Beating the Status Quo Ranking on Shorthead Queries

On the way: building a Learning to Rank Model for OTTO Search

MICES 2022 | Team Ranking @ OTTO | Felix Rolf

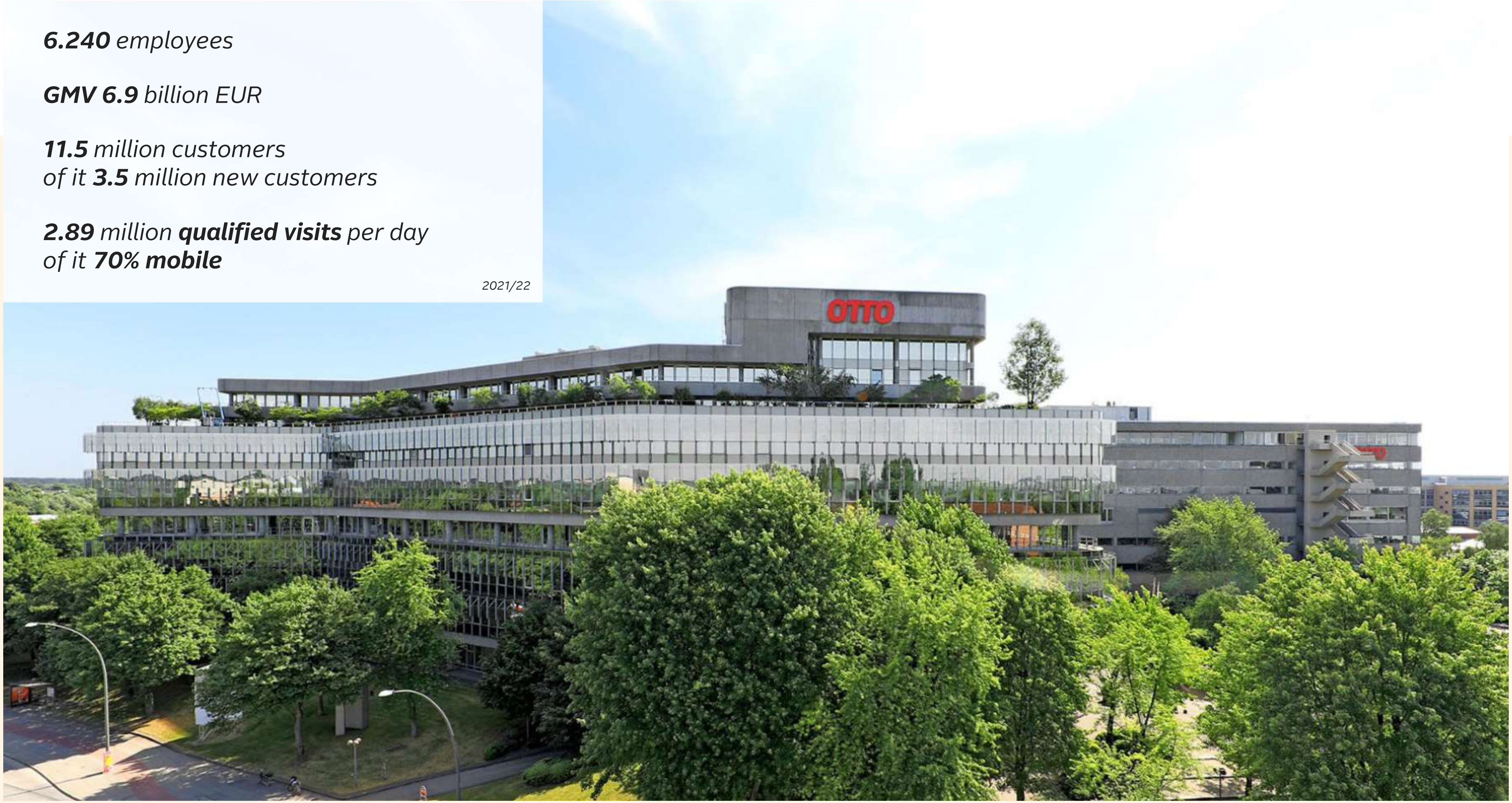
6.240 *employees*

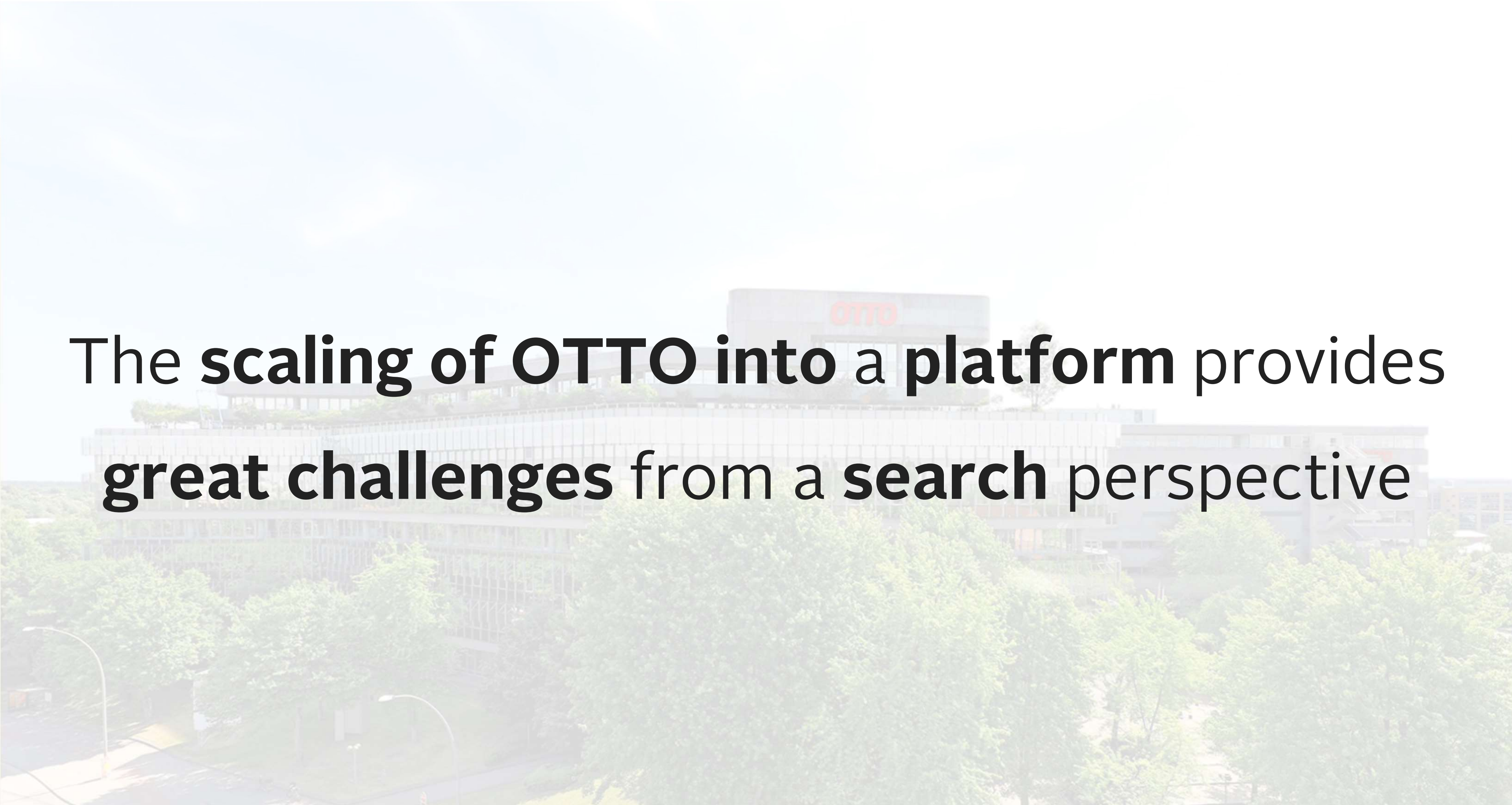
GMV 6.9 *billion EUR*

11.5 *million customers*
of it ***3.5*** *million new customers*

2.89 *million* ***qualified visits*** *per day*
of it ***70% mobile***

2021/22

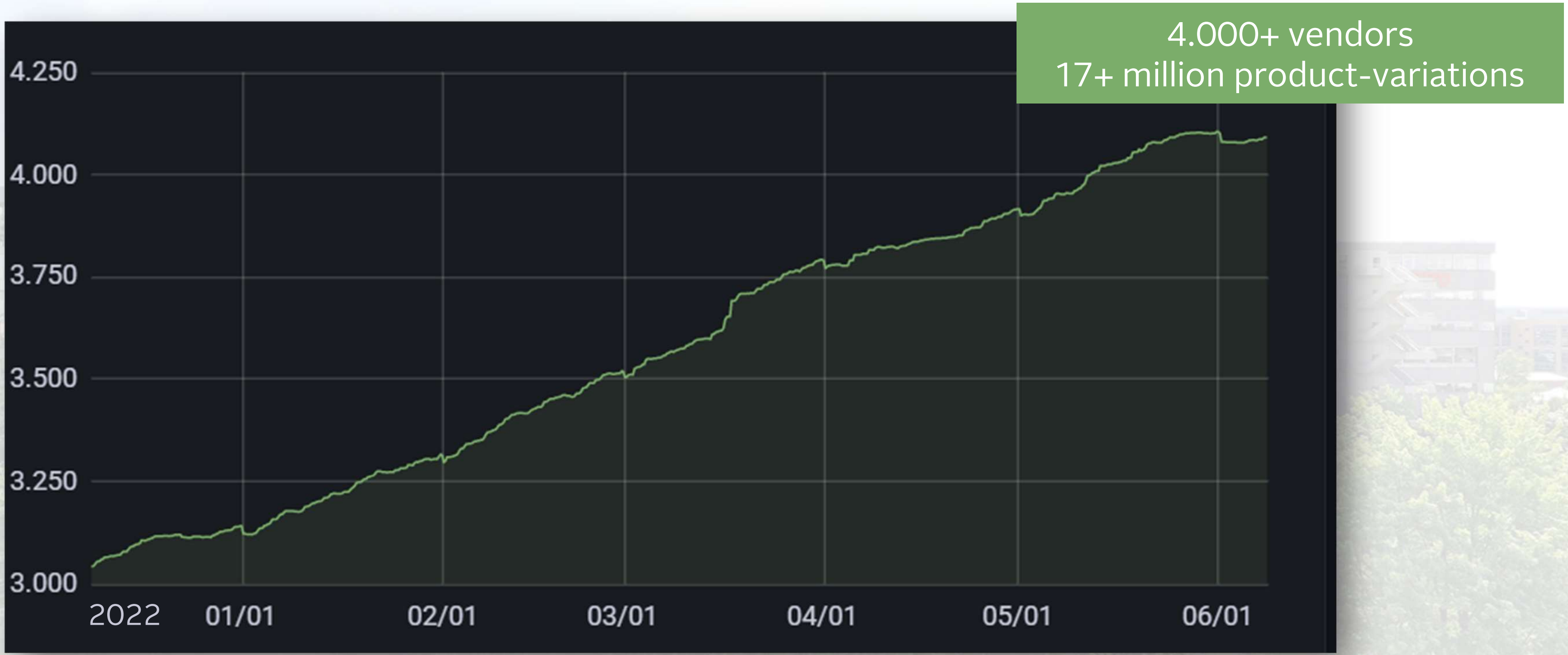




The **scaling of OTTO into** a **platform** provides **great challenges** from a **search** perspective

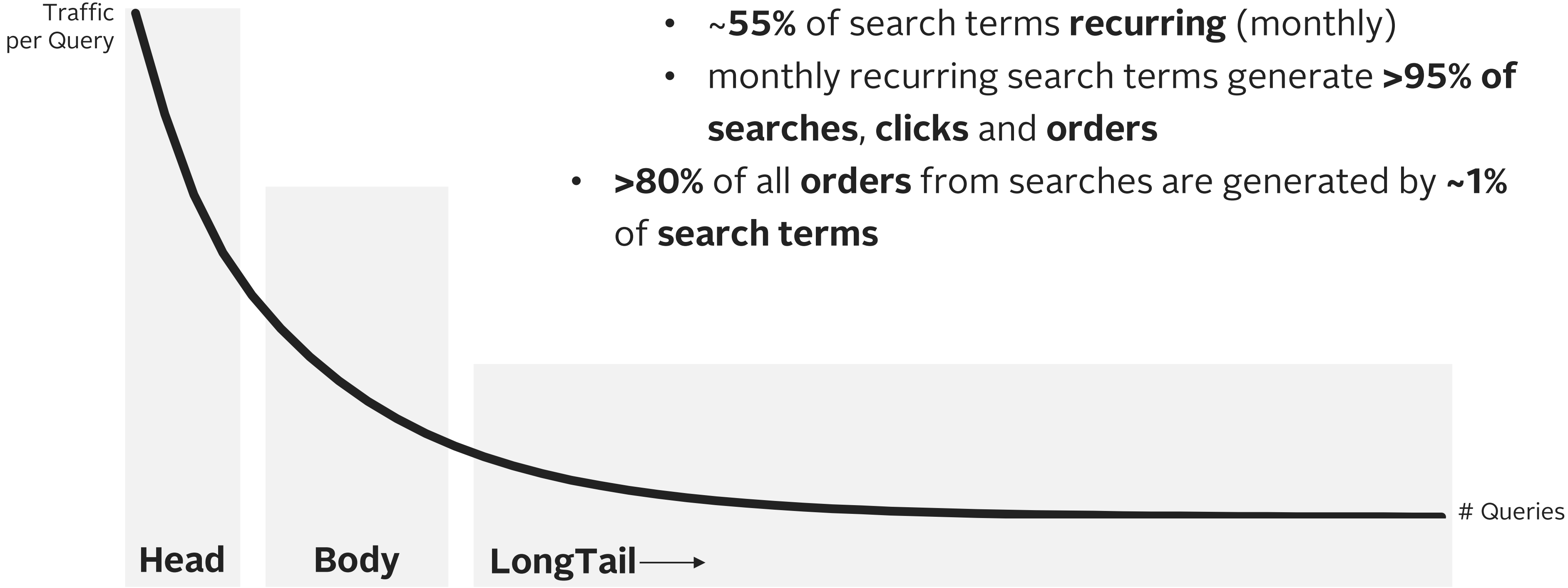
1. **Increase** in **number of vendors & products** available at OTTO
2. Heterogenous **product data quality** due to products from different sources
3. **Optimization** for **customer relevance**, not business KPIs
4. Increased number of **body-** and **longtail searches**

1 **Increase** in **number of vendors & products** available at OTTO



4 Increased number of **body-** and **longtail searches**

- ~**32 million** unique search terms per year
- ~**5.5 million** unique search terms per month
    - ~**55%** of search terms **recurring** (monthly)
    - monthly recurring search terms generate **>95% of searches**, **clicks** and **orders**
- **>80%** of all **orders** from searches are generated by **~1%** of **search terms**



Currently we are **ranking** all search results based on a **rule-based**, **linear combination** of three **main ranking factors**:

- **Linguistic relevance**
- **Availability**
- **Popularity** (Units/Revenue)

Learning to Rank algorithms **enable** a ranking based on **true relevance** from a **customers** perspective

Training data with perfect product ordering per query

Judgements
Give the perfect ordering of products per query for historic data

Features
Abstraction/Encoding of product, query and product-query-matching information

LTR Model

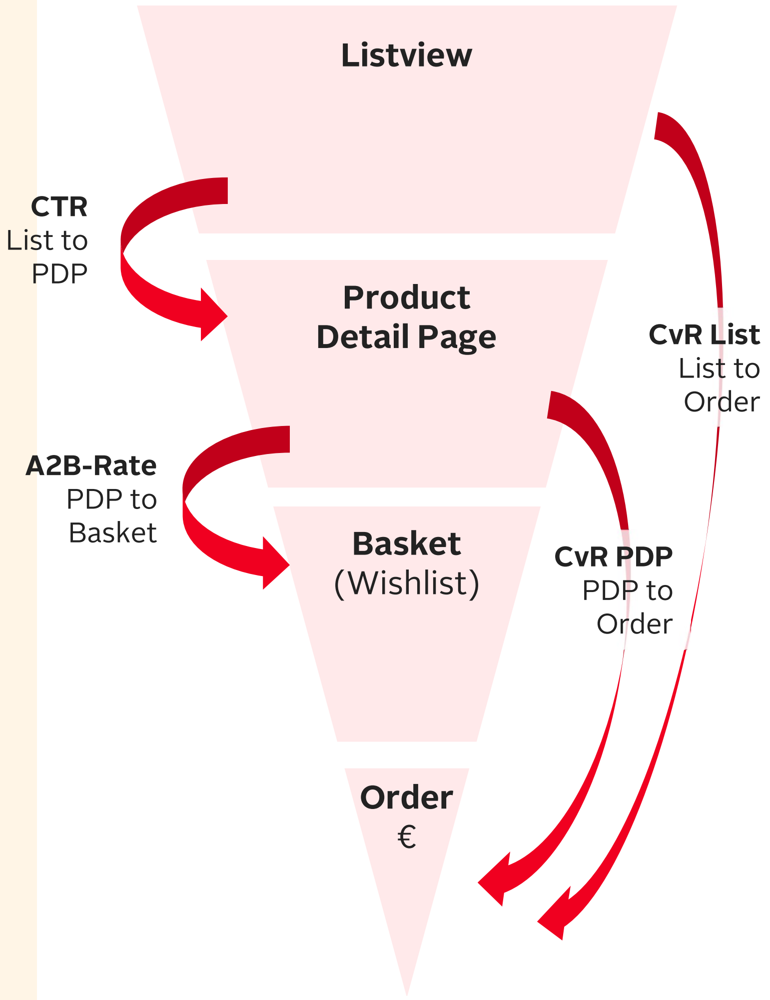
Training
Find patterns in the data and understand relationships between features and relevance of products

Product 1 is more relevant than product 2 given query x

Ranking of any given list of products

Trained ML Model
Can apply identified patterns to new data and deduce relevancy of unseen products/queries

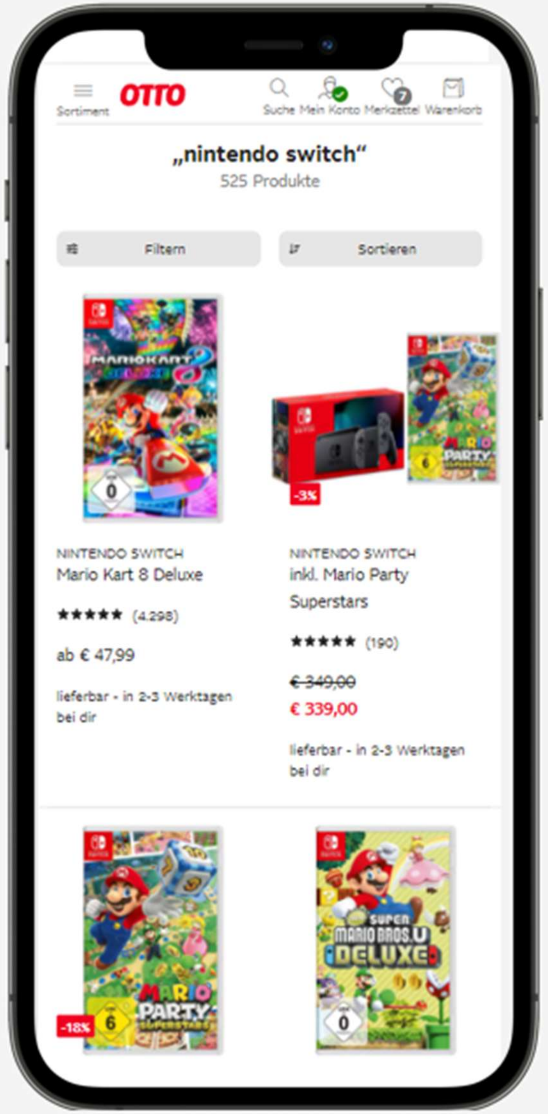
A Product with features like xyz is more relevant than a product with features abc given a query with features like 123



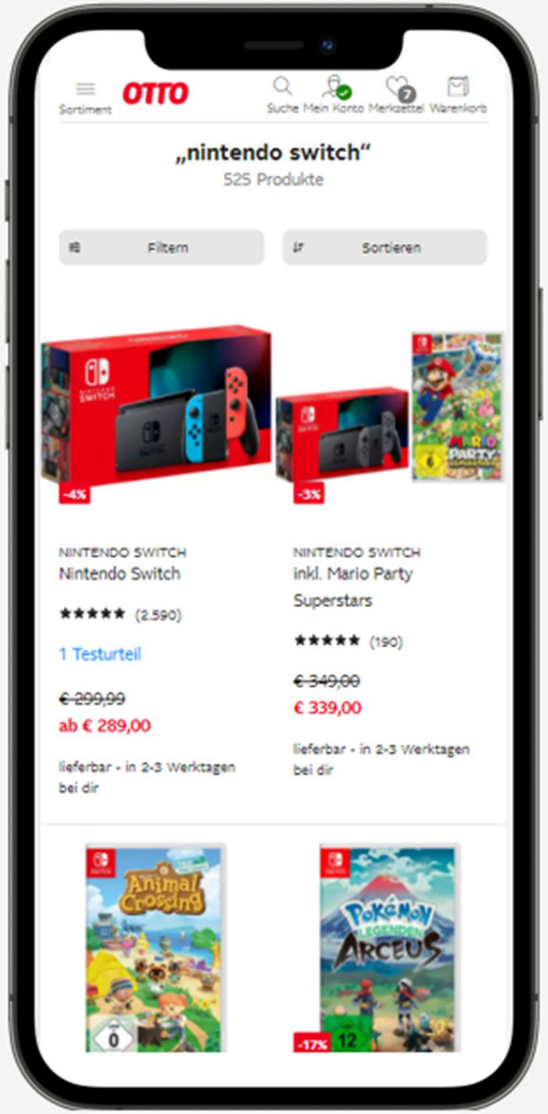
To **find** the **best ordering** of products, we **rank** our **shorthead** queries based on **customer interaction KPIs**

- 10 **onsite experiments** (A/B test)
- Shorthead: **~5.000 queries**
- Status Quo (A) against Judgements (B), reranked queries based on customer interaction KPIs
- Onsite experiments included Judgements ranked by:
  - **4x** on **CTR**
  - **1x** on **A2B-Rate**
  - **1x** on **CvR PDP**
  - **4x** on **CTR** and **CvR PDP**

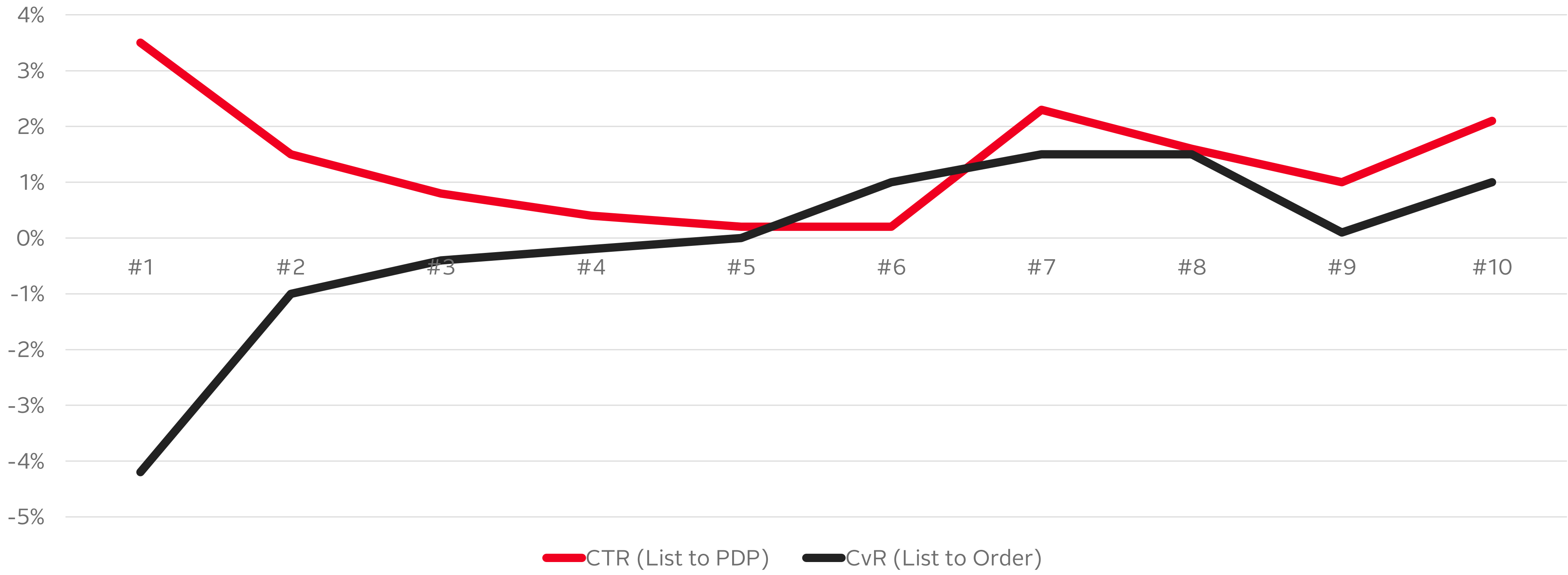
Group A
Status Quo



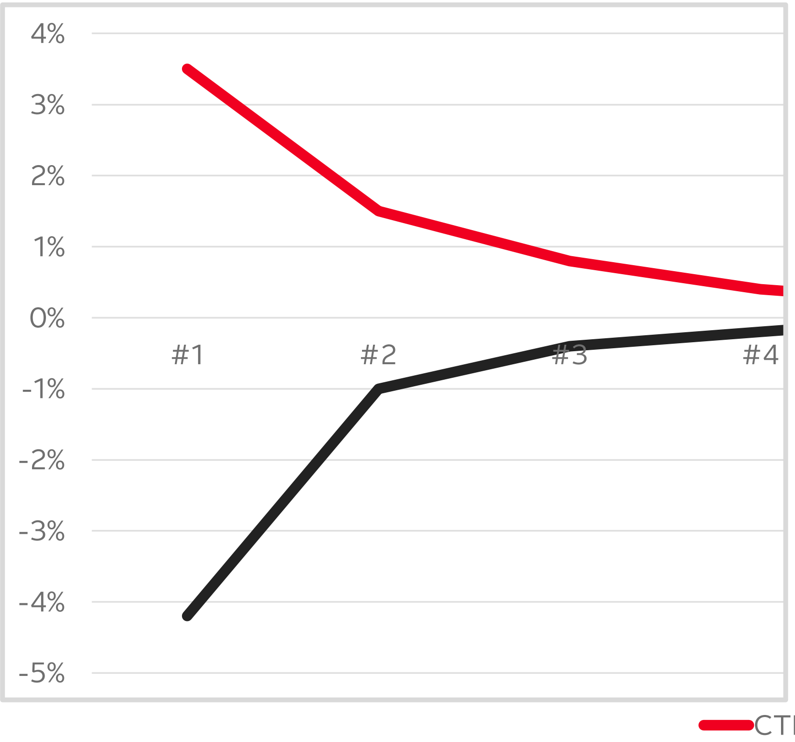
Group B
Judgement



Within 10 A/B tests we **iteratively improved** the **CTR** and **CvR** performance of our **Judgement** against the **Status Quo**

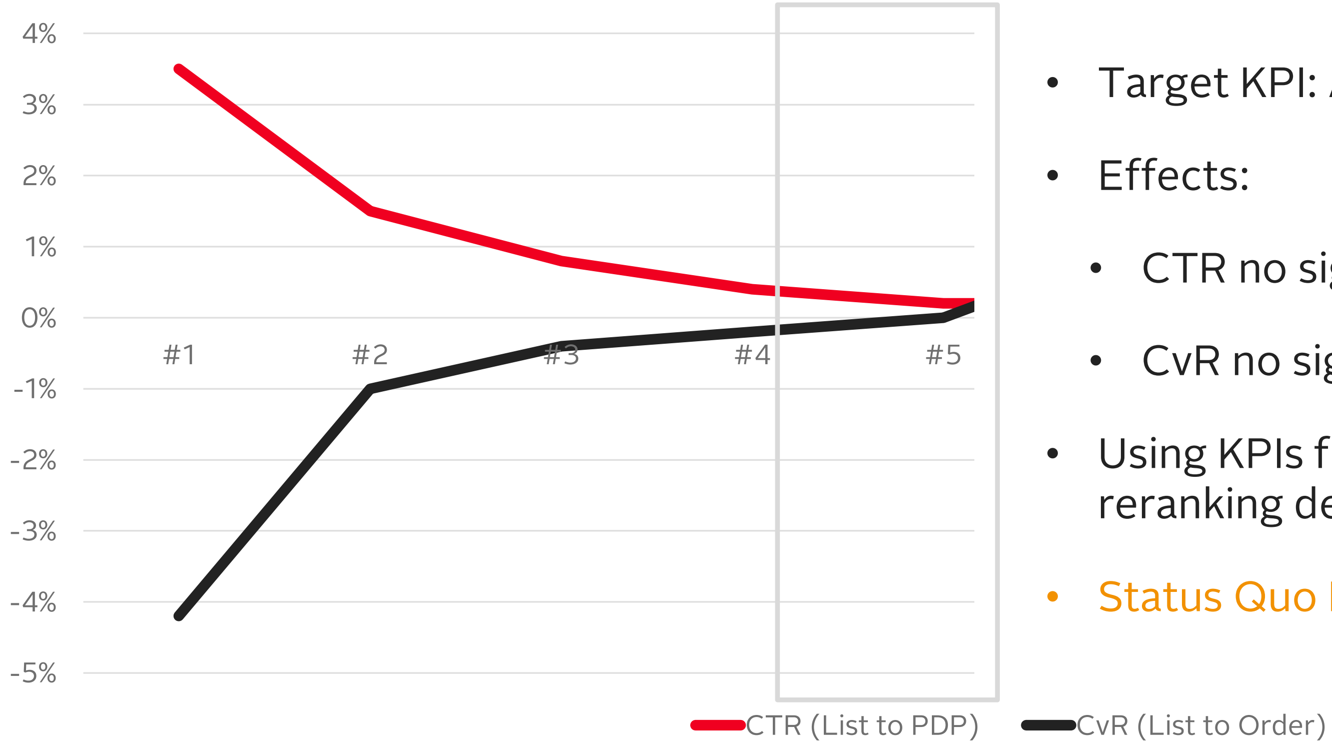


Our **first A/B tests** showed **great CTR uplifts** – yet **poor CvR performance**



- Target KPI: historic CTR of products per query
- ~5.000 queries reranked
- Effects
  - CTR significantly higher (+3.5%)
  - CvR significantly lower by (-4.2%)
- Usage of several filters (e.g. availability, minimum number of clicks per product) reduced positive and negative effects during experiment iterations

Reranking products by their **historic A2B-rate** showed **more promising results** – SQ ranking equalized

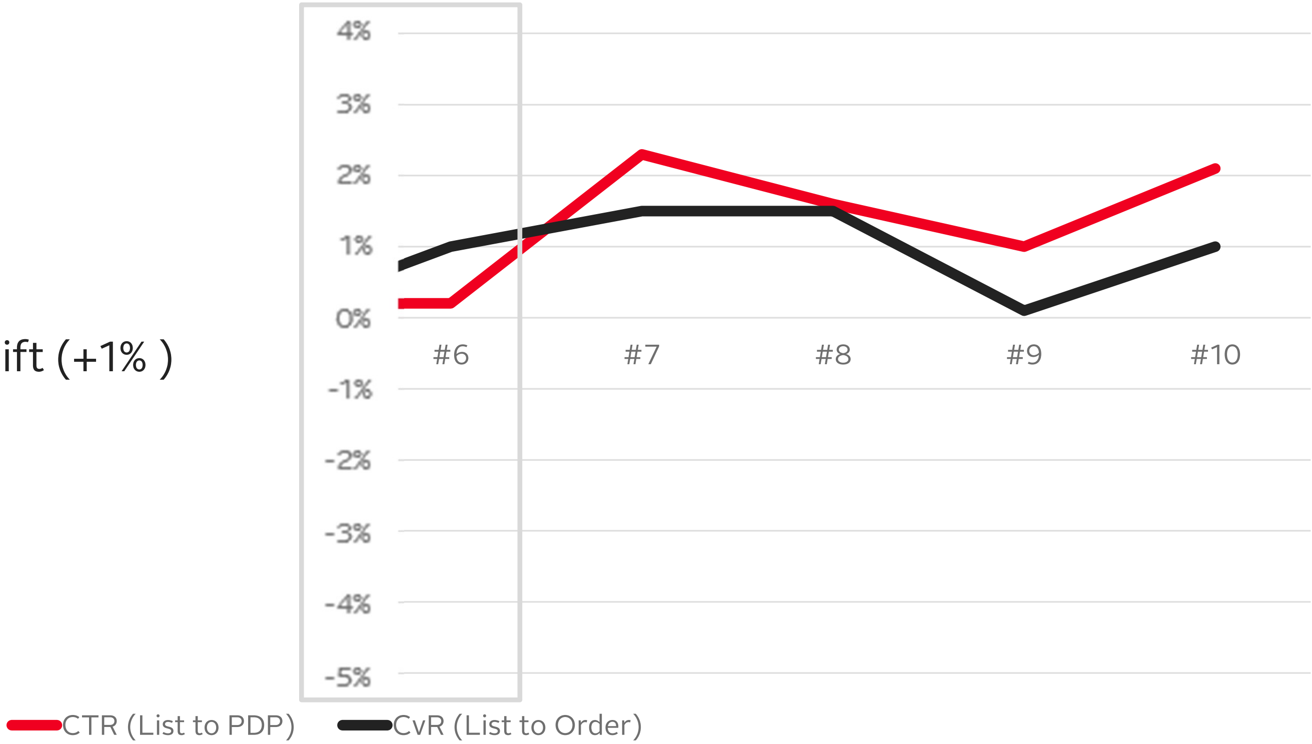


- Target KPI: A2B-rate (PDP to Basket)
- Effects:
  - CTR no significance (+0.2%)
  - CvR no significance (+0%)
- Using KPIs further down the search funnel for reranking delivers promising results
- Status Quo Ranking equalized, not yet beaten.

CvR based **target KPIs** lead to the **breakthrough** – Judgement significantly better than Status Quo

- Target KPI: CvR PDP to Order
- Effects:
  - CTR no significance (+0.2%)
  - CvR List to Order significant uplift (+1%)

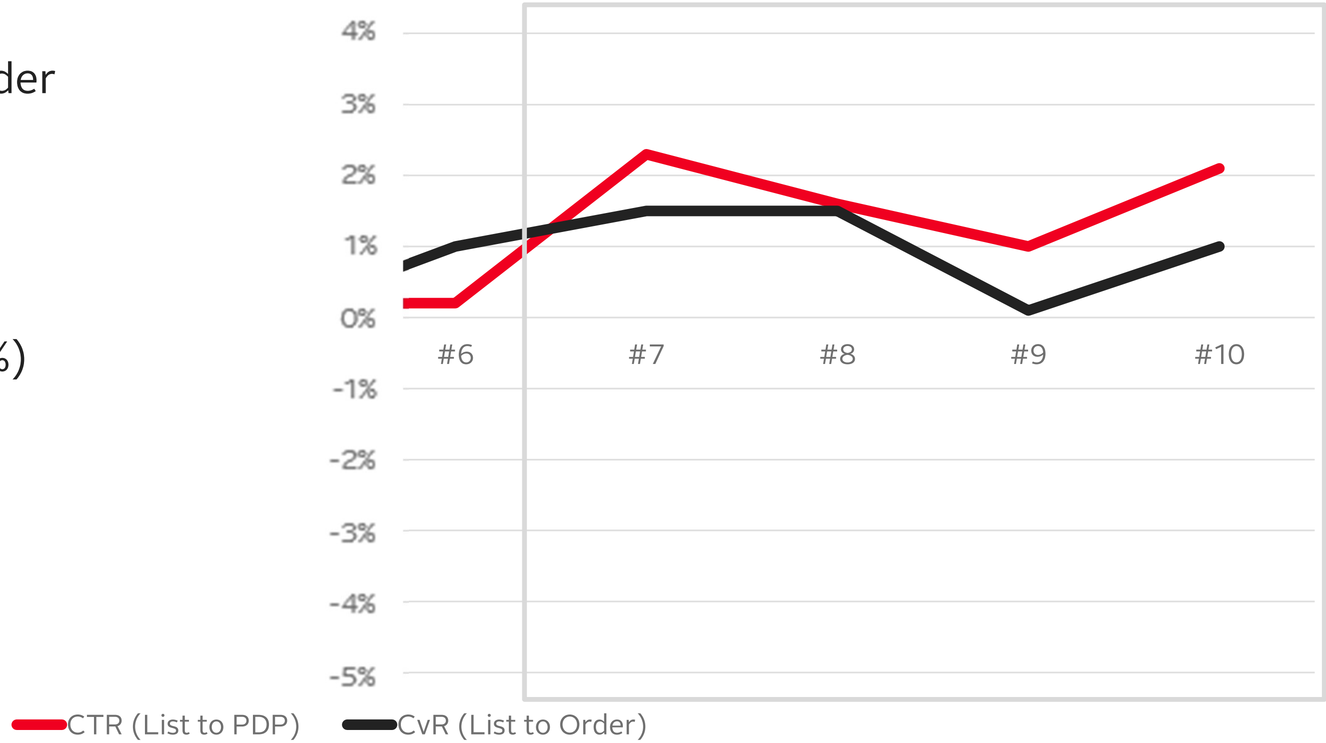
- Status Quo ranking beaten. ☺



CvR based **target KPIs** lead to the **breakthrough** – Judgement significantly better than Status Quo

- Target KPI: CTR x CvR PDP to Order
- Effects
  - CTR significant uplift (+2.1%)
  - CvR List significant uplift (+1.0%)

- Status Quo ranking beaten. ☺

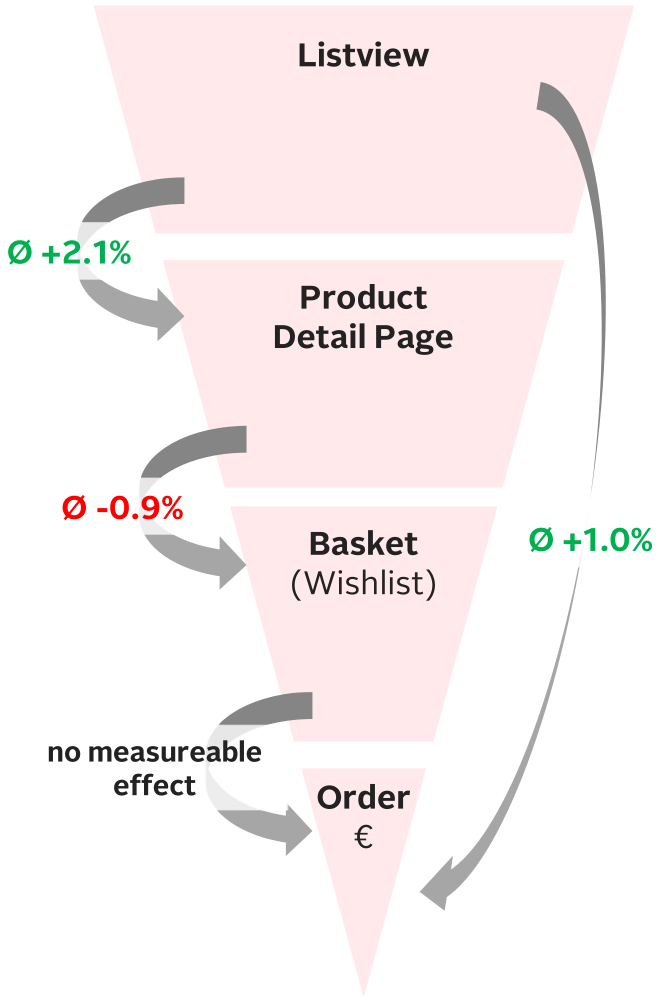


CvR based **target KPIs** lead to the **breakthrough** – Judgement significantly better than Status Quo

- Target KPI: CTR x CvR PDP to Order
- Effects
  - CTR significant uplift (+2.1%)
  - CvR List significant uplift (+1.0%)
- Status Quo ranking beaten. ☺

Combination of **CTR** and **CvR PDP** delivers best results

Status Quo ranking **beaten** 🎉



In detail: **strong CTR uplift** over-compensates a **weaker A2B-rate**

Search Funnel KPIs

- CTR uplift of Ø +2.1% *(95% confidence interval between +1.8 to 2.4%)*
- Increase in CTR compensates weaker A2B-rate
- Significant CvR improvement by Ø +1.0% *(95% confidence interval between +0.4 to +1.6%)*

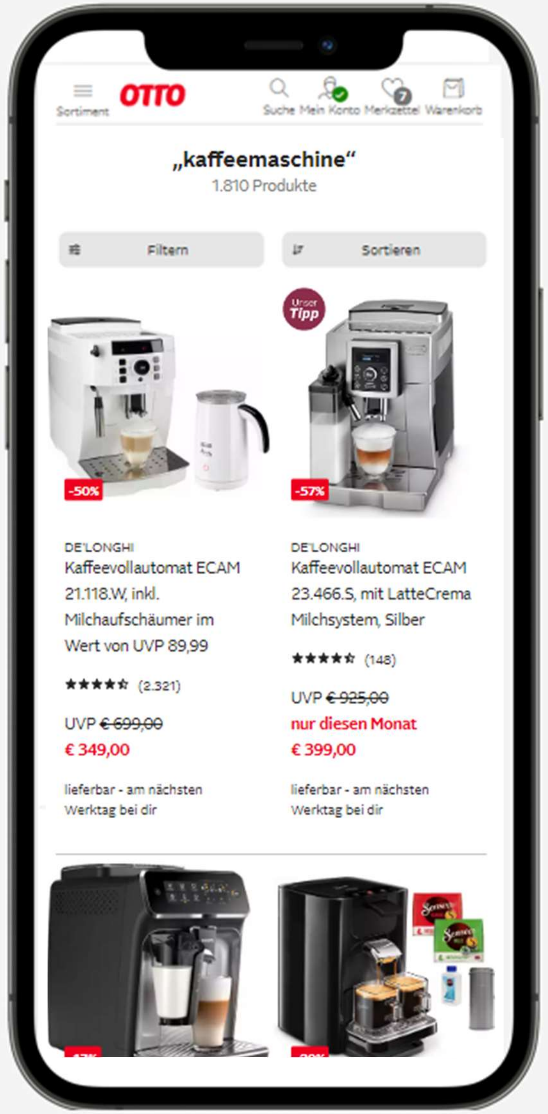
Ranking Change

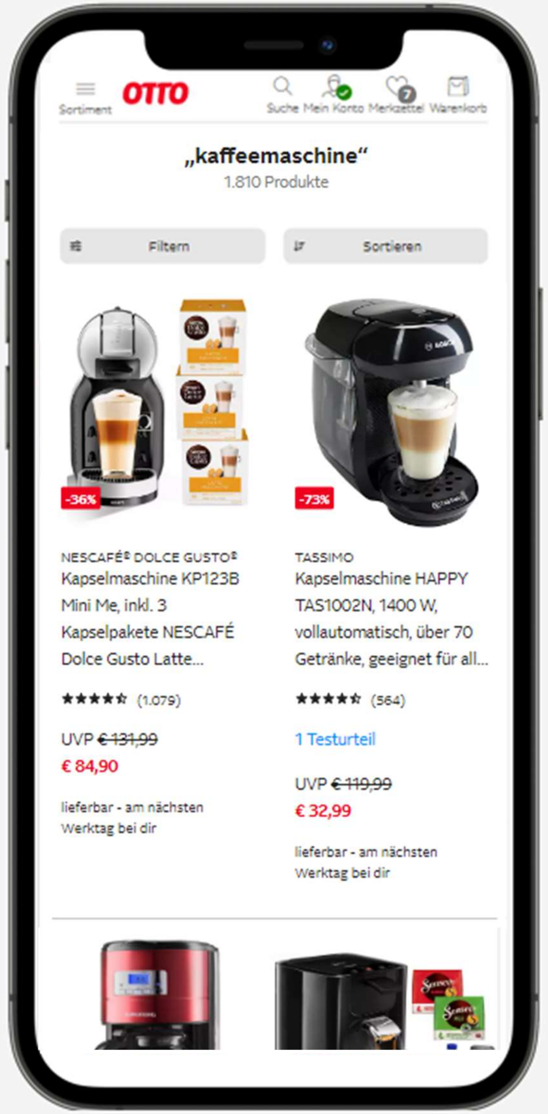
- cheaper products on top 20 positions (-6.7% average price)
- fewer customers
  - use facets (-0.8%)
  - change sorting (e.g. -2.5% sort price asending)

Economic KPIs

- more orders, yet lower AOV → no measurable effect on revenue

Group A
Status Quo





Group B
Judgement
CTR x CVR PDP

Conclusion: Judgement based on combination of funnel KPIs, namely *CTR* x *CVR (PDP to Order)*, improves ranking on shorthead

- **CTR** and **CvR List** significantly improved on ~5.000 queries
- Target KPIs that include **conversions** seem to work best for eCom ranking optimization

- **Next Steps:**
  - Go-live permanently with Judgements until LTR Model can transfer the uplift potential to all queries (Shorthead, Body, Longtail)
  - Use Judgements as training target for LTR models, also selectively test new Judgement ideas (e.g. sort by CvR List or Orders)
  - Iterate further on Models and API-infrastructure (e.g., scalability, ABC-test setup)

Thanks!