The Data Lakehouse Symposium – February, 2022

FORESTRIM TECHNOLOGY

Hosted by - Bill Inmon and DataBricks – Feb 1- 4, 2022

The Data Lakehouse Symposium – February, 2022

FORESTRIM TECHNOLOGY

Text in the Data Lakehouse

David Rapien

Partner in ForestRim Technology

Associate Professor – Lindner College of Business

University of Cincinnati

Let’s Look at the Major Historical Changes in Data Collection, Storage, and Usage

- **1980’s** - The Data Warehouse allowed us to hold a single version of the truth and make enterprise wide decisions.
- **2010** - The Data Lake allowed us to collect all of our “data” in one place.
- **2020**- The Data Lakehouse marries the two by adding governance and metadata to data going into the Data Lake so that it can be separately transitioned into a Data Warehouse AND consumed by decision makers and analysts.

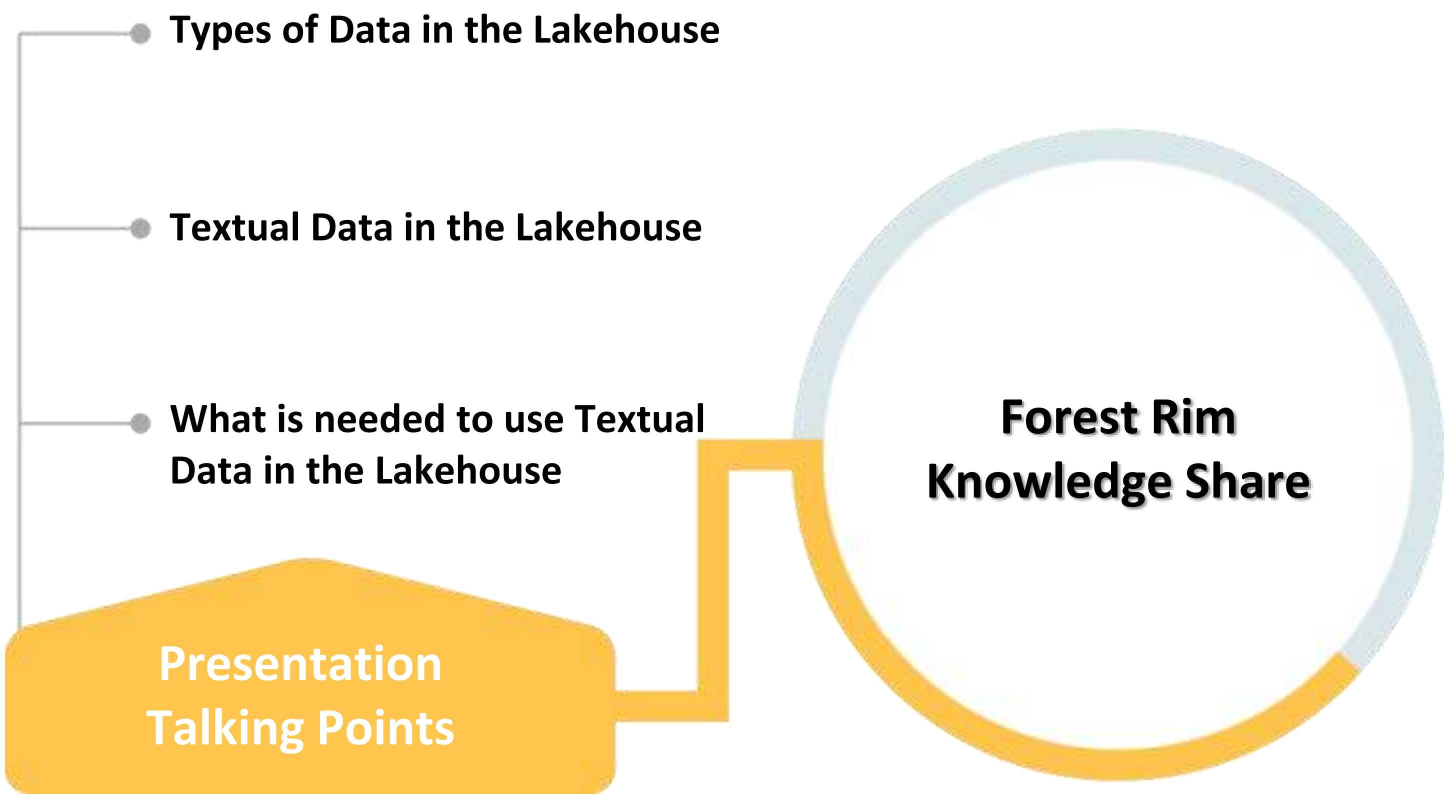
Where is your company’s data focus?

- Data collection for “future use”?
- Business decisions?
- Analysis and research?
- We have none!

What types of data does your company collect and store?

- Transactional Data from customer interactions?
- Machine generated data?
- Emails, blogs, customer reviews, medical records, contracts?
- Images, videos, scans, audio files?

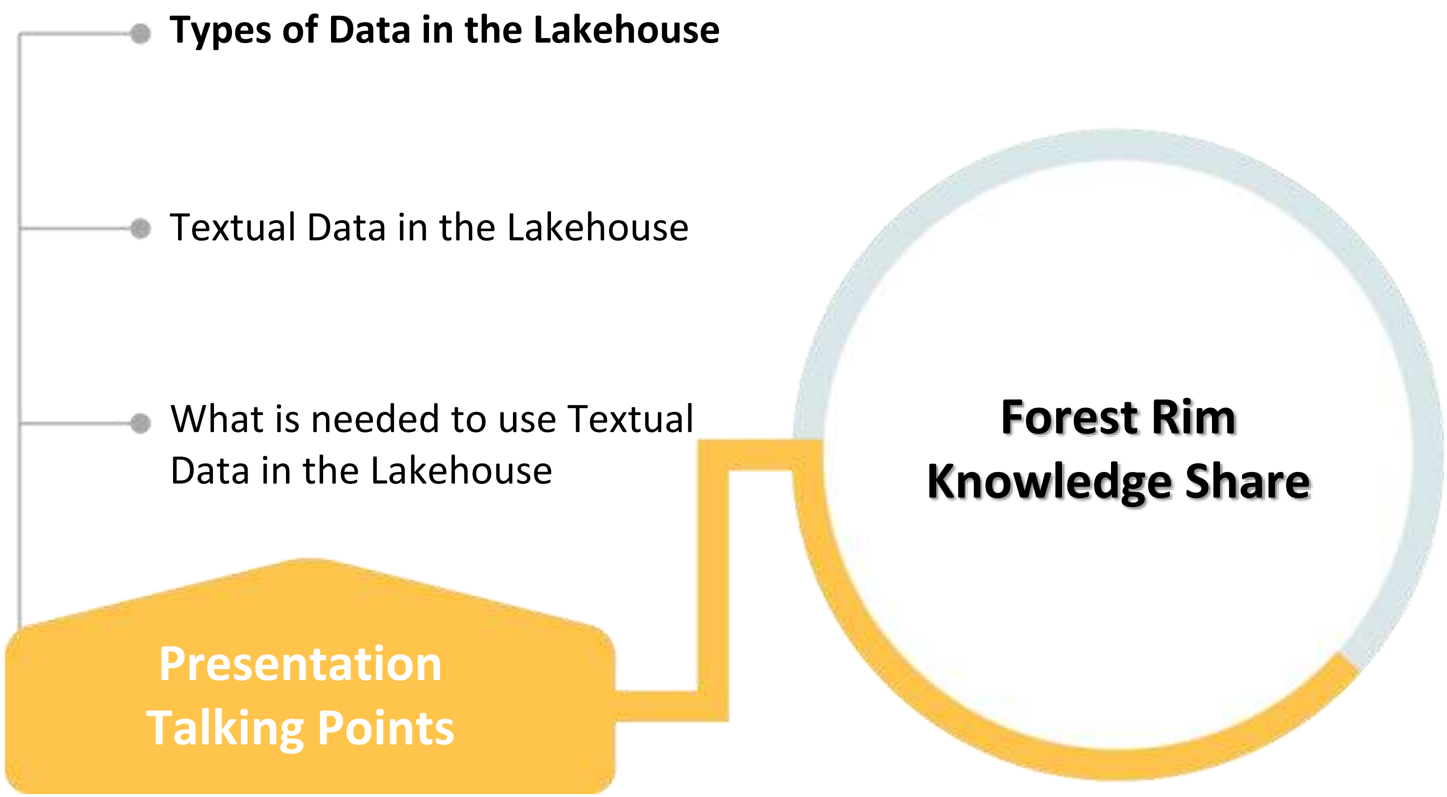
What We will Discuss in Today’s Presentation



- Types of Data in the Lakehouse
- Textual Data in the Lakehouse
- What is needed to use Textual Data in the Lakehouse

Presentation Talking Points

Forest Rim Knowledge Share

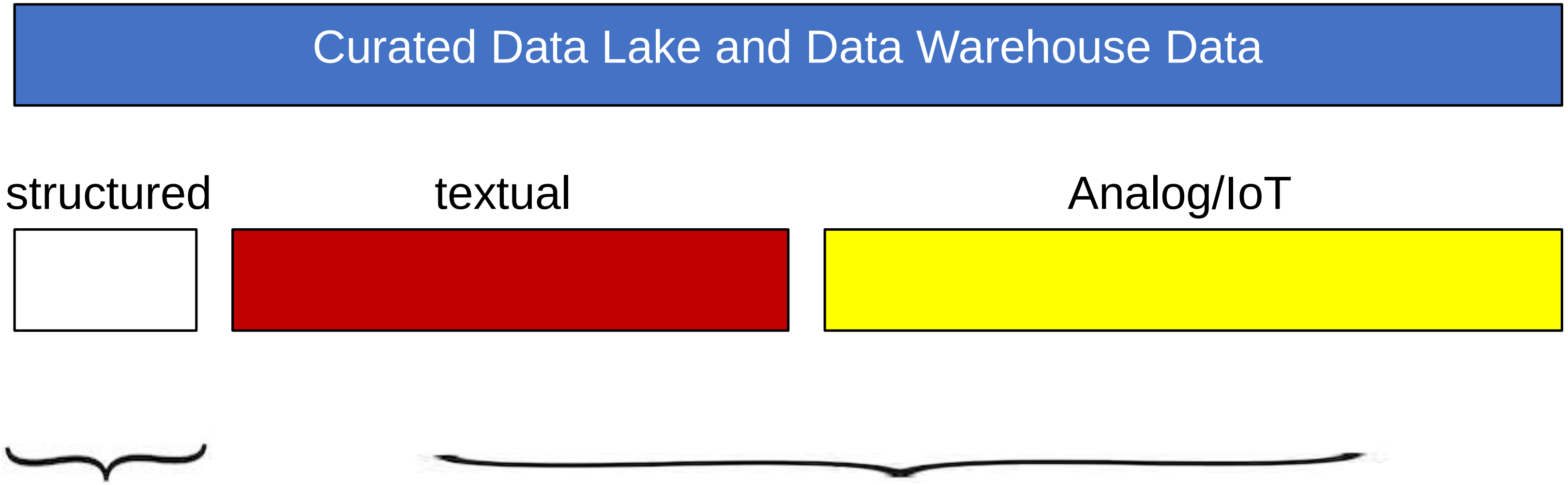


- **Types of Data in the Lakehouse**
- Textual Data in the Lakehouse
- What is needed to use Textual Data in the Lakehouse

Presentation Talking Points

Forest Rim Knowledge Share

All Corporate Data in the Lakehouse



Curated Data Lake and Data Warehouse Data

structured

textual

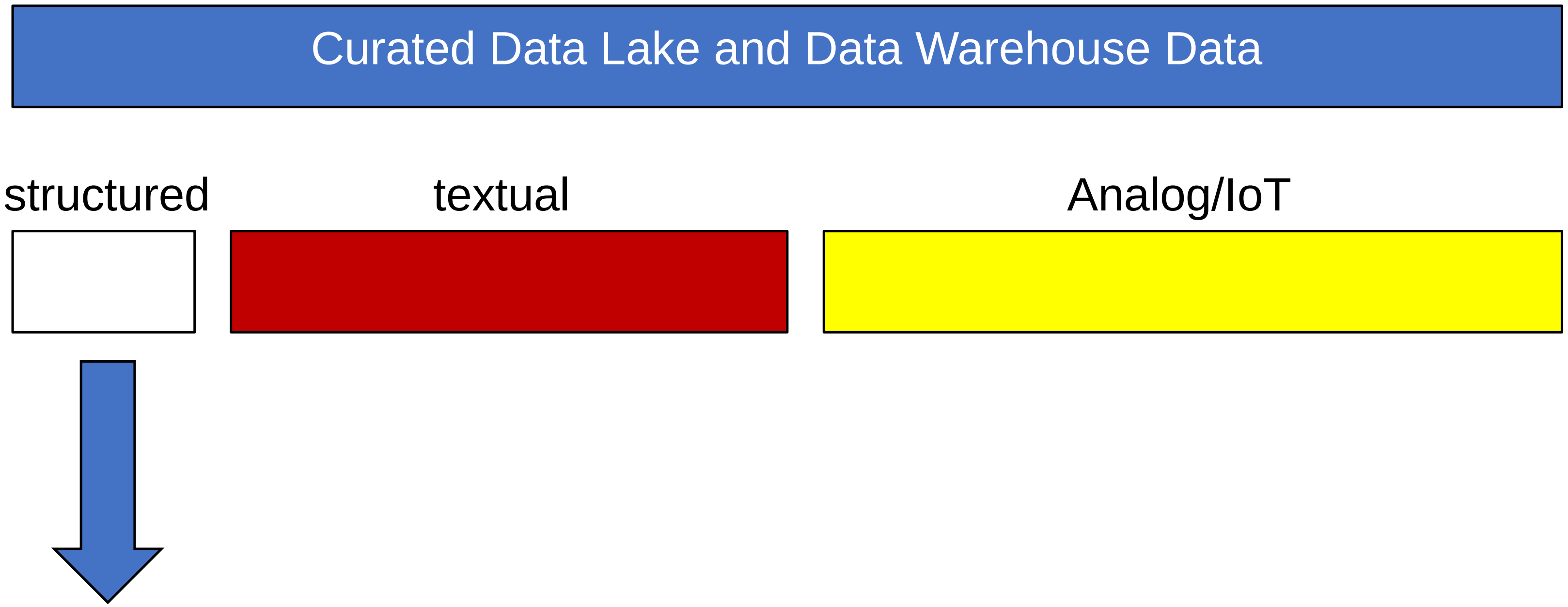
Analog/IoT

Amount of Data: ~ 20-% ~ 80+%

Pareto’s Law Holds True

Data Used for Decision Making: ~ 80+% ~ 20-%

All Corporate Data in the Lakehouse



Curated Data Lake and Data Warehouse Data

structured

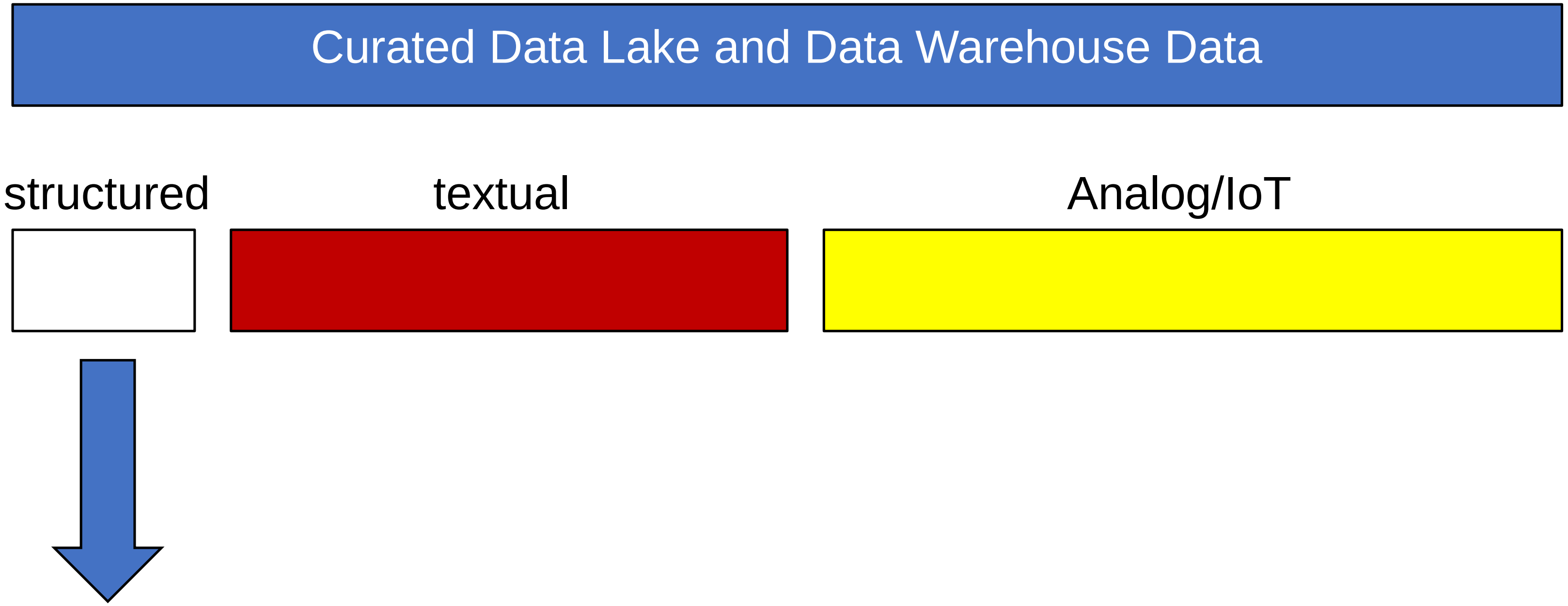
textual

Analog/IoT

~ 80-90% of business decisions are made on less than 20% of the data.

Is there something wrong here?

All Corporate Data in the Lakehouse



Curated Data Lake and Data Warehouse Data

structured

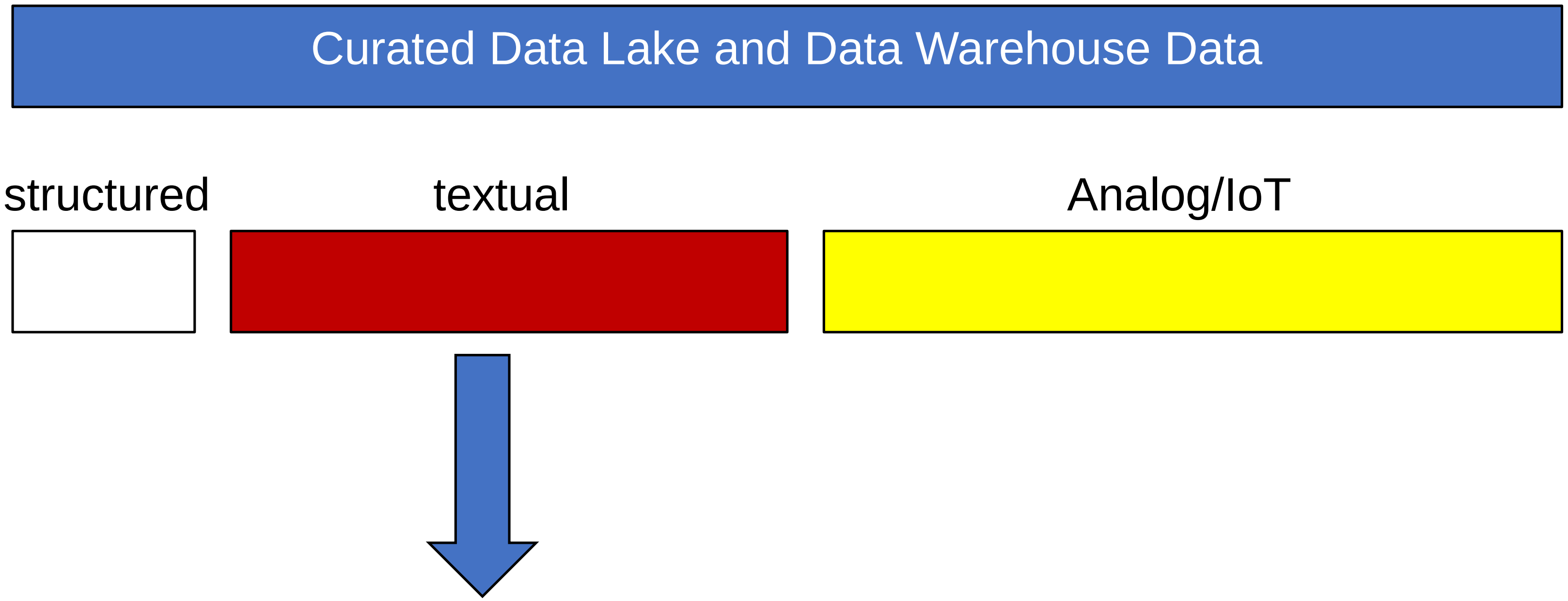
textual

Analog/IoT

Data Warehouse Data

- Physical models
- Tables
- Aggregated
- Scrubbed
- Additional Metadata
- Additional Data Governance

All Corporate Data in the Lakehouse



Curated Data Lake and Data Warehouse Data

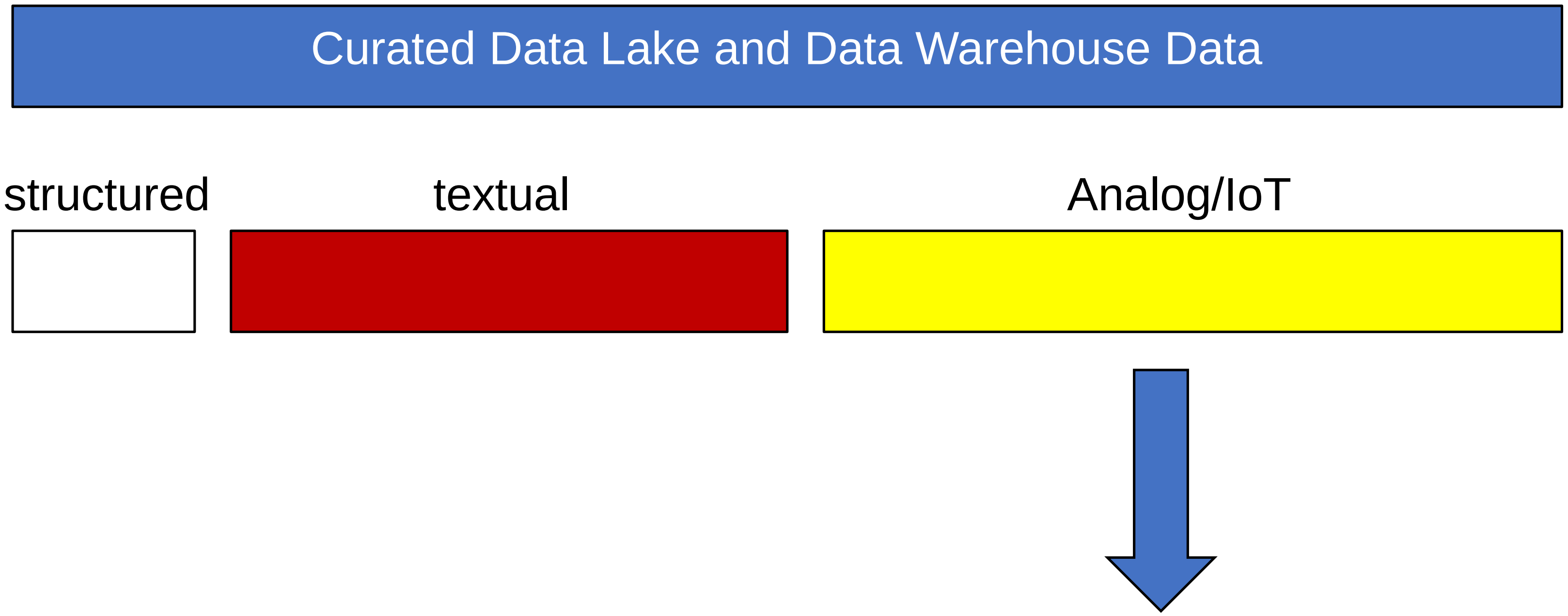
structured

textual

Analog/IoT

- Documents
- Emails
- Contracts
- Medical Records
- Voice of the Customer
- Insurance Claims
- Call Center …
- Other???

All Corporate Data in the Lakehouse



Curated Data Lake and Data Warehouse Data

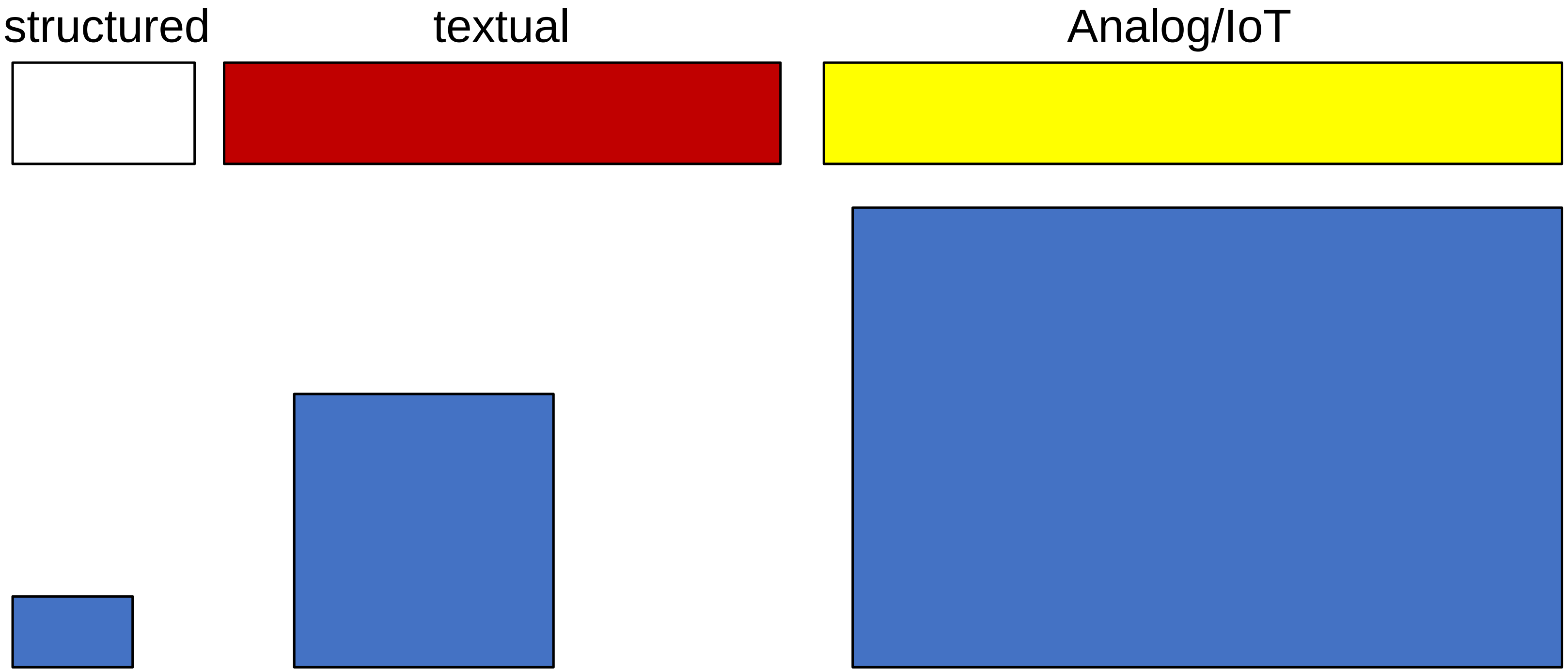
structured

textual

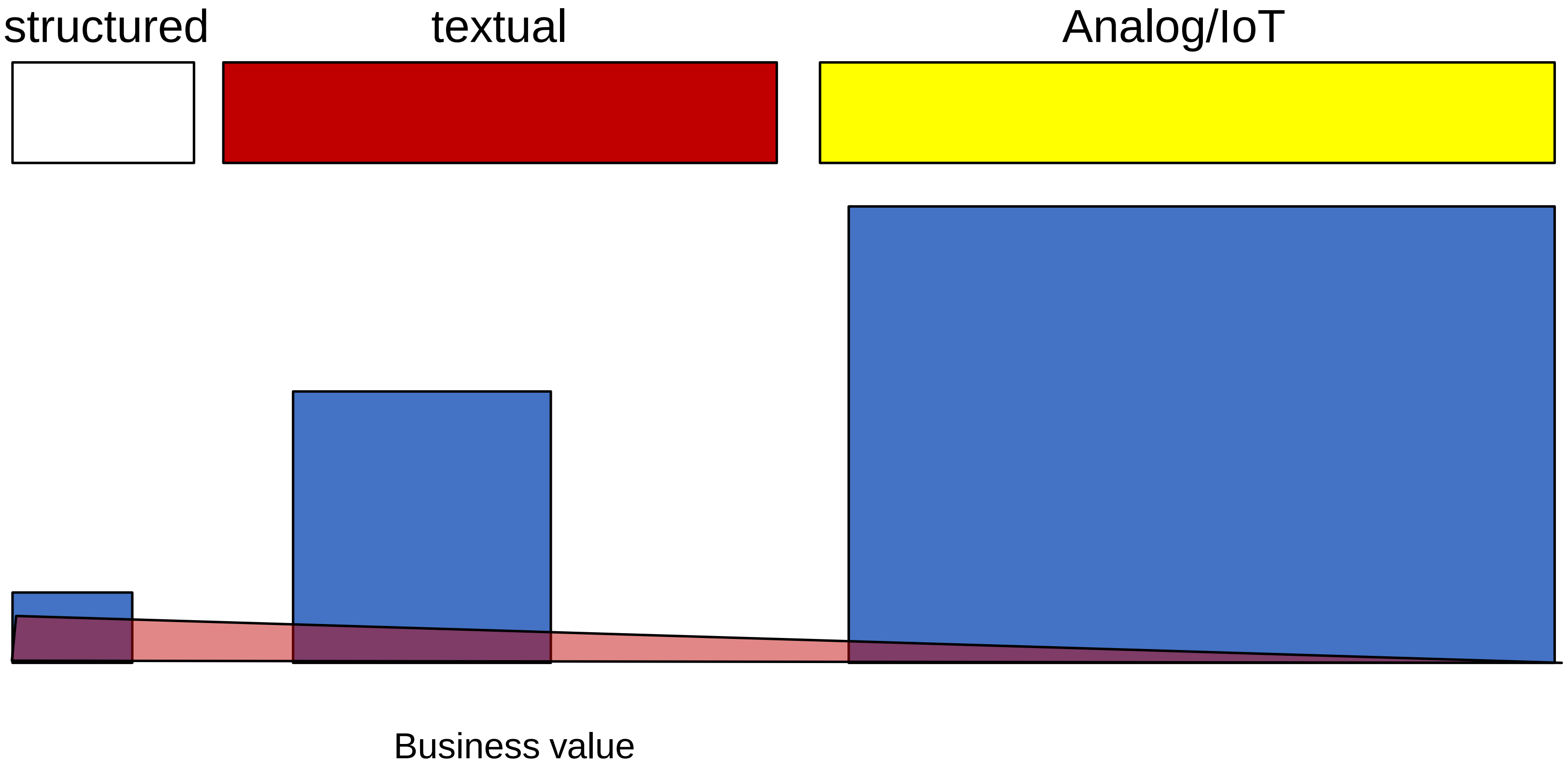
Analog/IoT

- Status Data
- Automation Data
- Location Data
- Clickstreams
- Sensor Data
- Images
- Audio / Video files

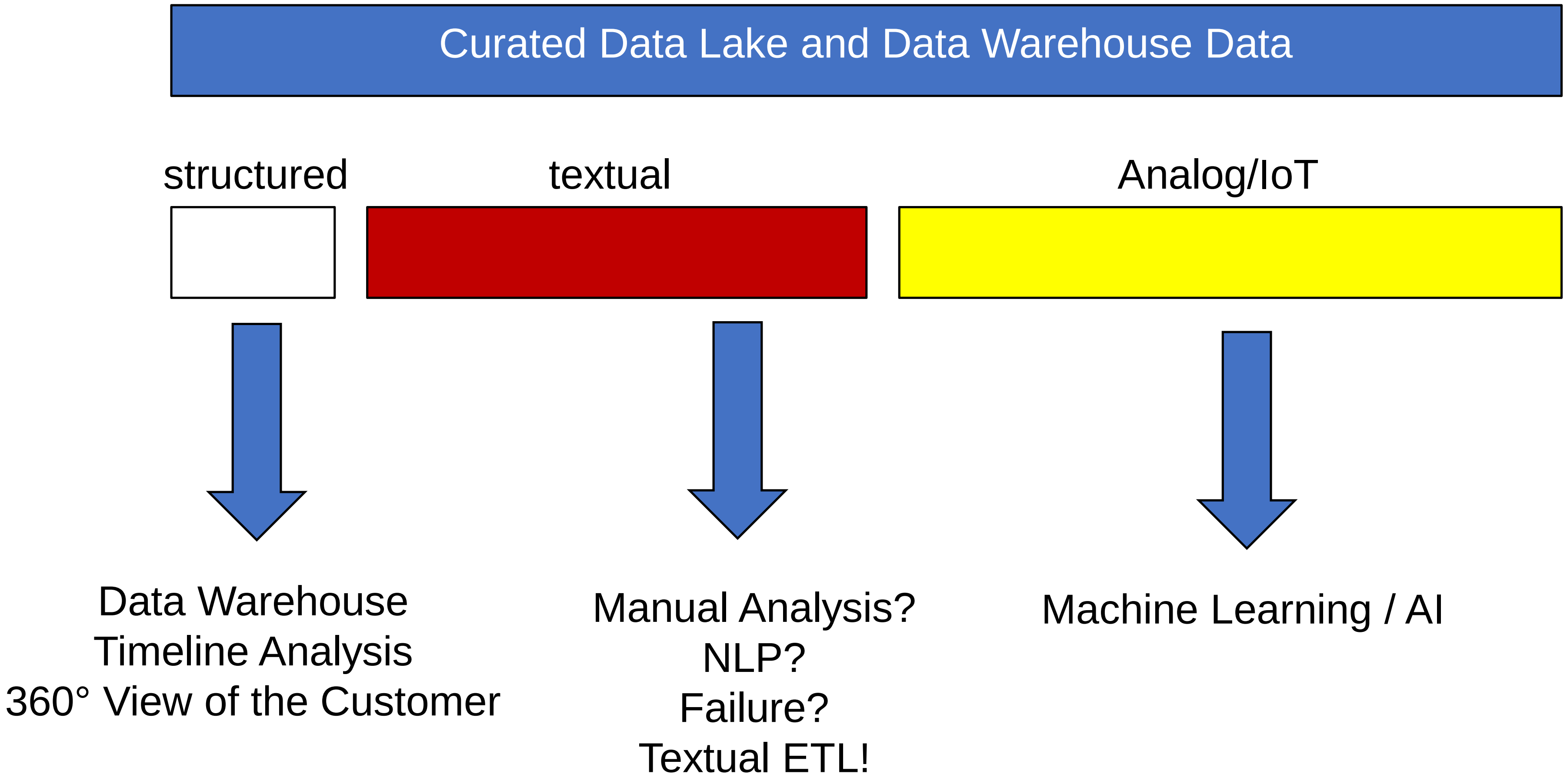
The relative volumes of data



The relative amount of business value to be found in the different sectors



How do we currently USE different types of data



Curated Data Lake and Data Warehouse Data

structured

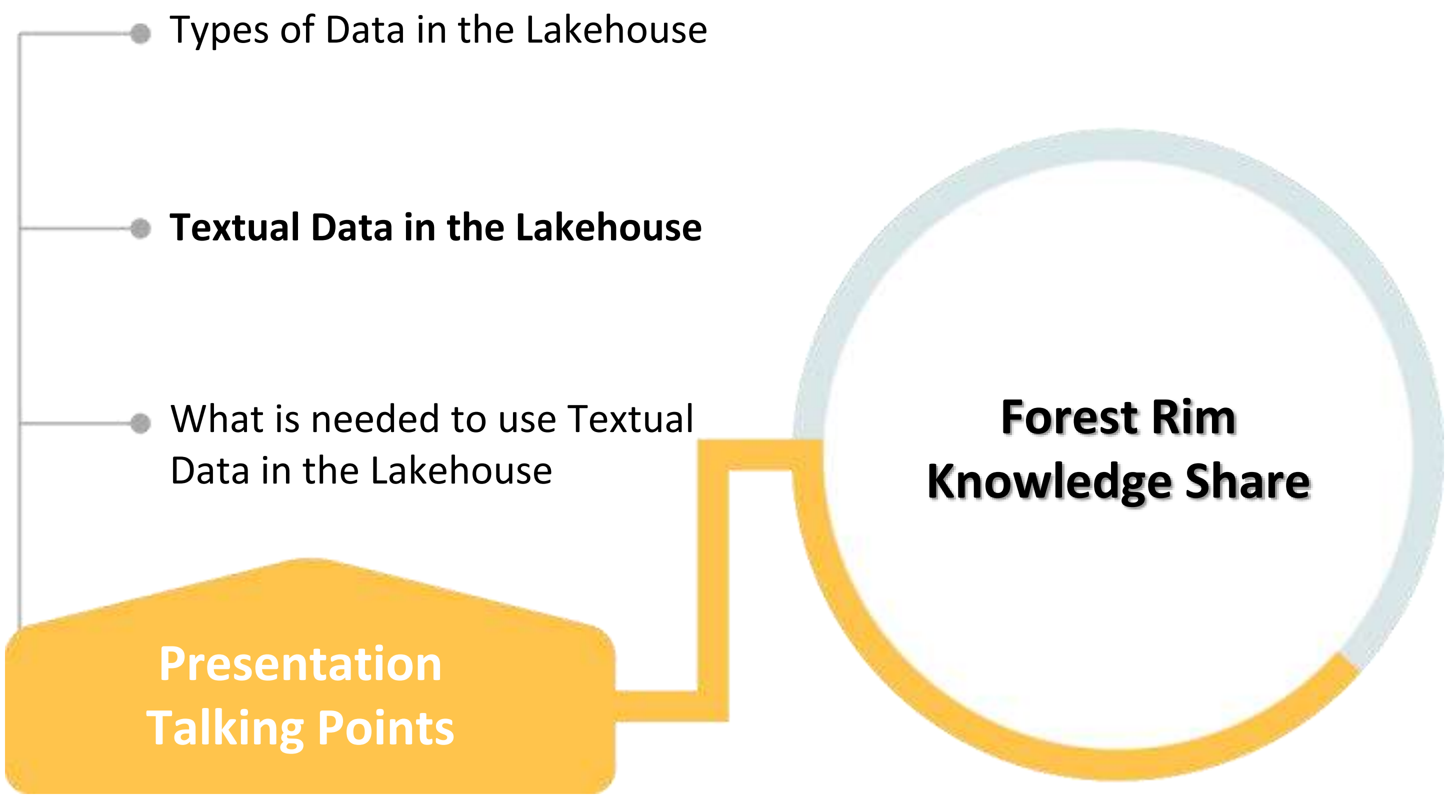
textual

Analog/IoT

Data Warehouse
Timeline Analysis
360° View of the Customer

Manual Analysis?
NLP?
Failure?
Textual ETL!

Machine Learning / AI

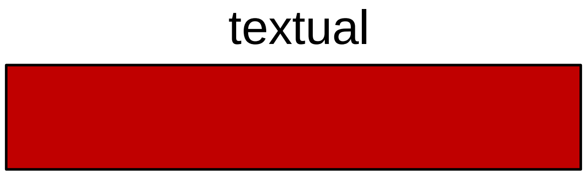


- Types of Data in the Lakehouse
- **Textual Data in the Lakehouse**
- What is needed to use Textual Data in the Lakehouse

Presentation Talking Points

Forest Rim Knowledge Share

What data you are missing in your analysis?



Contracts

Call center

Emails

Medical records

Warranties

Insurance claims

Yelp Reviews

Parquet Files

Real estate deeds/sales

Internet

Dictations

The voice of the customer

Voice mails

Document scans

Transcriptions

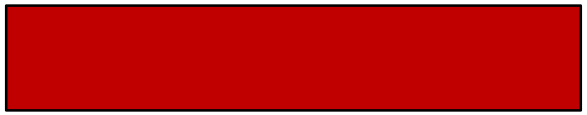
PDFs

CSVs

Word documents

What is similar about most of this textual data?

textual



It is stream of thought
It is different document by document
It does not have Primary Keys or Foreign Keys
It has little format
It is DIRTY DATA!

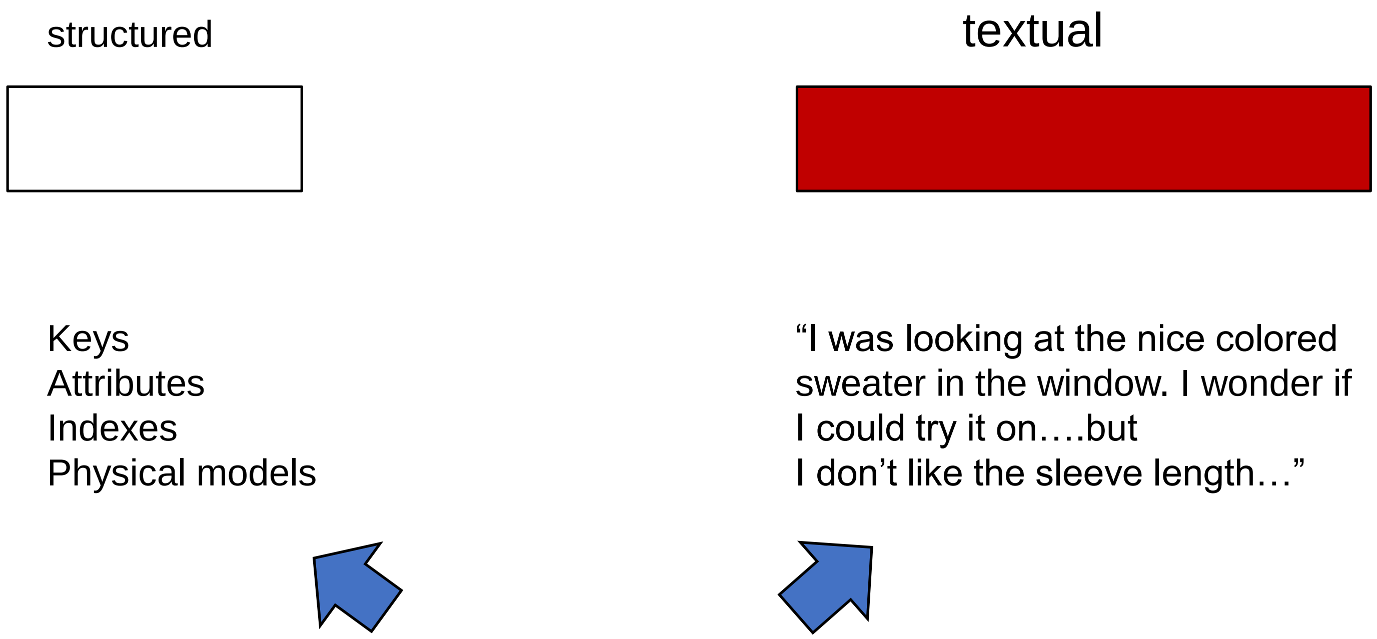
The Issue:

The modelling and design techniques that worked in the **Structured** world do not work in the world of **Text**.

Why?

Because people do not write or talk the same way that is found in the structured world

Think About it:



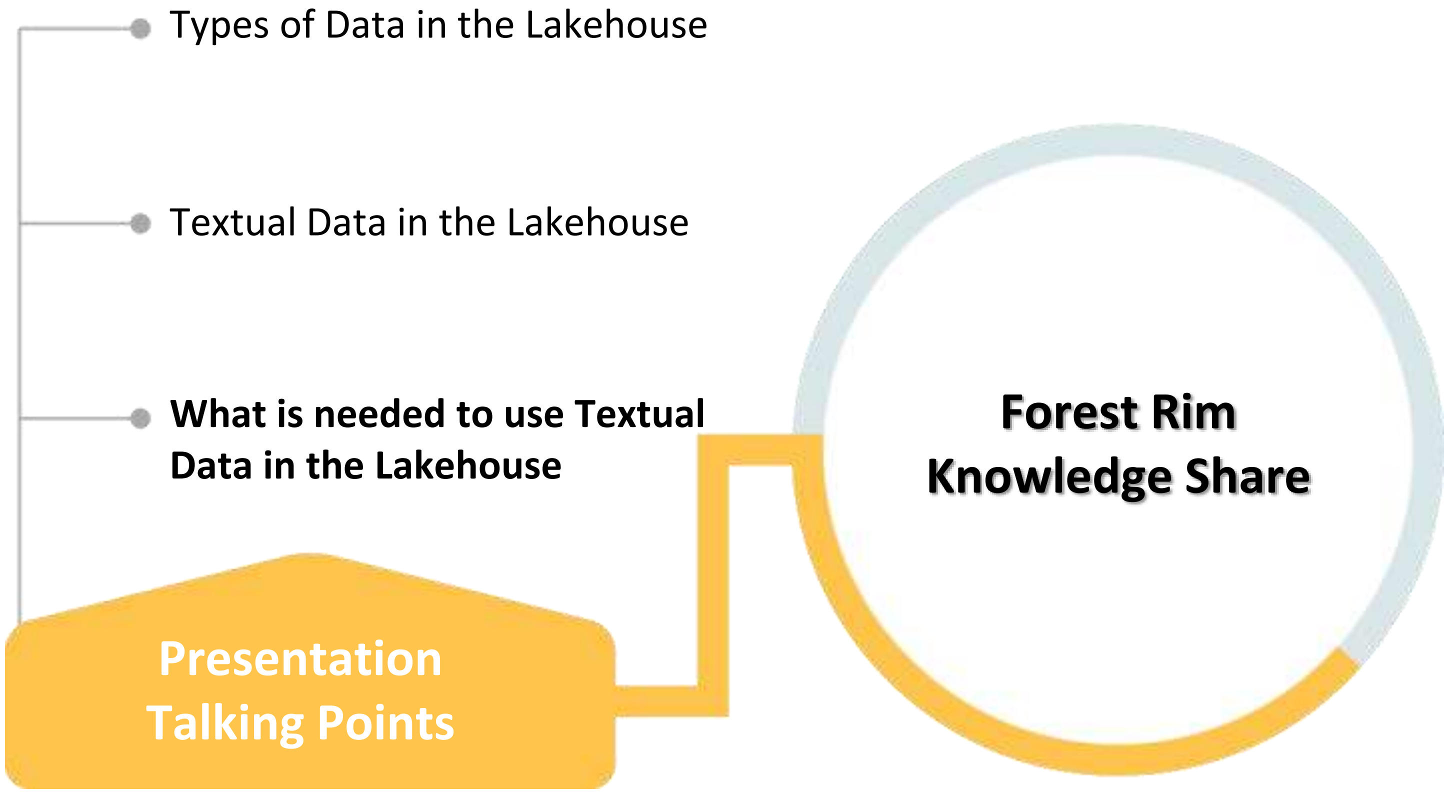
structured

textual

Keys
Attributes
Indexes
Physical models

“I was looking at the nice colored sweater in the window. I wonder if I could try it on….but I don’t like the sleeve length…”

These worlds are incompatible.
In order to address text you need a *completely different approach*

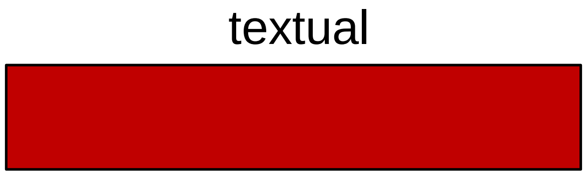


- Types of Data in the Lakehouse
- Textual Data in the Lakehouse
- **What is needed to use Textual Data in the Lakehouse**

Presentation Talking Points

Forest Rim Knowledge Share

We consider the types of text that we are storing and NOT USING?



Contracts

Call center

Emails

Medical records

Warranties

Insurance claims

Yelp Reviews

Parquet Files

Real estate deeds/sales

Internet

Dictations

The voice of the customer

Voice mails

Document scans

Transcriptions

PDFs

CSVs

Word documents

textual

You need to organize everything and convert each into a standard text format

General Documents	**Tabular Data**	**Audio Data**	**“Some Format”**	**Mixed**
Emails	Yelp Reviews	Voice mails	Contracts	PDFs
Word documents	CSVs	Dictations	Warranties	Document scans
Call center	Parquet files	Transcriptions	Insurance claims	
Internet	The voice of the customer		Real estate deeds/sales	
			Medical records	
USE: **Converters and Formatters**	**Set “Textual” Columns**	**Transcription (Dragon)**	**Inline Contextualization**	**OCR and Converters**

textual

Converters and Formatters

Set “Textual” Columns

Transcription (Dragon)

Inline Contextualization

OCR and Converters

Convert to a Common Textual Format

Now WHAT do we do with this data?

textual

Converters and Formatters

Set “Textual” Columns

Transcription (Dragon)

Inline Contextualization

OCR and Converters

Common Textual Format

Deidentify – Redact Personal Data

Apply <u>Context</u>!

textual

If you are going to address text you MUST have a handle on both **text** AND **context**.

It is not sufficient to merely address text.

Furthermore, most of the context that is needed lies OUTSIDE of the text.

You can analyze the text until you are blue in the face and never find the relevant context of the text

Text is relatively simple. **Context is 90% of the battle.**

So what is the purpose of all of this?

By Properly Applying Context

You can convert your Unstructured Textual Data into Structured Data!

This allows you to use your Textual Data for Structured Analysis!

What is Meant by “the Context” of Textual Data?

It has different meanings in different areas

Consider the word “Trust”

In **Friendship** – It is the ability to believe in the word and actions of another

In **Finance** – It is a legal vehicle used to pass and allocate assets to another

In **Networking** – It allows one computer to communicate and share with another

What is Meant by “the Context” of Textual Data?

It has different meanings in SIMILAR areas

Consider the word “Cervical” in the medical field

It could mean: pertaining to the **neck**

- cervical vertebra

It could mean: pertaining to the lowest segment of the **uterus**,

- cervical cancer
- cervical hemorrhage

What is Meant by “the Context” of Textual Data?

It has different meanings in Related areas

Consider the word “Dermatome” in the medical field

It means an area of the **skin** supplied by a specific nerve root

It is also a **surgical instrument** used to cut the skin

What is Meant by “Adding Context” to Textual Data?

It has different meanings in different areas

1. Extraction of key elements and phrases for categorization
2. Aggregation of terms into layered categories
3. Similar to Data Governance with Data Warehouse Data

- Requires subject matter experts
- Requires understanding of what dimensions you want for analysis
- Can be Highly Political between Departments
- It is controlled by BUSINESS, not IT or Data Analysts!

What is the Process of Adding Context to Textual Data?

It matters what analytics you want to perform on your text

1. Data Conversion (Maybe)
2. Data Redaction (Maybe)
3. Data **E**xtraction

- Identification of “Important” phrases or areas (Nexus)
- Running through an Engine to pair the Nexus with the text

4. Data **T**ransformation

- Classification of the matched Nexus phrases
- Adding Metadata
  - Dates, Sentiment, Sentence Information, Byte Location,
  - Batch #s, Business, Nexus, Customer, …

5. Data **L**oading
   - Data Warehouse, Data Mart, Parquet Files

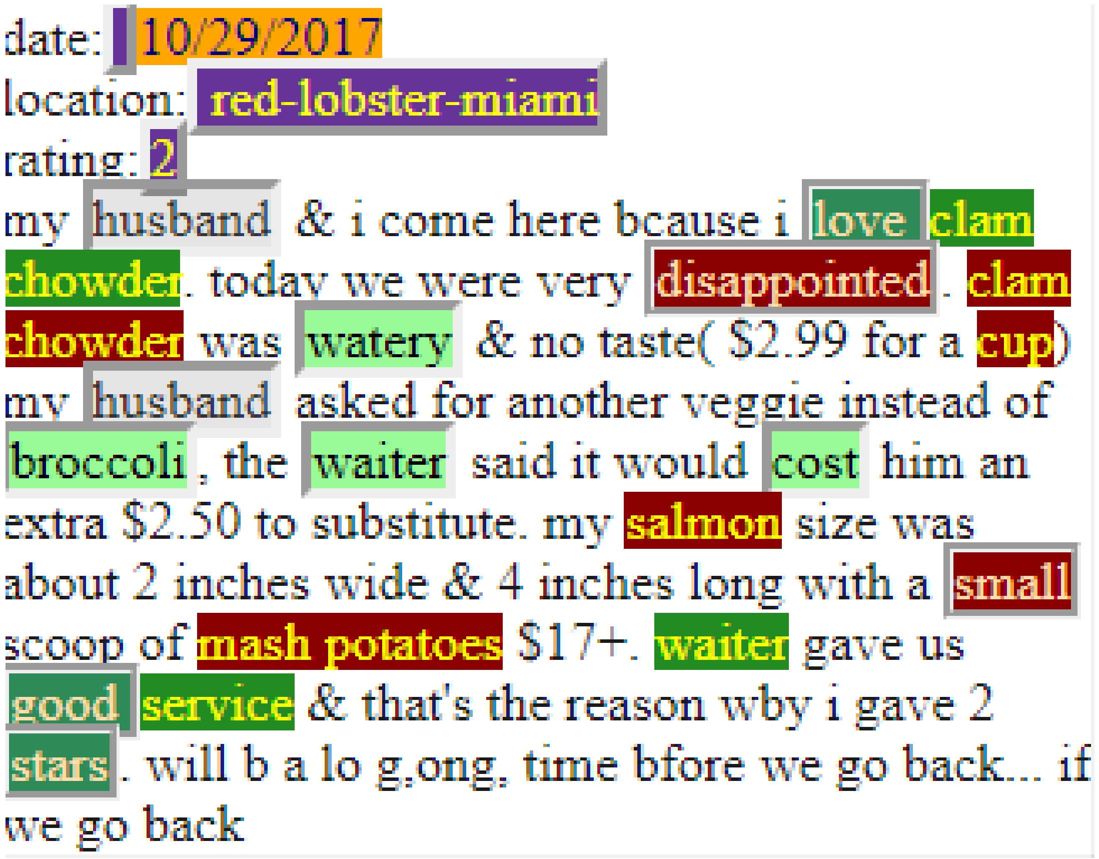
What Can Be Done with Contextualized Data?

We can do Structured Data Analysis

1. Document Markup
   - Visually identifies parts of the document
2. Sentiment Analysis
   - Gives feeling and degrees of feeling to parts of document
3. Inline Contextualization
   - Reverse Mail Merge – Pull out set of terms that have value
4. Document Classification
   - Give context to the areas of the document for correlation or basket analysis

What is Document Markup?

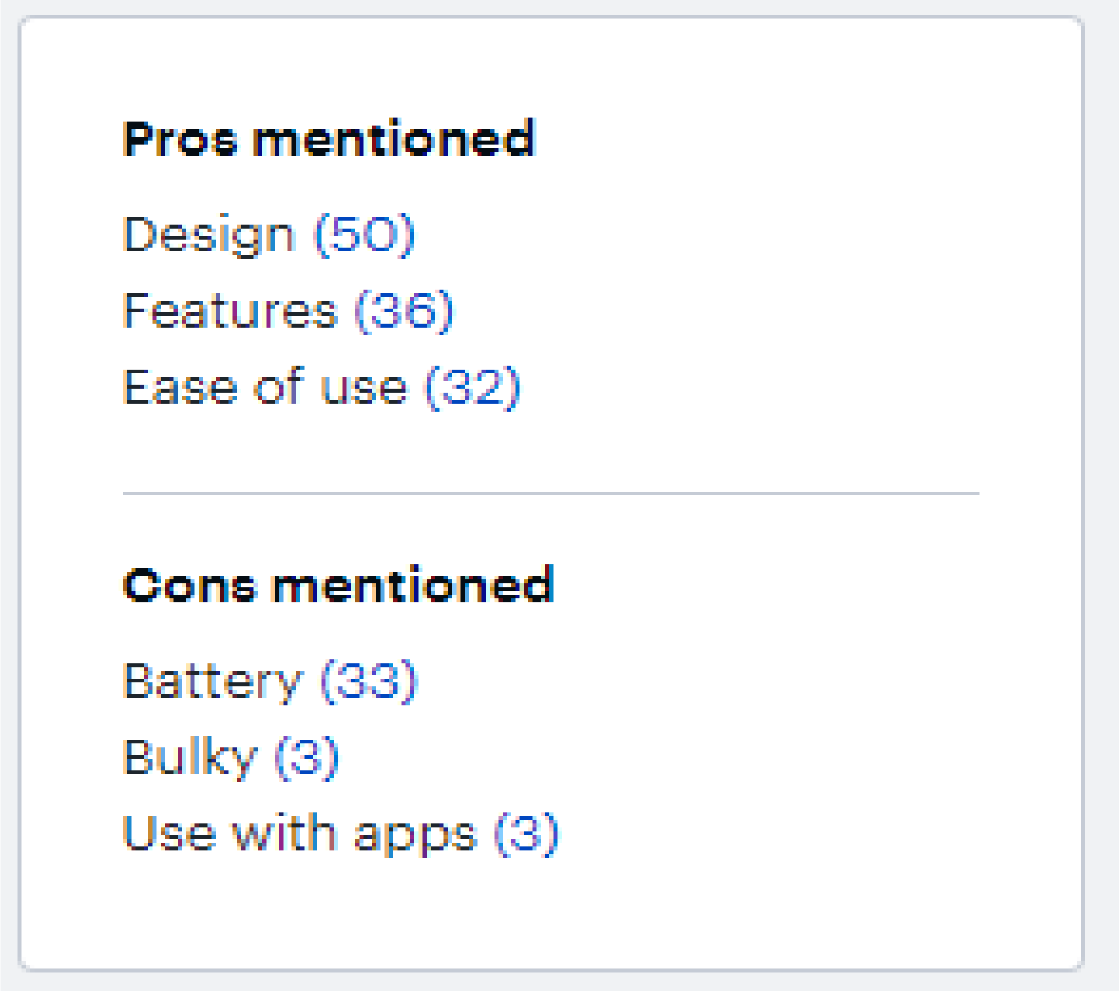
1. Data Visualization
   - Color coded
   - Draws the eyes
2. Used document by document
3. Great for “spot” review
4. Irrelevant and impractical for analyzing Big Data



date: 10/29/2017
location: red-lobster-miami
rating: 2
my husband & i come here bcause i love clam chowder. today we were very disappointed . clam chowder was watery & no taste($2.99 for a cup) my husband asked for another veggie instead of broccoli , the waiter said it would cost him an extra $2.50 to substitute. my salmon size was about 2 inches wide & 4 inches long with a small scoop of mash potatoes $17+. waiter gave us good service & that's the reason wby i gave 2 stars . will b a lo g,ong, time bfore we go back... if we go back

What is Sentiment Analysis?

1. Assigns Feeling to words
   - Color coded
   - Draws the eyes
2. Tries to identify and categorize opinions stated in some text
3. Great for Comments
4. A BASIC requirement for Voice of the Customer Analytics



Pros mentioned

Design (50)

Features (36)

Ease of use (32)

Cons mentioned

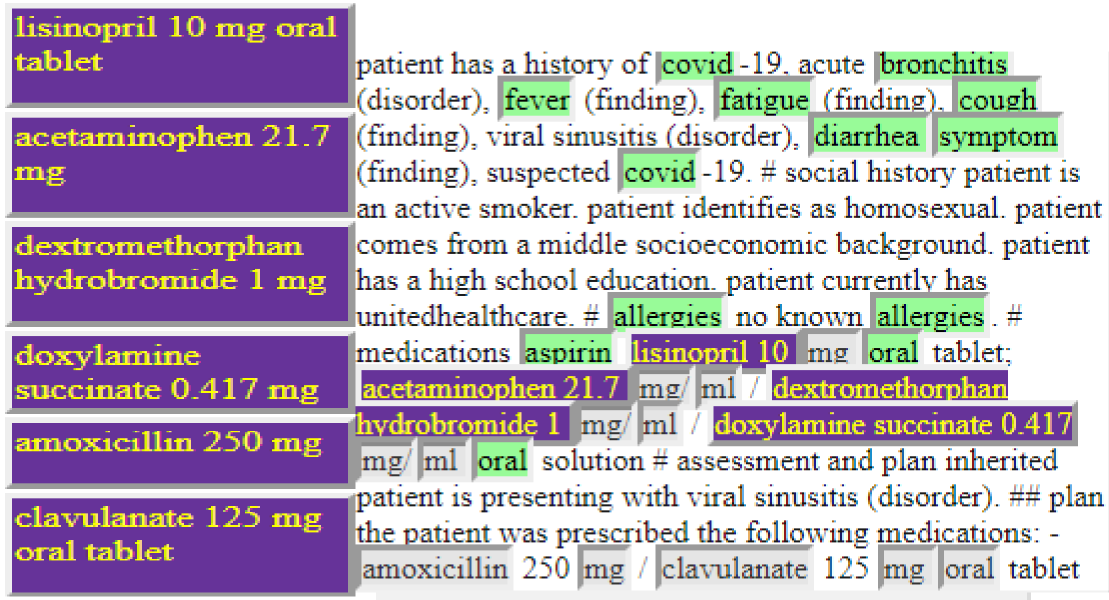
Battery (33)

Bulky (3)

Use with apps (3)

What is Inline Contextualization?

1. Reverse Mail Merge
2. Pull out set of terms that have value
   - Names
   - Contract Dates
   - Ratings
3. Useful for Contracts
4. Needed for Redaction
5. Needed for Document Separation
   - Medical Visits
   - Combined repeat visits
6. Needed for retrieval of grouped data from blocks of text



lisinopril 10 mg oral tablet

acetaminophen 21.7 mg

dextromethorphan hydrobromide 1 mg

doxylamine succinate 0.417 mg

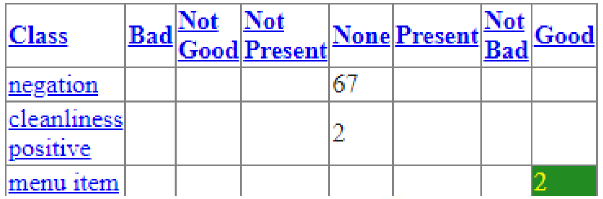
amoxicillin 250 mg

clavulanate 125 mg oral tablet

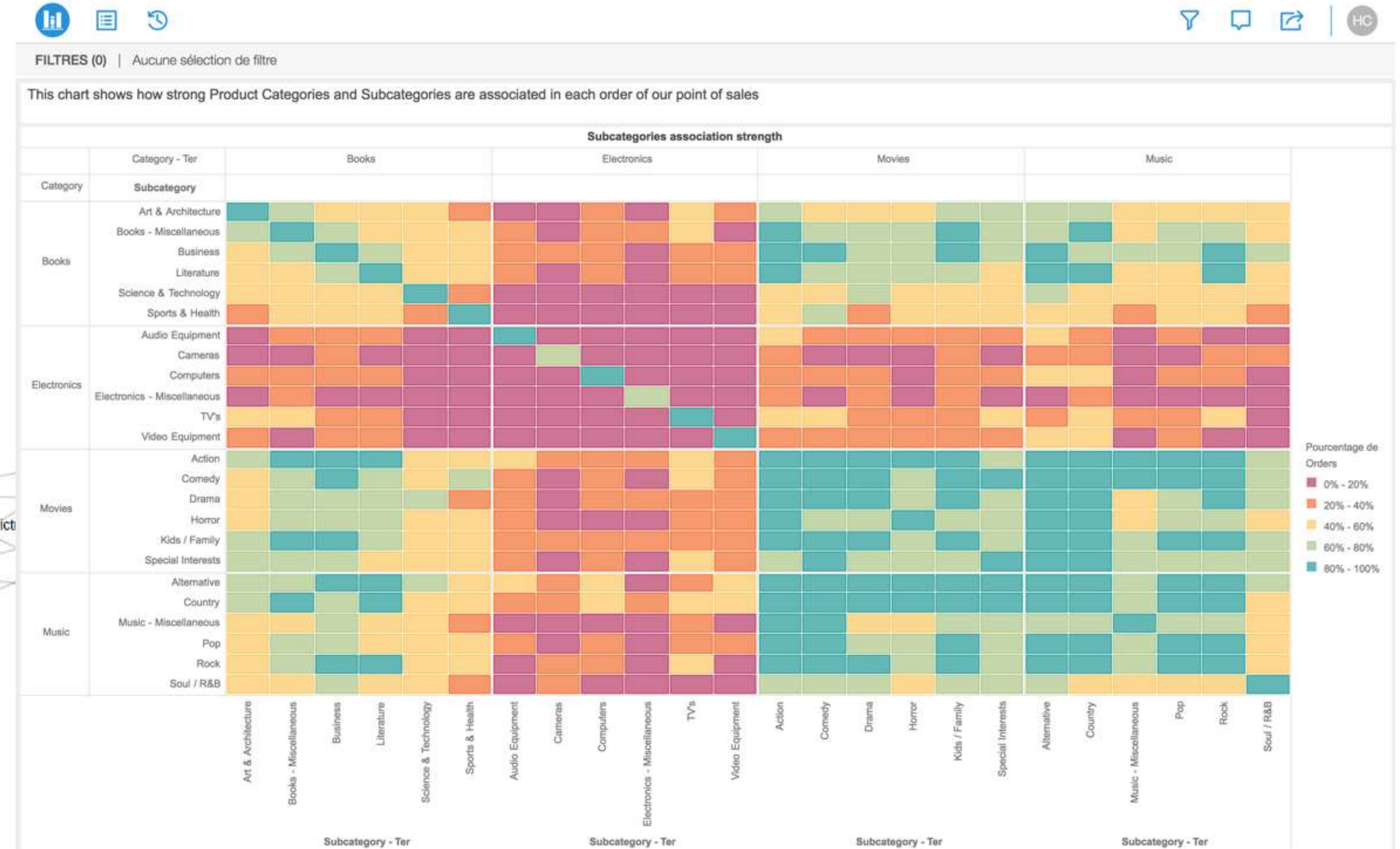
patient has a history of covid -19. acute bronchitis (disorder), fever (finding), fatigue (finding), cough (finding), viral sinusitis (disorder), diarrhea symptom (finding), suspected covid -19. # social history patient is an active smoker. patient identifies as homosexual. patient comes from a middle socioeconomic background. patient has a high school education. patient currently has unitedhealthcare. # allergies no known allergies . # medications aspirin lisinopril 10 mg oral tablet; acetaminophen 21.7 mg/ ml / dextromethorphan hydrobromide 1 mg/ ml / doxylamine succinate 0.417 mg/ ml oral solution # assessment and plan inherited patient is presenting with viral sinusitis (disorder). ## plan the patient was prescribed the following medications: - amoxicillin 250 mg / clavulanate 125 mg oral tablet

What is Document Classification?

1. Give context to the areas of the document
2. Correlation Analysis
3. Basket Analysis
4. Mind Maps
5. Knowledge Graph

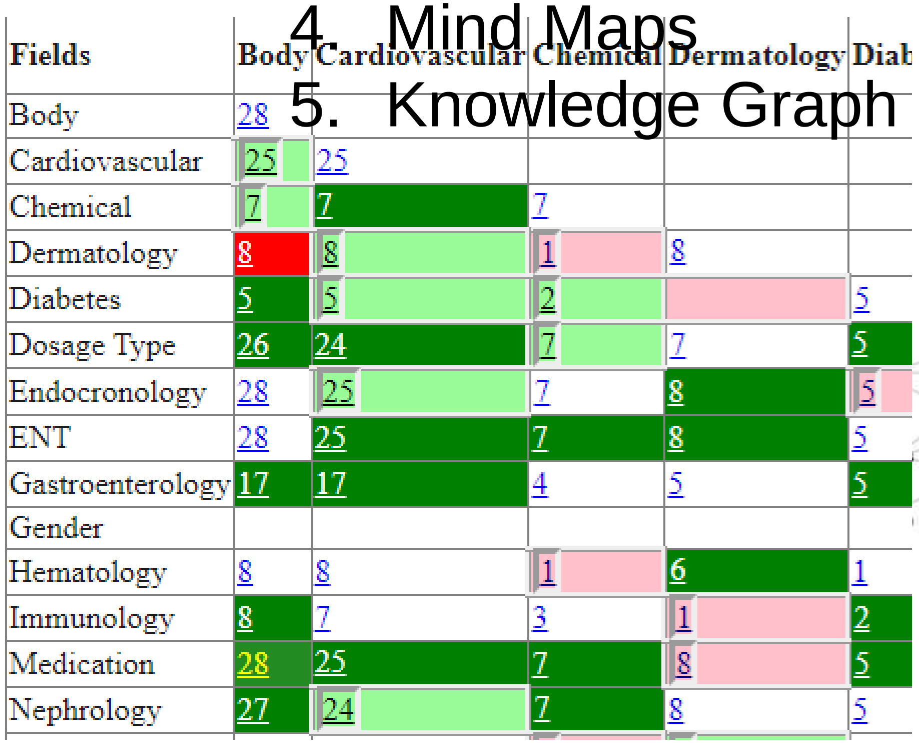


Class	Bad	Not Good	Not Present	None	Present	Not Bad	Good
negation				67			
cleanliness positive				2			
menu item							2



FILTRES (0) | Aucune sélection de filtre

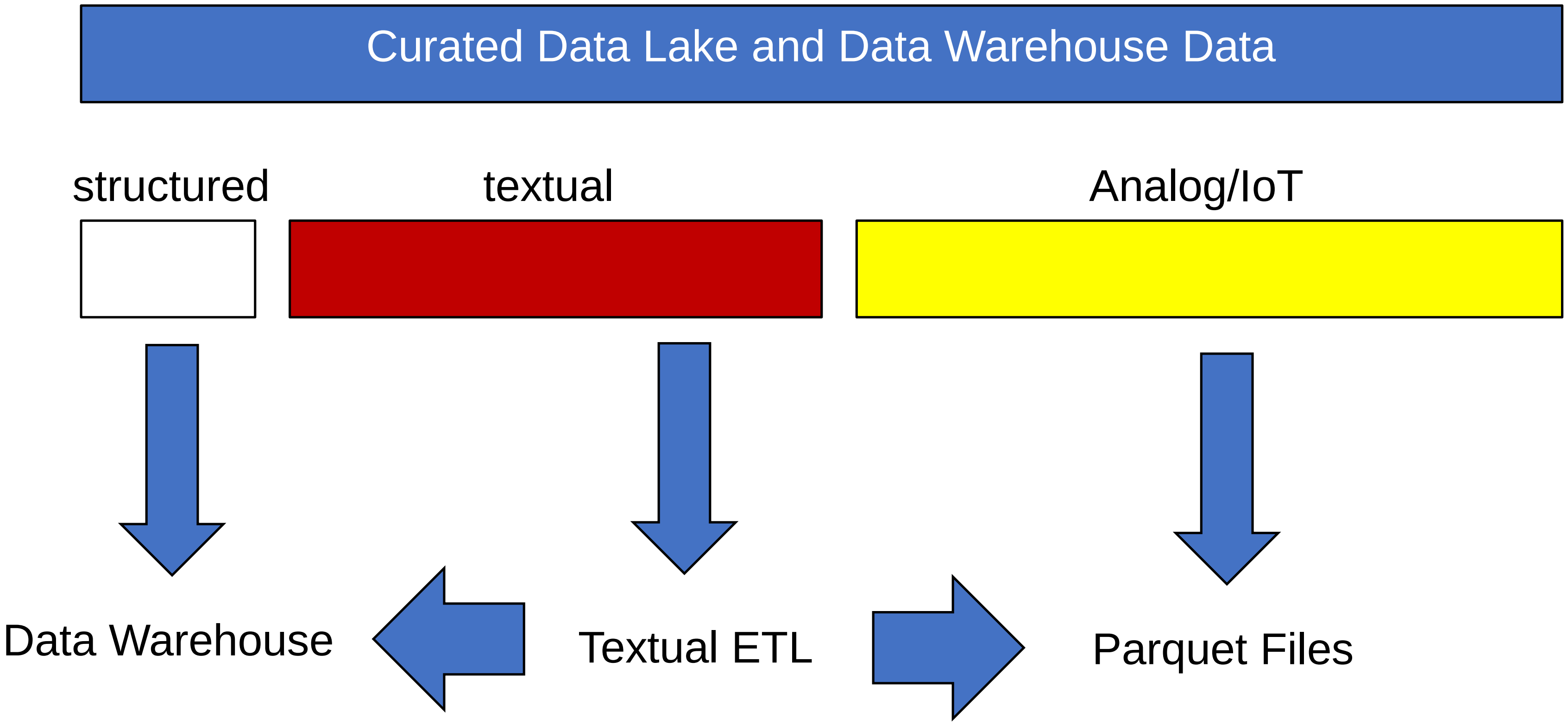
This chart shows how strong Product Categories and Subcategories are associated in each order of our point of sales



Fields	Body	Cardiovascular	Chemical	Dermatology	Diab
Body	28				
Cardiovascular	25	25			
Chemical	7	7	7		
Dermatology	8	8	1	8	
Diabetes	5	5	2		5
Dosage Type	26	24	7	7	5
Endocronology	28	25	7	8	5
ENT	28	25	7	8	5
Gastroenterology	17	17	4	5	5
Gender					
Hematology	8	8	1	6	1
Immunology	8	7	3	1	2
Medication	28	25	7	8	5
Nephrology	27	24	7	8	5

Review

There are many types of data in a Data Lakehouse


Curated Data Lake and Data Warehouse Data

structured

textual

Analog/IoT

Data Warehouse

Textual ETL

Parquet Files

Review

Sort your textual data documents by types

Convert your textual data to a common format

Deidentify data if you are going to store it

Apply Context to your textual data!

Using context, you can convert your **Unstructured Data** into **Structured Data**!

Review

This conversion allows for Structured Data Analysis

1. Document Markup
2. Sentiment Analysis
3. Inline Contextualization
4. Document Classification
5. Plus many others…

Questions

Gracias.

FORESTRIM TECHNOLOGY

https://www.forestrimtech.com/

info@forestrimtech.com

References and Sources

- Bill Inmon – Slides and Conversations
- Inmon, B. (2021). *Building The Data Lakehouse.* Technics Publications LLC.
- https://www.snowflake.com/guides/what-iot
- https://medicalterminologyblog.com/homonyms-medical-language-2/
- Andrea and Amanda Rapien – Format and Additional Clarifying Material