A live model for product type mapping

Michael Hardegger
27. 1. 2020

Contributions by Elvis Murina, Beate Sick and Helmut Grabner, all from ZHAW

About Digitec Galaxus AG

Founded as a start-up in 2001, with two of the founders still present

1’14 Mio.
Revenues 2019

32 Years
Average Age of Employees

16%
Growth 2019

> 1300 Employees
Logistics, Category Mangagement, Customer Service, etc.

> 150 Software Engineers
Focus on Online Shop and ERP

> 1.5 Mio.
Active Customers

5 ML Engineers
Team founded in 2019

> 3 Mio Products
~100’000 on stock

Part 1:
ML Use Cases at Digitec Galaxus AG

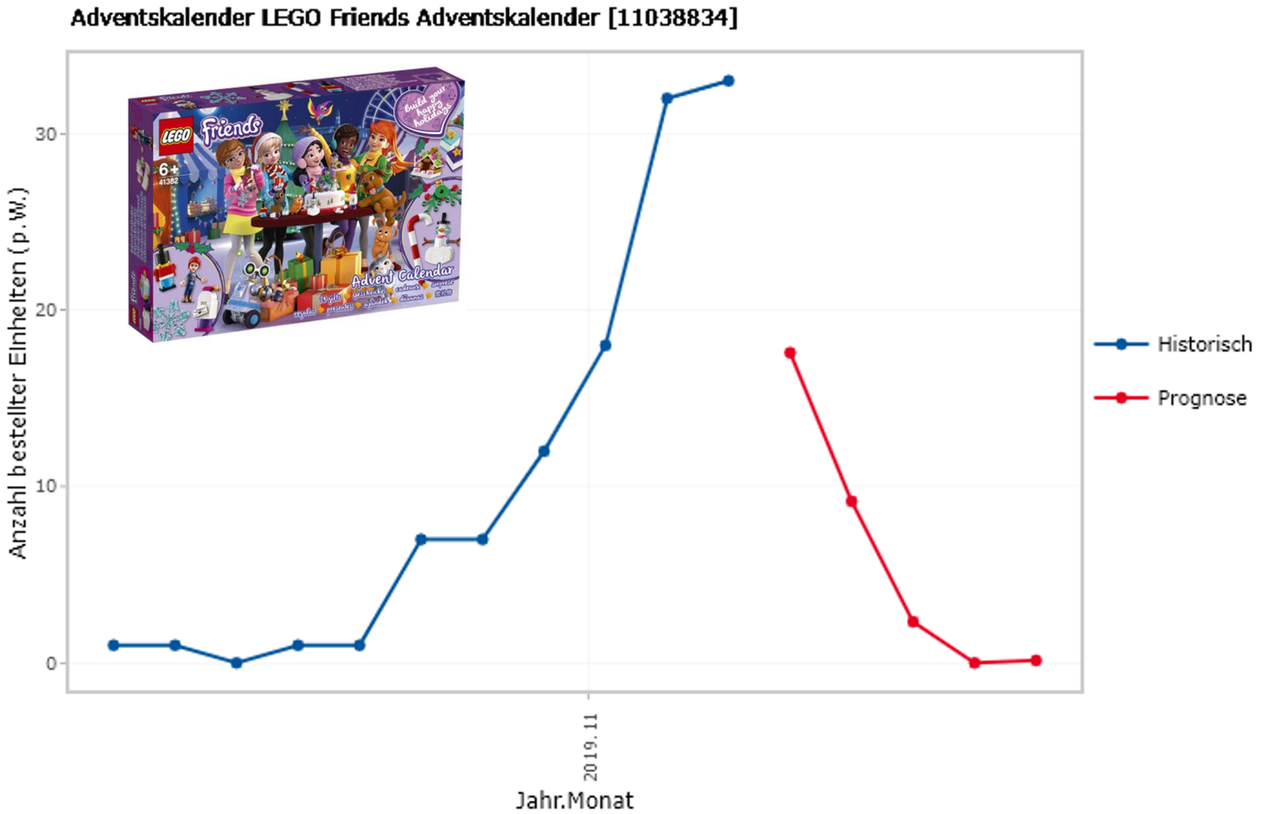
A customer’s journey through Galaxus’ ML world

Initial demand → *Finding the product* → *Is the price okay?* → *Checkout* → *Waiting for delivery* → *Complaining with Customer Service*

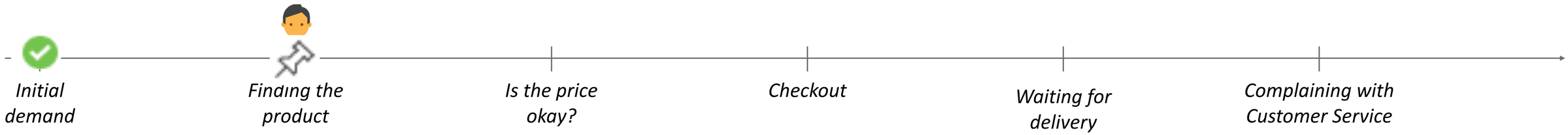
To fulfill this demand, we have to have the products on stock!

80% of the articles we have on stock were bought based on demand forecast models!

Method:
Time-Series-Models
+ ML-Stacker



A customer’s journey through Galaxus’ ML world



Oft zusammen gekauft mit

175.83
Bowflex Selecttech

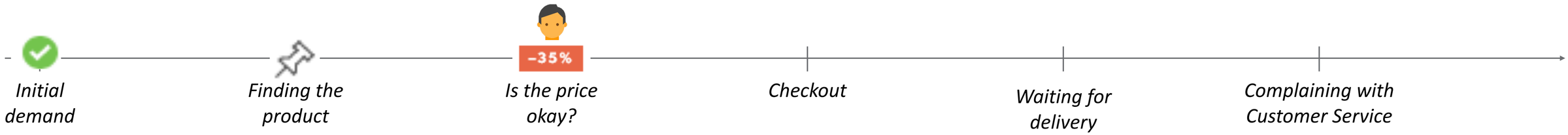
225.05 statt 280.–
Bowflex 4.1 Bench

21.26
Tunturi Fitnessmatte (15mm)

Recommenders speed up navigation and make customers aware of unconscious demands.

Models in productions:
- Meta Prod2Vec
- Affinity Profiles
- Many heuristic models

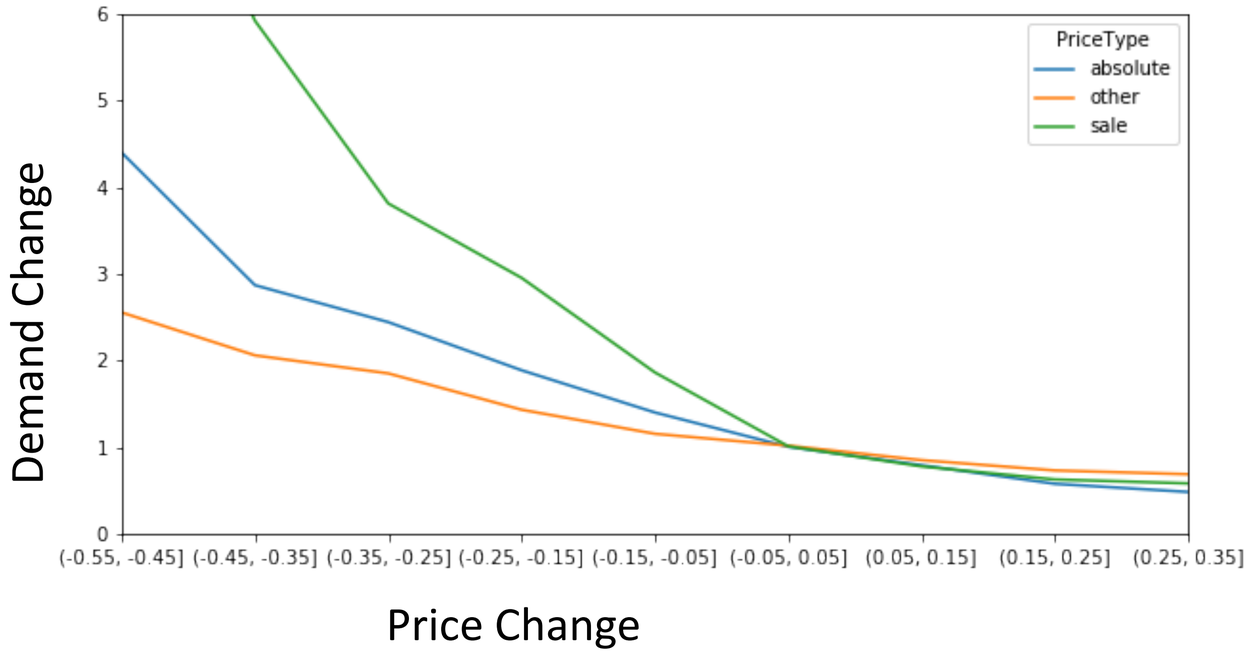
A customer’s journey through Galaxus’ ML world



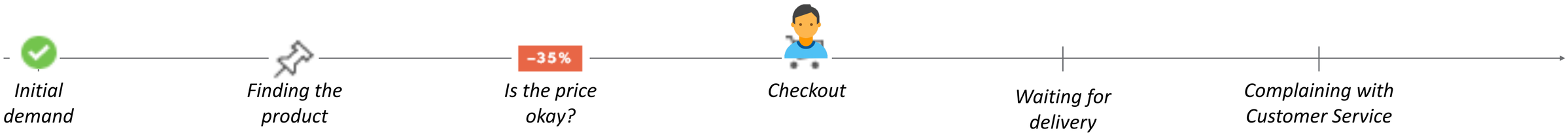
-35%

187.69 statt 289.–

ML-based price elasticity models tell us by how much we need to reduce a product to generate sales.

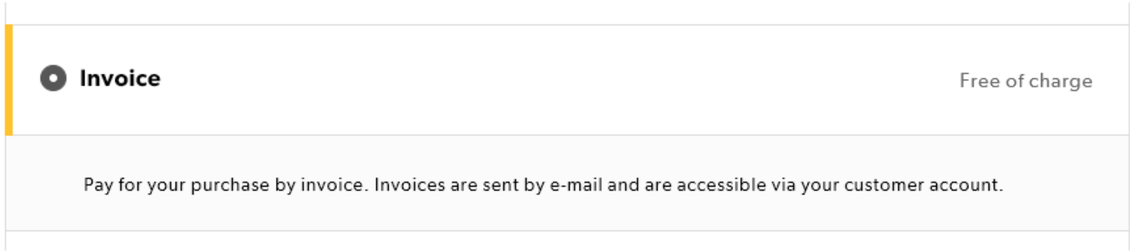


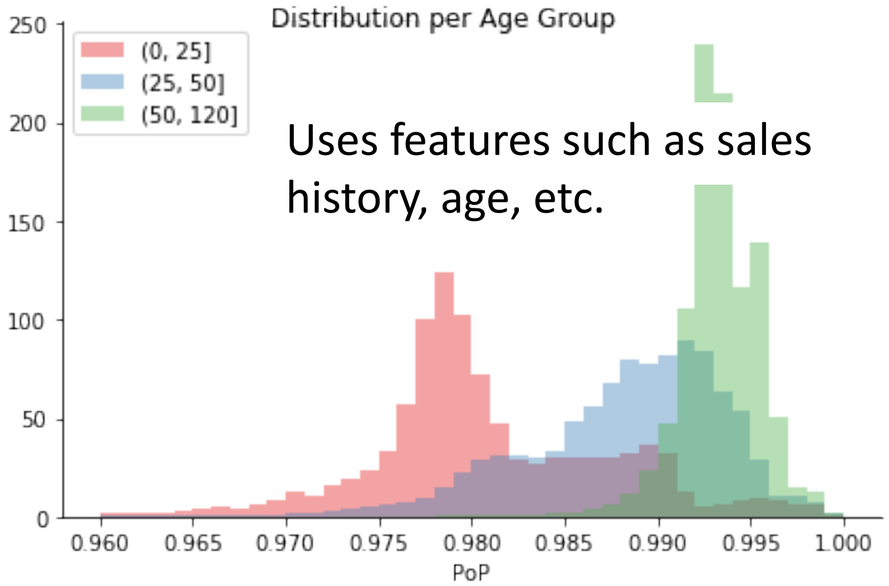
A customer’s journey through Galaxus’ ML world



ML-based credit rating

«Probability of Payment» learned from previous orders and defaults





Uses features such as sales history, age, etc.

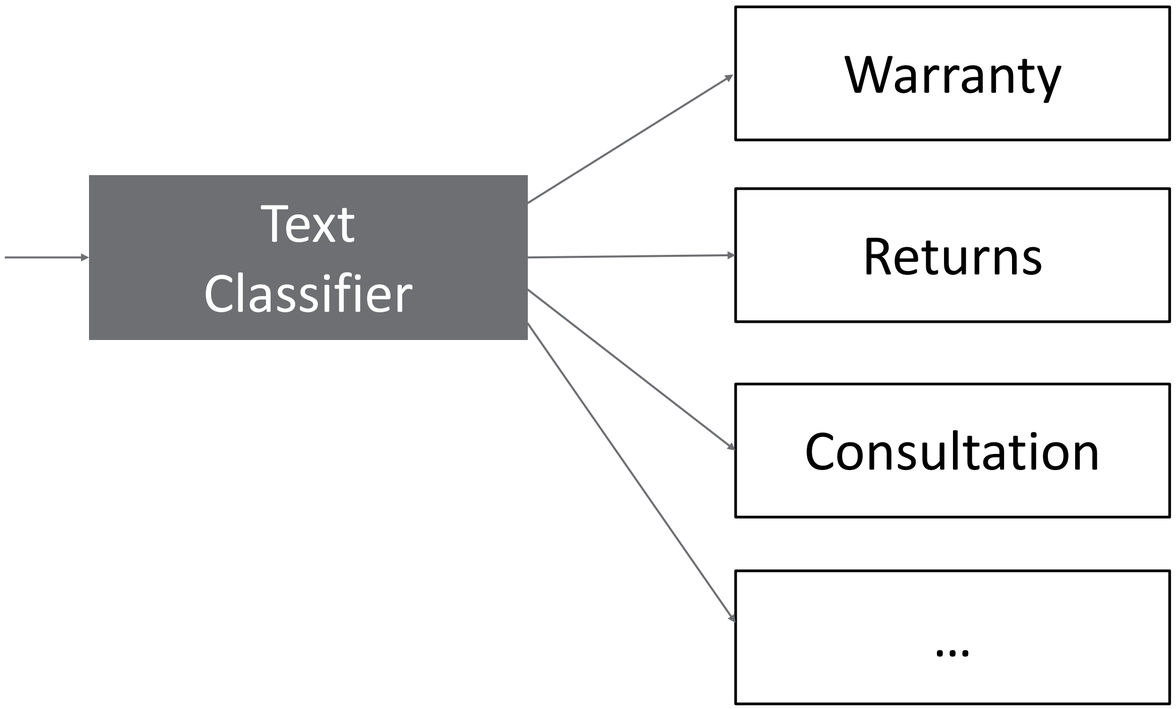
A customer’s journey through Galaxus’ ML world

Some logistics use cases:

- Predicting the delivery date
- Process Mining
- Placement of articles in the warehouse
- …

A customer’s journey through Galaxus’ ML world

«Dear Digitec,
I realized that the dumbbells are just too heavy for me.
Could I return them tomorrow?
Cheers!»

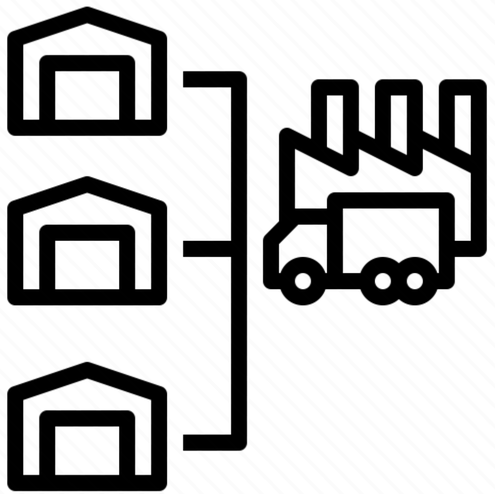


Part 2:
Product Type Mapping

Where do we have our product data from?

Product data quality is super important for us …

… but we are not really in control of it



- >100 suppliers send us data about new products
- Data formats and quality vary widely
  - Some suppliers don’t care enough about us to improve data quality
  - Others are not able to do it
- As a result, dozens of employees clean up data manually
- Even that is just barely enough to handle bestsellers, the long tail of products is never touched

Data Cleaning Process

Provider Record

Image:

Name: Flux S Smart T2900

Brand: Takx

Category: ...;Sport;Radsport;Zubehör

Product Description:

«Der Flux S Smart hat alles, was Sie sich von einem Smart-Trainer mit ...»

Product Properties

{Weight: 21 kg, Connectivity: [BT, ANT, ANT+, ...], Max. Resistance: 1.5kW, ...}

What arrives at our interfaces ...

... and how it ends up in the shop

Brand/ProductType/Property Mapping Process

...
Apple
Nike
Tacx
HP
...

...
Rollentrainer
Hockeyschläger
Spiele
Notebooks
...

ProductType

Rollentrainer

623.– Brand

Tacx Flux S Smart T2900

★★★★½ 8

Der Flux S Smart hat alles, was Sie sich von einem Smart-Trainer mit Direktantrieb wünschen. Zuverlässig, genau, geräuscharm, leistungsstark, und er vermittelt ein gutes Fahrgefühl...

Die wichtigsten Spezifikationen auf einen Blick

Max. Widerstand	1500 W
Gewicht	21 kg
Rollenbremsentyp	Elektromagnetischer Widerstand
Konnektivität	Bluetooth Smart, ANT+, ANT+ FE-C

↓ Mehr anzeigen

Property &Value

Data Cleaning Process

Provider Record

Image:

Name: Flux S Smart T2900

Brand: Takx

Category: ...;Sport;Radsport;Zubehör

Product Description:

«Der Flux S Smart hat alles, was Sie sich von einem Smart-Trainer mit ...»

Product Properties

{Weight: 21 kg, Connectivity: [BT, ANT, ANT+, ...], Max. Resistance: 1.5kW, ...}

«Product Type Mapping»

ProductType

Rollentrainer

623.–

Tacx Flux S Smart T2900

★★★★½ 8

Der Flux S Smart hat alles, was Sie sich von einem Smart-Trainer mit Direktantrieb wünschen. Zuverlässig, genau, geräuscharm, leistungsstark, und er vermittelt ein gutes Fahrgefühl...

Die wichtigsten Spezifikationen auf einen Blick

Max. Widerstand	1500 W
Gewicht	21 kg
Rollenbremsentyp	Elektromagnetischer Widerstand
Konnektivität	Bluetooth Smart, ANT+, ANT+ FE-C

↓ Mehr anzeigen

Naive Approach to PT-Mapping

Provider Record

Image:

Name: Flux S Smart T2900

Brand: Takx

Category: ...;Sport;Radsport;Zubehör

Product Description:

«Der Flux S Smart hat alles, was Sie sich von einem Smart-Trainer mit …»

Product Properties

{Weight: 21 kg, Connectivity: [BT, ANT, ANT+, …], Max. Resistance: 1.5kW, …}

ML-Model

ProductType

Rollentrainer

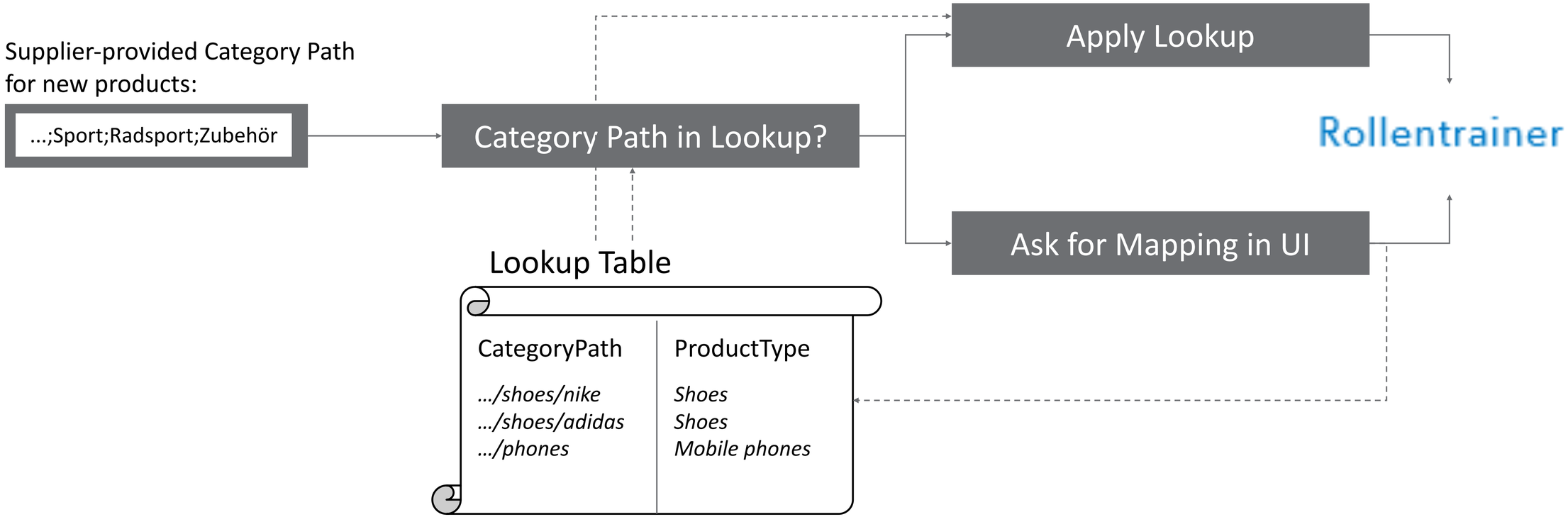
4’280’656 Products → 2’298 Product Types

Results:

Accuracy: 95% in test set (random splitting)
Most important features:

- «Category» provided by supplier
- InsertDate

Current Business Process explains the Result



The ML-Model just learned this lookup table!
It’s task is to help mapping new category paths

Think about the Train-Test-Split
We shouldn’t use category paths as a feature

Challenges with the PT-Mapping Problem

This is a really «dirty» problem!

- Many incorrect labels
- Very imbalanced classes (and features)
- Lots of NAs
- Mappings change over time (new PTs, PTs are combined, etc.)
- Evaluation / Train-Test-split by date of mapping
  - ➢ Test distribution can be very different from training
  - ➢ However, what we care about is real-life performance

Deep Neural Network for Product Type Mapping

Provider Record

Image:

Name: Flux S Smart T2900

Brand: Takx

Category: ...;Sport;Radsport;Zubehör

Product Description:

«Der Flux S Smart hat alles, was Sie sich von einem Smart-Trainer mit …»

Product Properties

{Weight: 21 kg, Connectivity: [BT, ANT, ANT+, …], Max. Resistance: 1.5kW, …}

CNN-Layer

Embedding

Embedding + CNN

OHE/Dense

Dense

Product Type Proposals:

Product Type	Score
Rollentrainer	0.73
Fahrradreifen	0.11
Fahrrad	0.07
Sportuhr	0.02
Notebook	0.01
…	…

Deep Neural Network + Stringmatching

Provider Record

Image:

Name: Flux S Smart T2900

Brand: Takx

Category: ...;Sport;Radsport;Zubehör

Product Description:

«Der Flux S Smart hat alles, was Sie sich von einem Smart-Trainer mit …»

Product Properties

{Weight: 21 kg, Connectivity: [BT, ANT, ANT+, …], Max. Resistance: 1.5kW, …}

...
Rollentrainer
Hockeyschläger
Spiele
Notebooks
...

String Matching

CNN-Layer

Embedding

Embedding + CNN

OHE/Dense

Dense

Merge Results

Product Type Proposals:

Product Type	Score
Rollentrainer	0.77
Fahrradreifen	0.11
Fahrrad	0.07
Sportuhr	0.00
Notebook	0.00
...	...

Controller for Assigning Product Type Mappings: High Confidence

CategoryPath provided by supplier

Proposed Product Type

KategoriePfad	Akzeptieren	Zuweisen	Ignorieren	Kategorien
Ausrüstung;Pflege & Hygiene;Schuhpflege;Schuhpflege;Schuhpflege	◉	○	○	Schuhpflege
...n;Baby;Babypflege & Bad;Babybadewannen & Zubehör;Badespielzeug	◉	○	○	Badewannenspielzeug
...pielwaren;Baby;Schlafen & Nuggi;Spieluhren & Nachtlichter;Spieluhren	◉	○	○	Spieluhr
Baby & Spielwaren;Baby;Spielsachen;Babyspielzeug;Holzspielzeug	◉	○	○	Bauklötze + Stapeln
...aren;Kinderfahrzeuge;Laufrad & Kinderfahrrad;Laufrad & Kinderfahrrad	◉	○	○	Kindervelo
...by & Spielwaren;Spiele & Puzzle;Spiele;Zubehör Spiele;Zubehör Spiele	◉	○	○	Sammelaufbewahrung
...aren;Spieltische & Spielwelten;Multimedia-Spielwaren;Roboter;Roboter	◉	○	○	Robotik Kit
Bekleidung;Hosen;Ski- & Snowboardhosen	◉	○	○	Skihose
...ng;Gewerbe- und Industrieleuchten;Bürostehleuchten;Bürostehleuchten	◉	○	○	Stehlampe
...telligente Beleuchtung;Intelligente Beleuchtung;Intelligente Beleuchtung	◉	○	○	Leuchtmittel

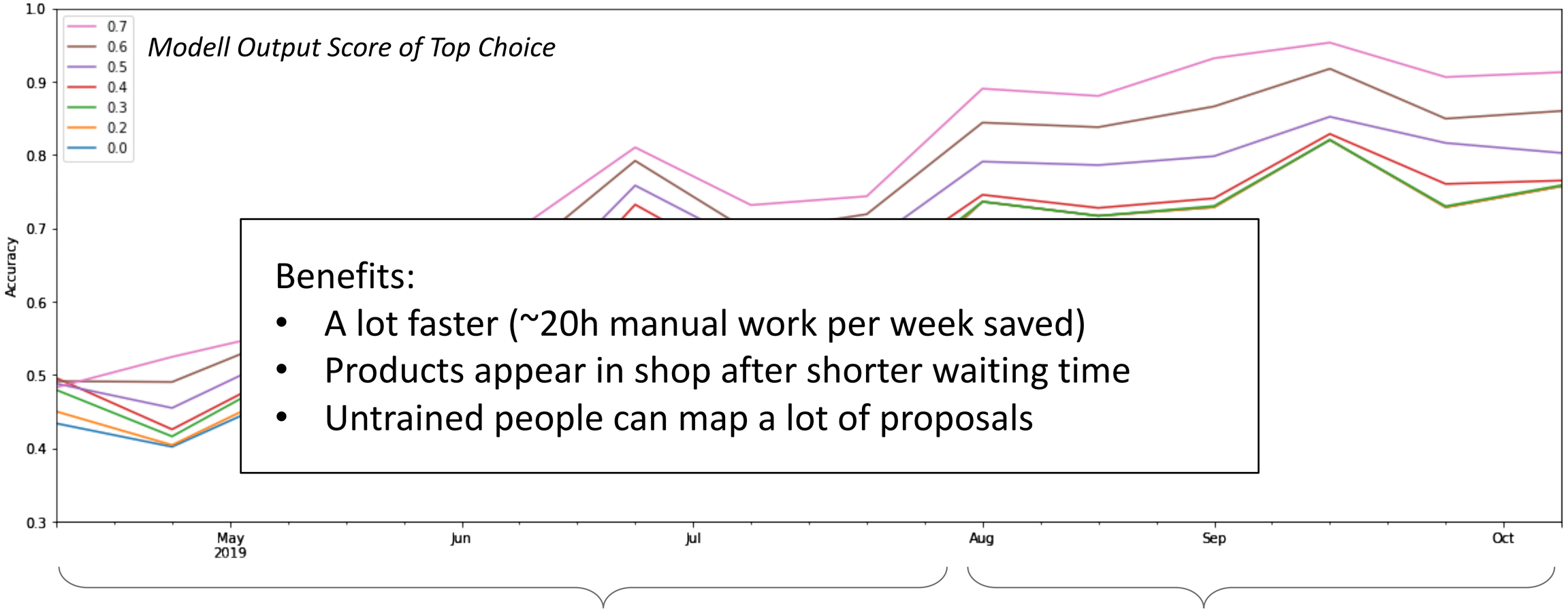
Controller for Assigning Product Type Mappings: Lower Confidence – Multiple Choice

CategoryPath provided by supplier

Proposed Product Type

Live Performance

Top 5-Accuracy: One of the five top proposals was used as mapping

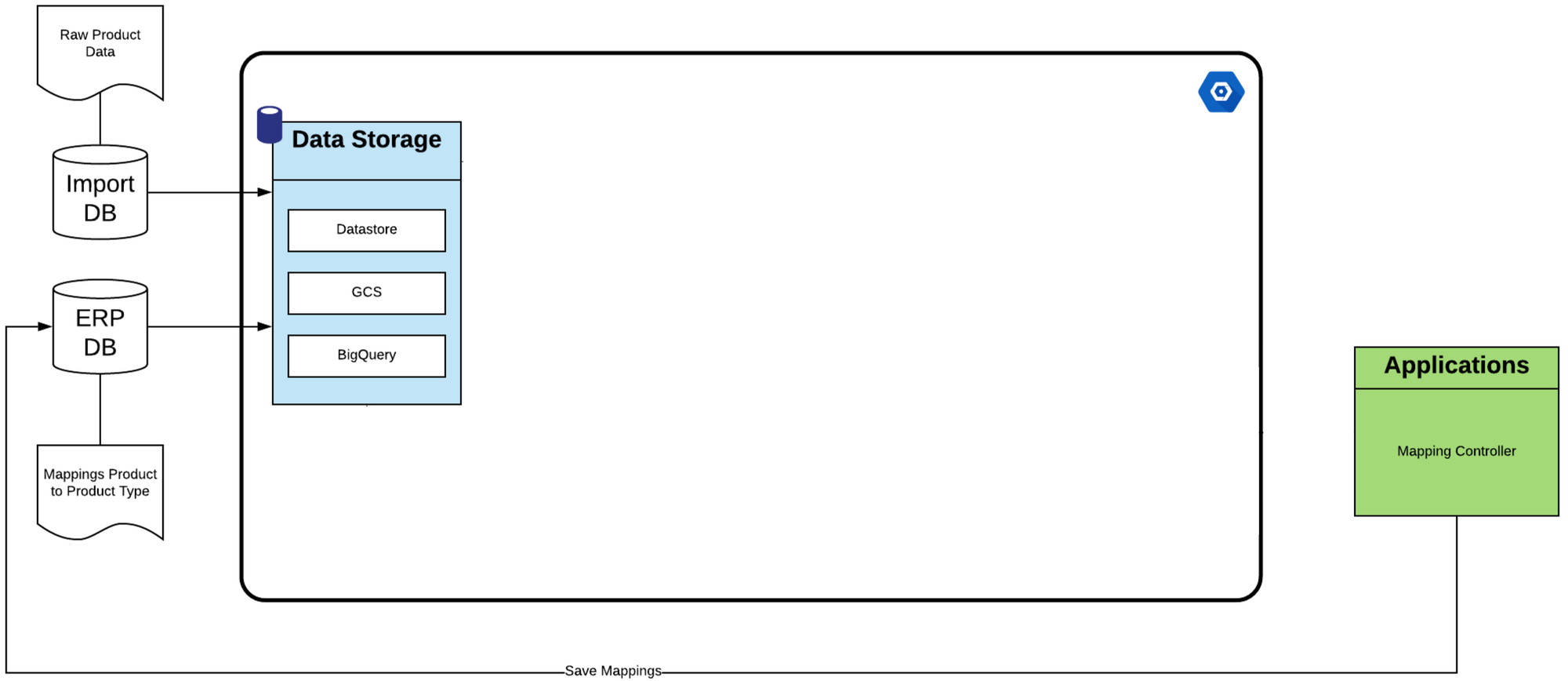


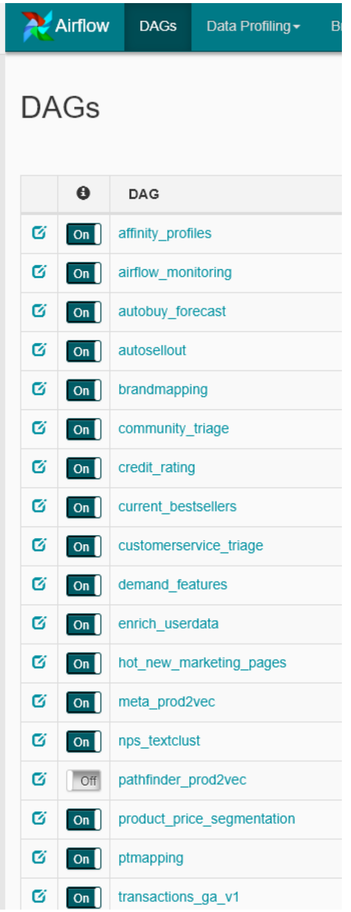
Benefits:

- A lot faster (~20h manual work per week saved)
- Products appear in shop after shorter waiting time
- Untrained people can map a lot of proposals

TF-IDF+XGB → (2nd model iteration) CNN

An overview of our ML module





Conclusions

Take home messages

- Online retailers are a great playing field for data scientists
  (if you have management support and a good infrastructure)
- Expect to work on really messy data sets
- Take your time to understand the business processes well in order to find out how ML can bring the most benefit
- By bringing solutions into production quickly, you learn fast and can iteratively improve

Appendix

A few findings from the modelling efforts

Performances in Test-Set
(Split by date; without string matching)

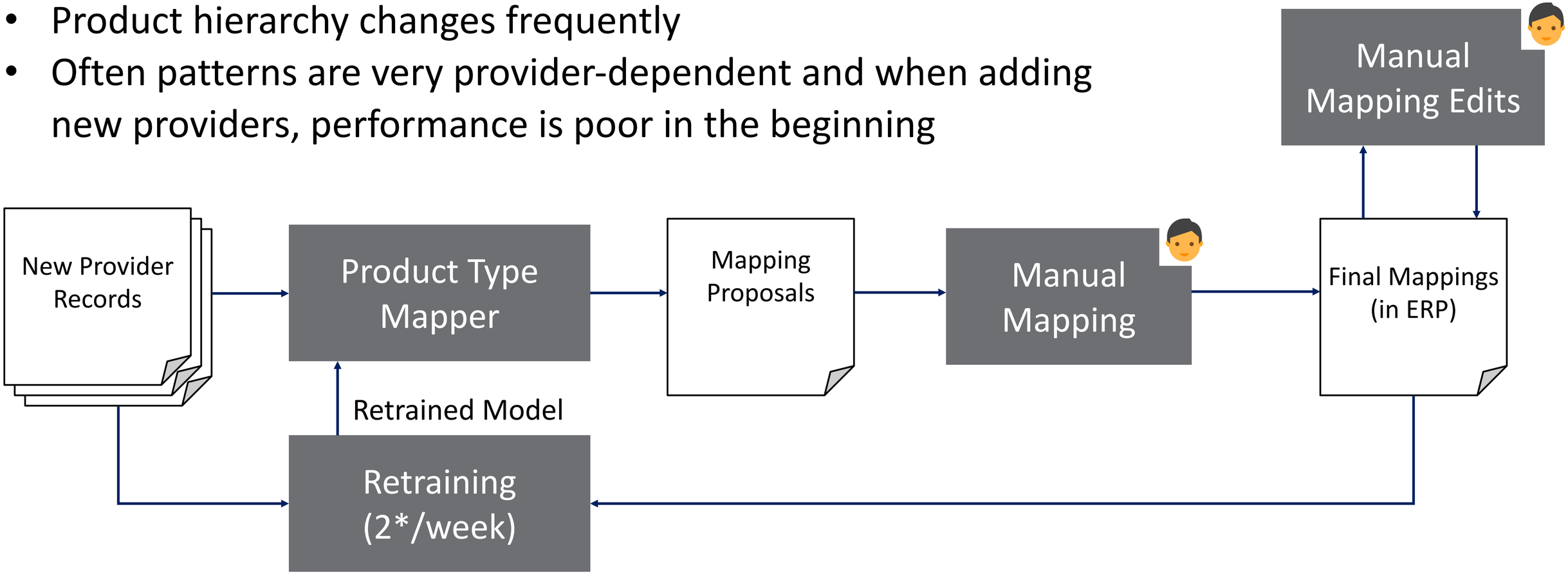
Modell	Top5-Accuracy
Text+Image	73.6%
Text+Img+Meta	72.3%
Text (stemming, …)	66.4%
Text (raw)	63.9%
Baseline (TFIdf+XGB)	51.5%
Metadata only	36.3%

- DNNs outperform baseline model
  - Memory-limitations for XGB (2300 values for each leaf)!
- Image+Text resulted in the best performances
- Other meta data did not improve quality
- Text Model:
  - LSTM or CNN did not make a difference
  - Preprocessing (stemming, etc.) reduced model complexity and improved performance
- Weekly Retraining required in production:
  - Product hierarchy changes frequently

Retraining

Required because:

- Product hierarchy changes frequently
- Often patterns are very provider-dependent and when adding new providers, performance is poor in the beginning



Live Performance

KPIs	Value
#Mappings/Week	~400
Top1-Accuracy	75.2%
Top5-Accuracy	91.3%

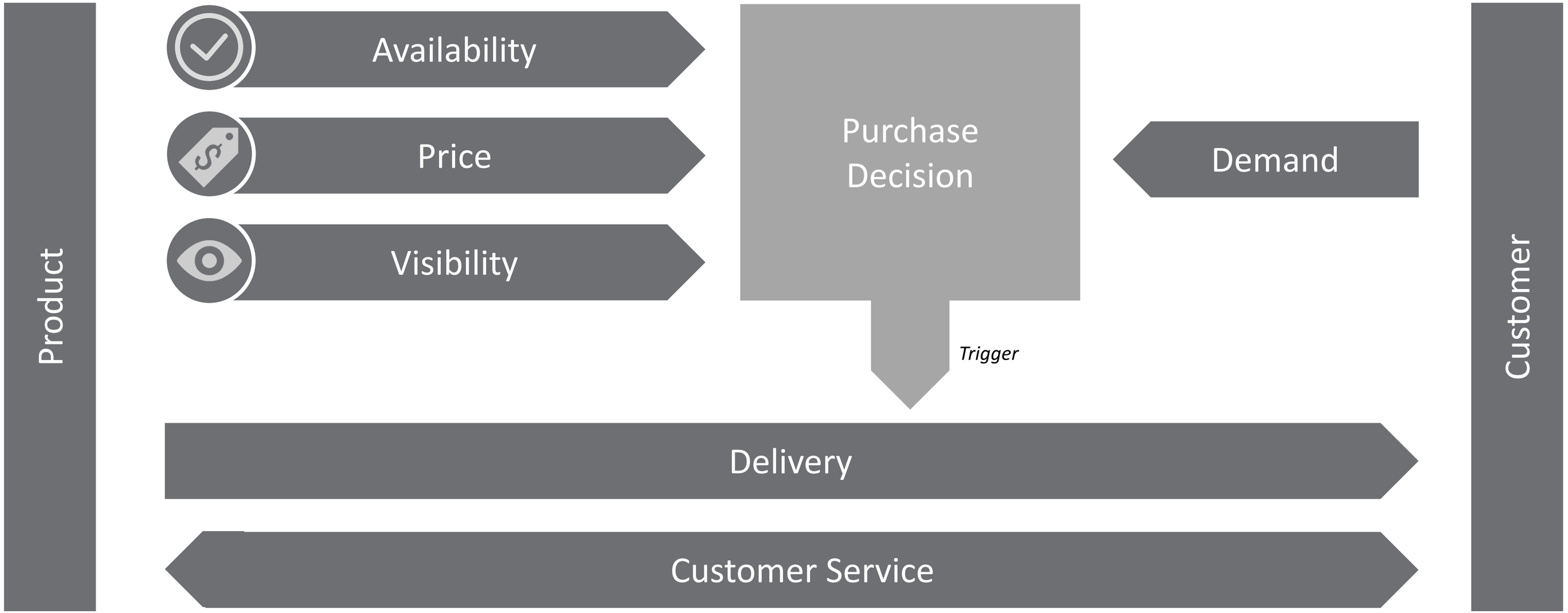
Benefits:

- A lot faster (before, 20-30h per week manual work)
- Products appear in shop after shorter waiting time
- Untrained people can map most proposals
- Full automation for cases with high score possible

Text preprocessing

	Step	Preprocessed Text
1	Raw text	Der Flux S Smart T2900 hat alles, was Sie sich von einem Smart-Trainer mit …
2	Lower case and remove stopwords	flux smart t2900 hat alles smart-trainer …
3	Replace string patterns with tokens	flux smart LETTERNUMBERTOKEN hat alles smart trainer …
4	Stemming	flux smart LETTERNUMBERTOKEN hat all smart train …
5	Tokenizing	171 5231 879 7 66 5231 9 …
6	Input to embedding layer (token to vector)	

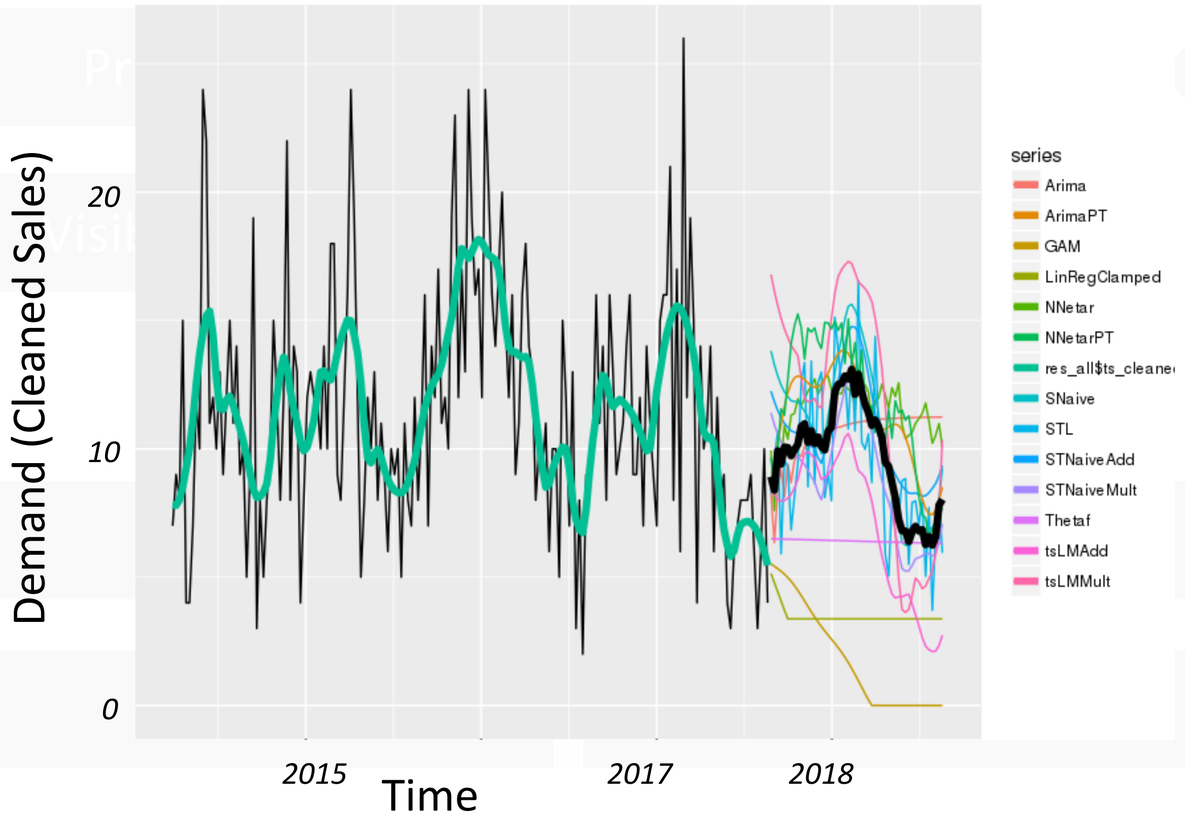
The core processes of an online shop



Machine Learning at Digitec Galaxus

Availability

Make sure we have the right products on stock



Machine Learning at Digitec Galaxus

Availability

Price

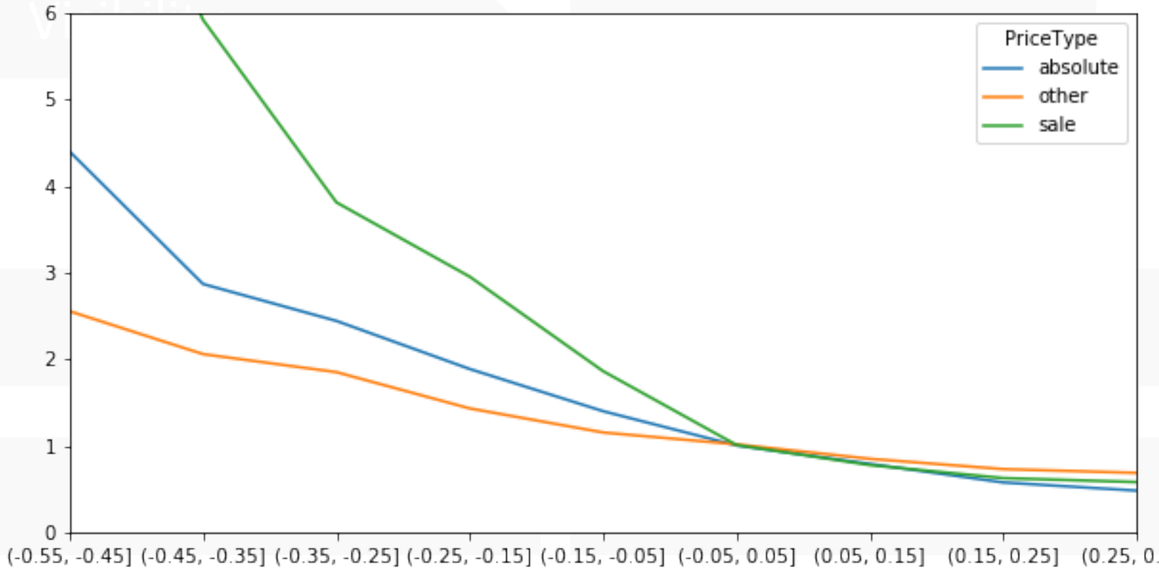
Make sure we have a competitive price while not loosing money …

Demand

Product

Customer

Price Elasticity Curves



Demand Change

Price Change

Machine Learning at Digitec Galaxus

Availability

Price

Visibility

Purchase Decision

Demand

Product

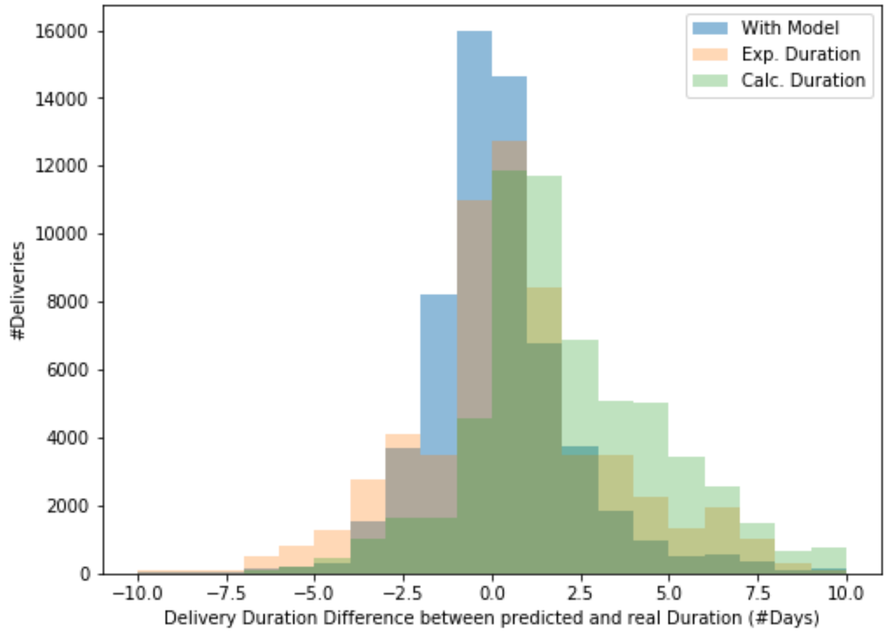
Customer

Make sure customers find our offers!

Recommenders:

Machine Learning at Digitec Galaxus

When will a delivery arrive at the customers home?



Product

Price

Visibility

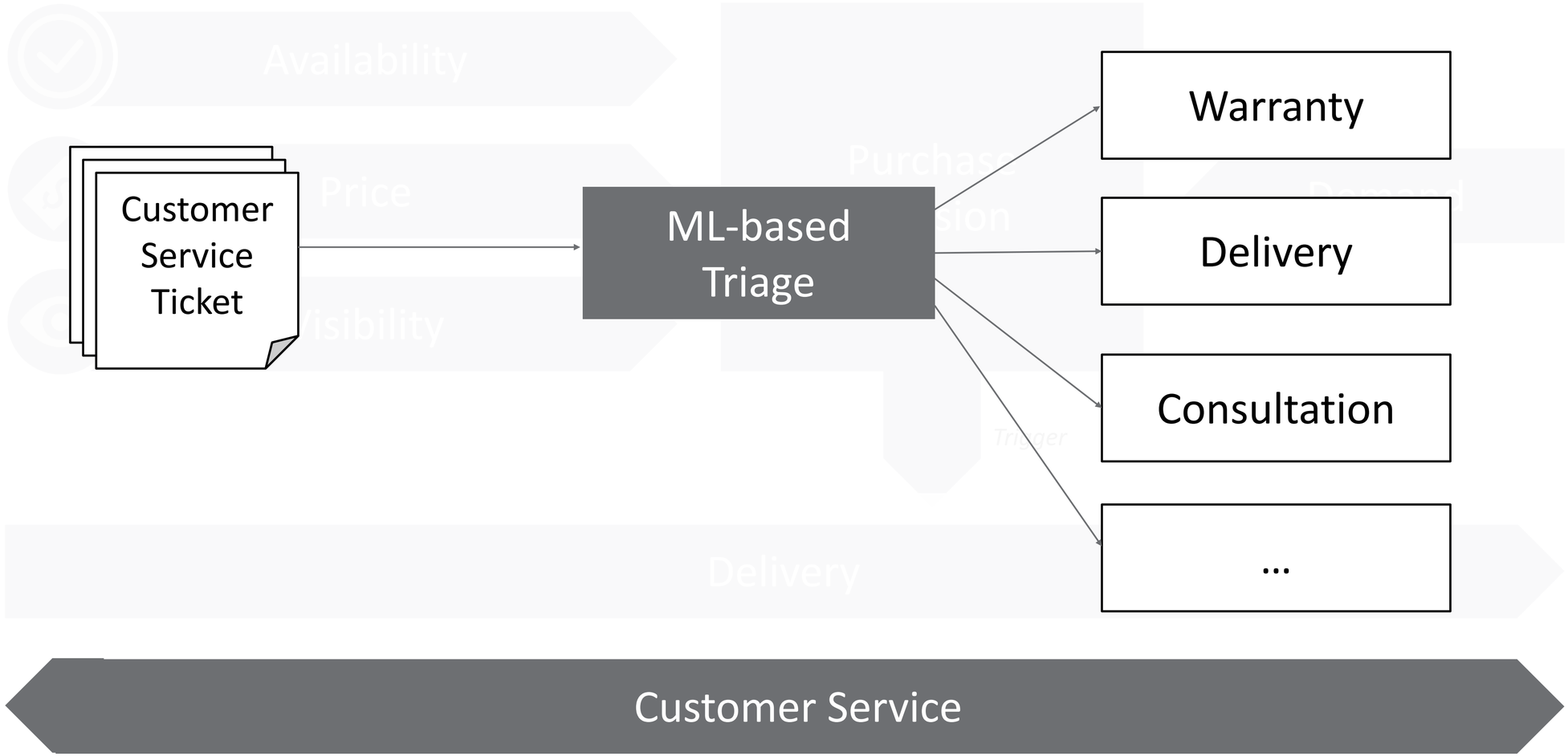
Demand

Customer

Delivery

Customer Service

Machine Learning at Digitec Galaxus



… what we'll be talking about next …