Smartricity

Catalogue deduplication

Team

Sebastian Schmidt, B.Sc.
CEO

Sales
Web Development

Andreas Donig, B.Sc.
CTO

Architecture
Development
Security

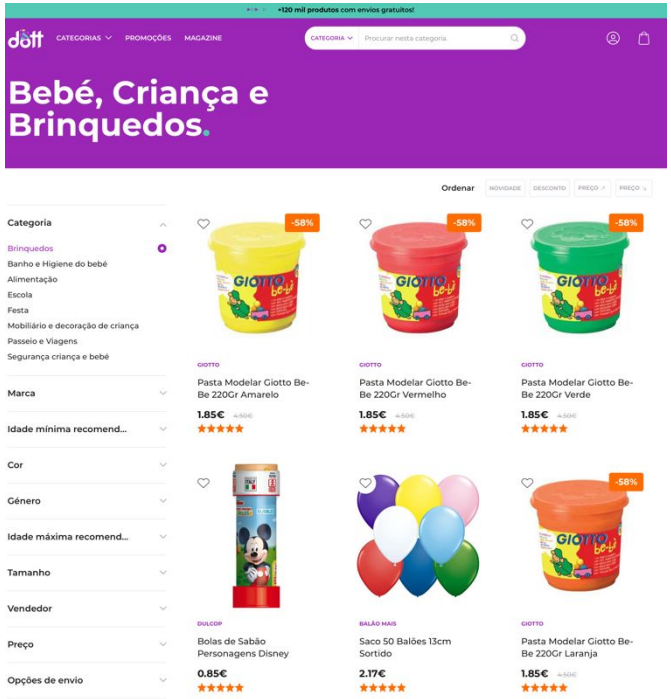
Michael Hasler, B.Sc.
COO

Operations / Finances
Sales

Dr. Florian Wahl
Data Scientist

Machine Learning
Data Analytics

Problem



- Heterogenity
- Highly dimensional data
- Natural sparsity

Solution

- F1: 92.31%
- Accuracy: 99.97 %
- ROC AUC: 94.99 %

	No match predicted	Match predicted
No match	8,590 *99.99 %*	1 *0.01 %*
Match	2 *10.00 %*	18 *90.00 %*

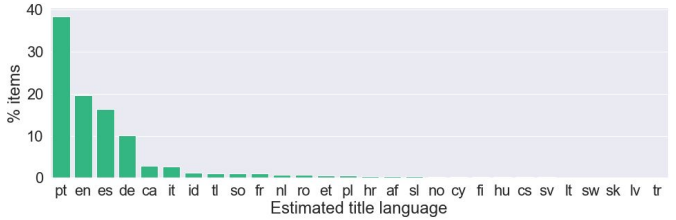
Scalability

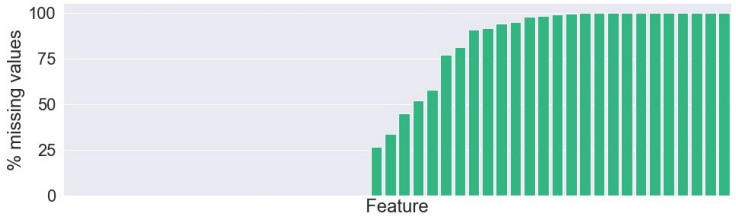
- F1: 98.61 %
- Accuracy: 99.53%
- ROC AUC: 98.78%

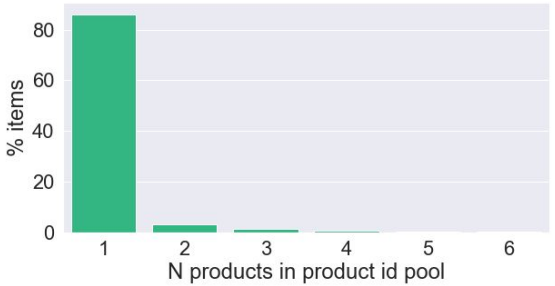
	No match predicted	Match predicted
No match	57,561 *99.85 %*	84 *0.15 %*
Match	238 *2.30 %*	10,094 *97.70 %*

Dataset

- Multilingual word embedding
- Attribute summarization with connected neural networks
- Hierarchical deep ML for matching inference



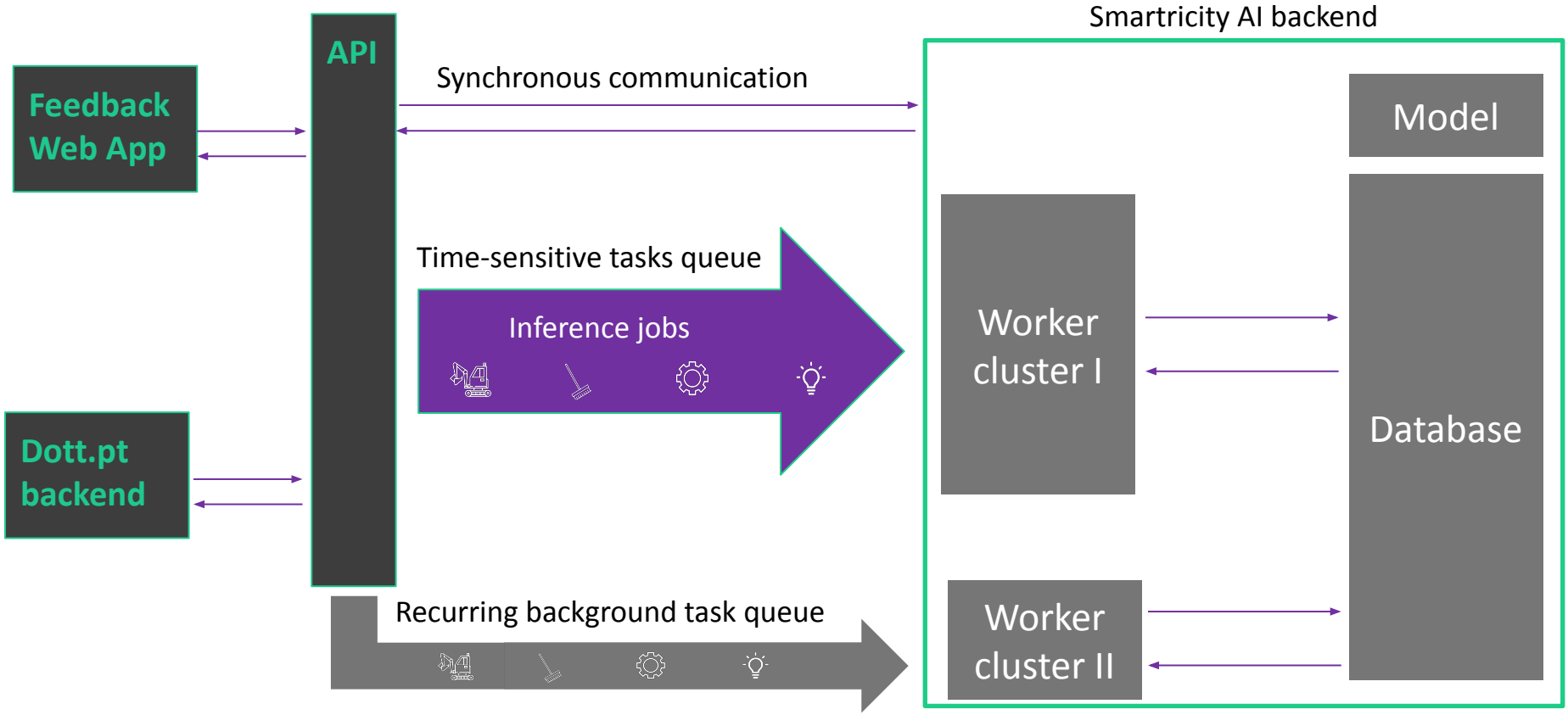




Methodology

- Flatten & split
- Rule based
- ML pipeline:
  - Overlap blocking
  - Word embedding
  - Attribute similarity computation
  - Classification
  - Inference

Product Architecture



Demo

https://reach-demo.vercel.app/

Traction, Business Model

Own website & app

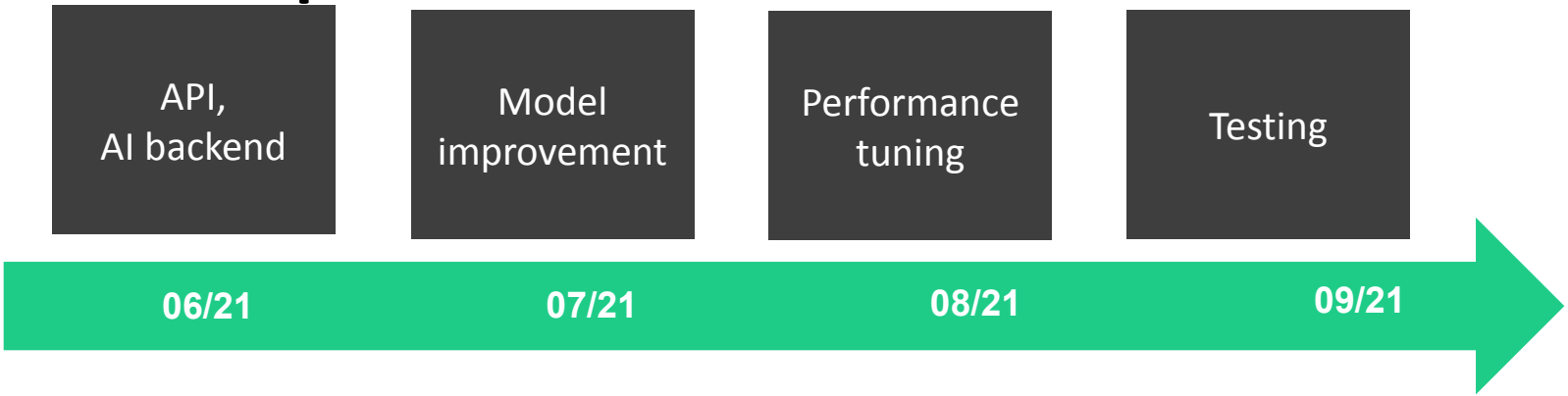
Integration: Energy provider apps

Data processing & deduplication for price comparison company

Pricing:

Base fee + variable share

Roadmap, Risks



- Privacy
- Throughput
- Unrepresentative sample data

SWOT

	Strenghts • Working model prototype • Team • Market proven architecture	**Weaknesses** • Prototype still extendible • Small company • No understanding of portuguese
Opportunities • Increased sales • Low risk, high reward • Additional benefits	➢ Very good results for test data ➢ Solution can improve UX and increase sales ➢ Market proven architecture and business model promise high rewards at low risk ➢ Architecture allows recurrent plausability checks and recurrent cleaning, correcting of existing data	➢ Prototype will be extended in the next phase ➢ Team will be enhanced in the future ➢ Language problems are solved by language processing and improved skills
Threats • Privacy • Throughput • Unrepresentative sample data	➢ Infrastructure GDPR compliant ➢ Architecture already in use: High throughput for other costumers ➢ Working model prototype looks very promising	➢ Focus on product data ➢ Tools for multilanguage processing reduce influence of human understanding ➢ Prototype will be improved

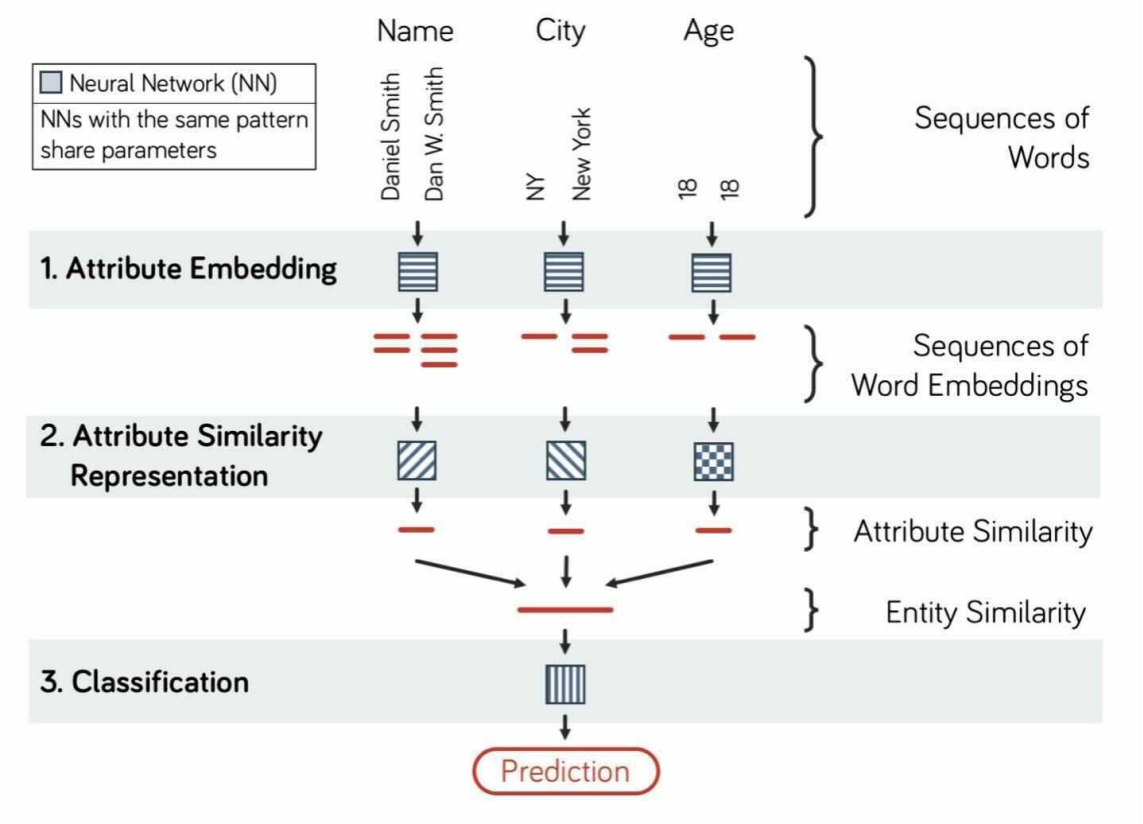
Benefits

- Scalable working proof of concept:
  92% F1 on dataset, 98 % on other data
- UI for monitoring, expert feedback and result explanation
- Recurrent cleaning, correcting, improvement of existing data in background tasks

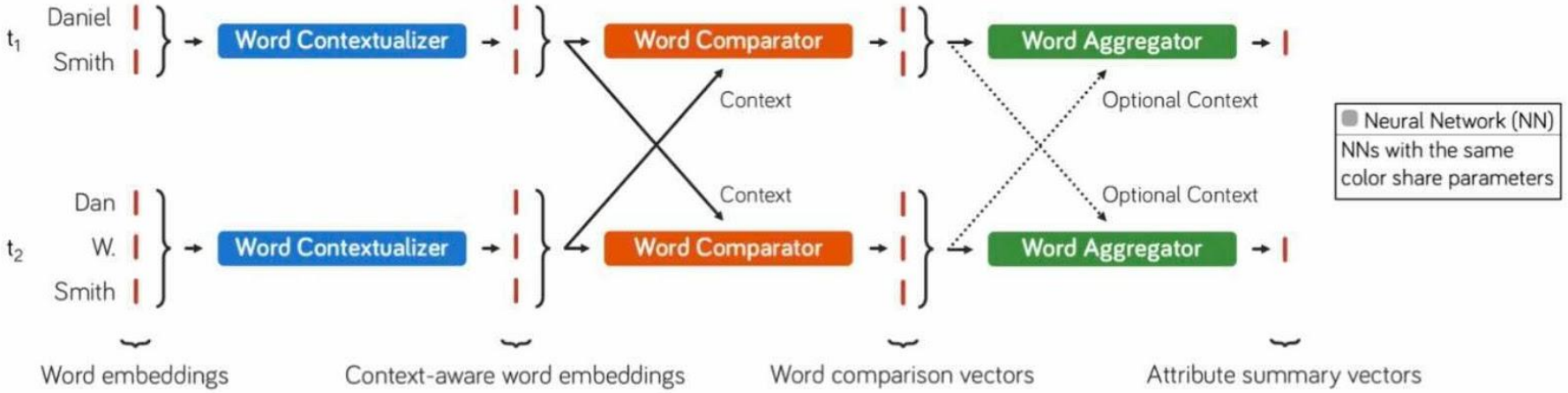
Smartricity GmbH
Bahnhofstraße 10, 94032 Passau
+49 1577 022 42 37
hello@smartricity.de

Contact:
Michael Hasler
+49 162 331 6968
michael.hasler@smartricity.de

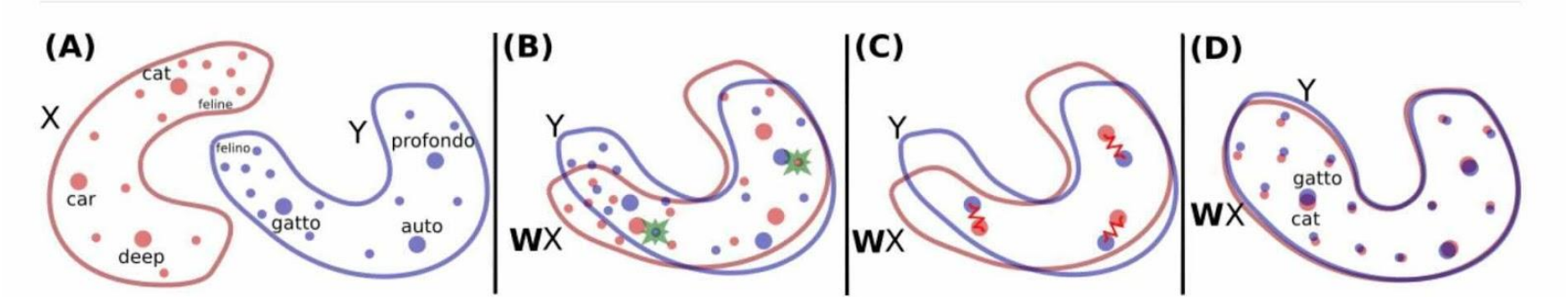
A1 – ML model


Name
City
Age
Neural Network (NN)
NNs with the same pattern share parameters
Daniel Smith
Dan W. Smith
NY
New York
18
18
Sequences of Words
1. Attribute Embedding
Sequences of Word Embeddings
2. Attribute Similarity Representation
Attribute Similarity
Entity Similarity
3. Classification
Prediction

A2 – attribute similarity



A3 – Muse


(A)
cat
feline
X
Y
profondo
felino
car
gatto
auto
deep
(B)
Y
WX
(C)
Y
WX
(D)
Y
gatto
WX
cat

A4 – dataset

- 1852 entries
- 43 columns after flattening
- Mixed languages for titles and descriptions: mainly, pt, es, fr, en, de

Top 5 in product name:

- pt 712
- en 361
- es 294
- de 191
- ca 51

Top 5 in description:

- pt 1717
- en 60
- tl 15
- es 13
- it 11

- No duplicates in terms of unique product identifiers:
  Groups of product variants (e.g. color, or variant, such as type of animal).
- Largest number of variants of one product is 7
- 89% / 1650 entries with valid GTINs (13 digits + checksum ok)
- Out of 43 features, 19 features have no missing values, 15 have more than 95% values missing.
  Those are mainly the flattened attributes.
- No matches were available by using the product id (variant) column.

A5 – attribute similarity computation

1. Computed considering their context using a single-layer gated recurrent unit

2. New items context compared with candidate items word and vice versa resulting in word comparison vectors.
   Comparison performed using attention based word comparator.

3. Word aggregation performed using attention and RNN based word aggregation

A6 – tools

- Numpy, pandas for data organisation and explorative analysis
- torch for deep learning.
- We use a framework for candidate blocking and matching
- magellan toolbox for data preparation and candidate blocking and deepmatcher for matching pairs.
- For explaining decisions, we use the SHAP library
- In the next phase
- extend the deepmatcher framework
- word embedding of multilingual data. A first step will be to use the MUSE framework as a custom word embedding mechanism in deepmatcher. Both utilize the fastText library.
- feedback tool: React based web app using Next.js and Redux. It is written in JSX and TypeScript.
- The API will be built on Python3 and Django3. \
- AI backend is built on Celery, RabbitMQ, Pandas, Numpy, Dask and PostgreSQL.
- horizontal scalability, container orchestration using kubernetes will be implemented.
- vertical scalability and load-dependent autoscaling, we rely on cloud infrastructure hosted in European data centers. Doing so ensures data will not leave the European Union.