

# AI Powered tools for name data harmonization

Yann Cherel  
Data Management and Governance Section  
12 May 2025

# Introduction / objectives

Typical problems that we are trying to address

Name harmonization achievements

# Typical problems that we are trying to address

- **Subsidiaries and affiliates**

Honda Motors USA vs Honda Motors vs Honda

- **Legal identifiers**

Inc, LLC, Ltd, Pvt, công ty TNHH, AB, SNC

- **Different transliterations across languages**

"محمد" can be "Mohammed" or "Muhammad"

- **Inconsistent formatting**

Coca-Cola, Coca Cola, CocaCola

- **Typographical errors and noise**

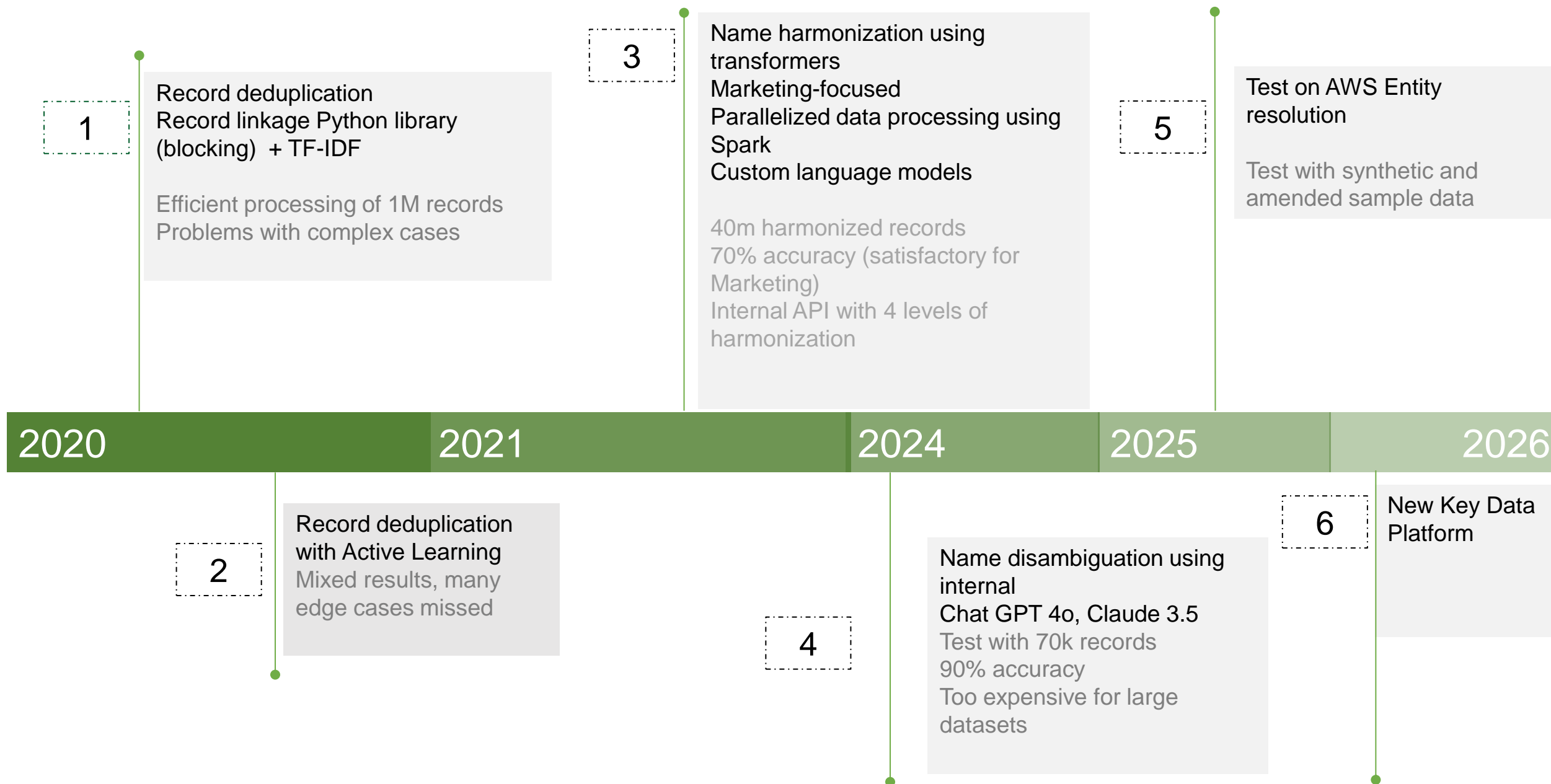
Amazoan Technologies, Phillip Moris

- **Acronyms, abbreviations, region-specific spellings**

IBM vs International Business Machines

Bayerische Motoren Werke AG vs BMW

# Name Harmonization achievements



# Name harmonization for Marketing - Process

	1 Producer	2 Landing Zone	3 Cleanse, control, enrich	4 Match & Dedupe	5 Mastering and survivorship
Activities & Rules		Filtering rules Ingestion	Normalization rules	Grouping rules Country-level grouping	Survivorship rules Global Grouping
References & tools			Language detection of non latin characters: Nakatani Shuyo  Neural translation  ER Model based on RoBERTa (Robustly Optimized BERT Approach)  Magerman, T., Van Looy, B. & Song, X: <a href="#">Name Harmonisation Eurostat – 2006</a>	TF-IDF (Term Frequency – Inverse Document Frequency)  Semantic distance model based on RoBERTa (Robustly Optimized BERT Approach)	TF-IDF  MinHashLSH (MinHash with Locality-Sensitive Hashing)



# Thank you! Questions?

© WIPO, 2025



Attribution 3.0 IGO  
(CC BY 3.0 IGO)

The CC license does not apply to non-WIPO content in this presentation.

Photo credits:

