Machine Learning & Data Fusion in three implementations

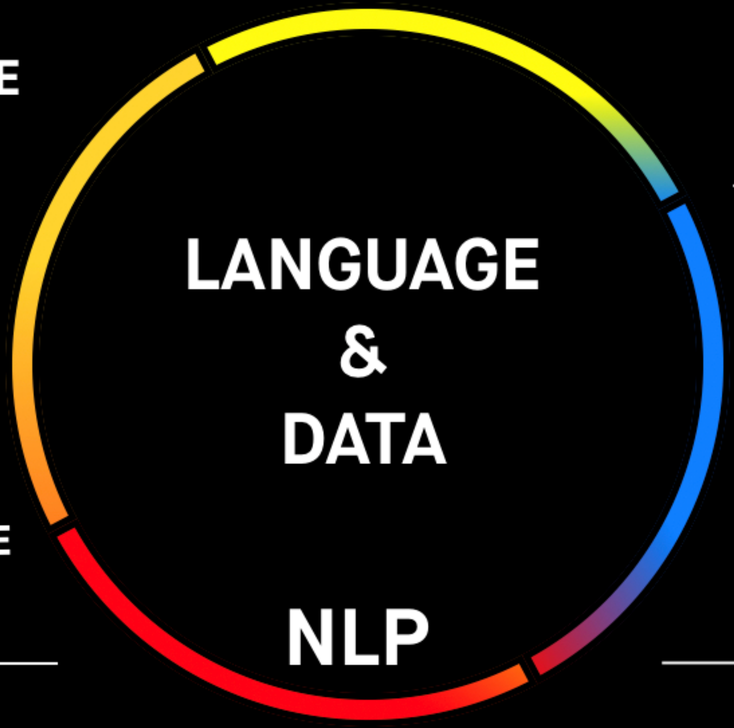
Francesco Tarasconi • Lead Data Scientist • francesco.tarasconi@celi.it

CELI LANGUAGE TECHNOLOGY

POLITECNICO DI TORINO 1859 1906

CELI – LANGUAGE TECHNOLOGY

45+ languages



DATA SCIENCE

Data Science BY CELI

- Machine learning
- Predictive/forecasting models
- Recommender systems
- Network analysis
- Deep learning
- Advanced mathematics
- Data fusion
- Market segmentation & Personas

ANALYTICS

SOPHIA analytics
SPEECH AND TEXT

- Text analytics
- Speech analytics
- Text mining
- Text classification
- Information extraction
- Sentiment analysis
- Opinion mining

SPEECH & VOICE

Voice First BY CELI

- Voice interaction design
- Voice assistants development
- Voice improvement (TTS & ASR)
- Voice evaluation (TTS & ASR)
- Natural Language Understanding
- Phonetization library for TTS

SEARCH

SOPHIA search
SEMANTIC SEARCH ENGINE

SOPHIA chatbot
CONVERSATIONAL INTERFACES

- Semantic search
- Conversational Systems
  - chatbot and voicebots
- Visual Search

NLP

- NLP software components
- Language resources
  (text and speech corpora, phonetic dicts, morphologica db, NSW collections, named entities,polarity lexicons)
- Data collection, cleaning and annotation
- Native speaker linguists for 45+ languages

CELI LANGUAGE TECHNOLOGY

Copyright © CELI S.r.l.

>500
PERSONE

>100
CLIENTI

>50M
FATTURATO

H-FARM is a platform where innovation, entrepreneurship and education coexist and influence each other.

Founded in 2005. First mover in the world to adopt a model that combines innovation investments, business advice and training programs dedicated to digital in one place.

Strategy, Brand, Content Transformation, Data Analysis, Artificial Intelligence, Training and Education are the components of the INNOVATION ecosystem.

CELI enables enterprises to unlock the full value within their unstructured and structured data via:

TECHNOLOGY

Technology for acquiring, enriching and presenting structured and unstructured data in search and analytics applications

CONSULTANCY

Strategy to understand structured and unstructured contents to drive actionable insights

PROJECTS

We combine innovative technologies such as machine learning and natural language processing with search and big data analytics to transform the way people work.

Data Science for Industries

We employ smart algorithms, Machine and Deep Learning to understand the relations across a vast number of KPIs, in order to develop complex forecasters (e.g. to predict sales across a vast product catalogue) and decision support systems (e.g. to identify the most dangerous situations and prescribe the best course of action)

The potential impact of AI

The impact of Deep Learning is greater on top line functions and bottom line ones, such as supply chain management and manufacturing

---

Last year’s production and sales data are not enough to understand and forecast a world in continuous evolution

---

Product sales are influenced by the WHOs, WHEREs and WHATs of each transaction, plus external factors: weather, events, ...

Forecasting based on underlying causal drivers of demand can improve forecasting accuracy by 10 to 20 percent

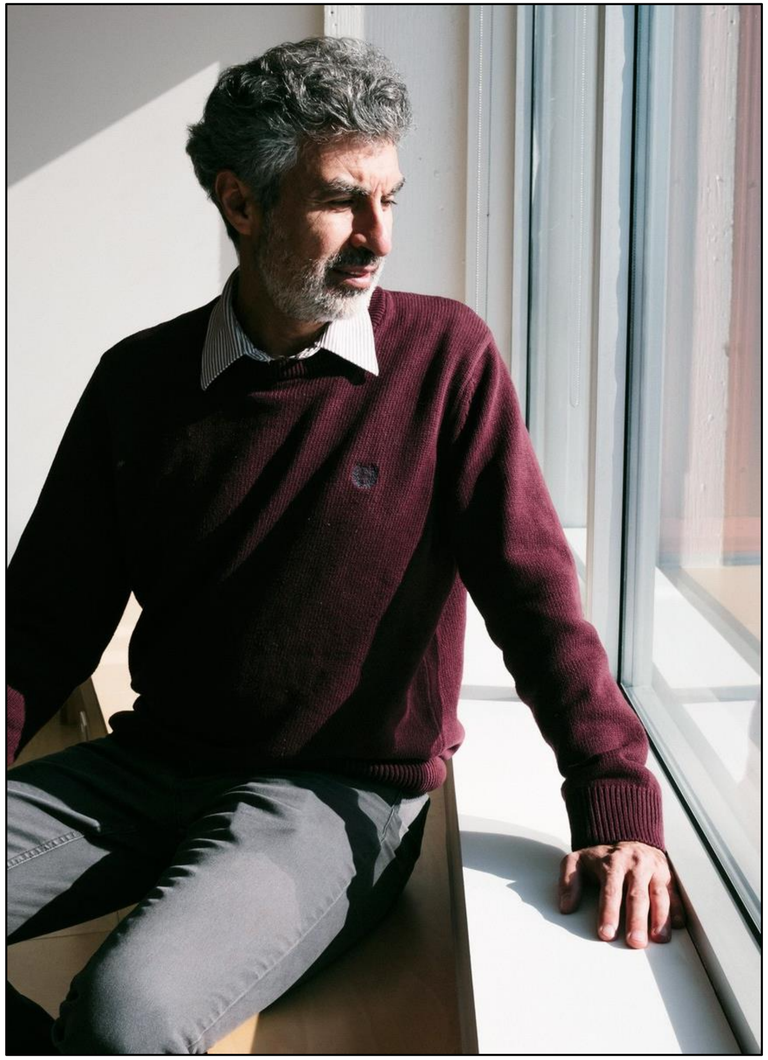
McKinsey & Company - Notes from the AI frontier: Applications and value of deep learning, April 2018

LIMITATIONS OF DEEP LEARNING

*...deep learning has demonstrated a particularly **impressive ability to recognize patterns** in data; it's been put to many practical uses, from spotting signs of cancer in medical scans to uncovering fraud in financial data.*

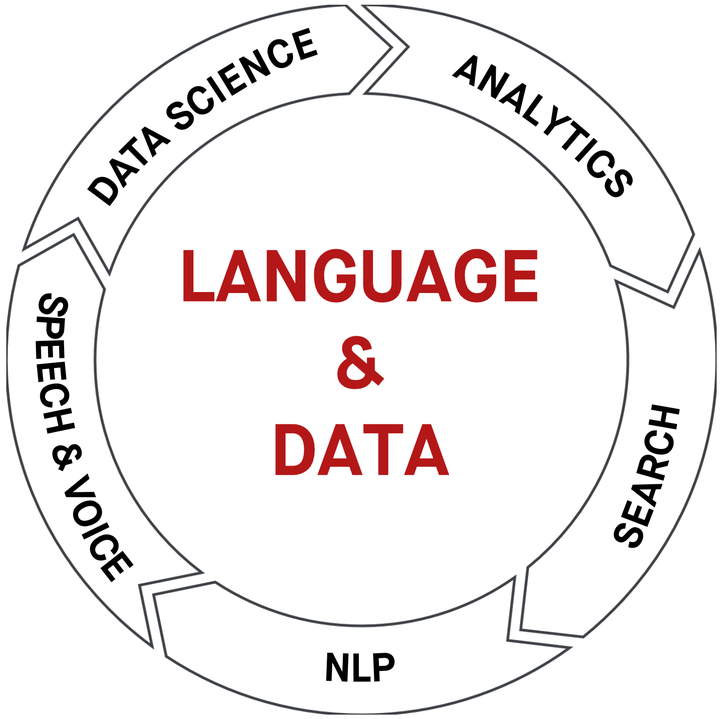
*But deep learning is fundamentally **blind to cause and effect**. Unlike a real doctor, a deep learning algorithm cannot explain why a particular image may suggest disease. This means deep learning must be used cautiously in critical situations.*

https://www.wired.com/story/ai-pioneer-algorithms-understand-why/amp



CELI
LANGUAGE TECHNOLOGY

OUR APPROACH



1. Identify the correct **role** of Human and the correct role of AI

2. Identify the data and the models that can help you pinpoint and measure the **causal drivers**

3. In many situations, must allow the human to review and **validate** the results of Machine Learning procedures

CELI
LANGUAGE TECHNOLOGY

THE ROLE OF HUMAN AND AI

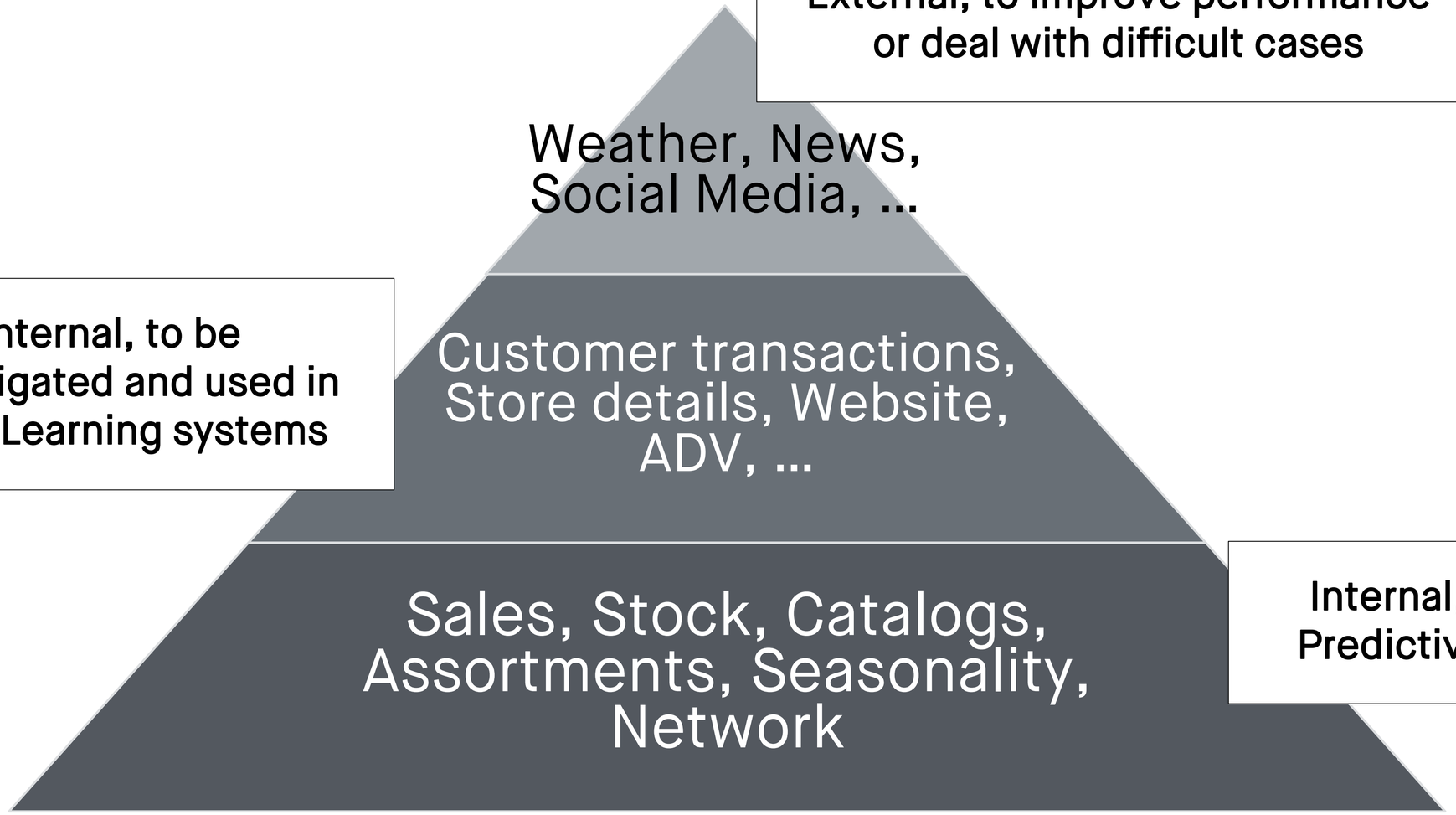
- The type of problem to be solved and the **environment** the problem will be solved in determine the responsibility of Human and AI

- This choice must consider factors of **risk**, but also current business practices and the degree of success of traditional methods in the task at hand

CELI
LANGUAGE TECHNOLOGY

DATA TYPES FOR AI PREDICTIVE SYSTEMS

External, to improve performance or deal with difficult cases

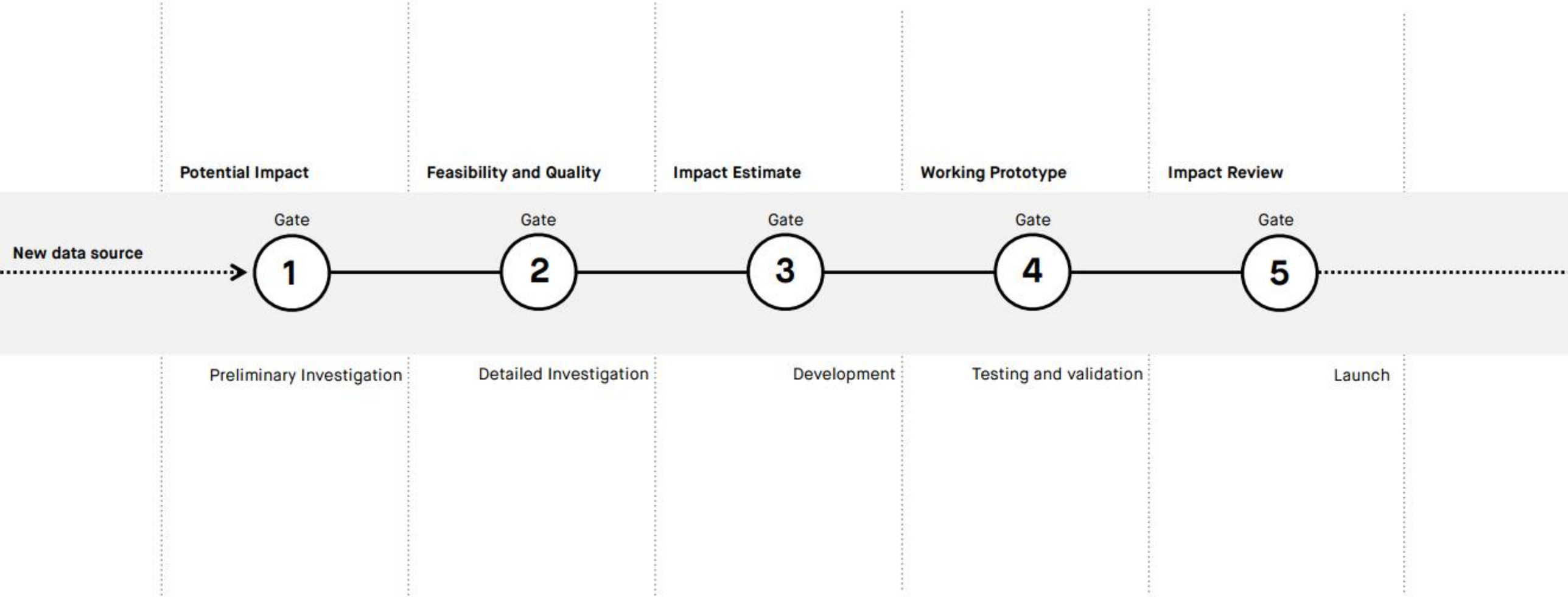


Internal, to be investigated and used in Deep Learning systems

Internal, used in most Predictive applications

CELI
LANGUAGE TECHNOLOGY

STAGE GATE PROCESS OF DATA FUSION



CELI
LANGUAGE TECHNOLOGY

THREE IMPLEMENTATIONS

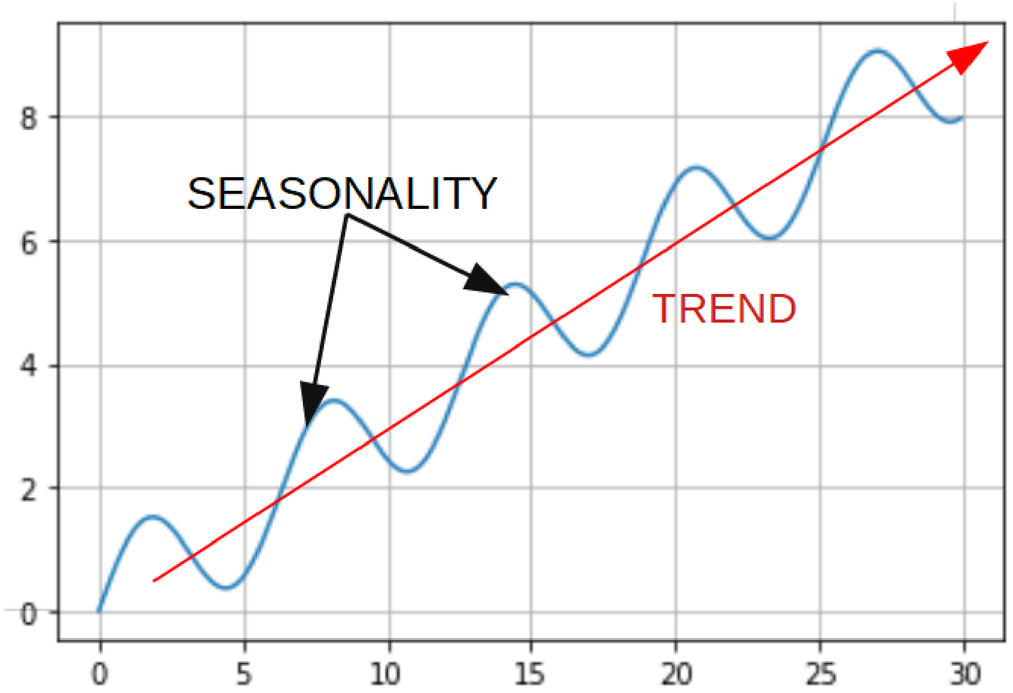
1. ML-BASED SHORT-TERM FORECAST
2. ML-ASSISTED FORECAST & ASSORTMENT
3. PREDICTIVE AI FOR HEALTH, SAFETY AND ENVIRONMENT

THREE IMPLEMENTATIONS

1. ML-BASED SHORT-TERM FORECAST
2. ML-ASSISTED FORECAST & ASSORTMENT
3. PREDICTIVE AI FOR HEALTH, SAFETY AND ENVIRONMENT

INTRODUCTION: COMMON APPROACHES TO A FORECASTING PROBLEM

- Several algorithms involve a form of trend/average/autoregression, seasonality/cycles and the possibility to include additional "exogenous" variables
- Most of these variables are proxy of actual causal drivers e.g. what we usually call "trend" represent a combination of several factors
- Models that associate a confidence to a forecast will involve some kind of hypothesis on the distribution: if the hypothesis does not hold, confidence will be meaningless
- In practice it is often effective to start with "simple" methods
- Testing an approach on real data is priceless



CELI
LANGUAGE TECHNOLOGY

CHALLENGE: MANAGING A WORLD WIDE STORE NETWORK

CELI
LANGUAGE TECHNOLOGY

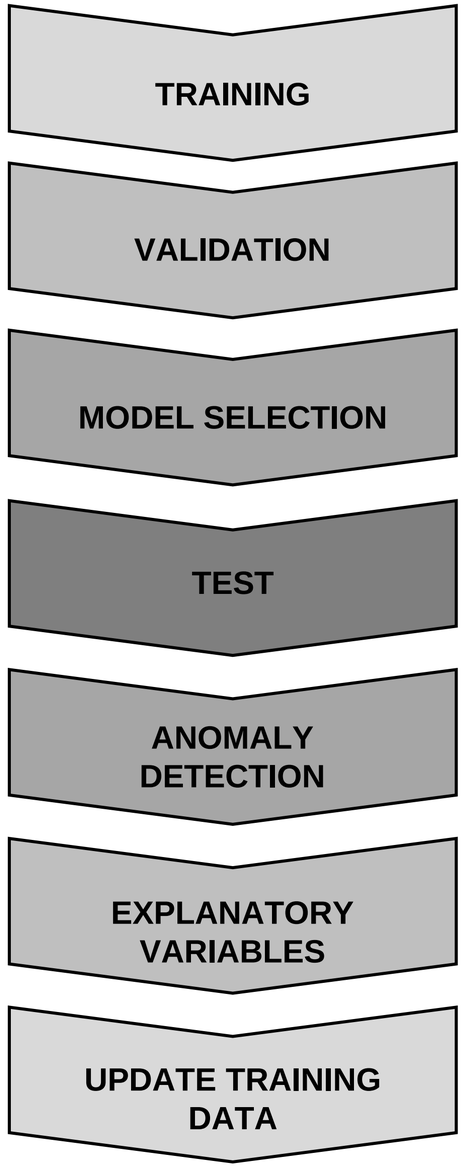
ML-BASED FORECAST FOR SHORT-TERM PLANNING & MONITORING

- Analyze historical sales data to detect key **sales patterns and cycles** in each store
- Integrate sales data with additional data that can have a **disruptive effect in the short term**, such as local weather, social media, news and events
- **Predict daily sales** in each store/region, highlighting deviations from original budget

CELI
LANGUAGE TECHNOLOGY

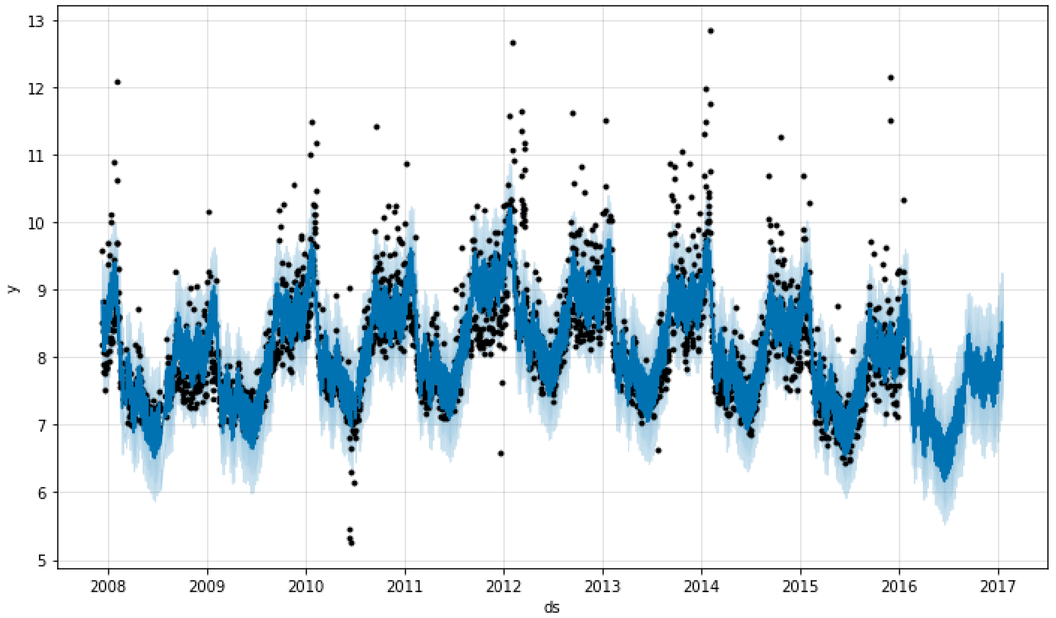
APPROACH

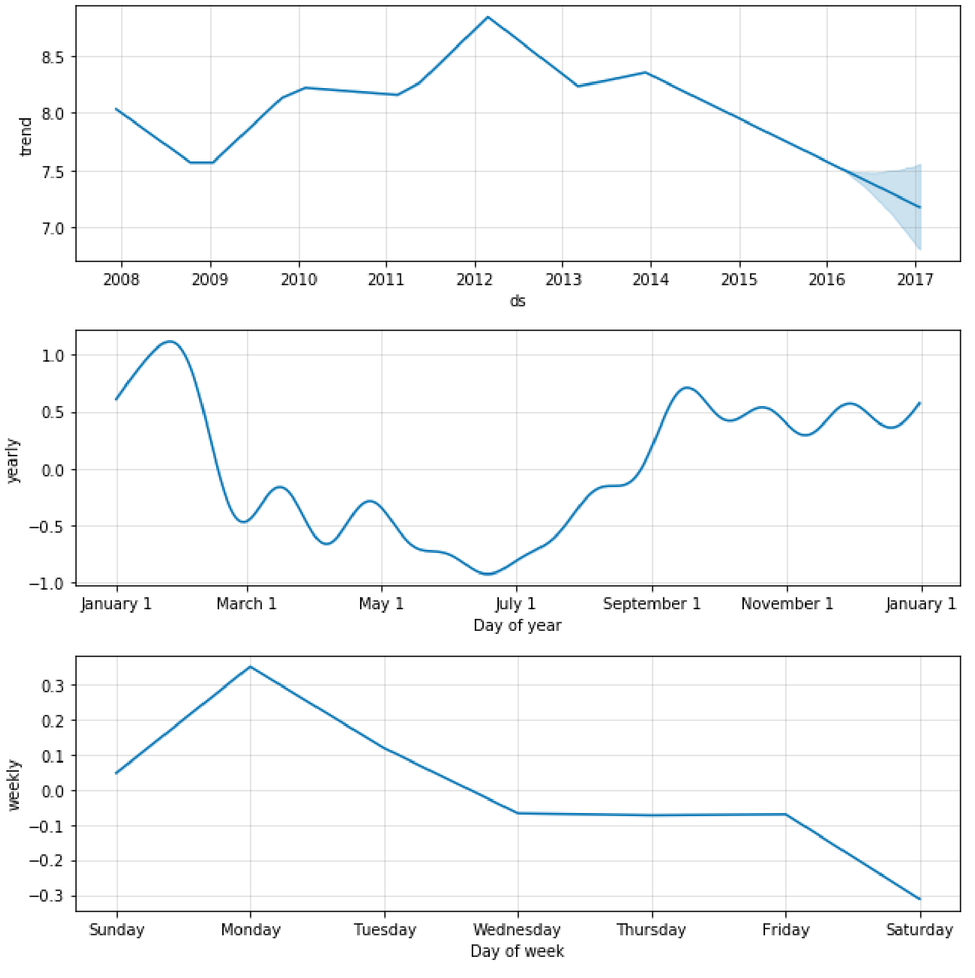
- We had **freedom** to experiment with several models, including non-linear ones, delivering **high frequency predictions with maximum accuracy**
- A scalable pipeline of **model selection** was built in order to find the best parameters and identify which locations (stores) were regularly impacted by **common weather phenomena** (e.g. rain, temperature, humidity)
- A subsequent **monitoring** component tested if anomalies in sales were correlated to **extreme** weather phenomena, news or events, in order to trigger future **alerts**



CELI
LANGUAGE TECHNOLOGY

PROPHET LIBRARY AND MODEL CLASS





Trend Changepoints

- Automatic changepoint detection in Prophet
- Adjusting trend flexibility
- Specifying the locations of the changepoints

Seasonality, Holiday Effects, And Regressors

- Modeling Holidays and Special Events
- Built-in Country Holidays
- Fourier Order for Seasonalities
- Specifying Custom Seasonalities
- Seasonalities that depend on other factors
- Prior scale for holidays and seasonality
- Additional regressors

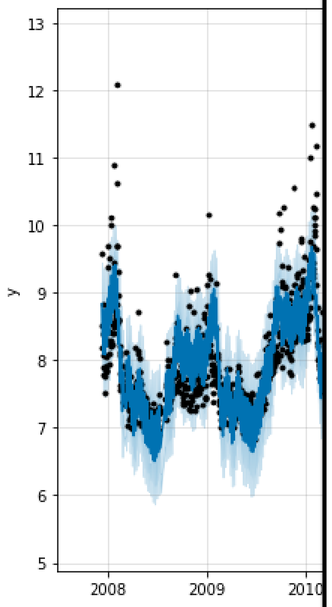
Multiplicative Seasonality

Uncertainty Intervals

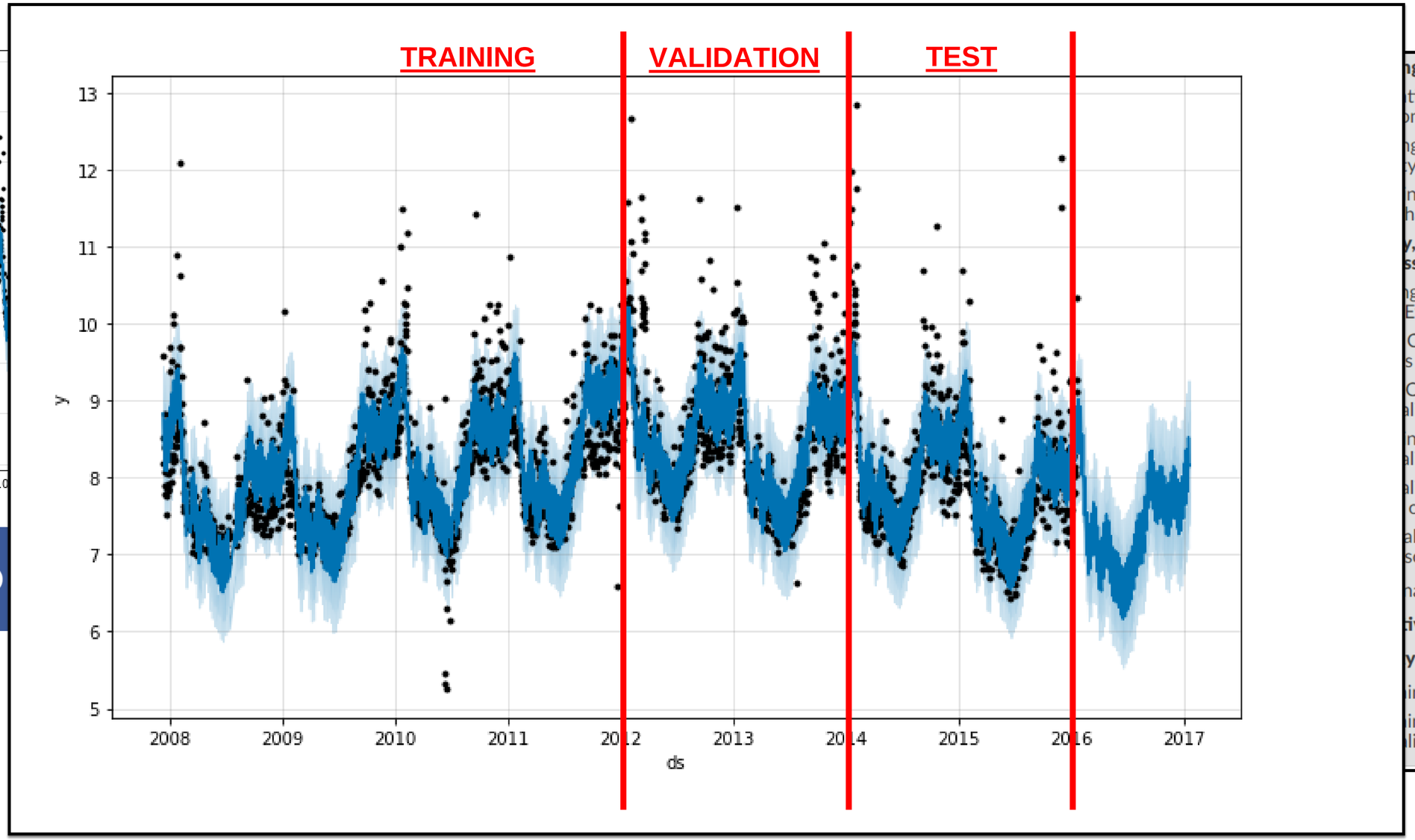
- Uncertainty in the trend
- Uncertainty in seasonality

https://facebook.github.io/prophet/docs/quick_start.html

PROPHET LIBRARY AND MODEL CLASS



PROP



ngepoints

tic changepoint
on in Prophet

g trend
y

ng the locations
hangepoints

y, Holiday Effects,
ssors

g Holidays and
Events

Country
s

Order for
alities

ng Custom
alities

alities that
on other factors

ale for holidays
sonality

nal regressors

tive Seasonality

y Intervals

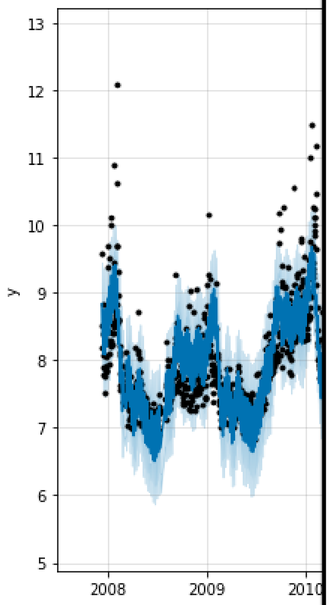
inty in the trend

inty in
lity

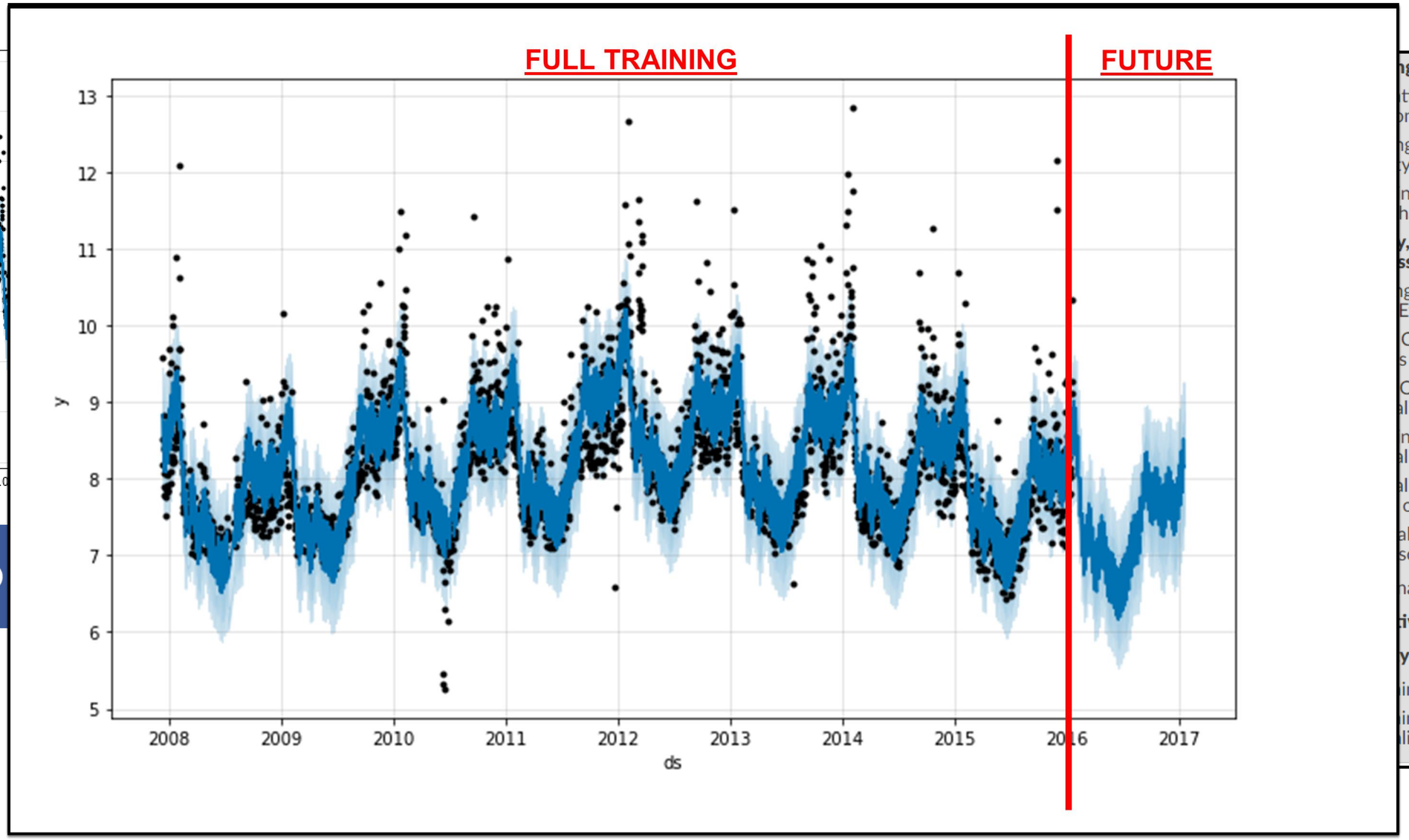
CELI
LANGUAGE TECHNOLOGY

https://facebook.github.io/prophet/docs/quick_start.html

PROPHET LIBRARY AND MODEL CLASS



PROP



ngepoints

tic changepoint
on in Prophet

ng trend
y

ng the locations
hangepoints

y, Holiday Effects,
ssors

ng Holidays and
Events

Country
s

Order for
alities

ng Custom
alities

alities that
on other factors

ale for holidays
sonality

nal regressors

tive Seasonality

y Intervals

inty in the trend

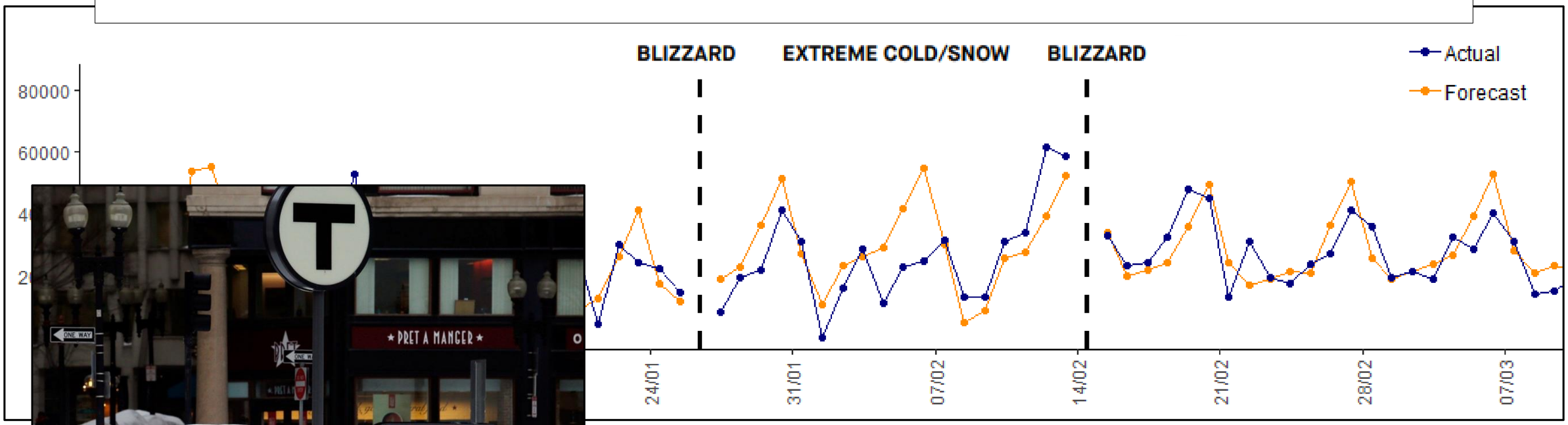
inty in
lity

https://facebook.github.io/prophet/docs/quick_start.html

CELI
LANGUAGE TECHNOLOGY

ADDED VALUE OF LOCATION INTELLIGENCE

Weather forecast can improve sales forecast accuracy by about 5%, and the impact is greater for tourist hotspots and places subject to extreme weather conditions



Boston (Winter 2015)

CELI
LANGUAGE TECHNOLOGY

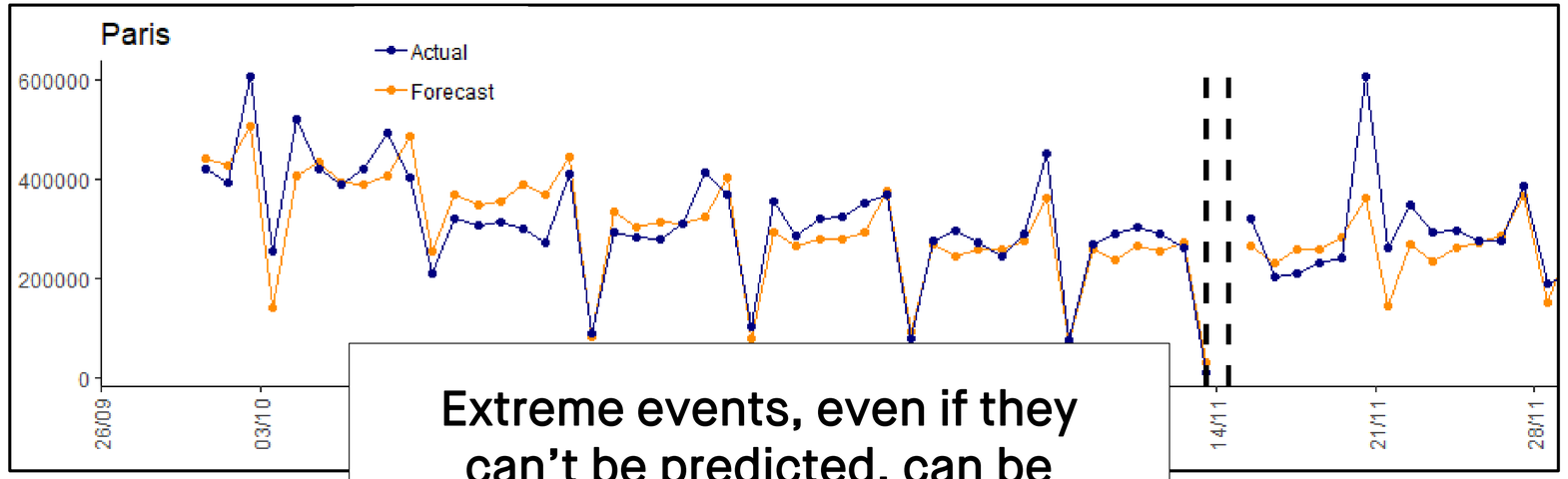
GDELT PROJECT

- GDELT provides **worldwide news data**, where articles on online media are aggregated into «Events» **enriched with geolocalization, sentiment analysis and geopolitical indicators** (e.g. Goldstein scale)

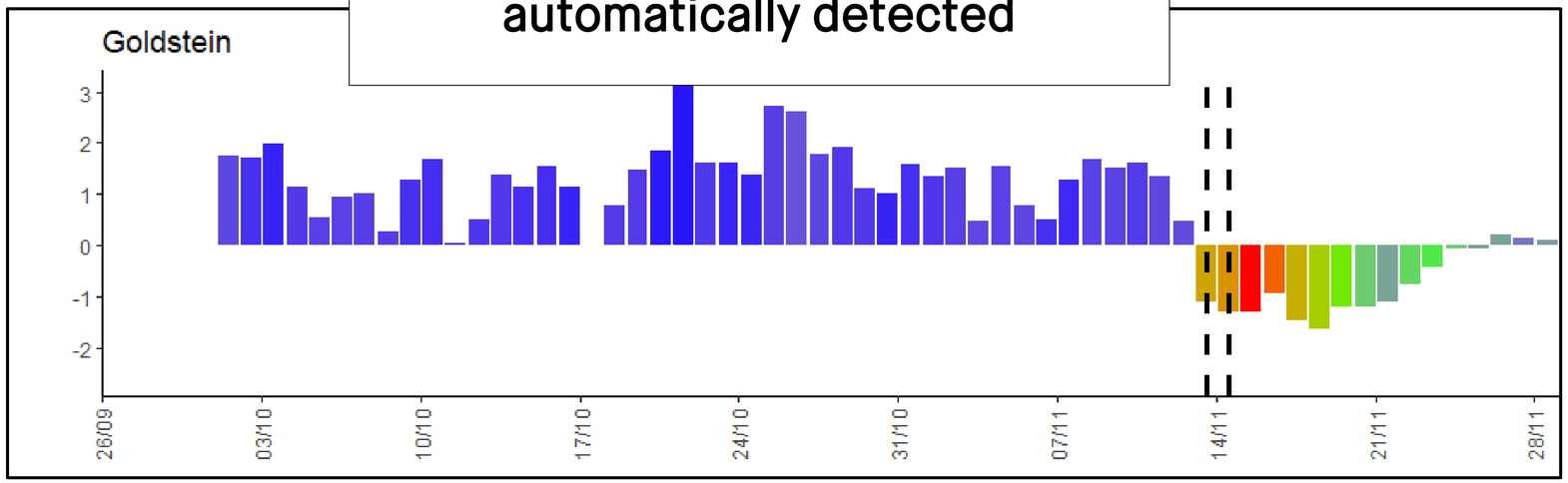
- Media coverage, Goldstein and sentiment data were compared with sales data and forecasts produced up to one year in advance

CELI
LANGUAGE TECHNOLOGY

ADDED VALUE OF NEWS & EVENTS MONITORING



Extreme events, even if they can't be predicted, can be automatically detected



Bataclan Episode (November 2015)

CELI
LANGUAGE TECHNOLOGY

THREE IMPLEMENTATIONS

1. ML-BASED SHORT-TERM FORECAST
2. ML-ASSISTED FORECAST & ASSORTMENT
3. PREDICTIVE AI FOR HEALTH, SAFETY AND ENVIRONMENT

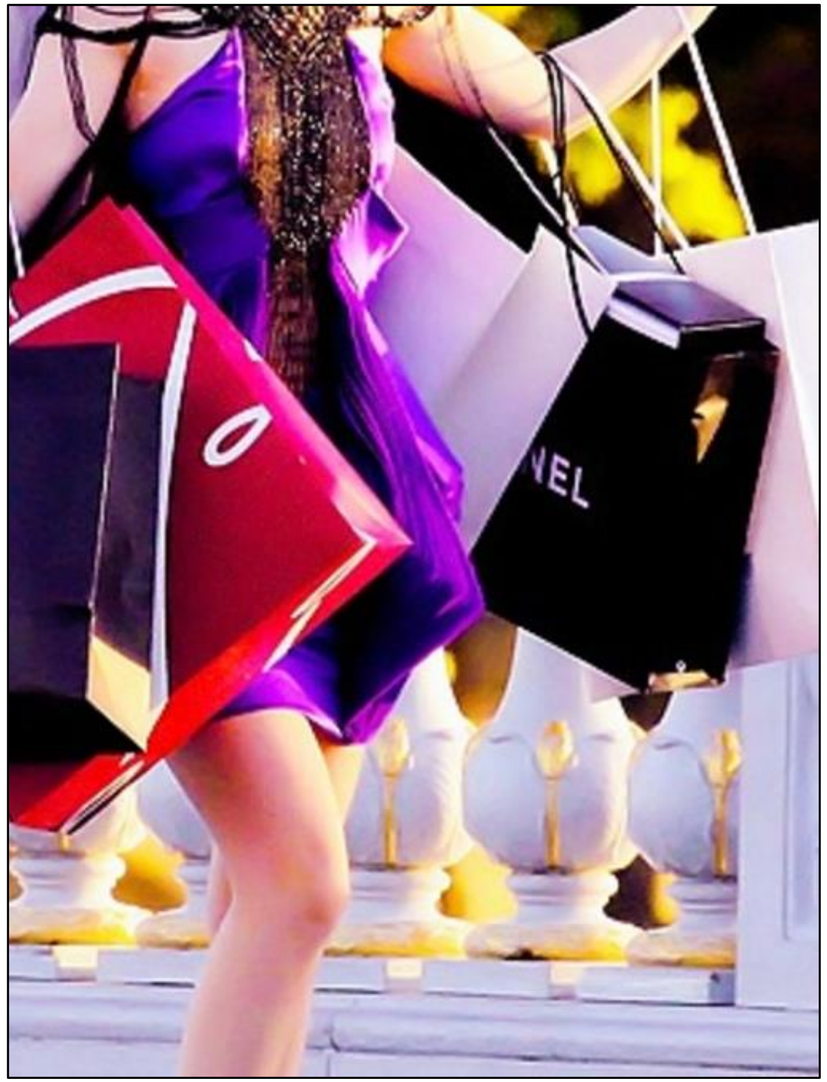
CHALLENGE: FORECASTING IN THE FASHION SECTOR AT PRODUCT LEVEL

- Data patterns are highly volatile and it is very difficult to analytically learn about the underlying pattern
- Life cycle of products can be quite short with **new collections** coming out every season
- Different horizons of forecast: long term (2 years) vs short term (few weeks).
- Quick Response: react to observed demand changes quickly.
- Seasonalities

CELI
LANGUAGE TECHNOLOGY

ML-ASSISTED FORECAST FOR LUXURY SUPPLY CHAIN

- Identify clusters of products and clusters of stores which exhibit similar sales behavior
- Find correlations and significant patterns which involve different types of data: information about store network, customers, product features and sales
- Forecast the demand for specific products/SKUs at store/region level or world-wide
- Provide first estimates for products which have never been sold



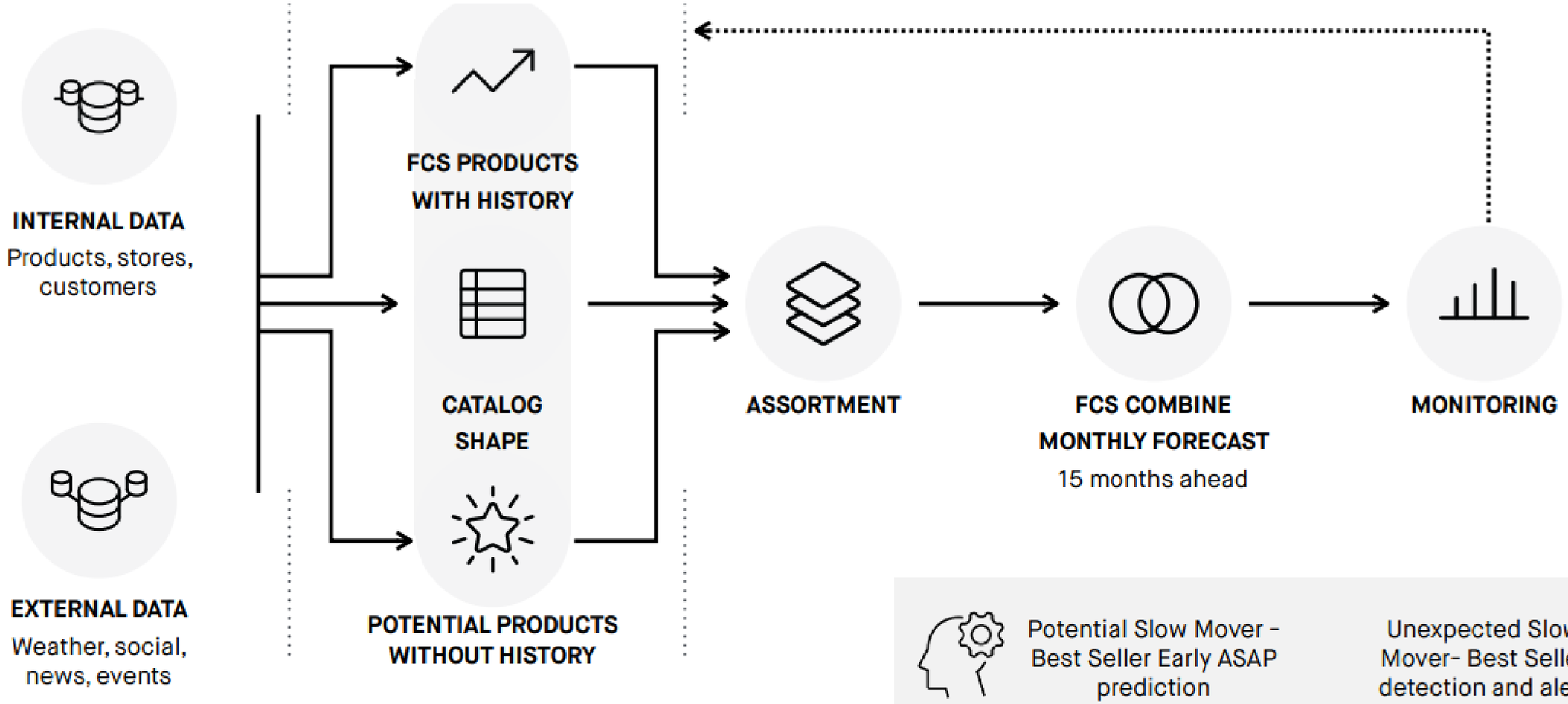
CELI
LANGUAGE TECHNOLOGY

ML-ASSISTED ASSORTMENT FOR MERCHANDISE

- Propose the most promising shop assortments, based upon past examples, future demand forecasts and preset constraints
- Expand or reduce the starting assortments coherently with budget requirements
- Suggest network expansions for best-selling products and phase-out for slow-movers

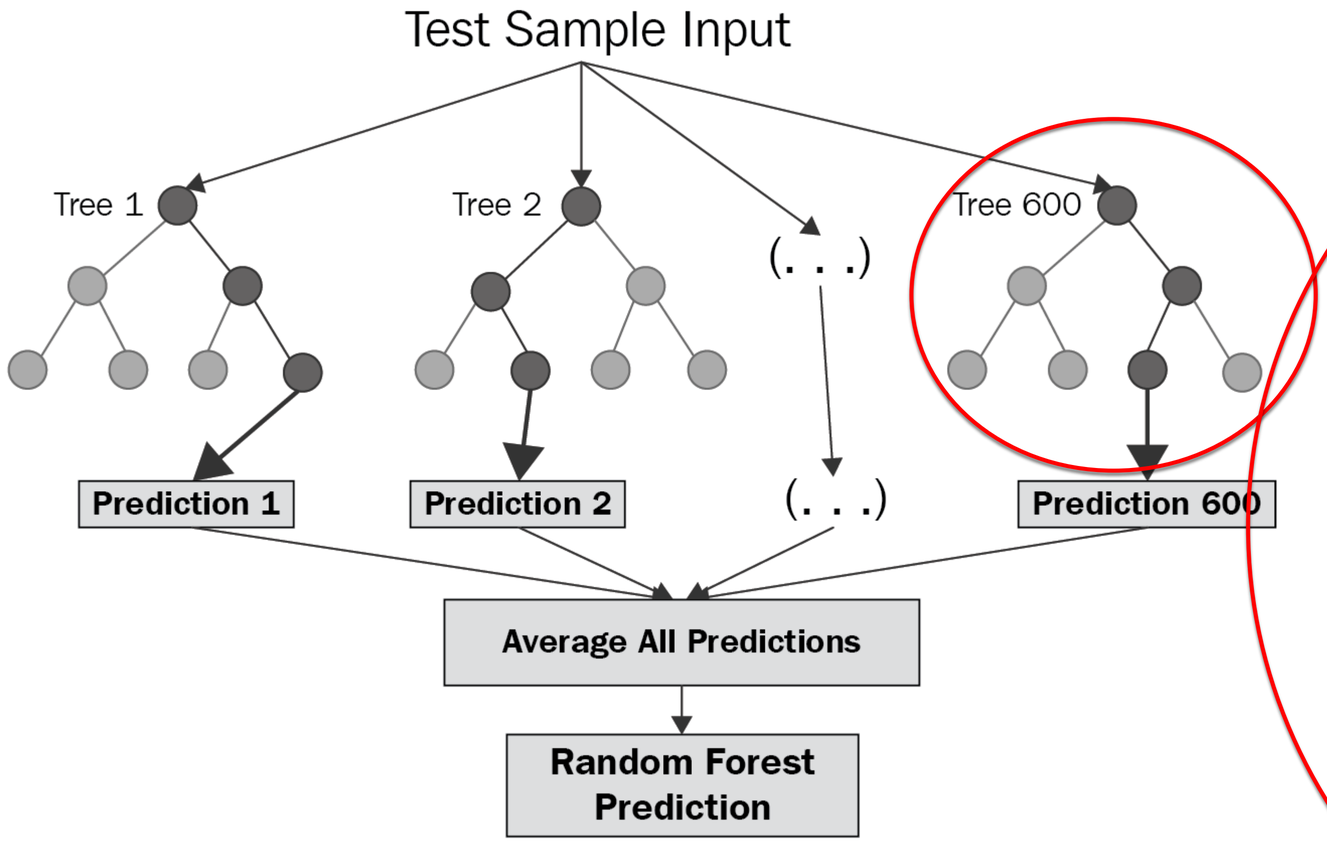
CELI
LANGUAGE TECHNOLOGY

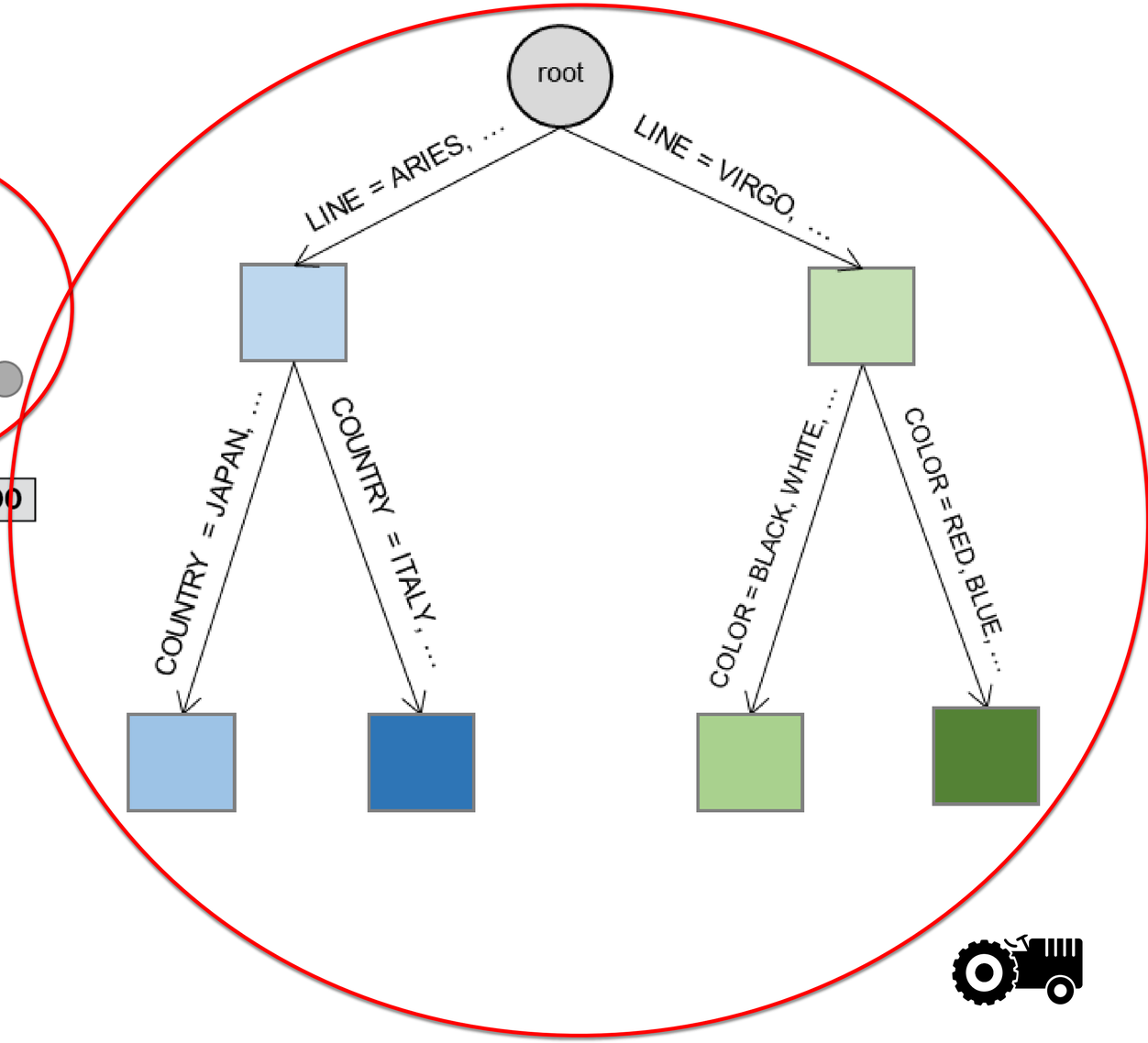
INTEGRATED SOLUTION FOR BUYERS AND SUPPLY CHAIN MANAGERS



CELI
LANGUAGE TECHNOLOGY

RANDOMFOREST REGRESSION MODEL FOR NEW PRODUCTS





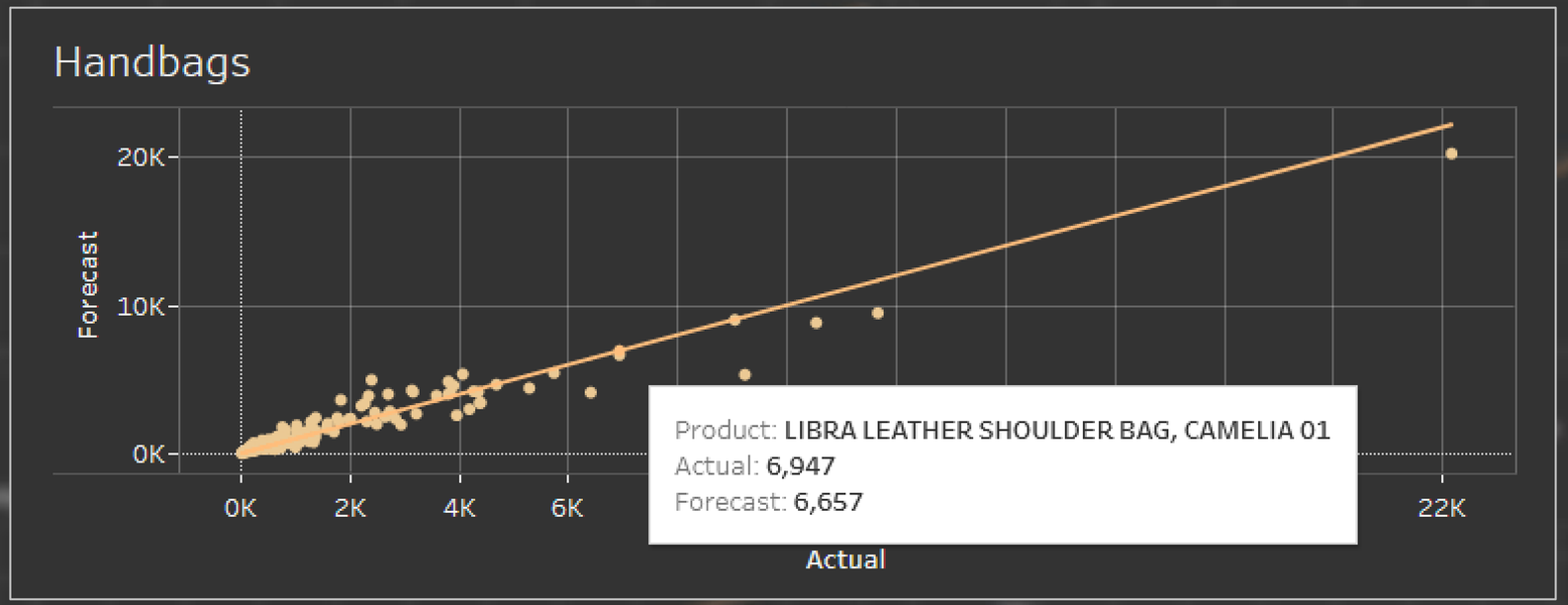
CELI
LANGUAGE TECHNOLOGY

EXPLAINING THE RANDOMFOREST MODEL FOR HANDBAGS

	PRODUCT 	STORE 	OTHER
MAIN VARIABLES	• COLOR	• AVERAGE HANDBAG SALES	
MAJOR CORRECTIONS	• LINE • PRICE	• STORE COUNTRY • SIZE	• SEASONALITY • NATIONALITY • AGE OF CUSTOMER
MINOR CORRECTIONS	• FUNCTION • MATERIAL	• POINTS OF INTEREST • CONCEPT STORE	

CELI
LANGUAGE TECHNOLOGY

Forecast Overview



THREE IMPLEMENTATIONS

1. ML-BASED SHORT-TERM FORECAST
2. ML-ASSISTED FORECAST & ASSORTMENT
3. PREDICTIVE AI FOR HEALTH, SAFETY AND ENVIRONMENT

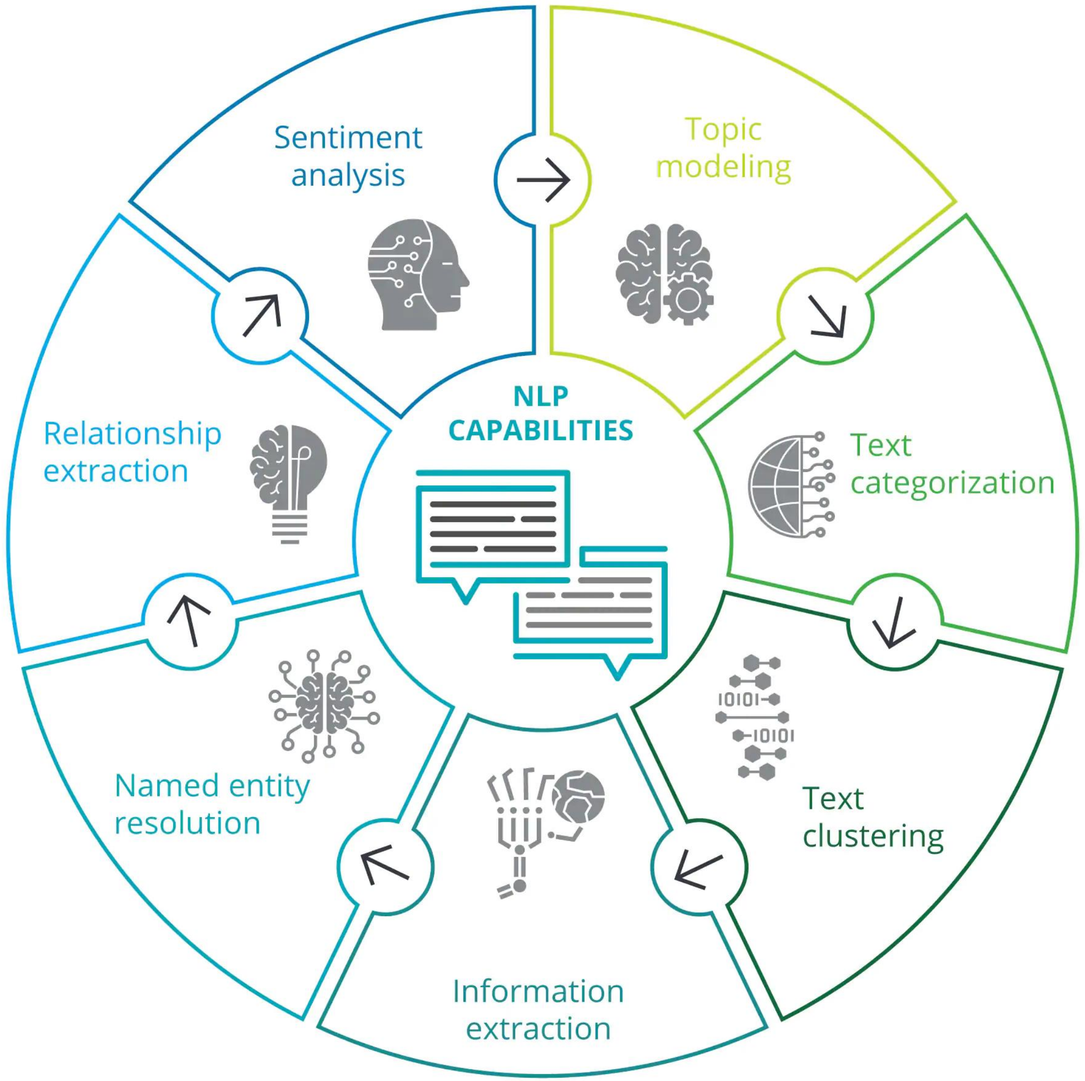
CONTEXT: OUR NLP CAPABILITIES

With recent technological advances, computers now can read, understand, and use human language.

These capabilities allows to recognize **patterns**, categorize **topics**, and analyze customer **opinion**.

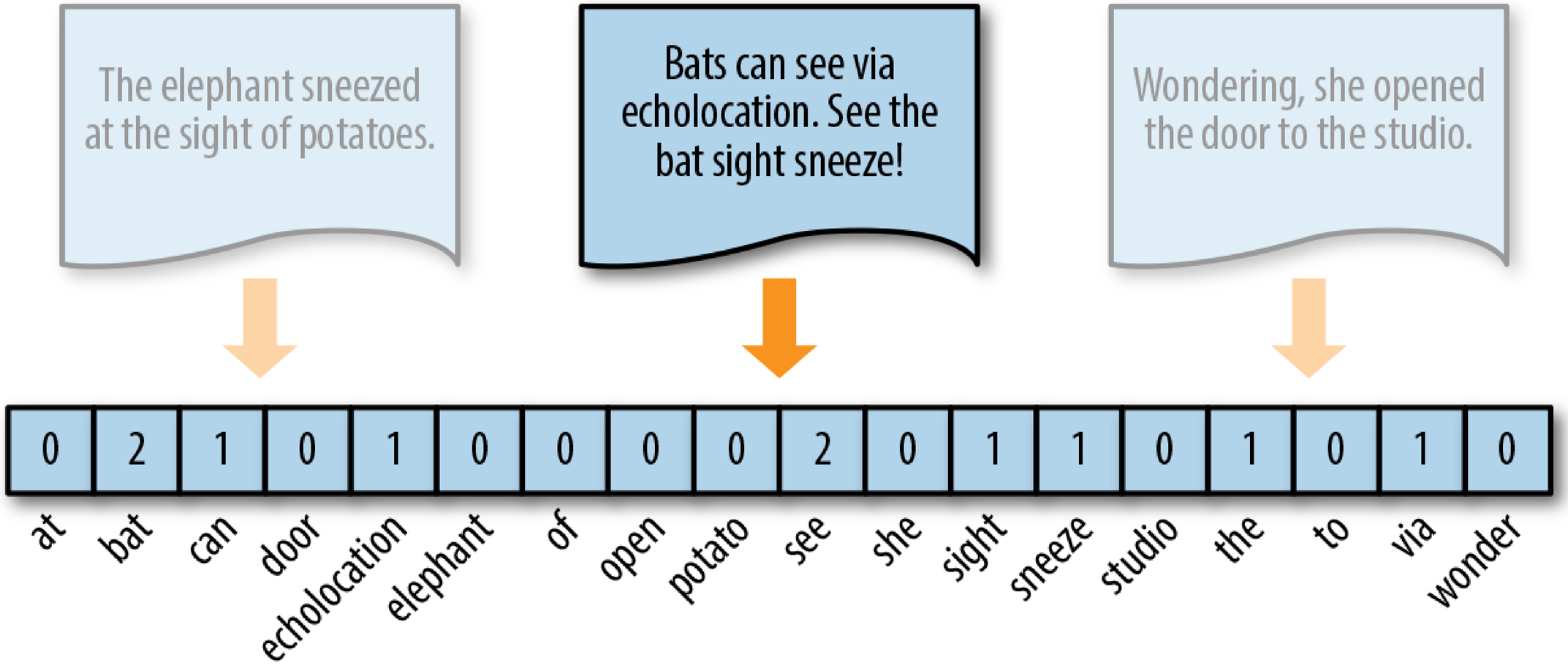
We combine deep learning algorithms with a powerful semantic engine.

We use cutting-edge **Natural Language Processing algorithms** (word and document embeddings methods, recurrent neural networks, attention models ...) to guide the user in their data exploration.



CELI
LANGUAGE TECHNOLOGY

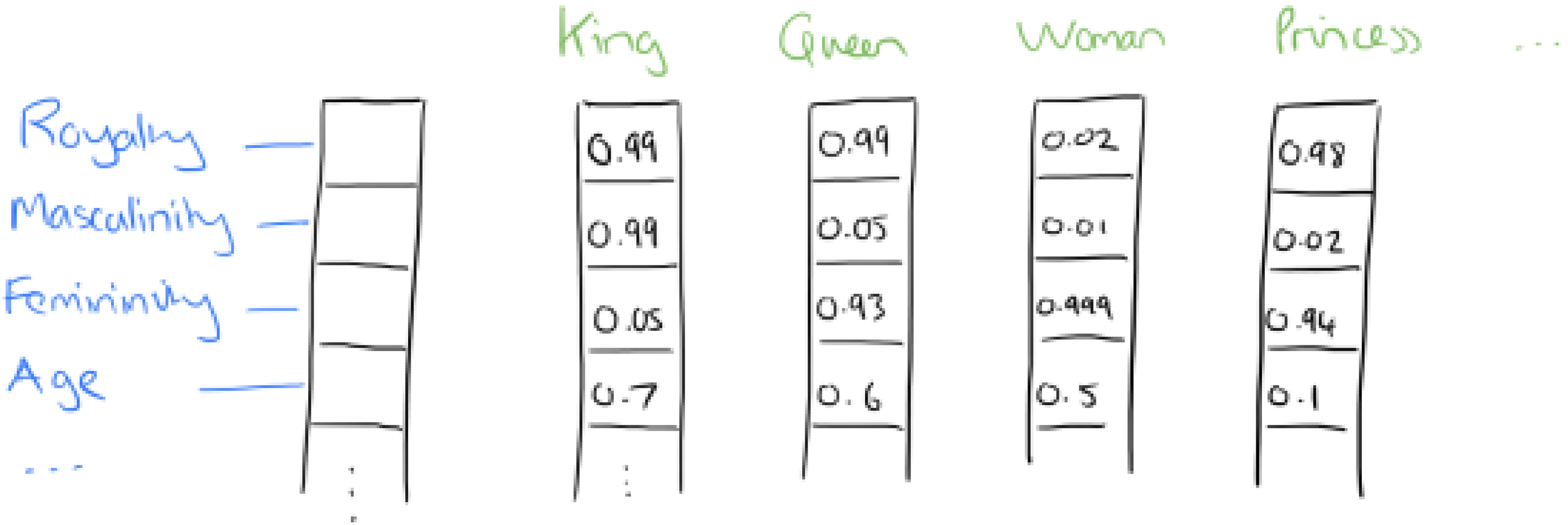
DOCUMENTS ARE BAG-OF-WORDS (1975)



WHAT ABOUT MEANING?

CELI
LANGUAGE TECHNOLOGY

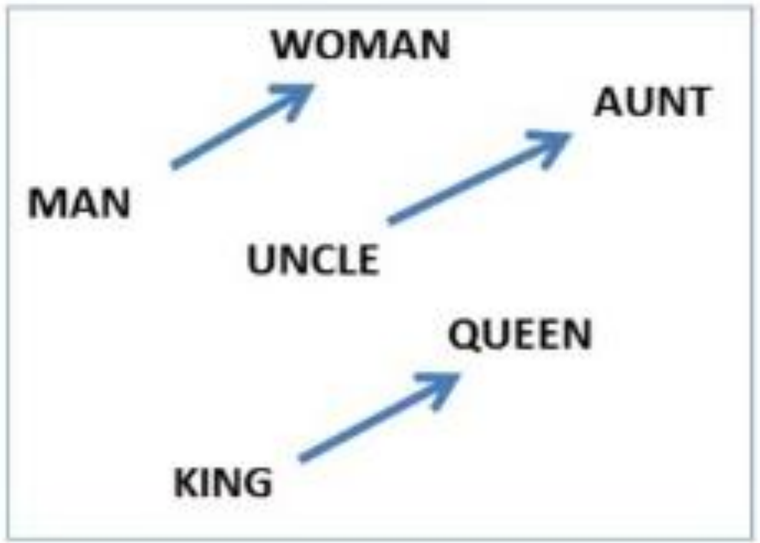
WORD EMBEDDINGS TO THE RESCUE

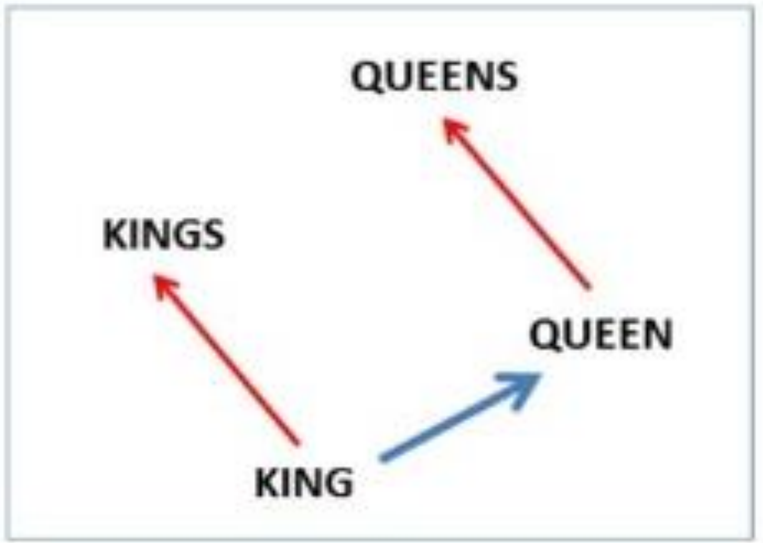


CELI
LANGUAGE TECHNOLOGY

THE DREAM IS TO REASON THROUGH ALGEBRA

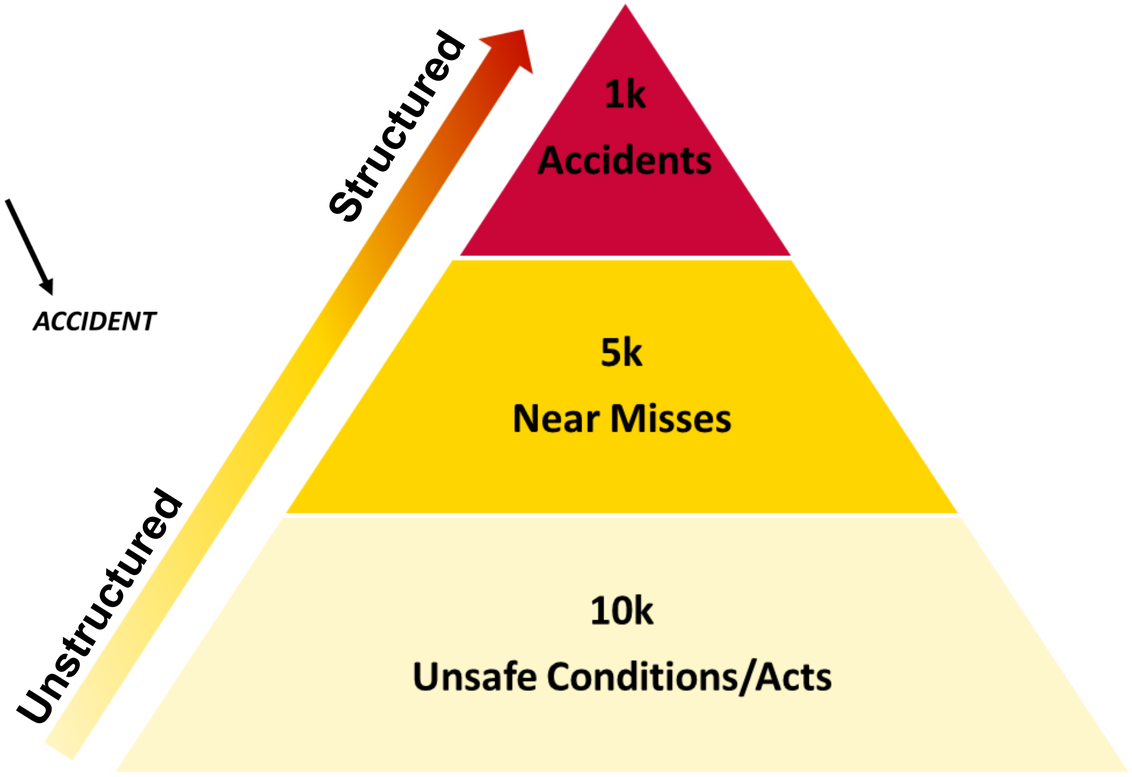
vec(king) – vec(man) + vec(woman) = vec(queen)





CHALLENGE: PREDICT DANGER IN ORDER TO SAVE LIVES

Dynamics/Type of Injury Cause Analysis



Event Description (multi-language, free text)

CELI
LANGUAGE TECHNOLOGY

PREDICTIVE AI FOR HEALTH, SAFETY AND ENVIRONMENT

- Create a **data-driven framework** able to identify the most **relevant and repeating causes and factors of incidents** concerning safety

- Leverage upon the large amount of **free text data** that has been collected, including shorter event descriptions, but also detailed investigation reports

- **Weak signs** hidden in minor incidents must be detected and used as premonitory signs of severe events, to perform scenario analysis and prescribe risk mitigation actions

CELI
LANGUAGE TECHNOLOGY

CHOSEN APPROACH: IN-DEPTH DATA & TEXT MINING

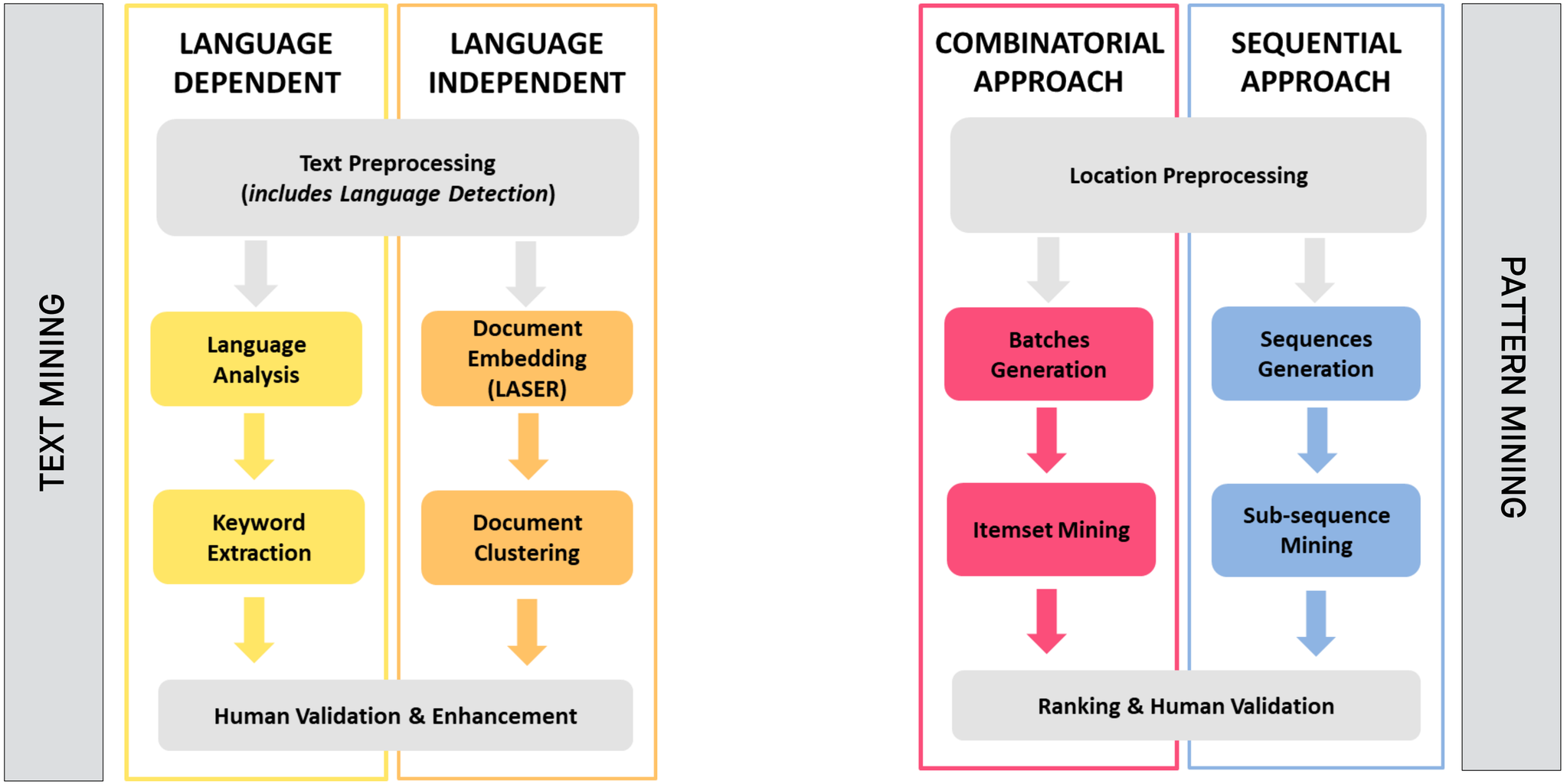
- Because real problem cases are few and Business has already well-established routines, collaborative analysis was undertaken in the first phase of this project

- Text Mining was used to generate text descriptors that could link a severe accidents to the preceding reports (i.e. the presence of misplaced objects could lead to a fall later on)

- Temporal Sequence Mining was conducted to find recurring and correlated patterns across Locations, then to be validated by Business and used as building blocks of subsequent project phases

CELI
LANGUAGE TECHNOLOGY

KEY DATA MINING MODULES FOR HSE



CELI LANGUAGE TECHNOLOGY

Natural Language Understanding for Safety and Risk Management in Oil and Gas Plants,
Abu Dhabi International Petroleum Exhibition & Conference, 11-14 November, Abu Dhabi, UAE
https://www.onepetro.org/conference-paper/SPE-197668-MS

LASER (Language Agnostic SEntence Representations)

Facebook AI Research (FAIR)

Sequence-to-sequence encoder-decoder architecture

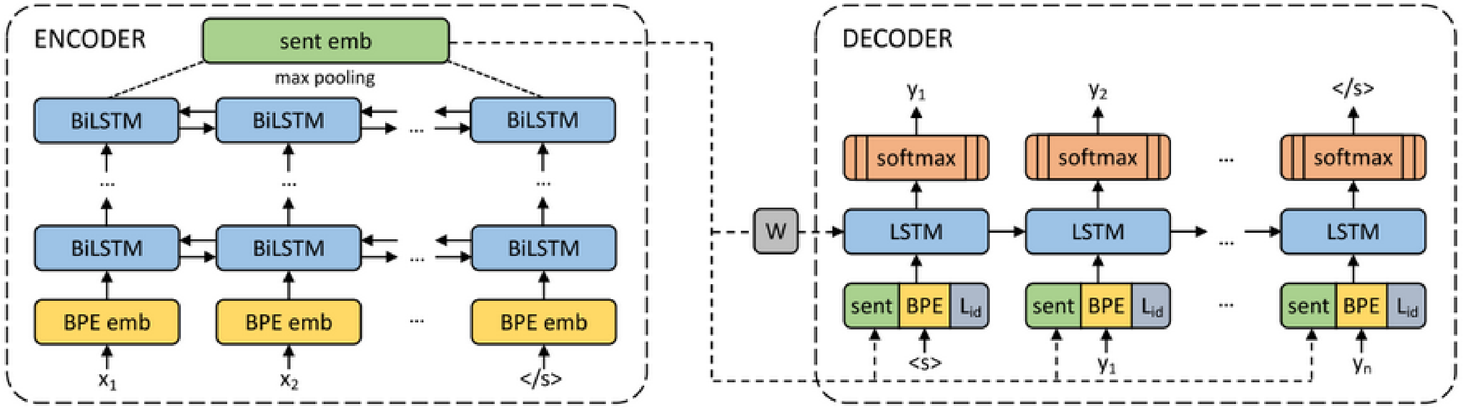


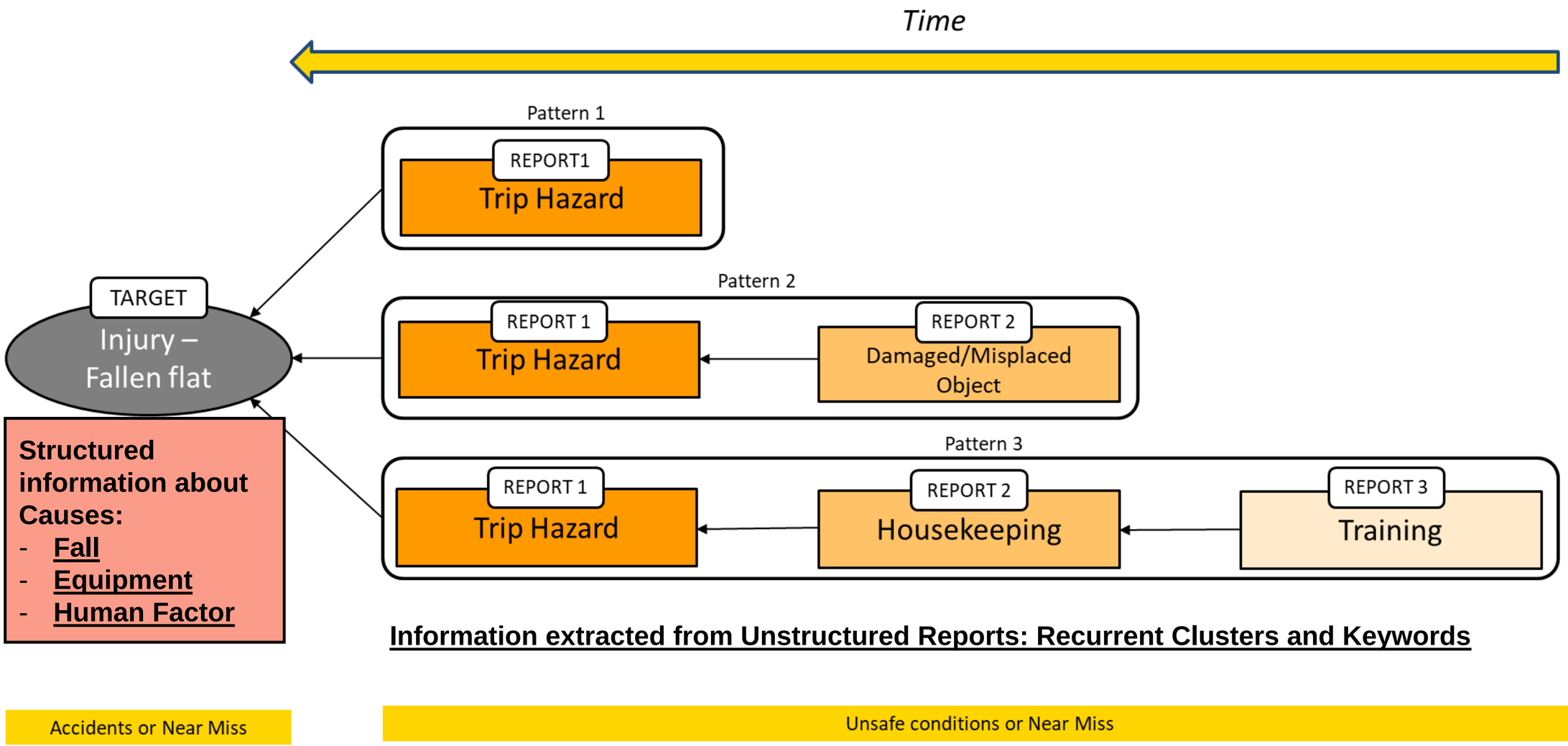
Figure 1: Architecture of our system to learn multilingual sentence embeddings.

Trained on parallel corpora in a way similar to neural machine translation:

- The encoder maps the source sequence into a fixed-length vector representation, which is used by the decoder to create the target sequence.
- This decoder is then discarded, and the encoder is kept to embed sentences in any of the training languages

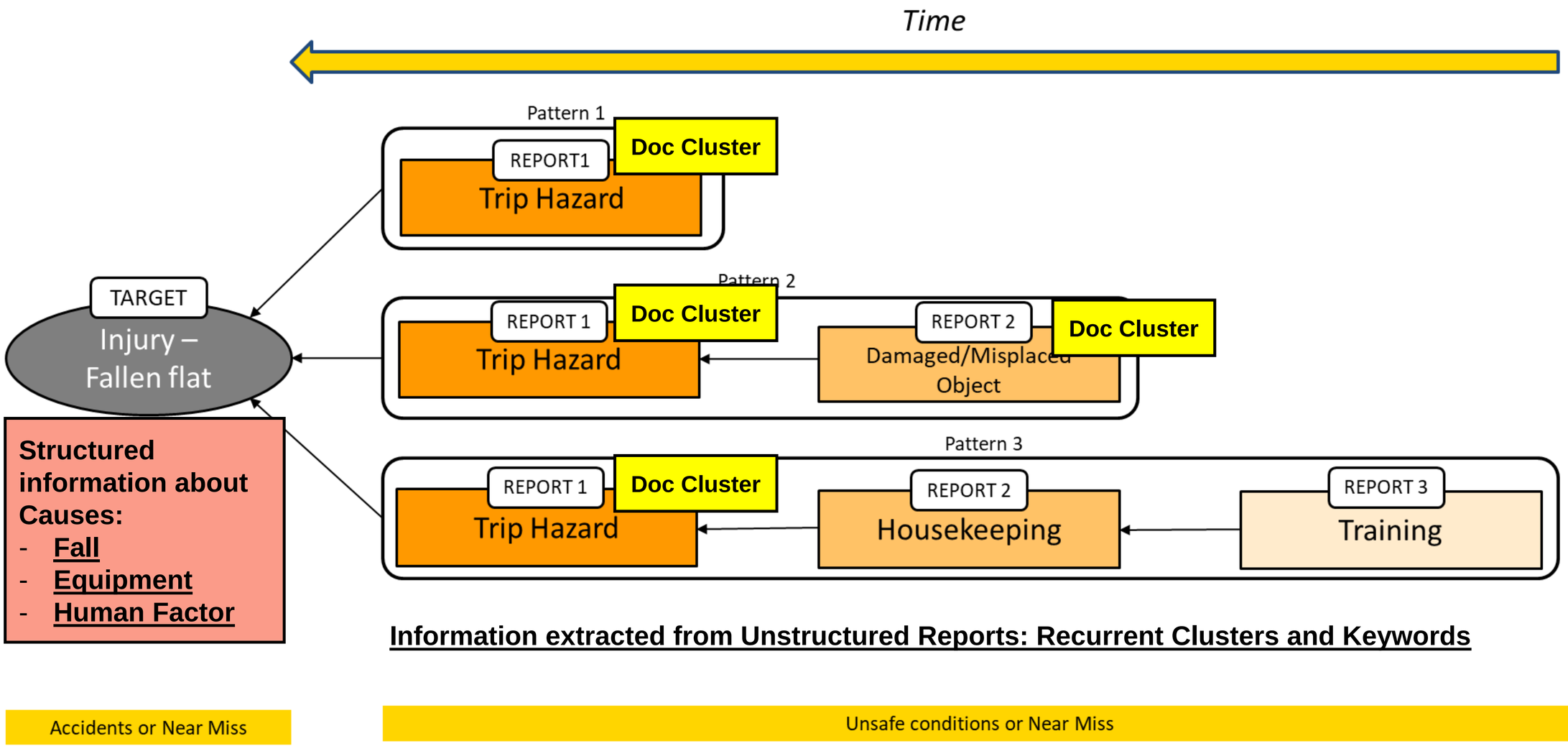
Mikel Artetxe and Holger Schwenk, Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond, 26 Dec 2018. - Published in ArXiv

ADDED VALUE OF UNSTRUCTURED TEXT REPORTS



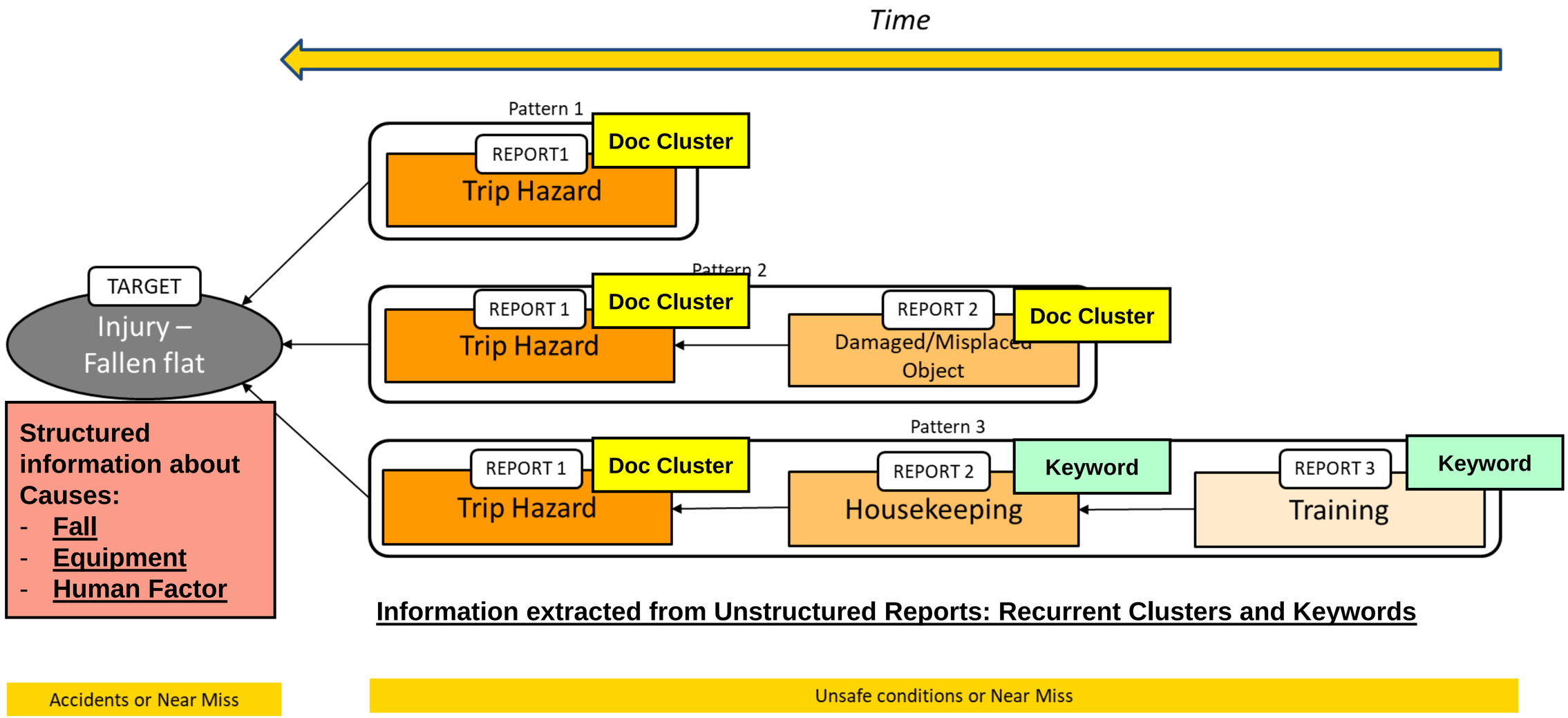
CELI
LANGUAGE TECHNOLOGY

ADDED VALUE OF UNSTRUCTURED TEXT REPORTS



CELI
LANGUAGE TECHNOLOGY

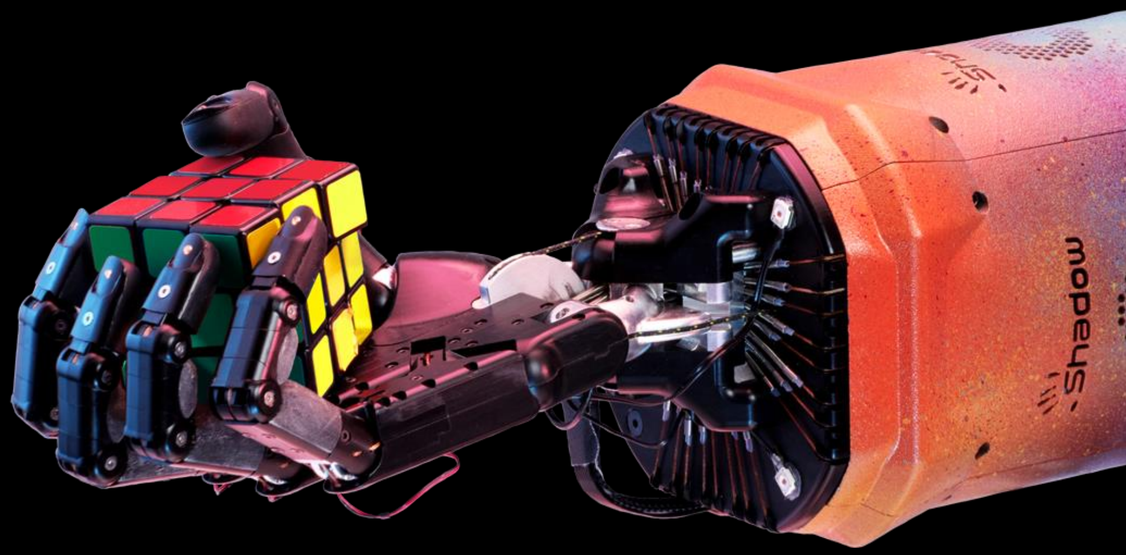
ADDED VALUE OF UNSTRUCTURED TEXT REPORTS



CELI
LANGUAGE TECHNOLOGY

LESSONS LEARNED

1. As of today, we have no doubt that AI can add value to operations in all industries. What makes the difference is how we apply AI techniques depending on context, and most importantly how we define the interaction between AI and Human, assigning correct roles to each of them.



2. Deep Learning has made great advances in recognizing patterns but it’s still far from understanding the cause-effect relationship that is crucial to business.

3. We presented three use cases where we helped to fill the gap between technology and business by defining the correct people-machine interaction that solved concrete problems.

LESSONS LEARNED

1. As of today, we have no doubt that AI can add value to operations in all industries. What makes the difference is how we apply AI techniques depending on context, and most importantly how we define the interaction betwee
assigning correct r

Get in touch!
francesco.tarasconi@celi.it

2. Deep Learning has made great advances in recognizing patterns but it’s still far from understanding the cause-effect relationship that is crucial to business.

3. We presented three use cases where we helped to fill the gap between technology and business by defining the correct people-machine interaction that solved concrete problems.

Big Shoes and Beautiful Mind