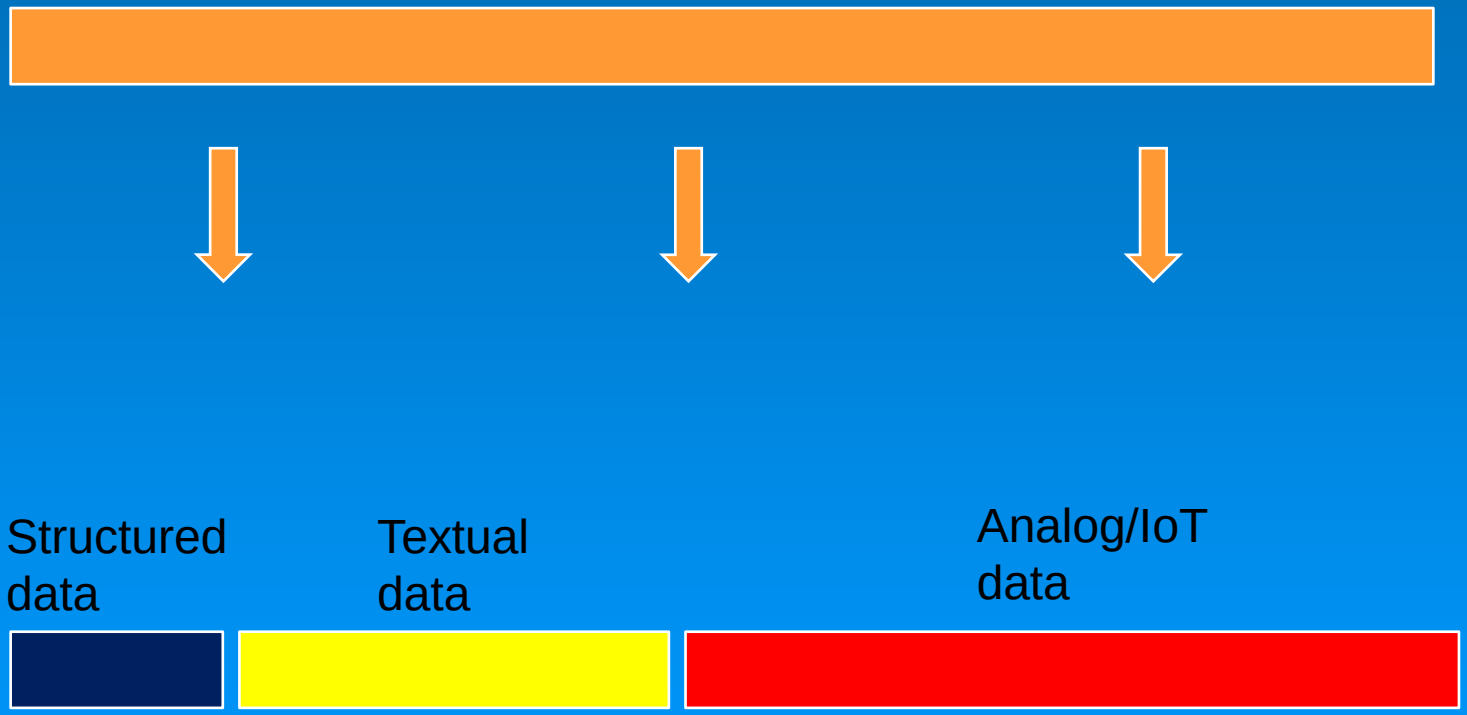
DATA LAKEHOUSE – THE BASIC ELEMENTS

A presentation by
W H Inmon

FOREST RIM TECHNOLOGY INC

All data in the corporation

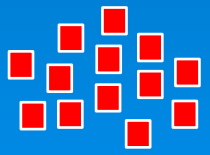


FOREST RIM TECHNOLOGY INC

Structured data

Textual data

Analog/IoT data



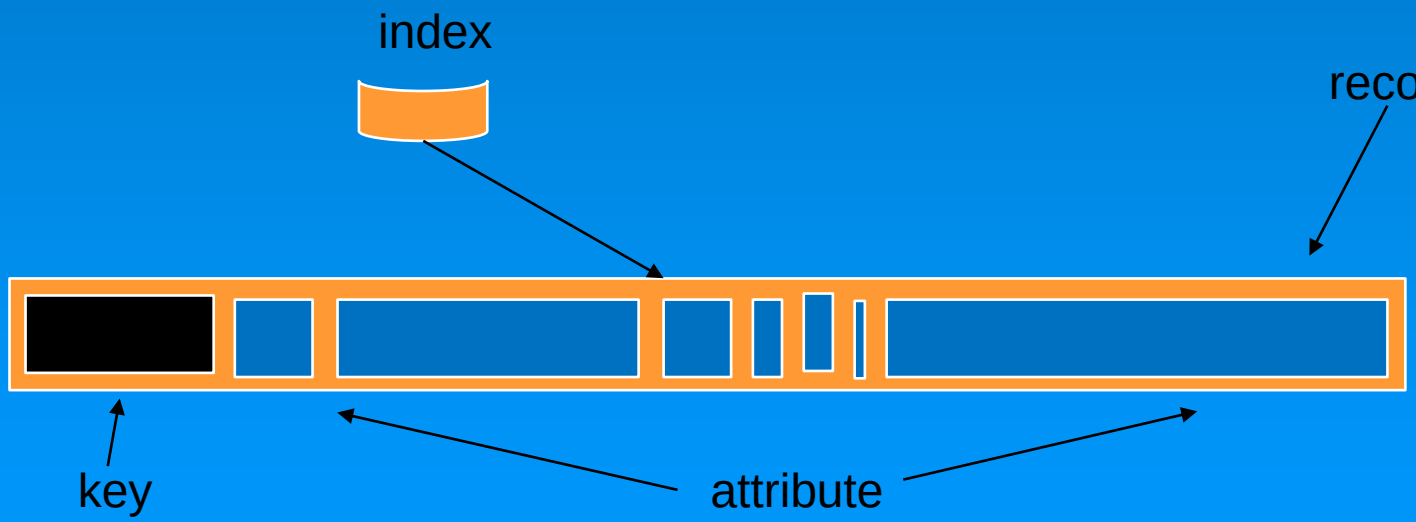
Each of the different types of data have their own unique characteristics

FOREST RIM TECHNOLOGY INC

Structured data

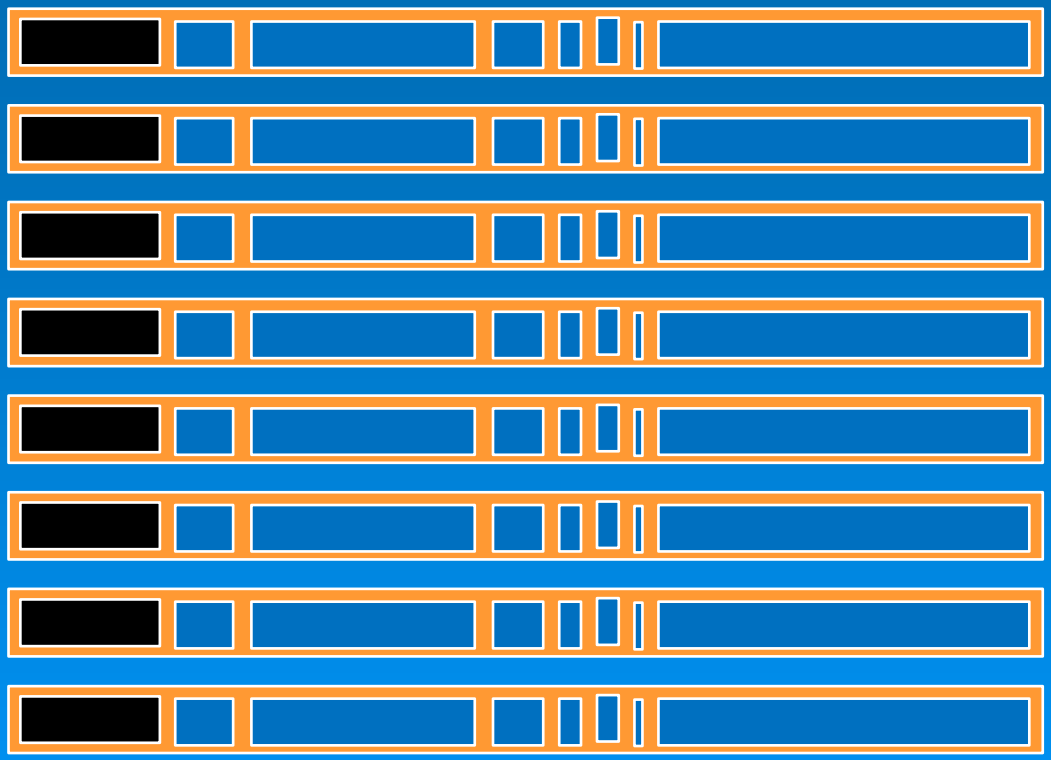
Usually transaction based

Bank transactions
Point of sale
Telephone call
Payments made
Payments received
……



FOREST RIM TECHNOLOGY INC

Structured data



The same record type is repeated
Each record has different contents

FOREST RIM TECHNOLOGY INC

Textual data

Text is found everywhere

Medical records
Contracts
Internet
Call centers
Warranty claims
Insurance claims
Email
……..

Voice
Written
Internet
Video
……..

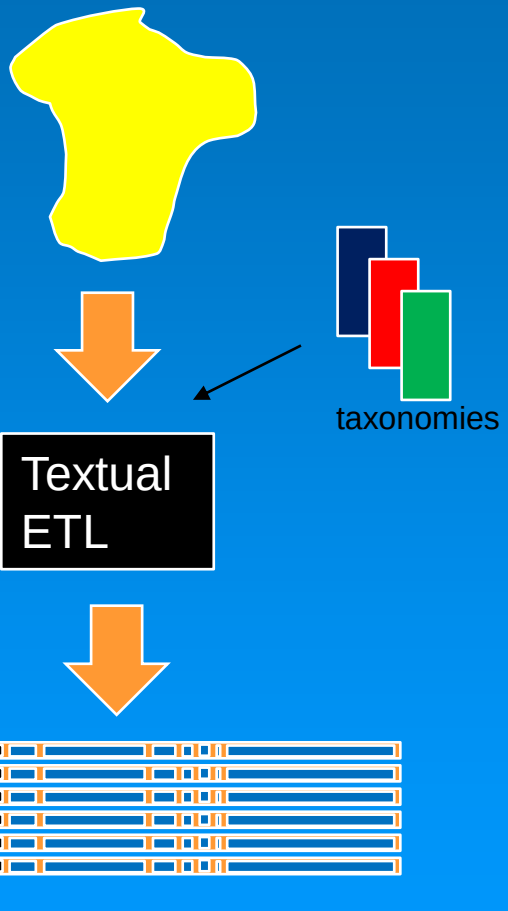
English
Spanish
Portuguese
French
Mandarin
Korean
German

Formal language
Slang
Acronyms
……..

FOREST RIM TECHNOLOGY INC

Textual data

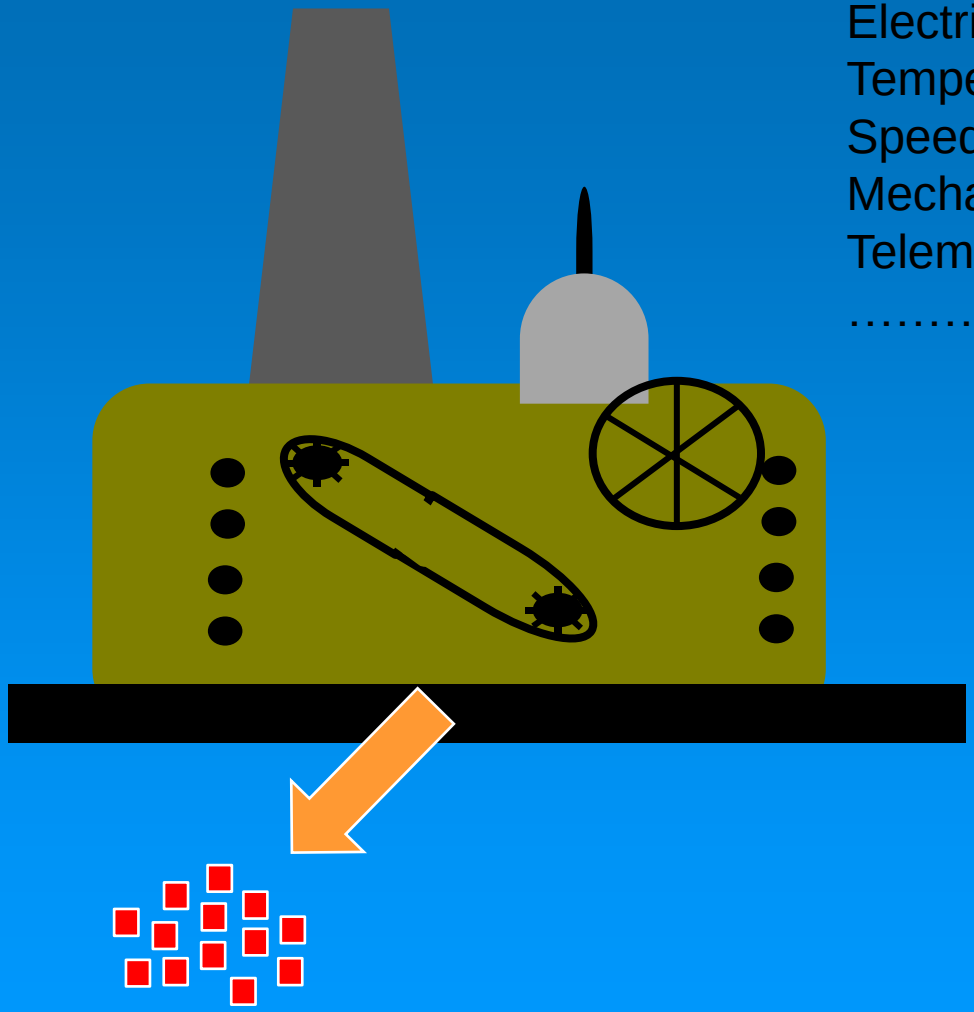
Text is transformed Into a structured format



FOREST RIM TECHNOLOGY INC

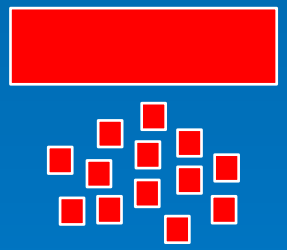
Analog/IoT

Machine generated



Drones
Electric eye
Temperature gauge
Speed
Mechanical
Telemetry
……

Analog/IoT



Surveillance camera

99.999% of data is not useful
.001% of data is really useful

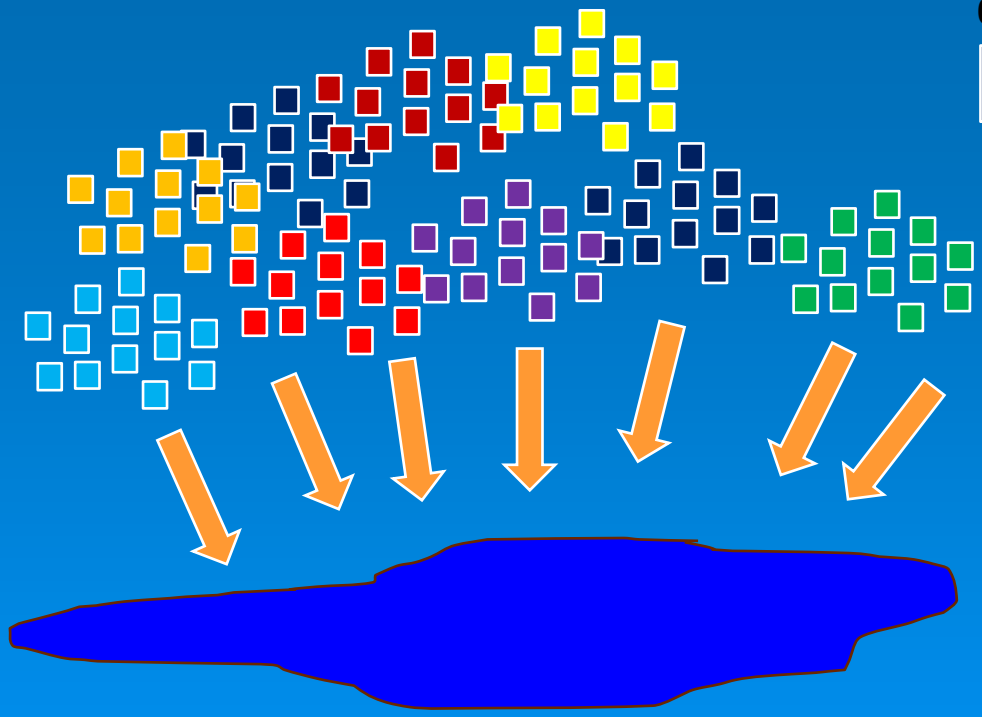
Surveillance data must be distilled before it is useful
You don’t want to store vast volumes of data that are useless

FOREST RIM TECHNOLOGY INC

Analog/IoT

Structured data

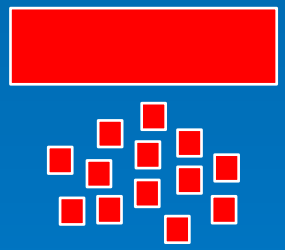
Textual data

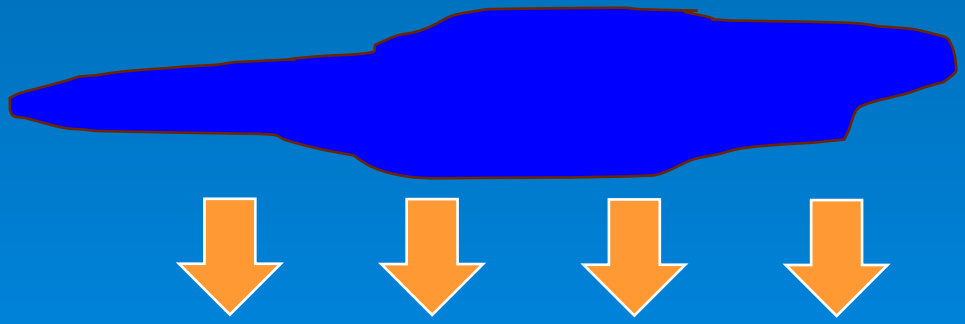


The data lake is created by throwing data into the lake

FOREST RIM TECHNOLOGY INC

Analog/IoT

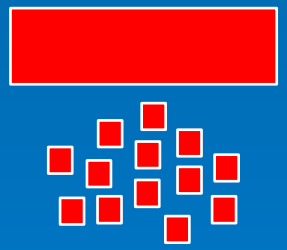




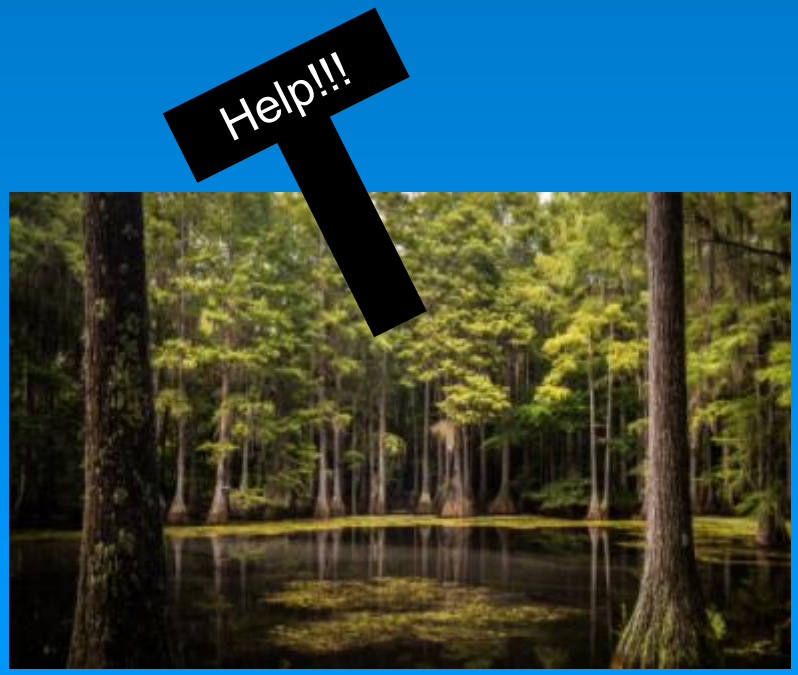
Soon the data lake
turned into a swamp

FOREST RIM TECHNOLOGY INC

Analog/IoT

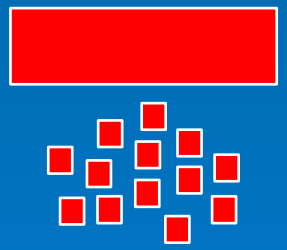


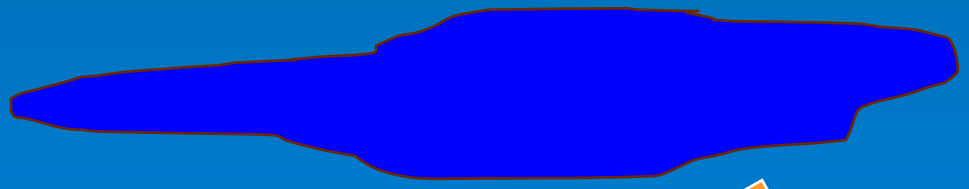
The data swamp was not good for anyone….



FOREST RIM TECHNOLOGY INC

Analog/IoT

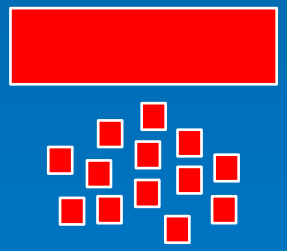


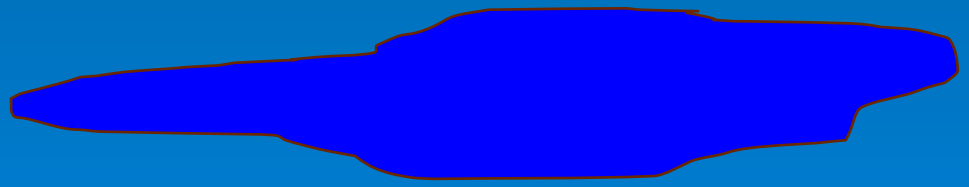


The data lake needs to be transformed into a lakehouse

FOREST RIM TECHNOLOGY INC

Analog/IoT



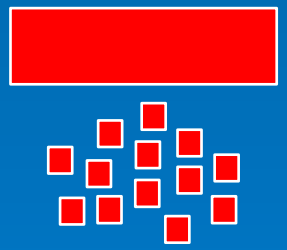


Data scientist

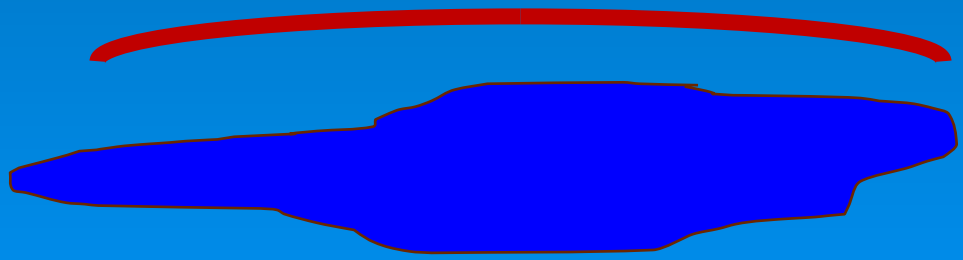
All this education and 95% of my job is being a data garbageman

FOREST RIM TECHNOLOGY INC

Analog/IoT



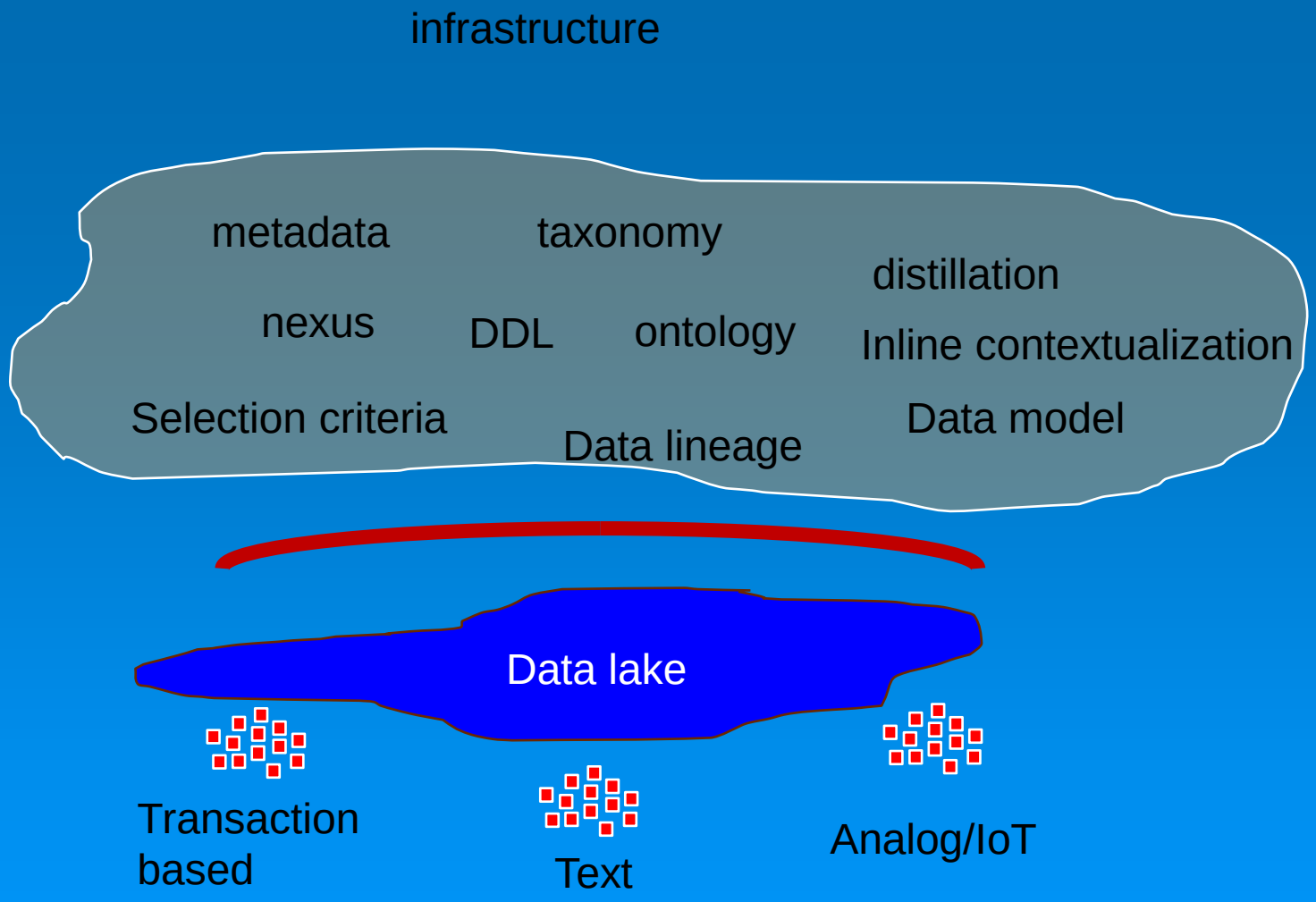
infrastructure



Data scientist

Ah, that’s more like it

FOREST RIM TECHNOLOGY INC



Turning the data lake into a data lakehouse

infrastructure

metadata taxonomy distillation

nexus DDL ontology Inline contextualization

Selection criteria Data lineage Data model

Transaction based

ETL

Data lake

Getting transaction based data into the data lake

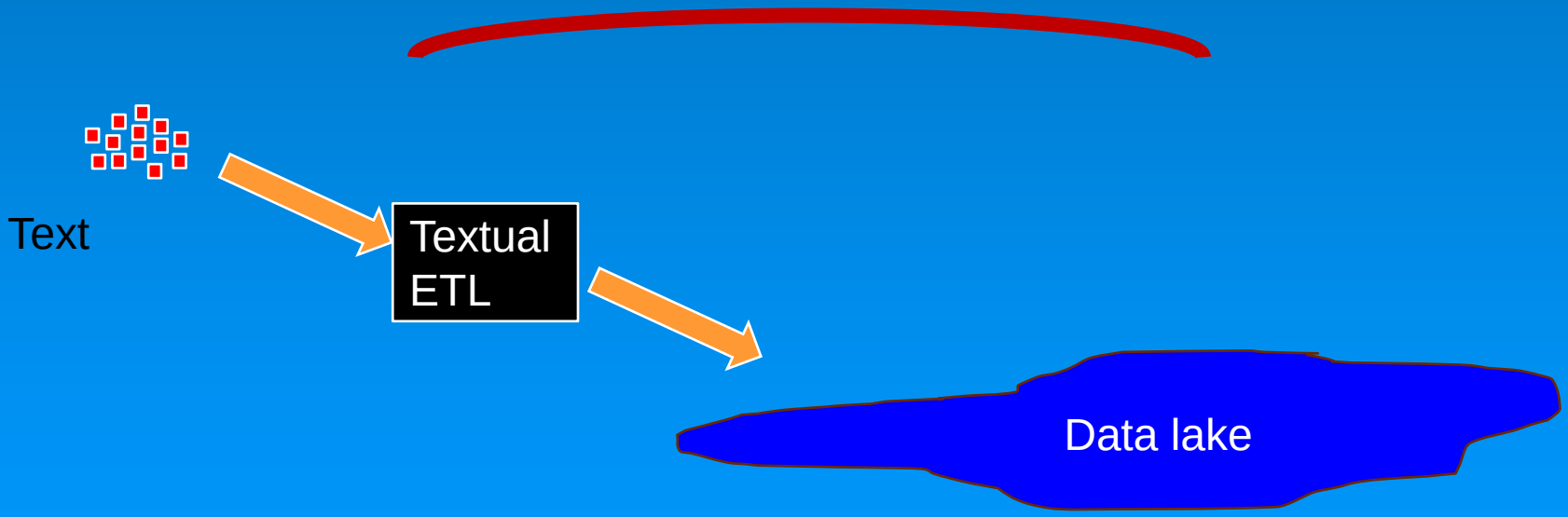
FOREST RIM TECHNOLOGY INC

infrastructure

metadata taxonomy distillation

nexus DDL ontology Inline contextualization

Selection criteria Data lineage Data model



Getting text based data into the data lake

FOREST RIM TECHNOLOGY INC

infrastructure

metadata taxonomy distillation

nexus DDL ontology Inline contextualization

Selection criteria Data lineage Data model

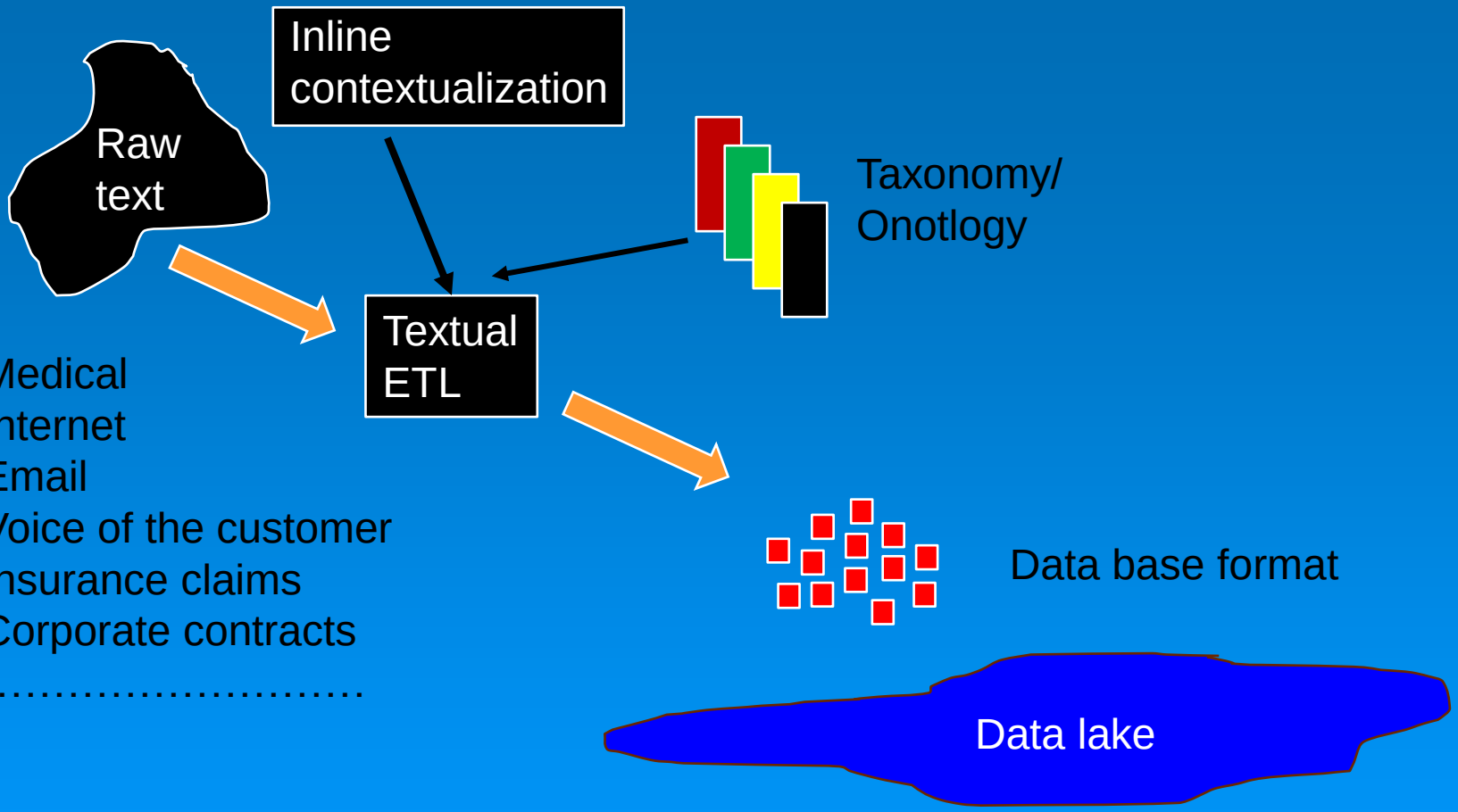
Analog/
IoT

Distillation/
refinement

Data lake

Getting analog/IoT based data into the data lake

FOREST RIM TECHNOLOGY INC

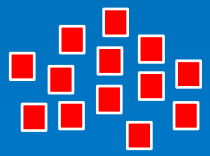


Bringing text into the data lakehouse

FOREST RIM TECHNOLOGY INC

Analog/IoT

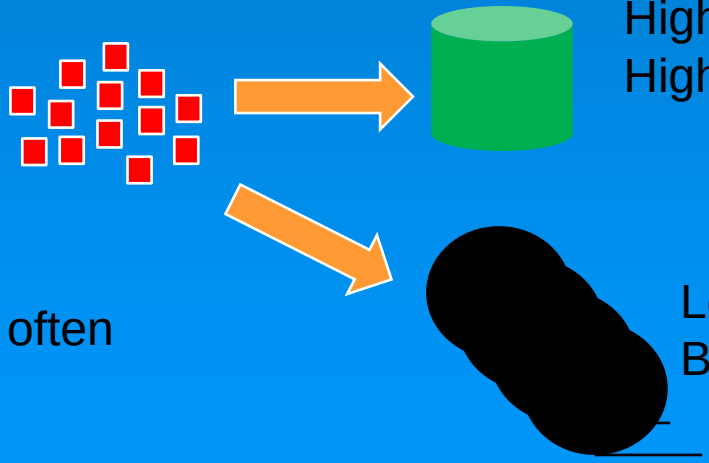
Machine generated



Basic, raw measurements

Time – 0912
Time – 0916
Time – 1002
Time – 1008
Time – 1017
…….

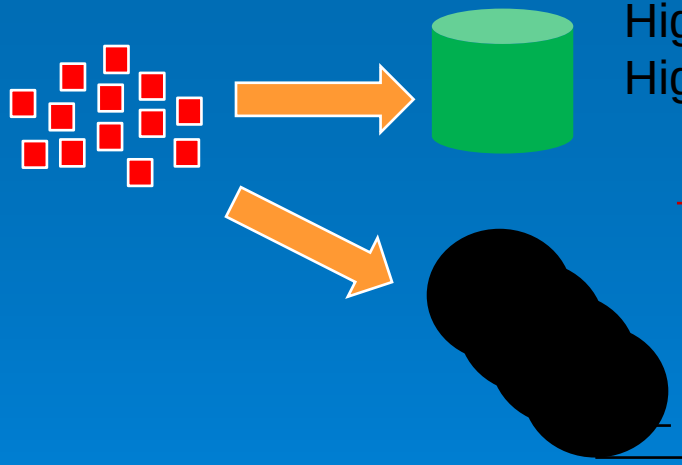
The distillation process



High probability
High performance

Low probability
Bulk storage

Analog/IoT data is often segmented



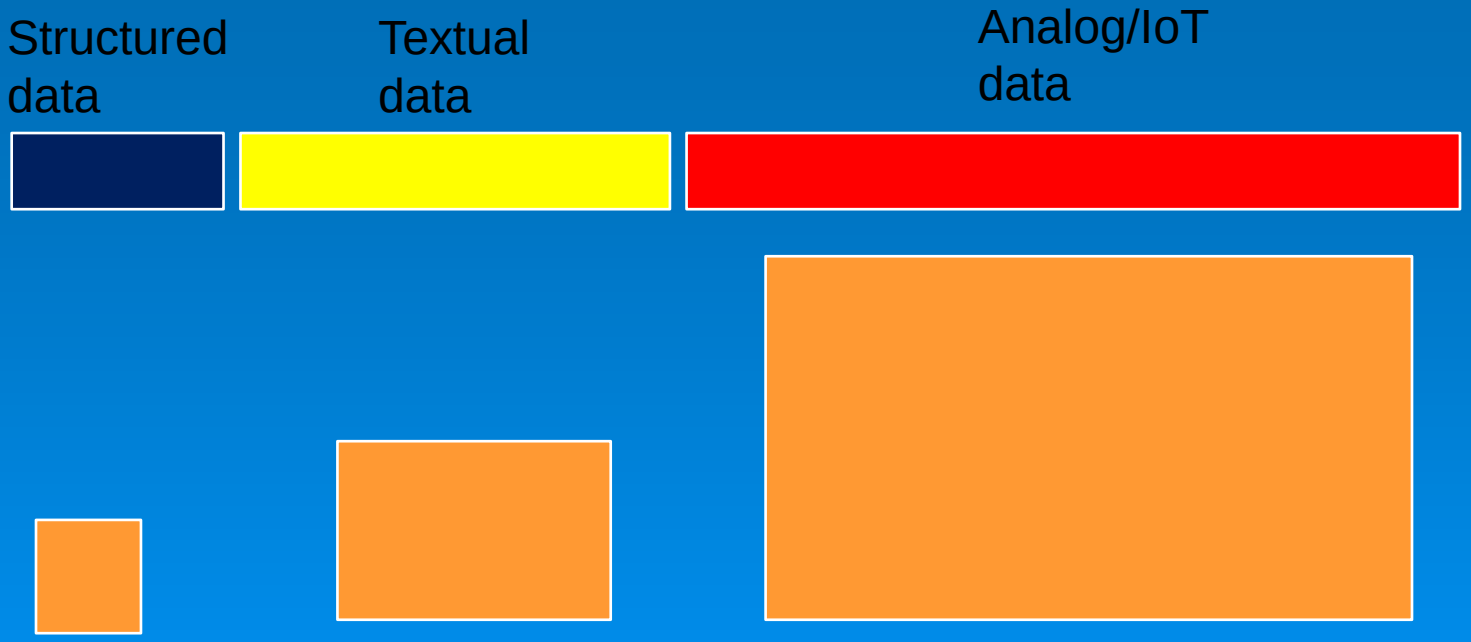
High probability
High performance

Robbery 6:31 am Tuesday
3.5 minutes
Accident 5:15 pm Thursday
1.0 minutes

Low probability
Bulk storage

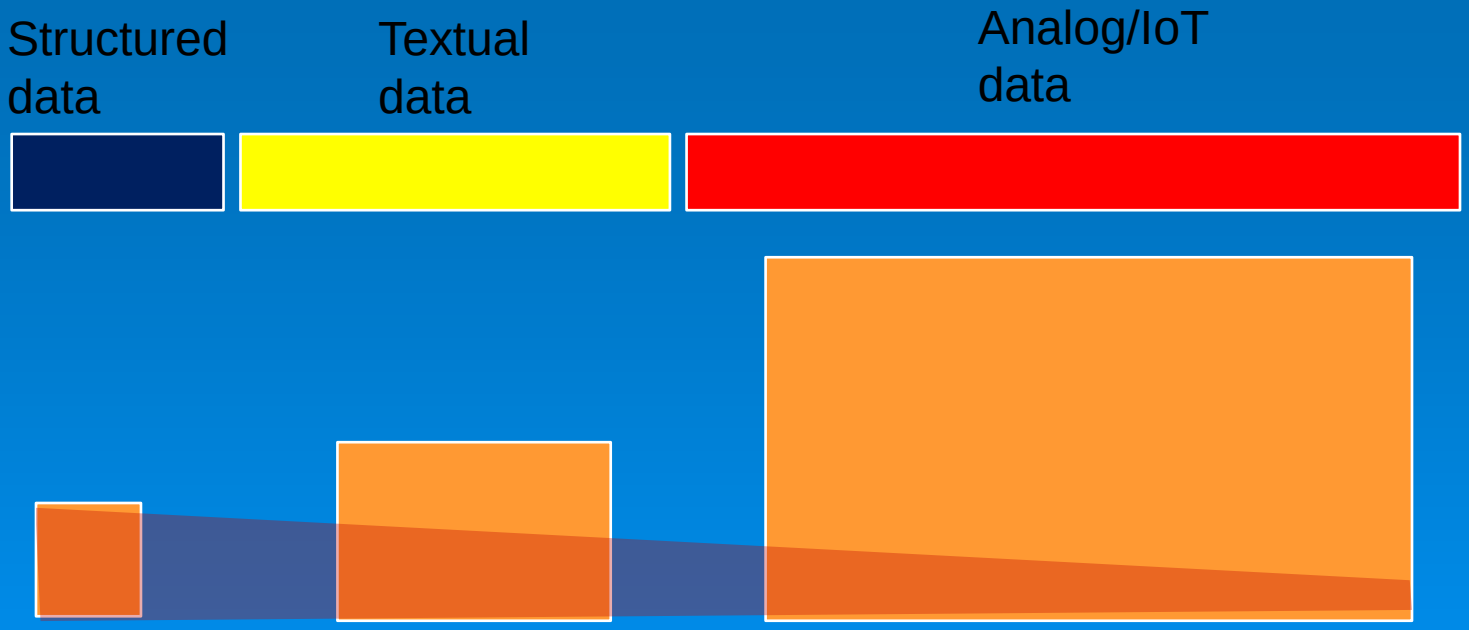
Image 5:19 am
Image 5:20 am
Image 5:21 am
……

FOREST RIM TECHNOLOGY INC



Relative volumes of data in each sector

FOREST RIM TECHNOLOGY INC



Business value and the volumes of data

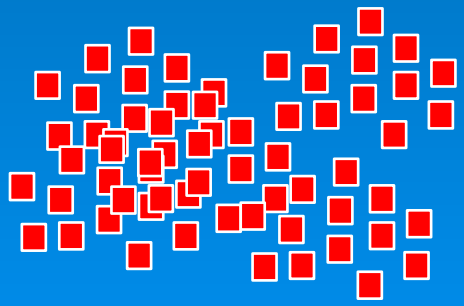
FOREST RIM TECHNOLOGY INC

Format compatibility

Structured data

Textual data

Analog/IoT data



Relational format

Raw data format

From a format standpoint, the structured and the textual environments are very different from the analog/IoT environment

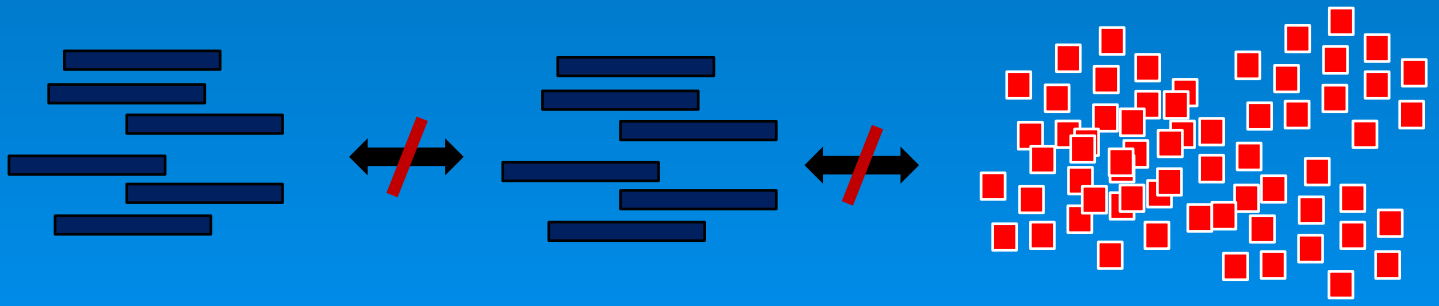
FOREST RIM TECHNOLOGY INC

Content compatibility

Structured data

Textual data

Analog/IoT data



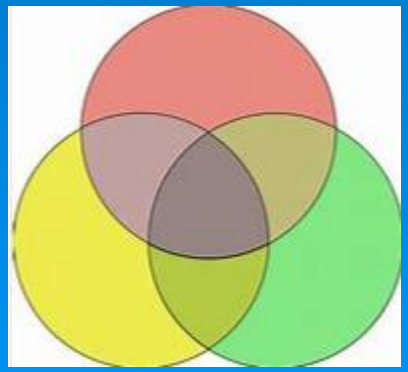
Key compatibility – very unintegrated

FOREST RIM TECHNOLOGY INC

Structured data

Textual data

Analog/IoT data



In order to do analytics, there must be some common data on which to do a comparison

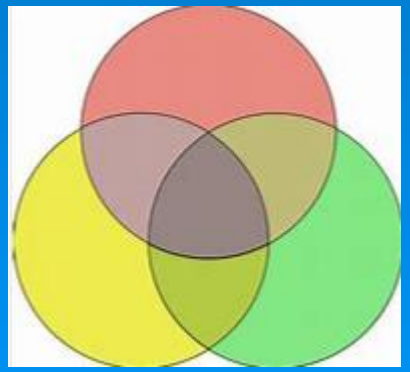
Without common data it is very difficult to do a meaningful comparison

FOREST RIM TECHNOLOGY INC

Structured data

Textual data

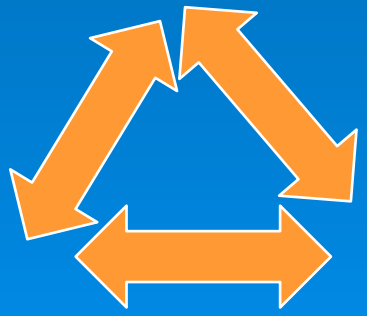
Analog/IoT data



The problem is that there may be no obvious, easy way to isolate common identifiers

FOREST RIM TECHNOLOGY INC

Textual data



Structured data

Analog/IoT data

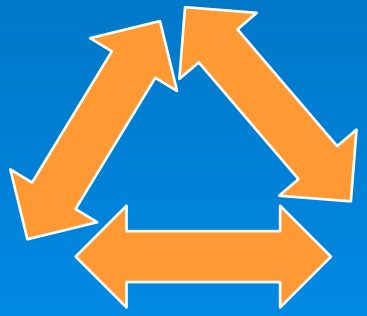
Fortunately there are such things as universal common connectors

Universal common connectors exist regardless of the way that data has been collected

FOREST RIM TECHNOLOGY INC

General common connectors

Textual
data



Structured
data

Analog/IoT
data

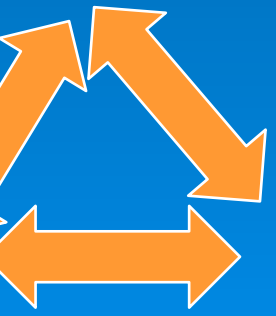
Universal common connector for anything

geography
time
dollar amount

FOREST RIM
TECHNOLOGY INC

Common connectors for humans

Textual
data



Structured
data

Analog/IoT
data

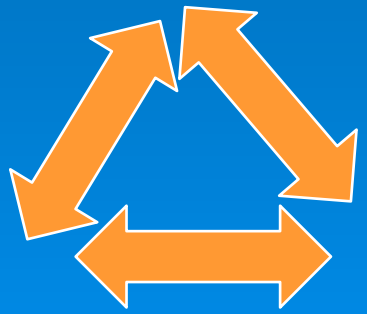
Universal common connector for humans

gender
age
race

FOREST RIM
TECHNOLOGY INC

Common connectors for objects

Textual
data



Structured
data

Analog/IoT
data

Universal common connector for physical objects

weight
color
cost
size
shape

FOREST RIM
TECHNOLOGY INC

Common connectors for objects

Textual data

Structured data

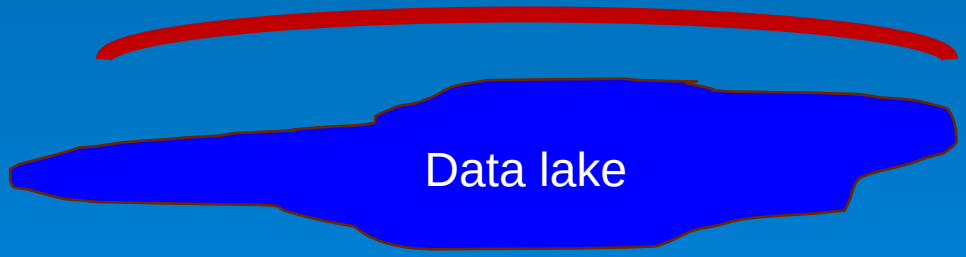
Analog/IoT data

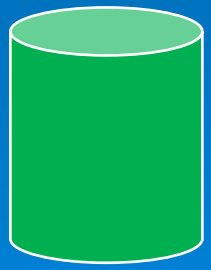
All things considered, the universal connectors are a weak link for trying to connect and relate data

FOREST RIM TECHNOLOGY INC

A frequently asked question…..

Is a data warehouse the same thing as the data lakehouse?





Data Warehouse

Are two cousins the same person?

FOREST RIM TECHNOLOGY INC

With analytics from the data lakehouse, you can improve the lives and livelihood of many people

FOREST RIM TECHNOLOGY INC

If you want to learn more – for FREE –
A book on bringing text into your lakehouse –

Go to www.forestrimtech.com, ebook, and tell us where to send it to

FOREST RIM TECHNOLOGY INC