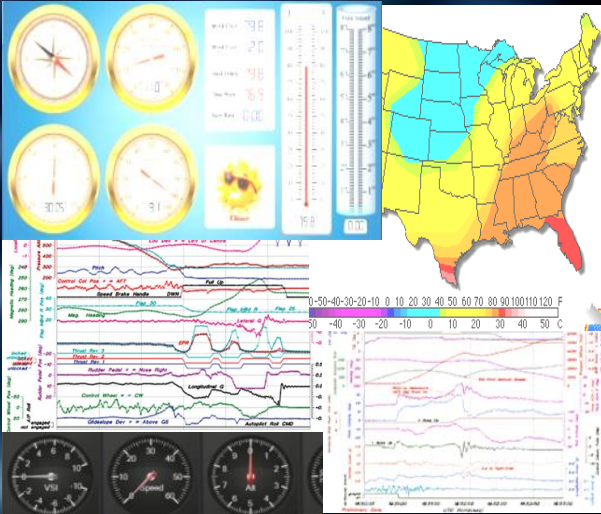
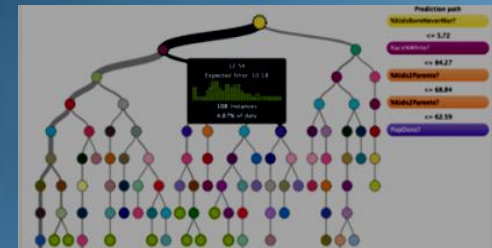
Evaluate Multiple Learning Models

Prediction on New and Unseen Data

DATA



Insights, Trends and Informed Decisions





**Key Take Away #2: Algorithm Design and Choice – *Horses
for Courses!***

AI Driven Analytics

Unsupervised Concept Drift Detection: Early Detection of Change



Silent Attrition



Changing Lifestyle and Life Stages



Monitoring for Risk Profiles

ConTrack: An unsupervised concept drift detection method that can:

- (1) track K concepts $B_k^{(t)}$ that are shared amongst the N actors
- (2) track each of the N actors' participation $\theta_i^{(t)}$ in the K concepts

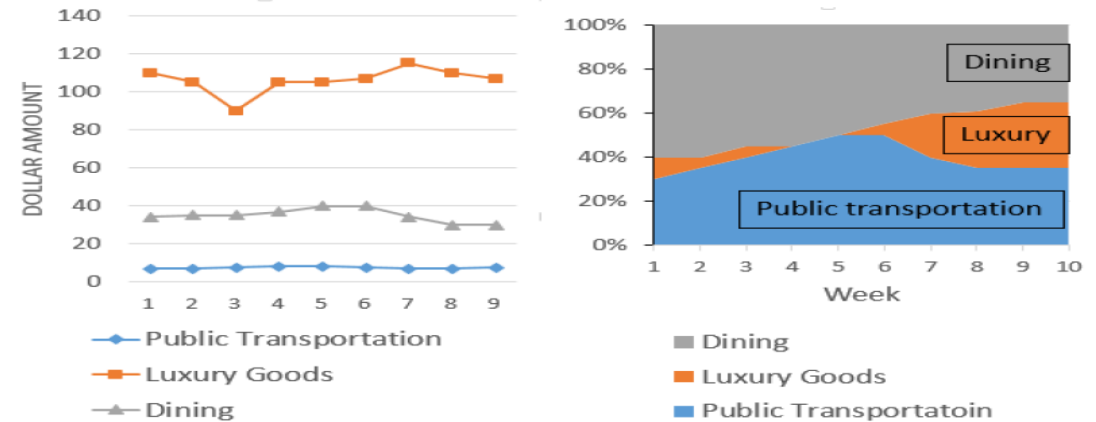


Figure 1: Visualization of changes in actor participations and concepts in banking data. **Left:** Tracking the changes in dollar amount dimension of public transportation, dining and luxury goods concepts over time. **Right:** Tracking the changes in actor's participation for public transportation, dining and luxury good concepts over time.

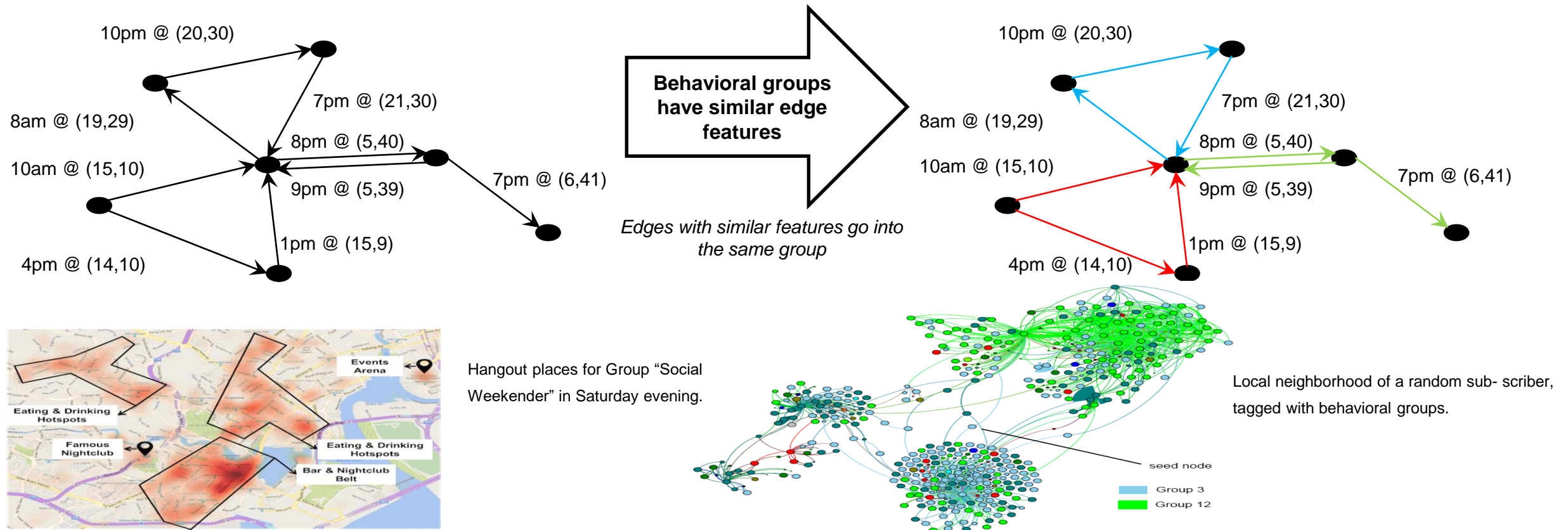
Identifying Behavioural Groups from Telco / CDR Data

Problem: Derive behavioral groups from CDR Data.

Solution: Behavioural Groups have similar Edge Features.

Distributed Graph Edge Clustering (DGEC) is an optimization model that discovers:

- (1) K behavioral groups within service/network graphs (which are directed, multi-edge, and have features on each edge)
- (2) which of the K behavioral groups each edge and node is affiliated with (where nodes can belong to multiple groups)





Key Take Away #3: Modelling a Problem in Machine Learning Terms – *Half the Battle!*

AI Driven Analytics

AI-CLAIMS: AI *Driven Analytics* for Claims Management

**Improving
Processes and
Customer
Experience**

Machine Learning for Information
Extraction & Predictive Claims
Processing

Augmented Intelligence for
Claims Management



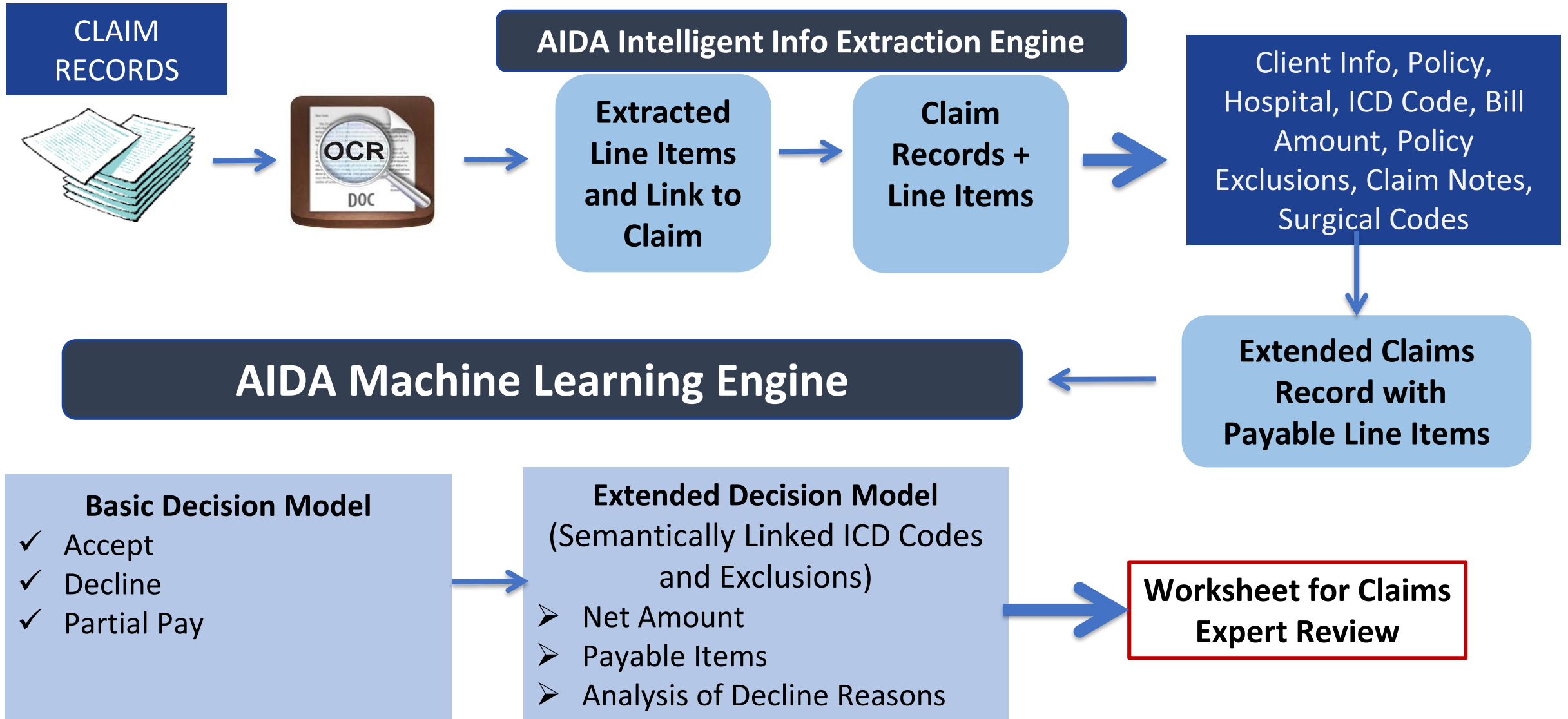
**Managing
Evolving
Risks**

Outlier Detection for Suspicious
Claims and Drift Detection for
Evolving Costs

**Increasing
Revenue and
Reducing Cost**

Predicting Modelling for
Propensity to Claim

AIDA AI-CLAIMS System



Deep Learning to Learn Contextual Connections

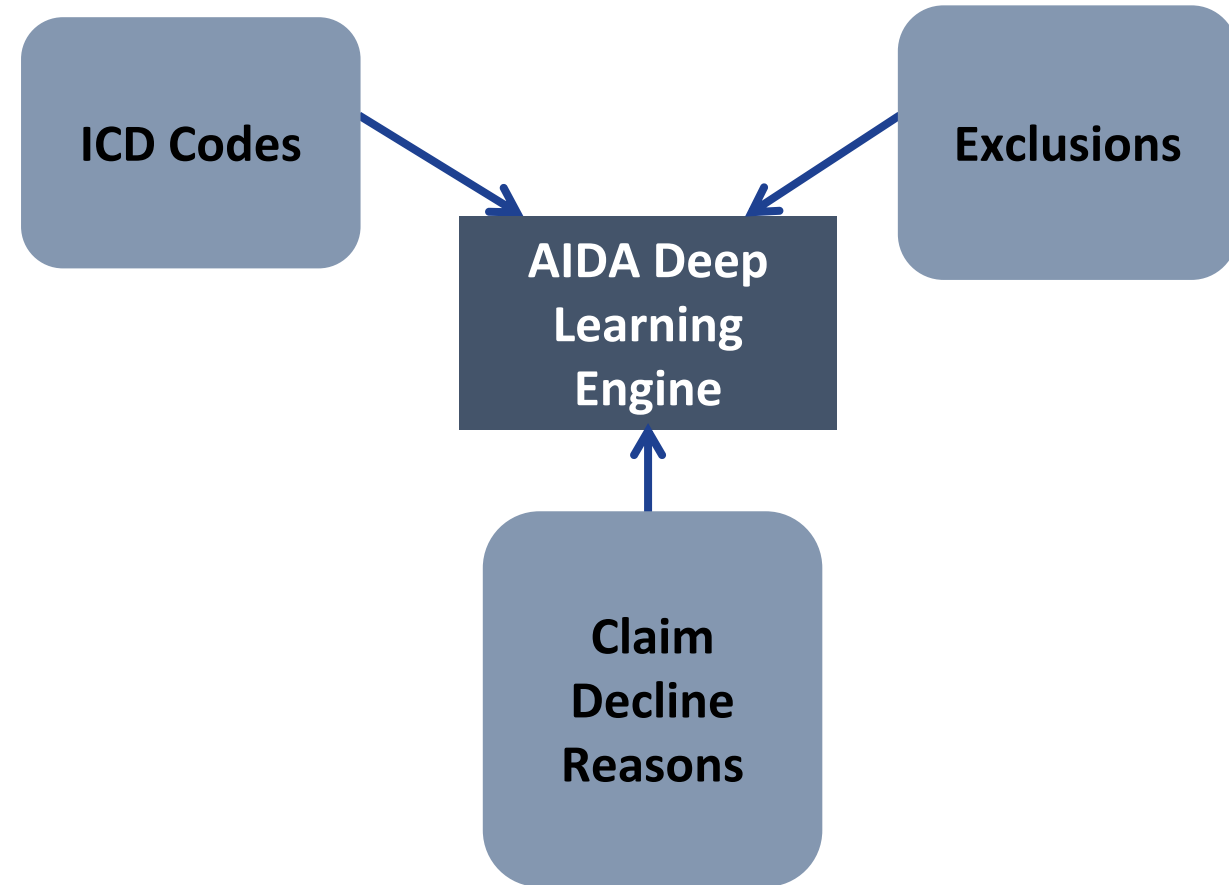
Contextual and Intelligent Linking of Text Data

Key Challenge

- ❑ ICD Codes and Exclusions are Not Semantically Linked
- ❑ ICD Codes are in Medical Terms (e.g. Obstetrics)
- ❑ Exclusions are in Layman's Terms (e.g. Maternity)

AIDA AI-CLAIMS Text Mining Engine Automatically Learns Contextual Connections from Text Data

- Connect Terms to ICD Codes
- Connect Terms to Exclusions
- Connect ICD Codes to Related ICD Codes
- Connect Exclusion Codes to ICD Codes





ICD CODES

EXCLUSIONS

TERM SEARCH

CLAIM NOTES

USER ▾

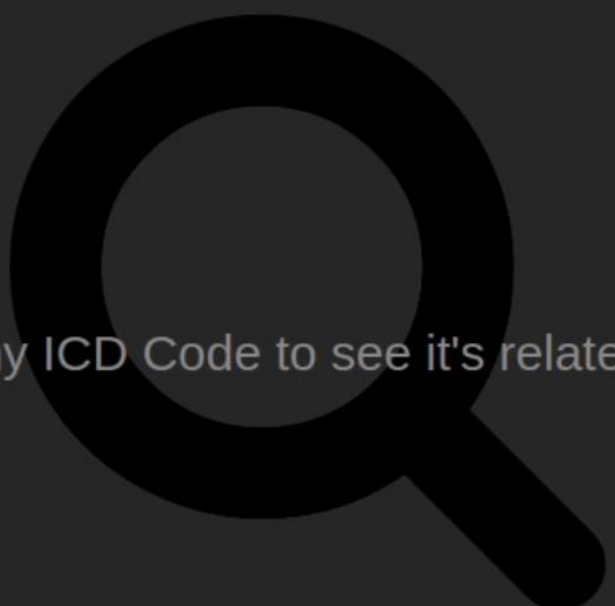
ICD CODES

Search By Related Terms

CODES ☒ TERMS

I

Search

Type any ICD Code to see it's related terms

Connecting...



9:05 AM



Key Take Away #4: Imbalance is a Very Real Challenge!

AI Driven Analytics

Profitability Modelling

Categories 2015-2016	Count	%
Total No. of Customers	XXXX	100
Total Number of Profitable Customers (Premium – Total Claims/ Premium) = 0	XXXX	91.x
Total No. of Medium Profitable Customers (0 < (Premium – Total Claims/ Premium) < 1	XXXX	3.x
Total Number of Non Profitable Customers (Premium – Total Claims/ Premium) < 0	XXXX	5.x

Revenue (\$)

Total Revenue : R

Total Claims Value: C

Total Profit : R-C

Comparison of Models

Approach	Model Accuracy	No. of Non-Profitable Cases	Number Correctly Identified	Total Cost (Value of Claims in \$)
Business As Usual	N/A	XXXX	None	X %
Predictive Model #1 (Naïve)	91%	XXXX	1	X %
Predictive Model #2 (With Imbalance Factored)	82%	XXXX	625	Approx. Claims Cost Reduction By 30 %



**Key Take Away #5: Not All Metrics Contribute to
Adoption and Usability!**

AI Driven Analytics

Machine Learning Metrics

- ✓ General Sense of Model Performance
- ✓ Useful for Model to Model Comparison

Predict X (X=Good Customer)	Actual X	
1	1	→ True Positive (TP)
1	1	
1	1	
1	1	How many did the model NOT CATCH ?
0	1	→ False Negative (FN)
1	1	
0	1	How many did the model GET WRONG ?
1	0	→ False Positive (FP)
0	0	
0	0	→ True Negative (TN)

ACCURACY: % of Instances Classified Correctly

Precision = True Positive / (True Positive + False Positive)

Recall = True Positive / (True Positive + False Negative)

F-Score: A combined measure that considers both *Precision* and *Recall*

Area Under the Curve (AUC): Different true positive/false positive rates using a threshold. As you decrease the threshold, you get more true positives, but also more false positives.

Lift (For Imbalanced Data):

What is the impact of the Model Vs Random?

E.G. 100 Retained Customers among the 10,000.

Randomly select 10% of population (1000), you can potentially identify 10 customers. If you select Top 10% of the Model, if you can identify 60 customers, then Lift is $60 / 10 = 6$.

How to Adopt Machine Learning: Business Metrics

Date	Customer ID	Machine Confidence	Prediction for Good Customer
1/1/2017	MUM001	0.85	Yes
1/1/2017	CHE008	1.0	Yes
1/1/2017	CHA009	0.2	No
4		0.4	No
5		0.5	No
6		0.6	No
7		0.7	Yes
8		0.8	Yes
9		0.85	Yes
10		0.9	Yes
12		0.9	Yes
13		0.82	Yes
14		0.1	No

1. What is the Acceptance Threshold ?

At What Threshold is the Machine Confidence Guaranteed to Get Most of it Right ? (e.g. 80% confidence, has 99 % accuracy, n false positives and m false negatives)

2. What is the Support/ Coverage (number of instances) above this Threshold ?

(e.g. 80% confidence has 50 % Support)

A Good Acceptance Threshold must have *high accuracy above the threshold* and *high support*.

Metrics for Model Decline and Data Changes ?





What's Next?

AI Driven Analytics

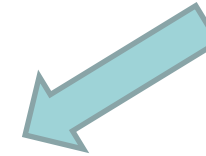
What's Next ?



History / Data



Feedback
from
User



Learning from Data + Knowledge
/ Context + Feedback



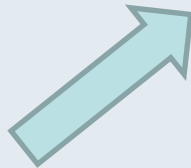
Learn 2 Learn



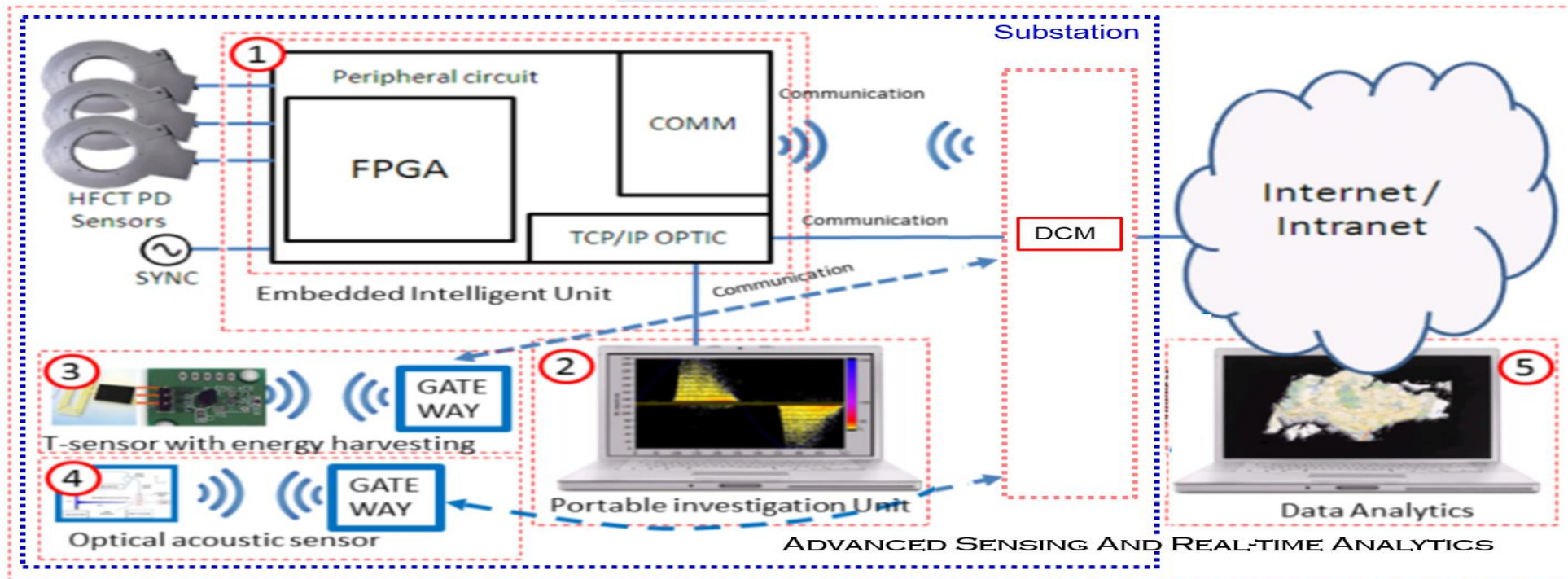
Domain Knowledge /
Semantics / Context



User/Activity/ Task
Observations



Distributed and Onboard Analytics: Many Open Challenges



Wattalyzer:
An Integrated Solution for Smart Grid Condition Monitoring through Advanced Sensing and Real-Time Analytics

Sensing

- **HFCT sensor** : HFCT PD Sensing
- **Temperature sensor**: Temperature Sensing with Energy Harvest
- **Fiber acoustic sensor**: Fiber Optic Acoustic Sensing

Analytics

- **On-board Analytics**
- **Backend Data Analytics**

Communication

- **DCM**: Data Communication Management

- Reliable and automated PD detection
- Integration of sensing and advanced real-time analytics
- Data driven trend analysis and visualization
- Intelligent and timely alerts

Multi-Modal Analytics



0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

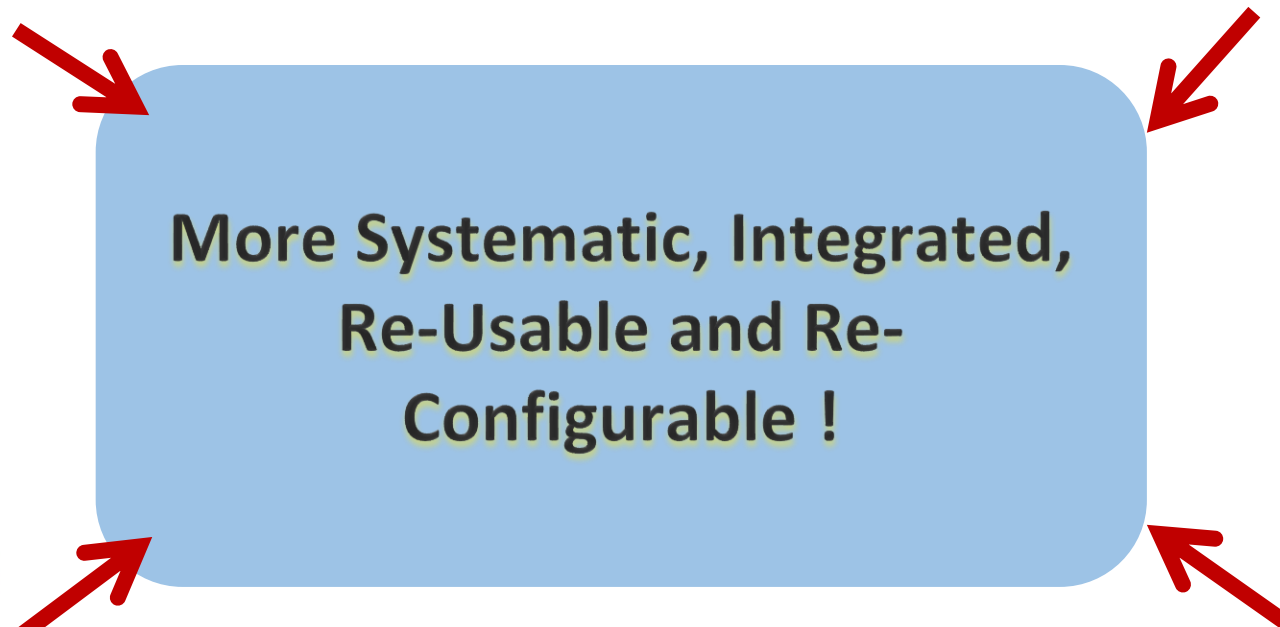
Structured Data



**Unstructured Data -
*Images***



**Unstructured Data
- *Text***



**Domain Knowledge
and Semantics**

Data, Data Everywhere!



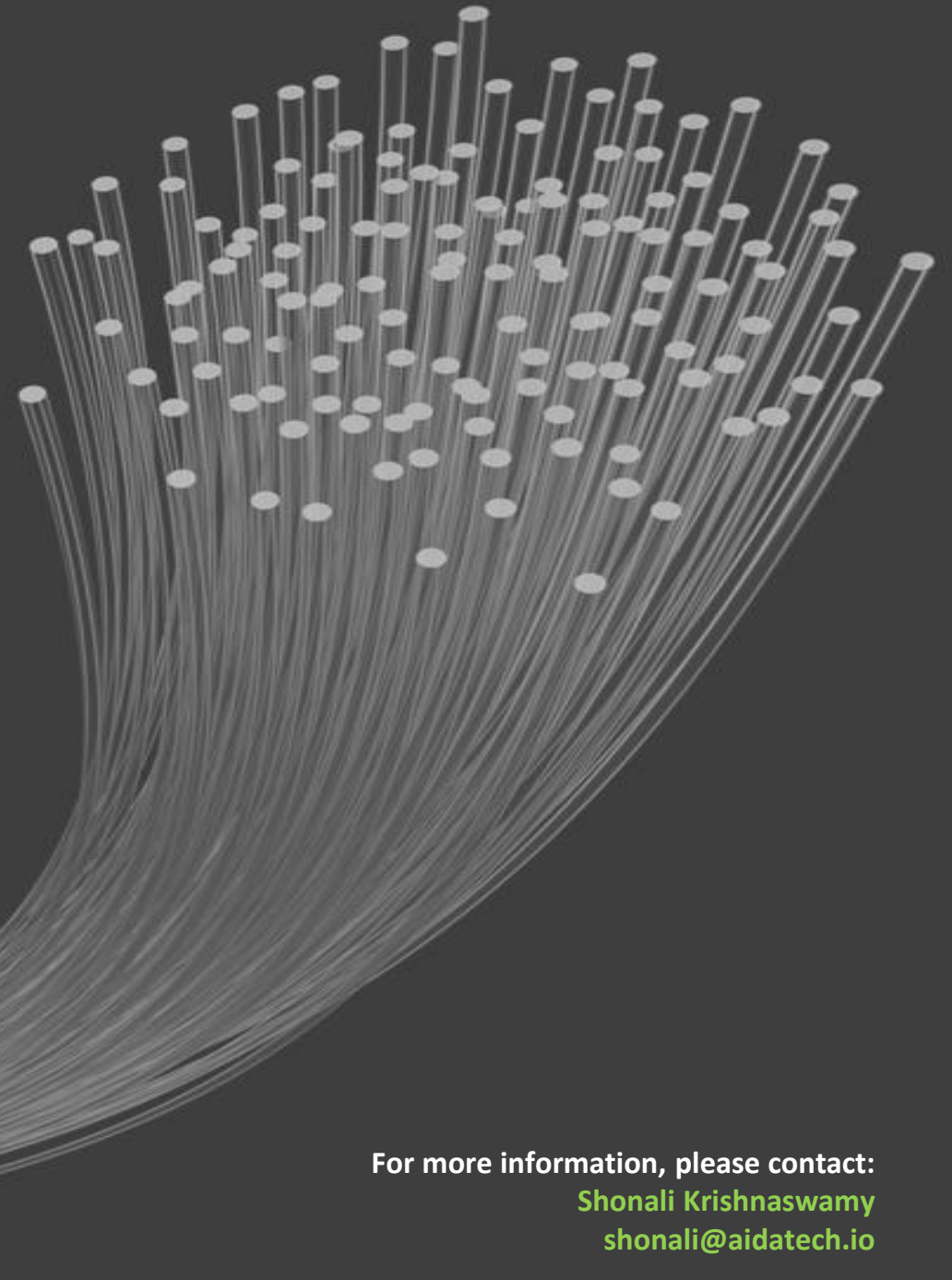
Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information?



The Rock By T. S. Eliot (1888-1965)

Thank **you!**



For more information, please contact:

Shonali Krishnaswamy

shonali@aidatech.io