# DataHalfLife

July 2019 White Paper

The philosophy of science has studied concepts related to **knowledge decay** for **centuries...**

Yet the information age of **big data** continues to spur **new challenges.**

Many scientists have referred to the analogy that **data**, or even **knowledge** itself, contains a **"half life."**

At DataHalfLife, we evaluate issues of data **reliability**, **validity**, and **evenness**—providing consumers with **optimal avenues** for analyzing the best data available.

We take this notion to the next step and offer a unique way to **evaluate existing datasets and identify their lifespans of reliable use**. In short, the aim is to turn data decay into data superiority.

# DataHalfLife

**THE GOAL: OPTIMIZE DATA ANALYSES**

Establishing data superiority involves at least three key elements. First, prior to data analysis, we conduct an analysis of the dataset. This ensures that the data points within the dataset are "even." The second step involves separating pertinent data from relevant data. The third step consists of evaluating data for decay.

**SOLUTIONS TO THE PROBLEM OF DATA UNEVENNESS**

*Pre-Analysis Data Analysis?*

A major problem in large-n or big datasets stems from "data evenness." This occurs most often when multiple coders build a dataset collectively and interpret coding rubrics differently and/or rely on unique sourcing practices. Ultimately, this ends in incongruent data points within a given dataset.

Thus, prior to data analysis, we conduct an analysis of the data. This involves a series of inter-coder reliability (ICR) tests and source detection and evaluation in order to make sourcing transparent. The core aim centers on confirming the bedrock of probabilistic analyses: reliable and valid data.

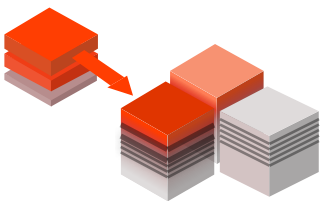
A follow-up procedure entails assessing specific consumer needs and matching pertinent (i.e. reliable and valid) data sources to meet those needs. We call this "data allocation optimization." However, this is not the end of the pre-analysis data analysis...

**SOLUTIONS TO THE PROBLEM OF DATA DECAY**

Unfortunately, all pertinent data is not necessarily equally relevant. Based on consumer needs for the model inputs, we rank the pertinent data. We can call this "ordering relevance within pertinent data." Over time, data remains fully pertinent but less relevant, as its predictive capacity diminishes as it becomes more dated.

Additionally, in terms of quality, all data is not created equally. We establish a hierarchy of data quality specific to the domain. (This is unrelated to the ordering of pertinent data.) We can call this step "data quality ranking," which requires a high-degree of domain expertise. Here, subject-matter experts (SMEs) are essential in accounting for data decay due to the notion of "splitting events."

Splitting events establish a before and an after. They usually mark significant shifts in human geography, demography, political institutions, and/or socio-economic structures. SMEs identify these splitting events to uncover markers for capturing data decay. These markers then inform the input weights for quantitative and computational modeling.

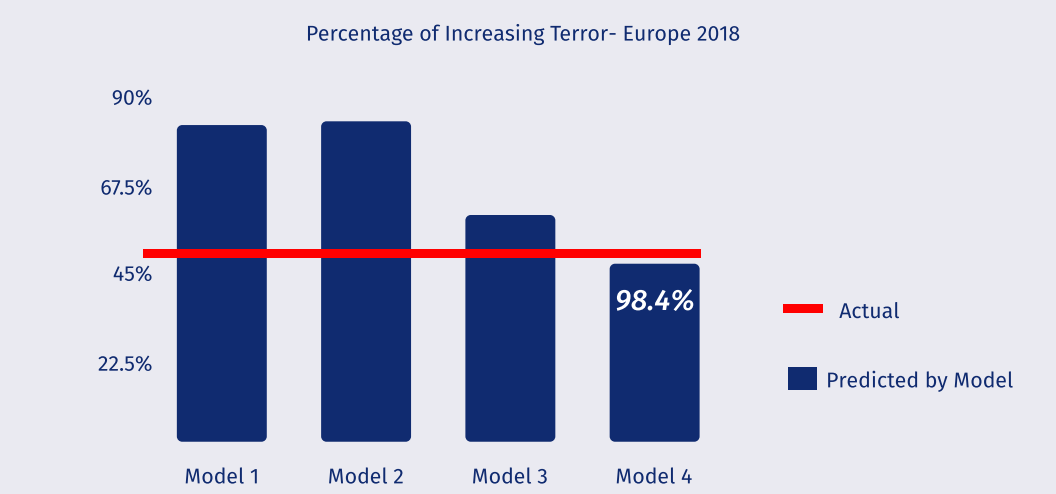


**1 Identify Pertinent Data**
**2 Order and Weigh Data by Relevance**
**3 Order and Weigh Data by Quality**

## Domain-specific stepwise approach for politics and conflict (first use case)

Example: Predicting terrorist attacks in Europe for the 2018 year, we applied our procedures to the Global Terrorism Database (GTD) (1970-2017).

Below, beginning from the left, the four models treat the GTD's data in varying fashions.




**Model 1** Analyzes each of GTD's decades of data (1970s, 1980s, 1990s, 2000s, 2010s) with equal degrees of relevance.

**Model 2** Groups the data from the 2010s and 2000s together, giving them twice the weight than the previous decades of data.

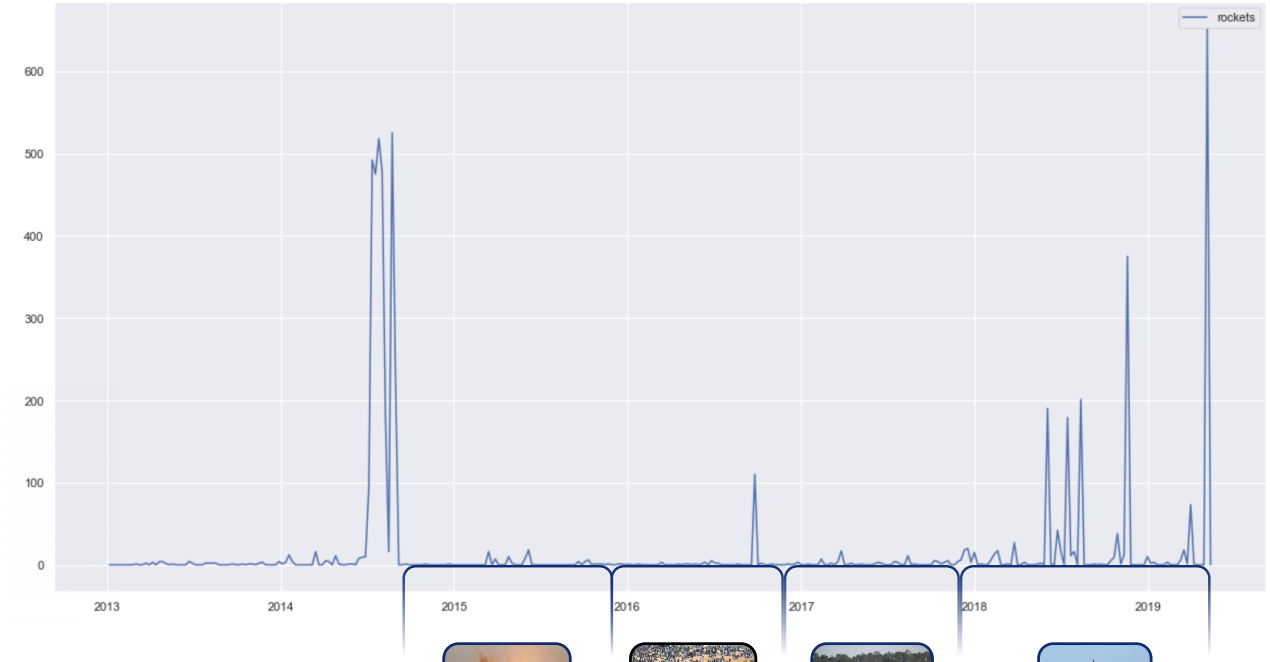
**Model 3** Weighs the data by decade increments, with each decade being twice as relevant than the previous.

**Model 4** Uses an approach with custom data decay steps that are congruent with shifts in terrorism strategy and the underlying changes in human geography that caused these shifts.

Each of the models produces a unique prediction. The fourth model, which fully follows our procedures, estimates the best predictive capacity—reaching 98.4% accuracy.

## Delving deeper into domain-specific stepwise approach for politics and conflict

Example: Understanding Gaza rocket & mortar attacks 2013-present (modeling in-progress)




**Events & changes in tactics**

(2014-)
Post-Op
Protective Edge

(2016-)
March of Return

(2017-)
Incendiary
Kites/Balloons

(2018-)
Iron-Dome
"teasing"

**Underlying geo-political shifts**

JCPA
("Iran deal")

US
Elections

US -JCPA
withdrawal

**Model hyper-parameters**

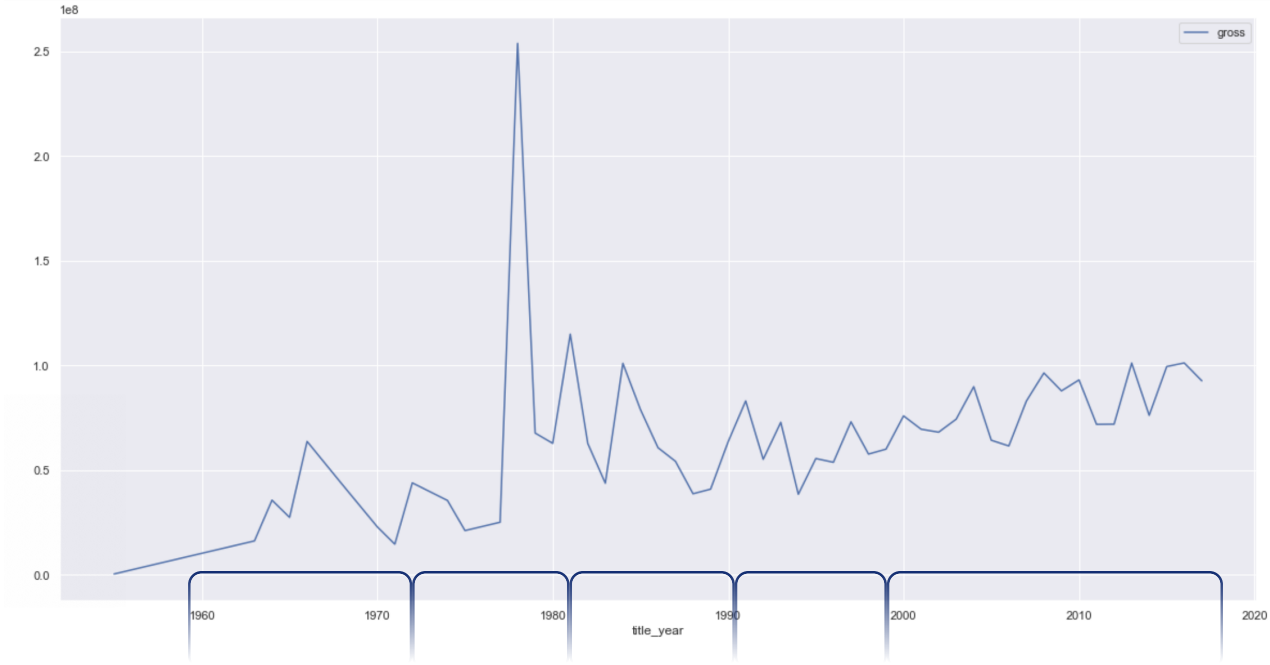
$\alpha_1$ $\alpha_2$ $\alpha_3$ $\alpha_4$

**How is this different from simple exponential smoothing?**

A stepwise & domain-specific approach eliminates the rule-of-thumb factor and introduces relevance.

## Domain-specific stepwise approach for entertainment

Example: US motion picture gross (adventure/action/scifi)




**Events and movies**
(non-comprehensive)


Spartacus


Star-Wars


Back to the Future


Jurassic Park

Netflix era

**Underlying geo-political shifts**
(non-comprehensive)

1960s
Vietnam War

1970s
Energy crisis

1980s
US GDP growth

1990s
-Fall of USSR
-WorldWideWeb

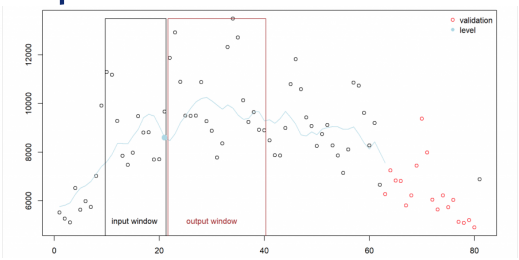
2000s
-War on Terror
-Google

**Model hyper-parameters** $\alpha_1$ $\alpha_2$ $\alpha_3$ $\alpha_4$

## How does our approach enhance modeling techniques and forecasting ability?

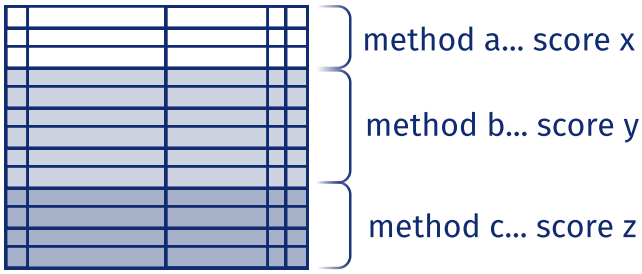
**1** Addressing the modeling paradigm of overfitting vs generalization which can enhance a wide variety of models in statistics and machine learning: regression, classification and clustering.

**Example**




**Hybrid-Exponential Smoothing RNN model**
Developed by Uber's 2018 M4 forecasting competition winners: Slawek Smyl, Jai Ranganathan, Andrea Pasqua

**2** Data pre-processing that takes into account changes in data collection.




**3** A framework to fuse static and dynamic data