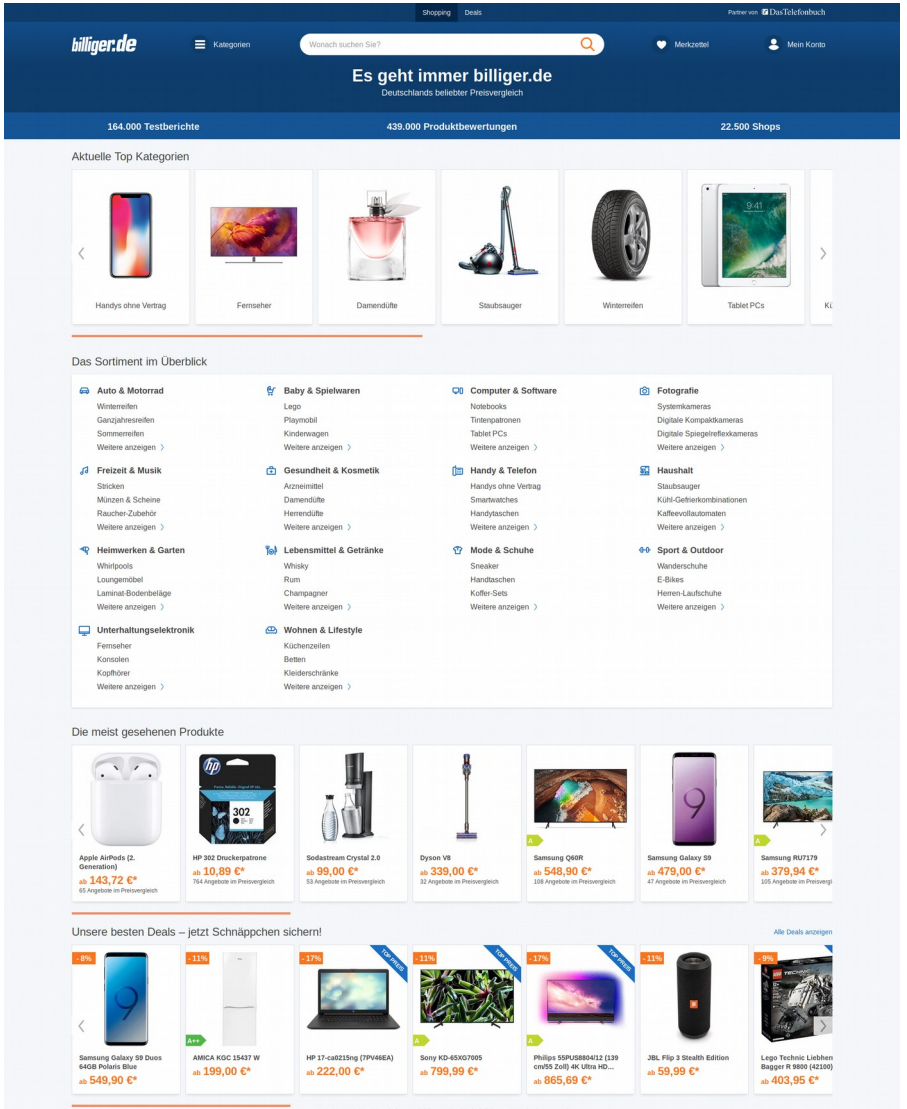
10 Years of Automated Category Classification for Product Data

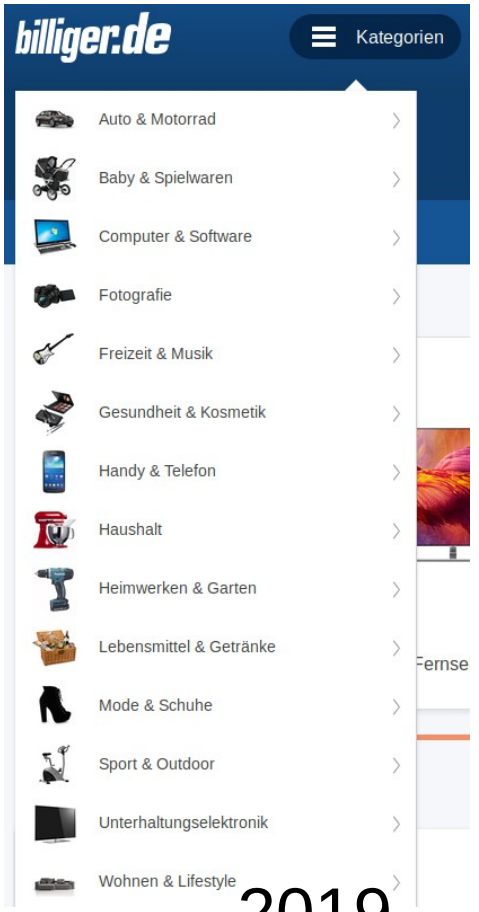
Johannes Knopp
jkn@solute.de

solute

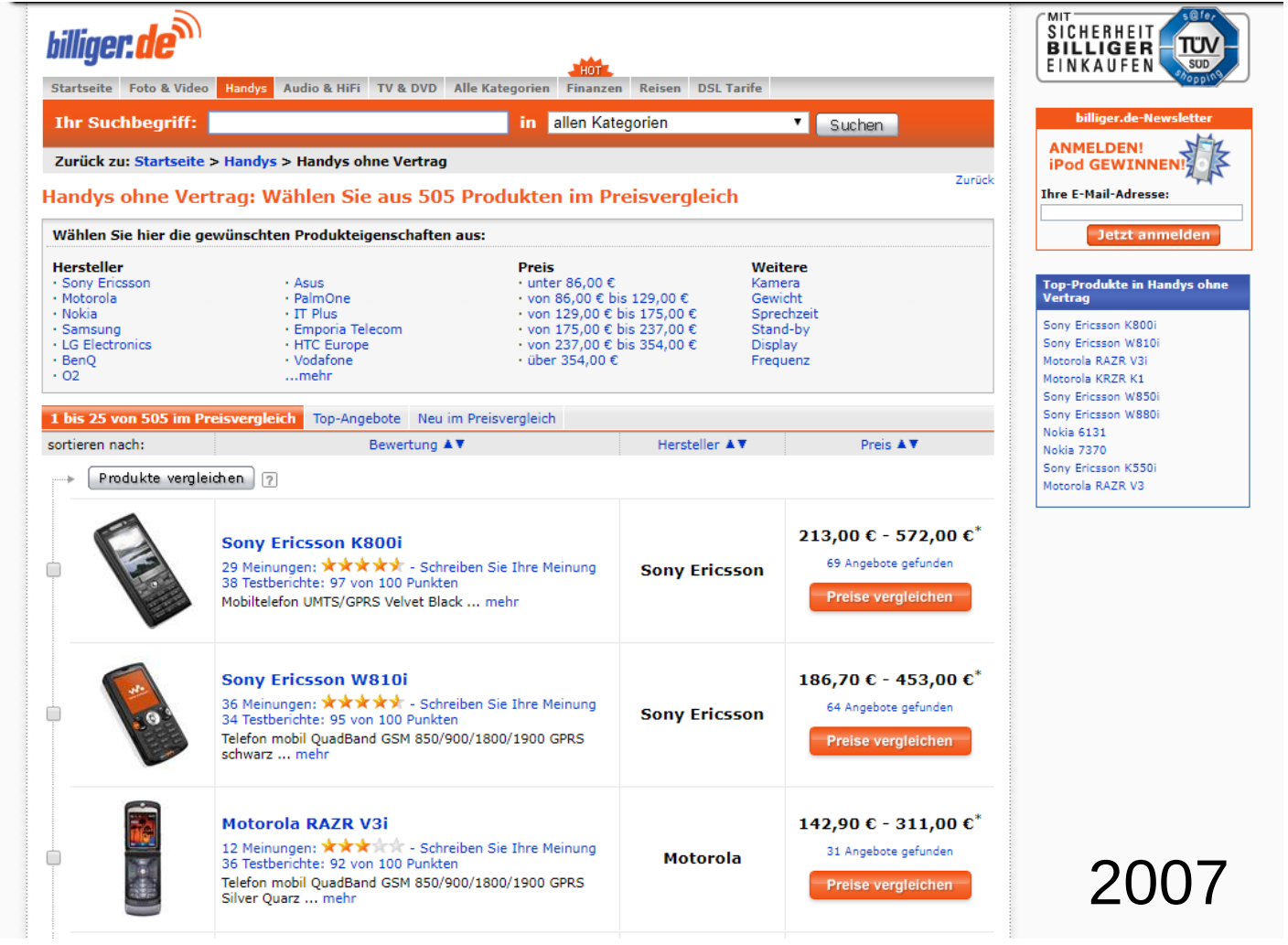
Founded in 2004

billiger.de SHOPPING.DE FRIENDS COMMUNICATION


billiger.de
Kategorien
Wonach suchen Sie?
Merkzettel
Mein Konto
Es geht immer billiger.de
Deutschlands beliebter Preisvergleich
164.000 Testberichte
439.000 Produktbewertungen
22.500 Shops
Aktuelle Top Kategorien
Handys ohne Vertrag
Fernseher
Damendüfte
Staubsauger
Winterreifen
Tablet PCs
Das Sortiment im Überblick
Auto & Motorrad
Wintereifen
Ganzjahresreifen
Sommerreifen
Baby & Spielwaren
Lego
Playmobil
Kinderwagen
Computer & Software
Notebooks
Tintenpatronen
Tablet PCs
Fotografie
Systemkameras
Digitale Kompaktkameras
Digitale Spiegelreflexkameras
Freizeit & Musik
Stricken
Münzen & Scheine
Raucher-Zubehör
Gesundheit & Kosmetik
Arzneimittel
Damendüfte
Herrendüfte
Handy & Telefon
Handys ohne Vertrag
Smartwatches
Handytaschen
Haushalt
Staubsauger
Kühl-Gefrierkombinationen
Kaffeevollautomaten
Heimwerken & Garten
Whirlpools
Loungemöbel
Laminat-Bodenbeläge
Lebensmittel & Getränke
Whisky
Rum
Champagner
Mode & Schuhe
Sneaker
Handtaschen
Koffer-Sets
Sport & Outdoor
Wanderschuhe
E-Bikes
Herren-Laufschuhe
Unterhaltungselektronik
Fernseher
Konsolen
Kopfhörer
Wohnen & Lifestyle
Küchenzeilen
Betten
Kleiderschränke
Weitere anzeigen
Die meist gesehenen Produkte
Apple AirPods (2. Generation)
ab 143,72 €*
HP 302 Druckerpatrone
ab 10,89 €*
Sodastream Crystal 2.0
ab 99,00 €*
Dyson V8
ab 339,00 €*
Samsung Q60R
ab 548,90 €*
Samsung Galaxy S9
ab 479,00 €*
Samsung RU7179
ab 379,94 €*
Unsere besten Deals – jetzt Schnäppchen sichern!
Alle Deals anzeigen
Samsung Galaxy S9 Duos 64GB Polaris Blue
ab 549,90 €*
AMICA KGC 15437 W
ab 199,00 €*
HP 17-ca0215ng (7PV46EA)
ab 222,00 €*
Sony KD-65XG7005
ab 799,99 €*
Philips 55PUS8804/12 (139 cm/55 Zoll) 4K Ultra HD...
ab 865,69 €*
JBL Flip 3 Stealth Edition
ab 59,99 €*
Lego Technic Liebherr Bagger R 9800 (42100)
ab 403,95 €*
Prozentuale Ersparnis im Vergleich zum durchschnittlichen Bestpreis der Vorwoche.

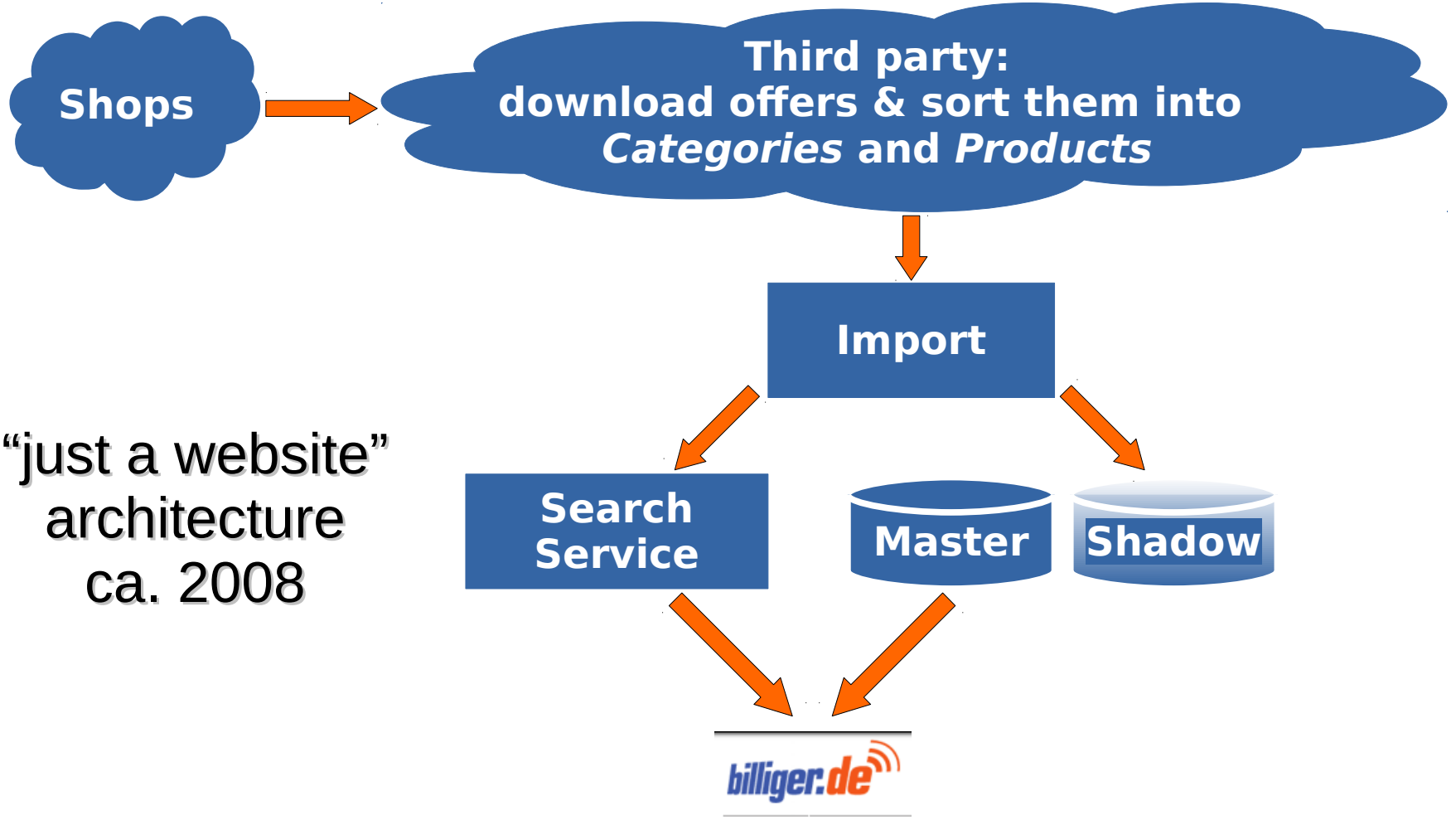

billiger.de
Kategorien
Auto & Motorrad
Baby & Spielwaren
Computer & Software
Fotografie
Freizeit & Musik
Gesundheit & Kosmetik
Handy & Telefon
Haushalt
Heimwerken & Garten
Lebensmittel & Getränke
Mode & Schuhe
Sport & Outdoor
Unterhaltungselektronik
Wohnen & Lifestyle

2019


billiger.de
Startseite
Foto & Video
Handys
Audio & HiFi
TV & DVD
Alle Kategorien
HOT
Finanzen
Reisen
DSL Tarife
Ihr Suchbegriff:
in
allen Kategorien
Suchen
Zurück zu: Startseite > Handys > Handys ohne Vertrag
Zurück
Handys ohne Vertrag: Wählen Sie aus 505 Produkten im Preisvergleich
Wählen Sie hier die gewünschten Produkteigenschaften aus:
Hersteller
· Sony Ericsson
· Motorola
· Nokia
· Samsung
· LG Electronics
· BenQ
· O2
· Asus
· PalmOne
· IT Plus
· Emporia Telecom
· HTC Europe
· Vodafone
...mehr
Preis
· unter 86,00 €
· von 86,00 € bis 129,00 €
· von 129,00 € bis 175,00 €
· von 175,00 € bis 237,00 €
· von 237,00 € bis 354,00 €
· über 354,00 €
Weitere
Kamera
Gewicht
Sprechzeit
Stand-by
Display
Frequenz
1 bis 25 von 505 im Preisvergleich
Top-Angebote
Neu im Preisvergleich
sortieren nach:
Bewertung
Hersteller
Preis
Produkte vergleichen
Sony Ericsson K800i
29 Meinungen: - Schreiben Sie Ihre Meinung
38 Testberichte: 97 von 100 Punkten
Mobiltelefon UMTS/GPRS Velvet Black ... mehr
Sony Ericsson
213,00 € - 572,00 €*
69 Angebote gefunden
Preise vergleichen
Sony Ericsson W810i
36 Meinungen: - Schreiben Sie Ihre Meinung
34 Testberichte: 95 von 100 Punkten
Telefon mobil QuadBand GSM 850/900/1800/1900 GPRS schwarz ... mehr
Sony Ericsson
186,70 € - 453,00 €*
64 Angebote gefunden
Preise vergleichen
Motorola RAZR V3i
12 Meinungen: - Schreiben Sie Ihre Meinung
36 Testberichte: 92 von 100 Punkten
Telefon mobil QuadBand GSM 850/900/1800/1900 GPRS Silver Quarz ... mehr
Motorola
142,90 € - 311,00 €*
31 Angebote gefunden
Preise vergleichen
MIT SICHERHEIT BILLIGER EINKAUFEN
TÜV SÜD
billiger.de-Newsletter
ANMELDEN!
iPod GEWINNEN!
Ihre E-Mail-Adresse:
Jetzt anmelden
Top-Produkte in Handys ohne Vertrag
Sony Ericsson K800i
Sony Ericsson W810i
Motorola RAZR V3i
Motorola KRZR K1
Sony Ericsson W850i
Sony Ericsson W880i
Nokia 6131
Nokia 7370
Sony Ericsson K550i
Motorola RAZR V3

2007

“just a website”
architecture
ca. 2008

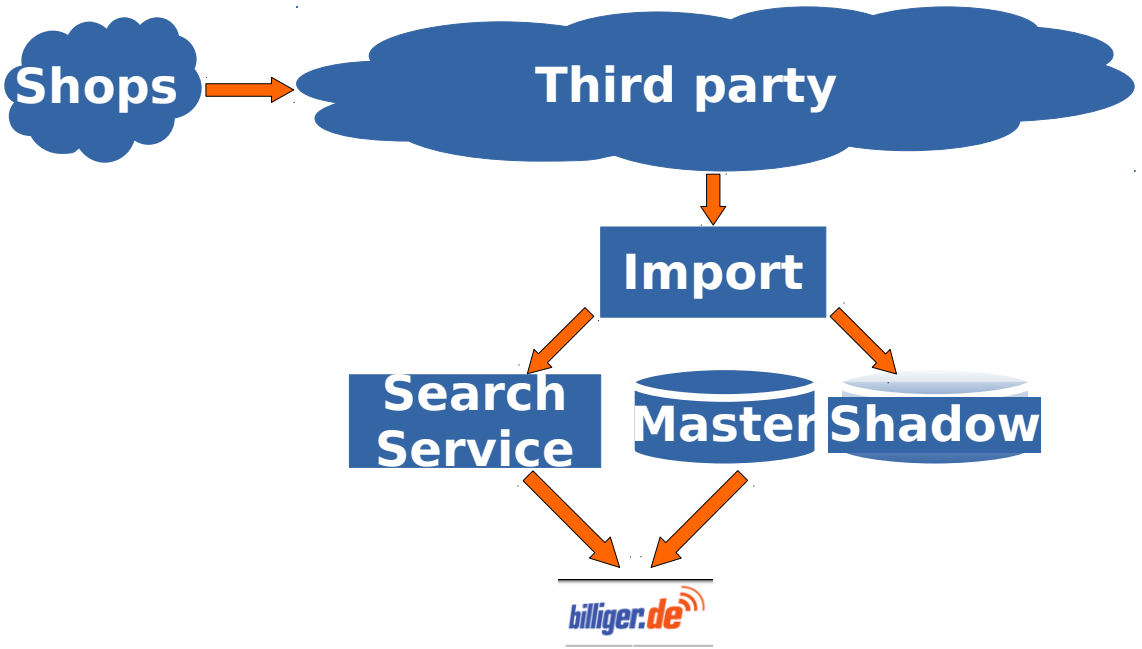


"just a website" architecture ca. 2008

Developer carrying out “the switch”

Source: https://www.youtube.com/watch?v=0gU35TgtImg

billiger.de ~2008

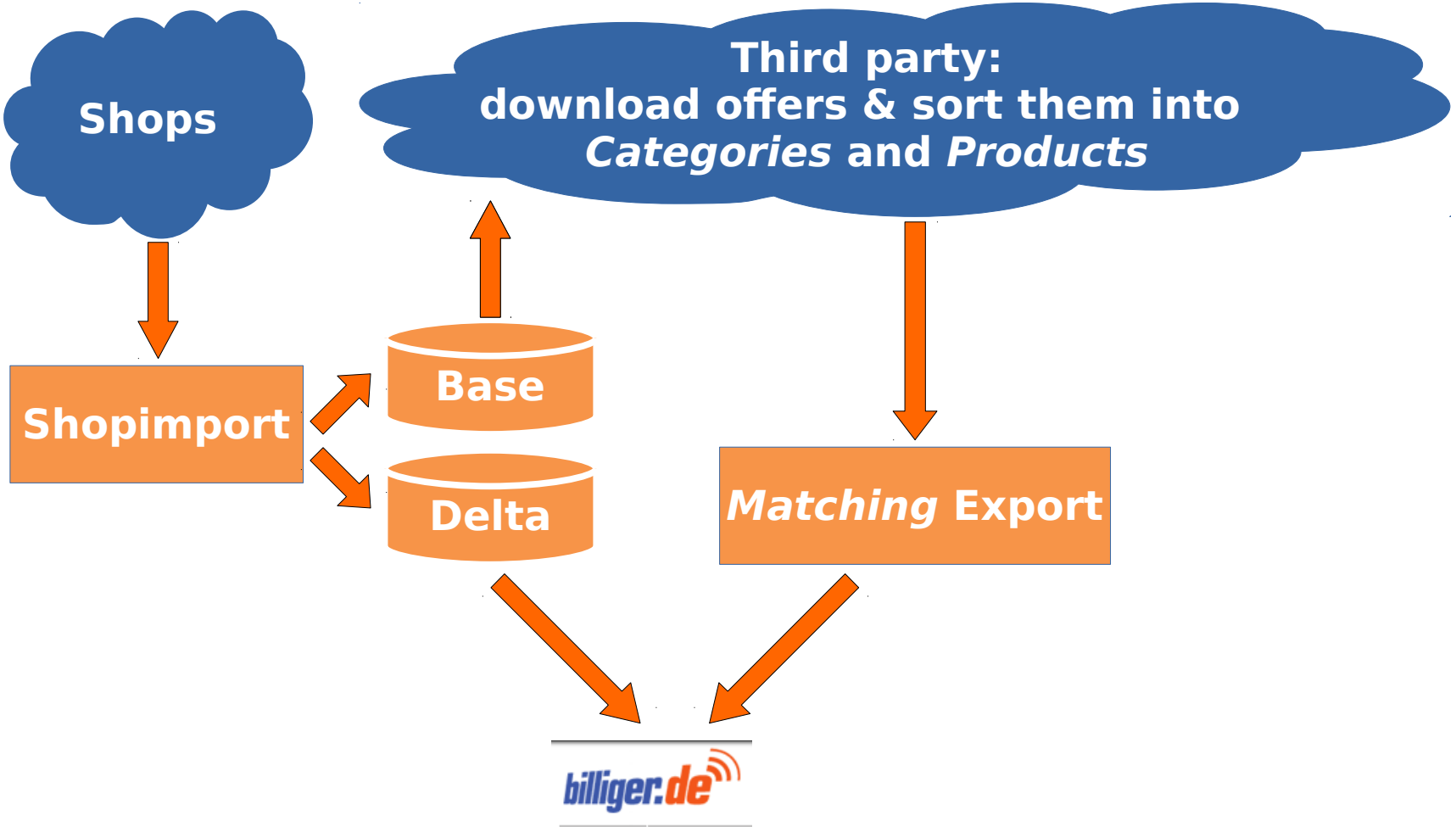


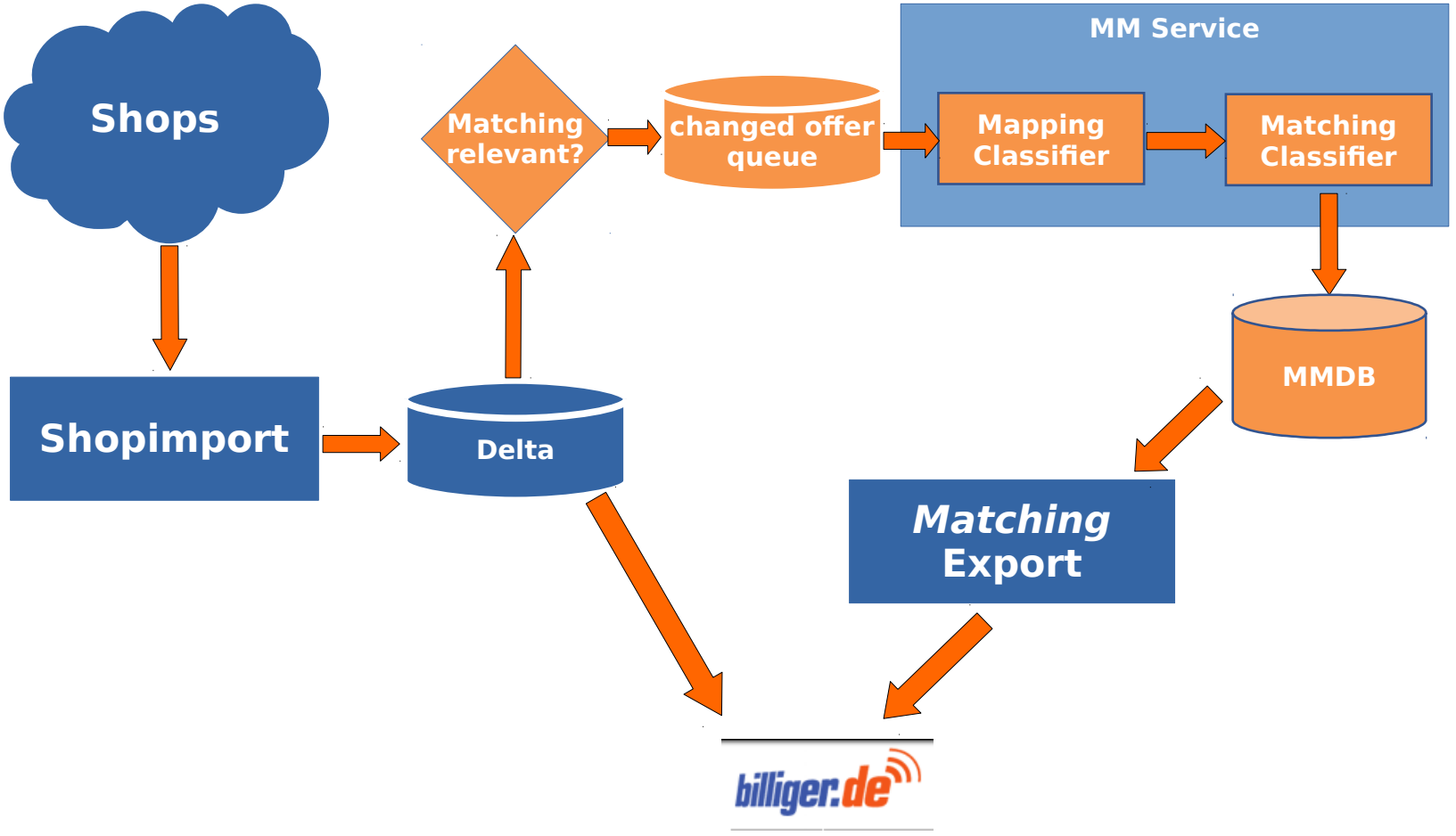
- huge lag: imports only every 4 hours
- hard to correct errors: No feedback API
- expensive: revenue share with third party

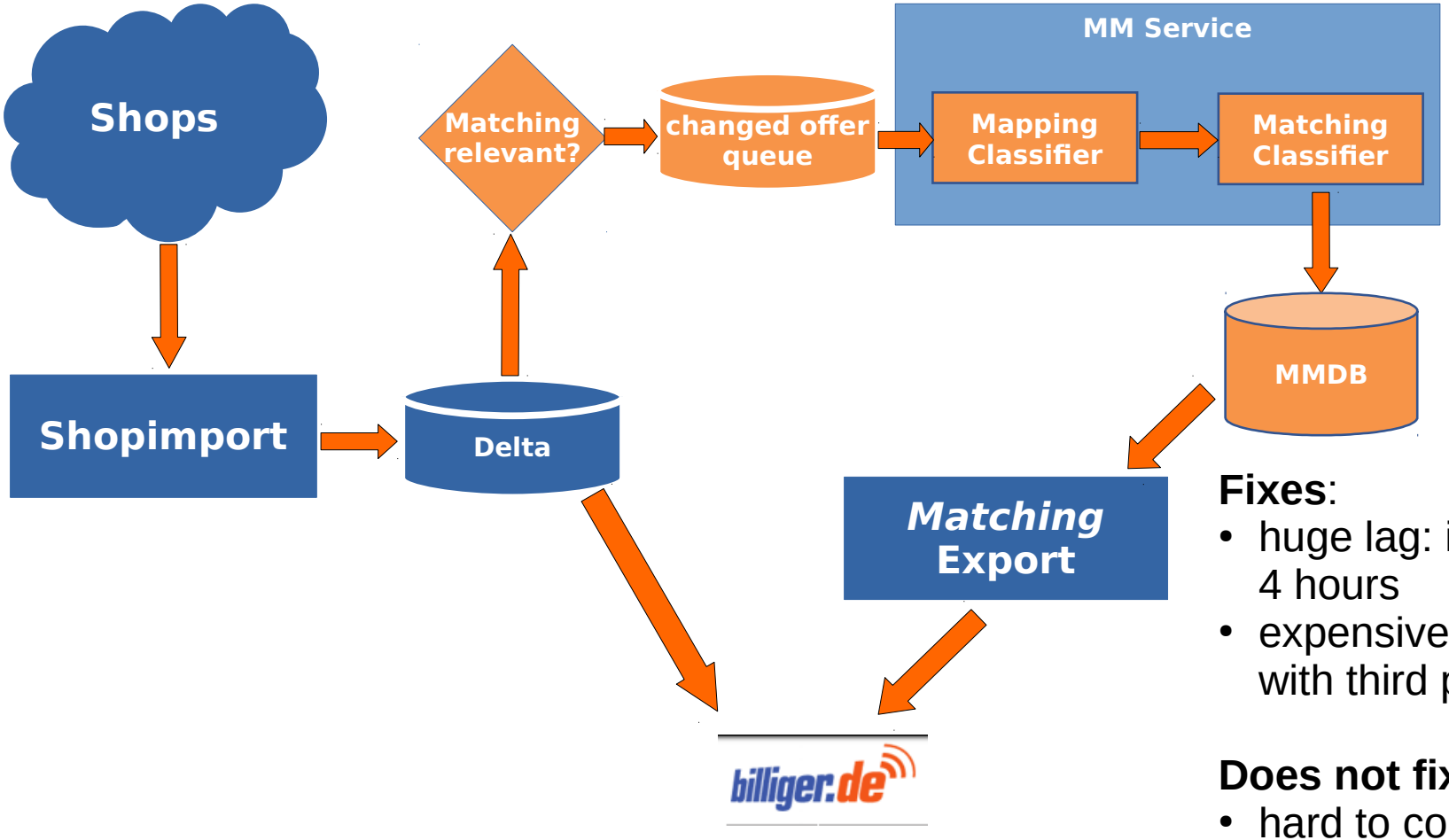
Decision: Inhousing

- We need…
  - a system that downloads shop data (shopimport)
  - to replace the 3rd party black box
    - categorization (*Mapping*)
    - sort offers into products (*Matching)*

➔ “Eigenes Matching Projekt” (EMP)


Shops
Third party:
download offers & sort them into
Categories and *Products*
Shopimport
Base
Delta
Matching Export
billiger.de


Shops
Matching relevant?
changed offer queue
MM Service
Mapping Classifier
Matching Classifier
MMDB
Shopimport
Delta
Matching Export
billiger.de

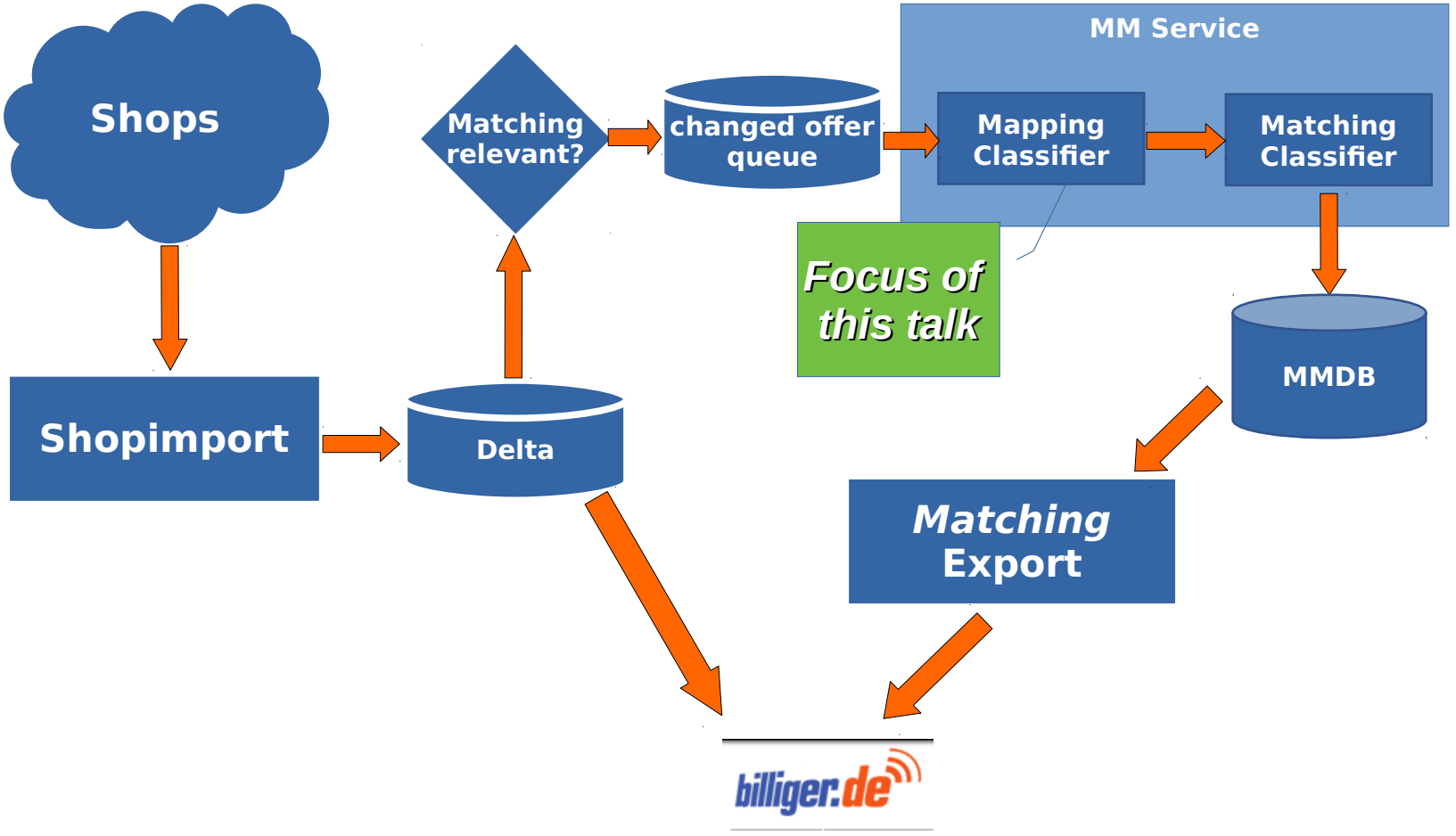


Fixes:

- huge lag: imports only every 4 hours
- expensive: revenue share with third party

Does not fix:

- hard to correct errors: No feedback API
  - ✓ At least we have the data in our hands now!


Shops
Matching relevant?
changed offer queue
MM Service
Mapping Classifier
Matching Classifier
Focus of this talk
MMDB
Shopimport
Delta
Matching Export
billiger.de

SVM Categorization

- Model categorization as a classification task
- Use the 3rd party's category labels as training data (> 2000 backend categories)
- State of the Art technology: Support Vector Machine (SVM)

SVM Categorization

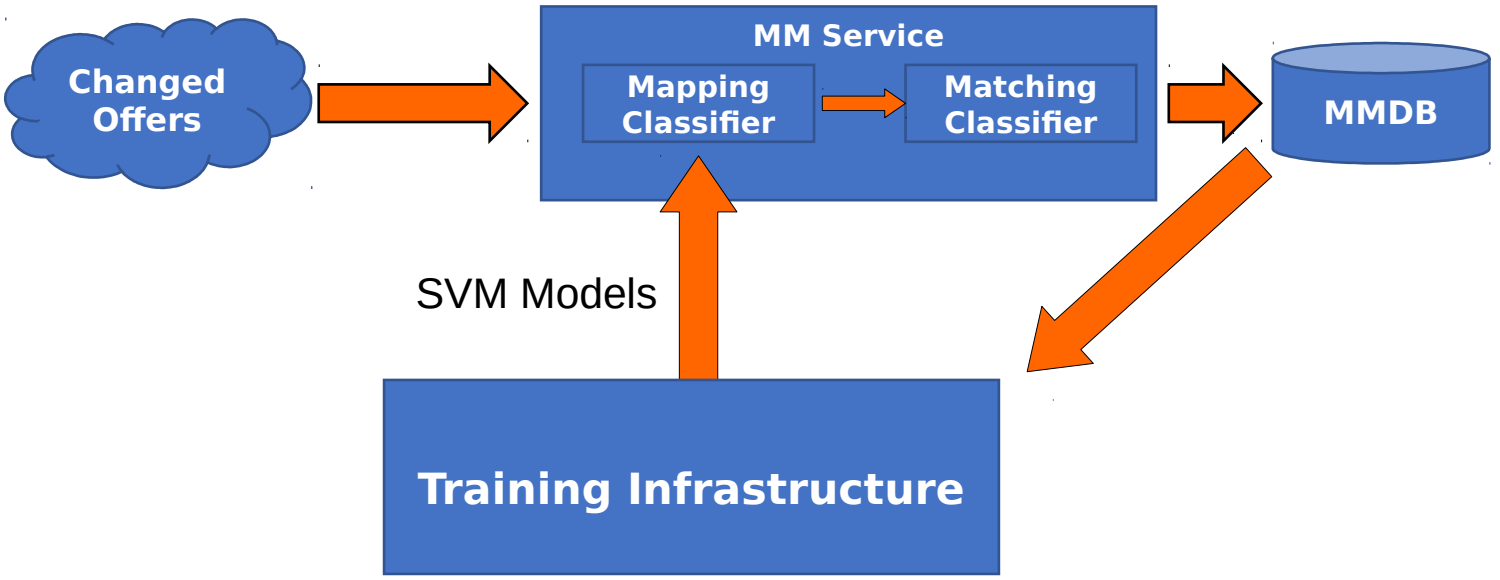
- Train one model per category (one vs. many)
- Classify an offer with each of the 2000+ models, pick highest score
- Acceptance threshold for each category
- Features: Bag of Words
  - Tokenized fields: *Name*, D*escription, Shop Category*
  - *Price bin* token (0..5€, 5..10€, …)

Challenges

- Python ecosystem hadn’t found its love for ML, yet.
- Use Liblinear, libsvm (C/C++)
- Preprocessing with nokia’s map reduce framework *disco*
- *“16 GB RAM was huge”*
  - *mmap custom binary files for parallelized training*

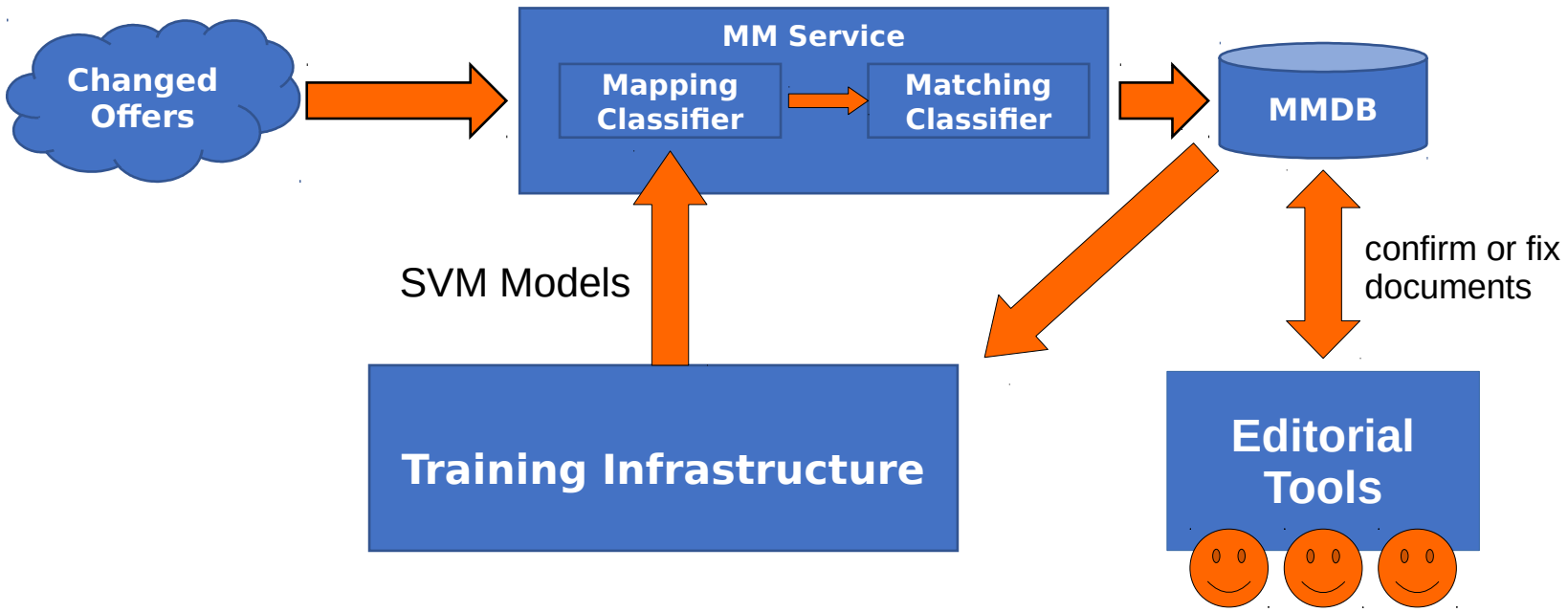
Ready, Set, Go

- Cross validation for each category (precision/recall)
- Adjusting thresholds for each category
  - False Positives vs False Negatives
- 2009: Mapping Classifier goes Live!

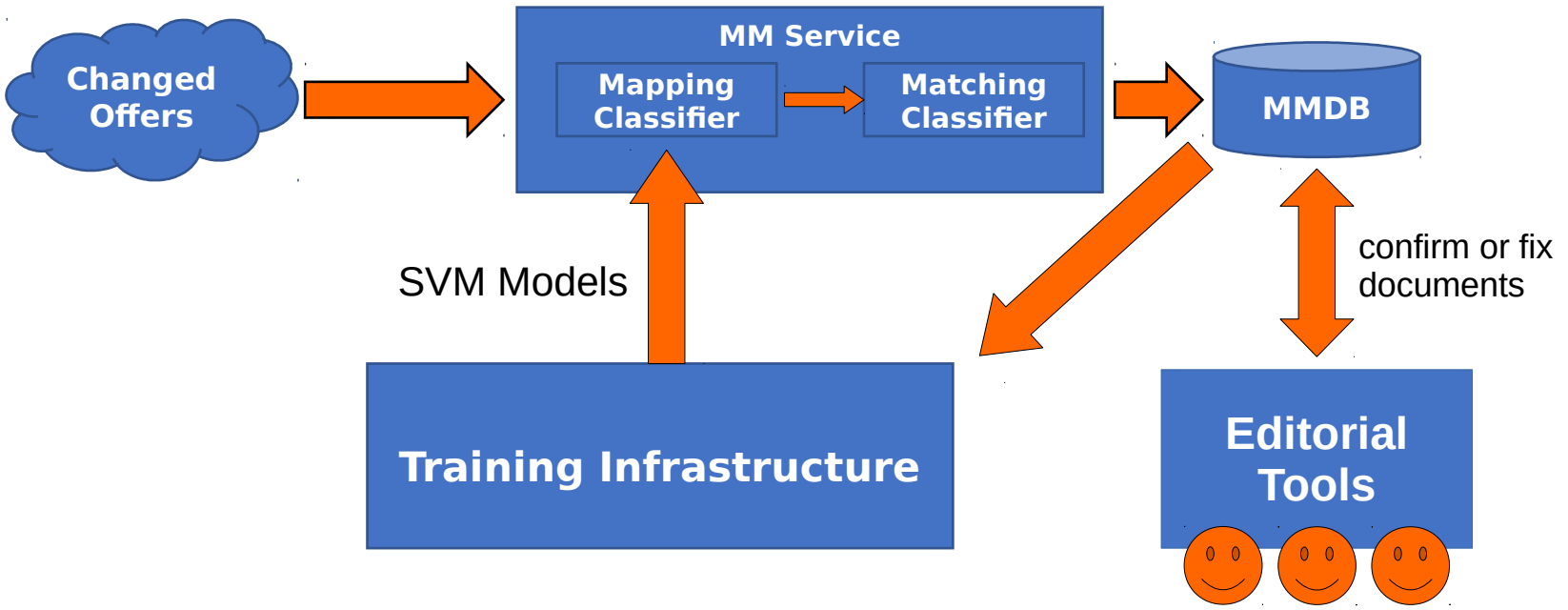

Changed Offers
MM Service
Mapping Classifier
Matching Classifier
MMDB
SVM Models
Training Infrastructure

Living Data

- Source of labeled data vanished
  - Need new source for training data, especially for new categories
- Fixing errors
  - needs cleanup imMeDiAtELy!!


Changed Offers
MM Service
Mapping Classifier
Matching Classifier
MMDB
SVM Models
confirm or fix documents
Training Infrastructure
Editorial Tools

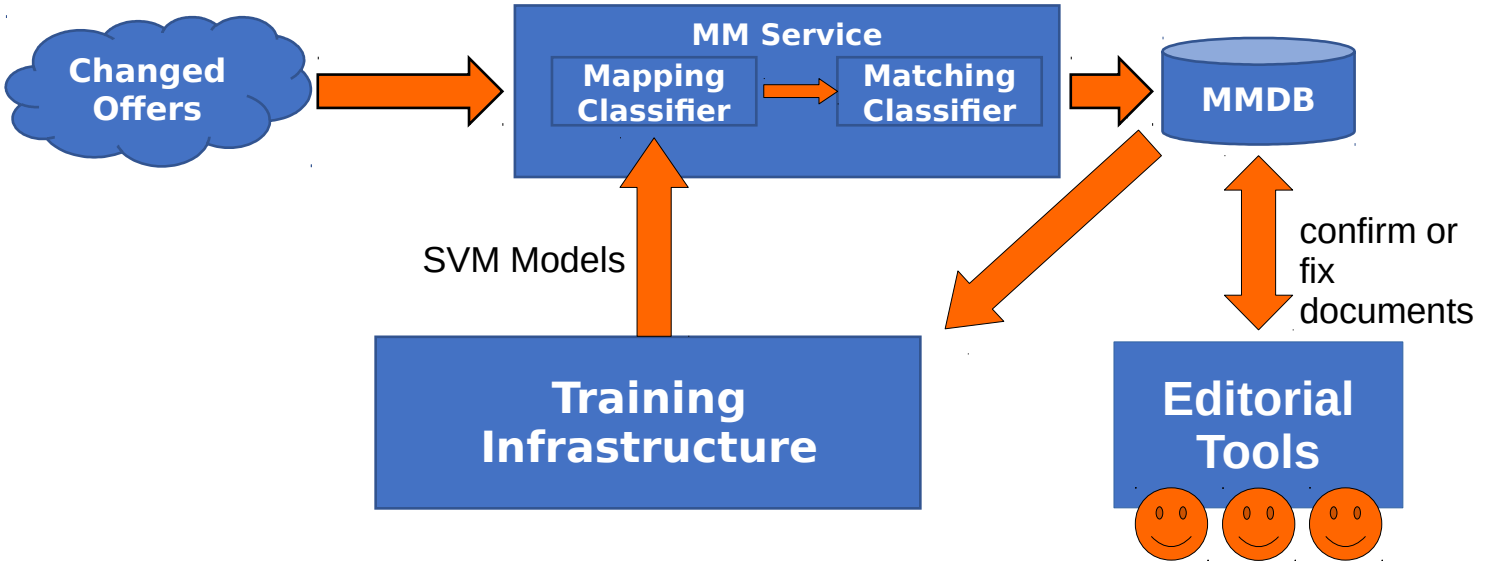
Steady architecture ~2009-2019

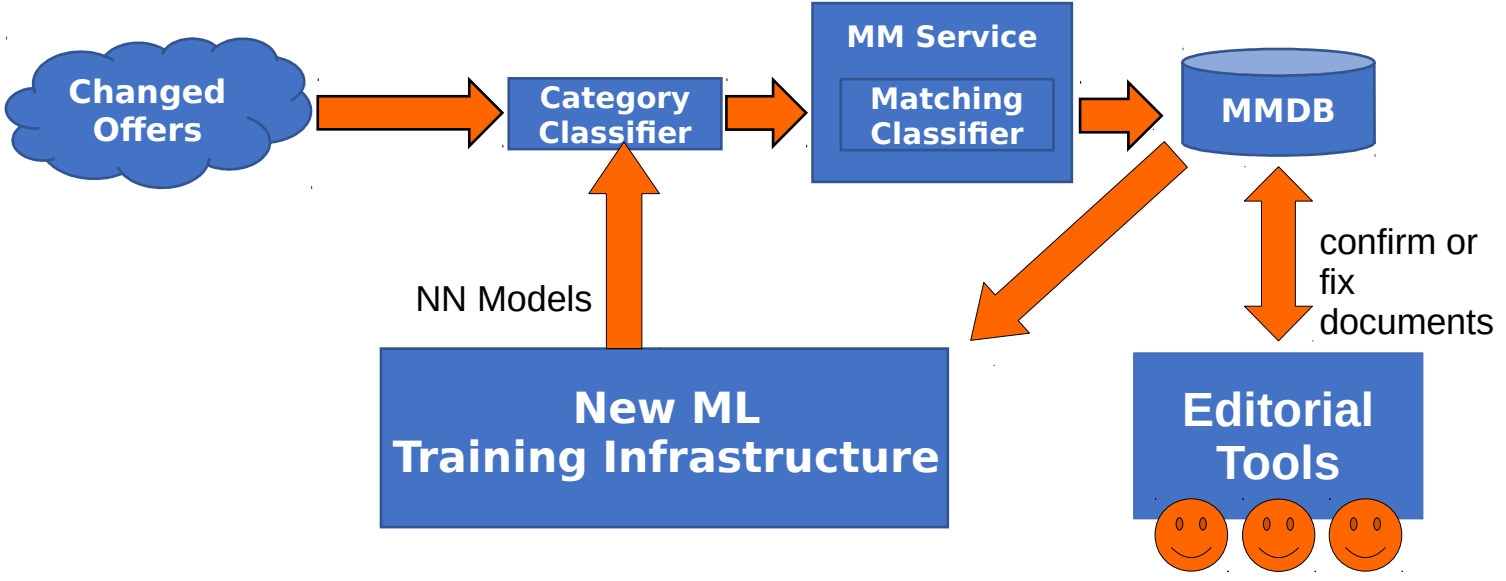


Nevertheless, much of the architecture around this system changed during that period!

Decision: Rewrite Mapping Classifier (2018)

- Replace unflexible infrastructure
  - so many ideas, so much work to integrate anything…
- Remove technical debt
  - map reduce framework Disco is unmaintained, doesn't run on recent Linux
  - replace Non-Python tools


Changed Offers
MM Service
Mapping Classifier
Matching Classifier
MMDB
SVM Models
confirm or fix documents
Training Infrastructure
Editorial Tools


Changed Offers
Category Classifier
MM Service
Matching Classifier
MMDB
New Design!
NN Models
confirm or fix documents
New ML Training Infrastructure
Editorial Tools

The new Category Classifier

- Learn trigram embeddings for each interesting field
  - *Name, Brand, Shop Category, Description*
- Include position information
  - no more BoW
- Price binning still useful
- Single result over all classes

Training Data Problems

- Wrongly labelled data

Training Data Problems

- Semantically overlapping categories
  - *Books > Spanish Books*
  - *Books > French Books*
  - *Books > Foreign Language Books (=non German)*

Training Data Problems

- *Dumping ground* categories
  - Mobile phone accessories:

    bag, cover, upscreen, holder, charger, stylus, gadgets…

Training Data Problems

- Imbalance of feature distribution within category
  - e.g. if the training data for a category are based mostly on one shop, the S*hop Category* will be the dominant feature for the prediction

Training Data Problems

- Imbalance of category distribution over whole dataset
  - Some categories are oversampled, others are undersampled compared to the real data distribution

We have editorial staff working on this, why is our training data such a mess?!

Misaligned goals for editorial staff

- Business: Fix data so results look nice on billiger.de
- Data Science: Constantly adapt category tree and maintain training data for each category

Insufficient Tooling

- Product centered tool: Business driven processes
  - Important categories are well maintained, others are not => non-random training data selection
- Category centered tool: coarse mass edits
- Category tree maintenance is complicated and cumbersome

Over an extended period of time, urgent things blocked important things

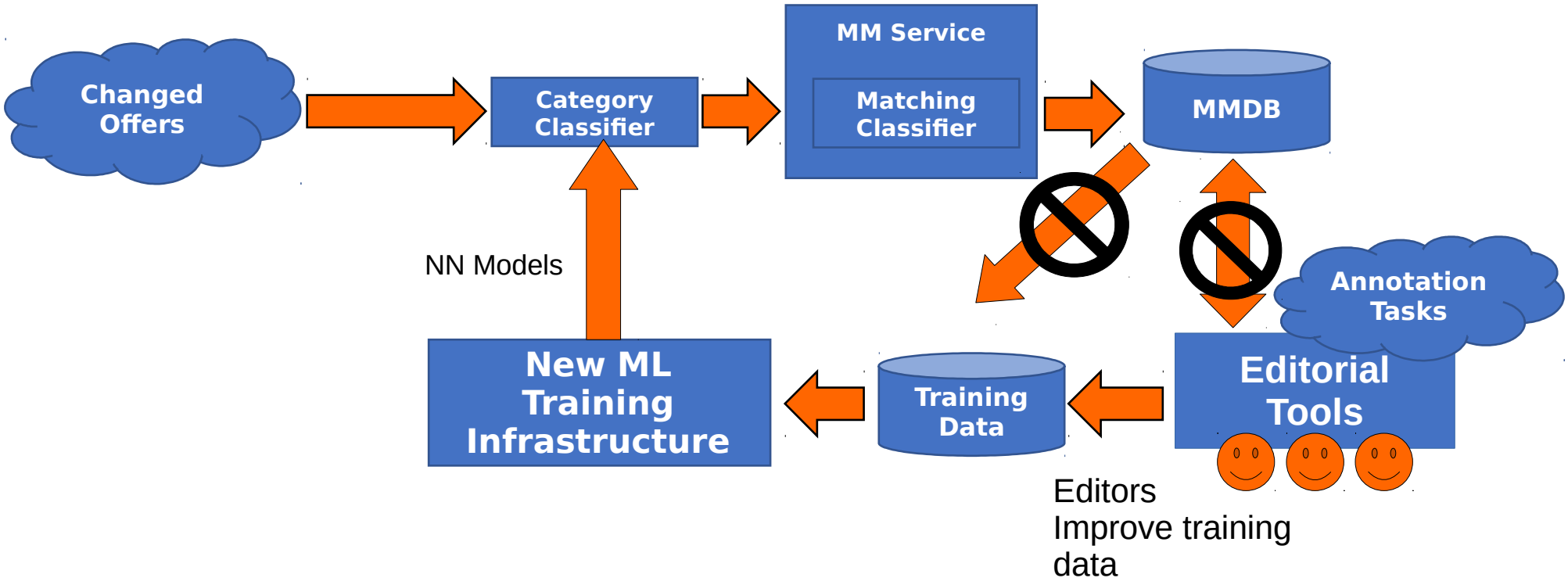
Towards a new ML infrastructure

1) Plan new ML infrastructure :-)
2) Rethink your whole training data management :-(

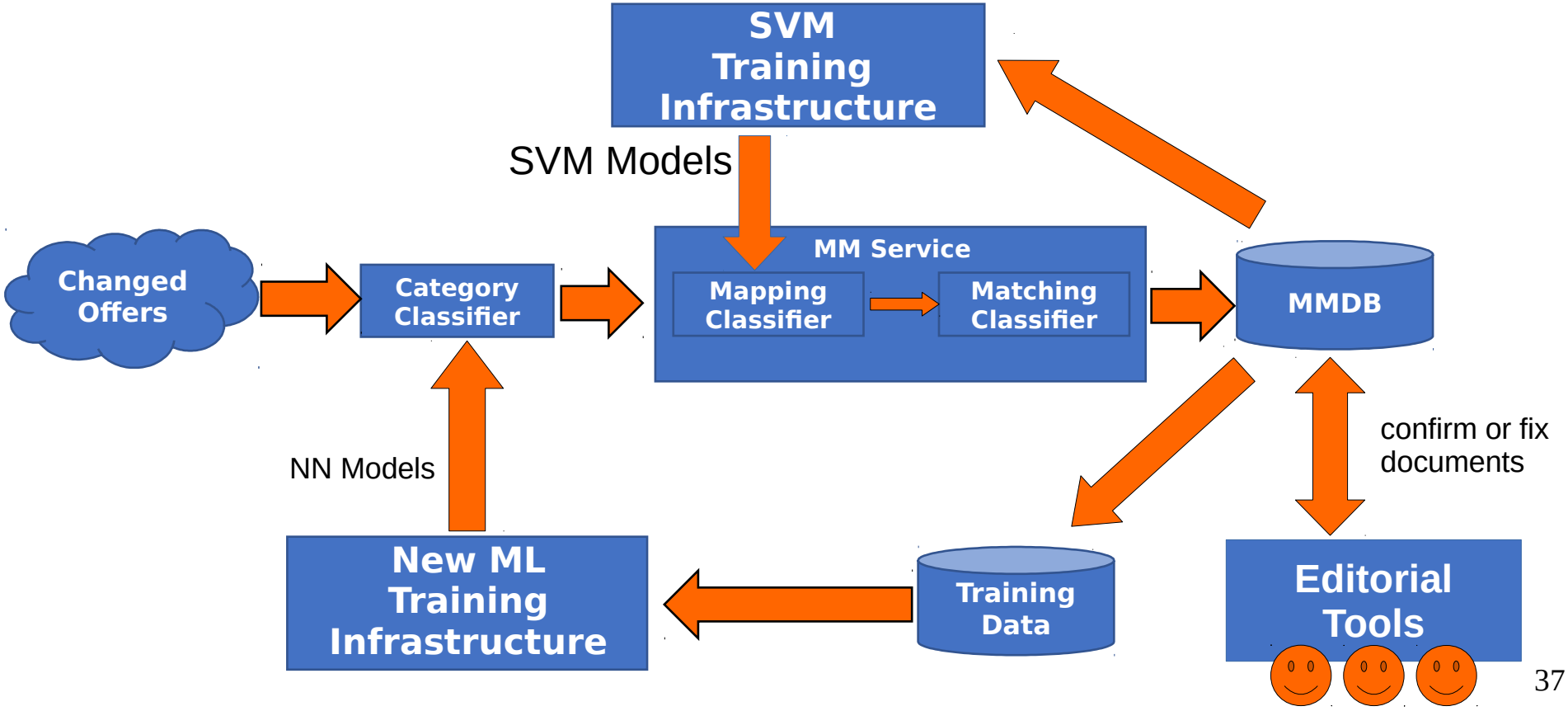
Rethinking Training Data Management

- Change editorial staffs' role: manage category tree and improve training data instead of fixing live data
- Generate annotation tasks
  - Monitor model quality to choose annotation tasks automatically
  - QA by annotator agreement
- Clean up training data that is getting old
- Separate live data from training data

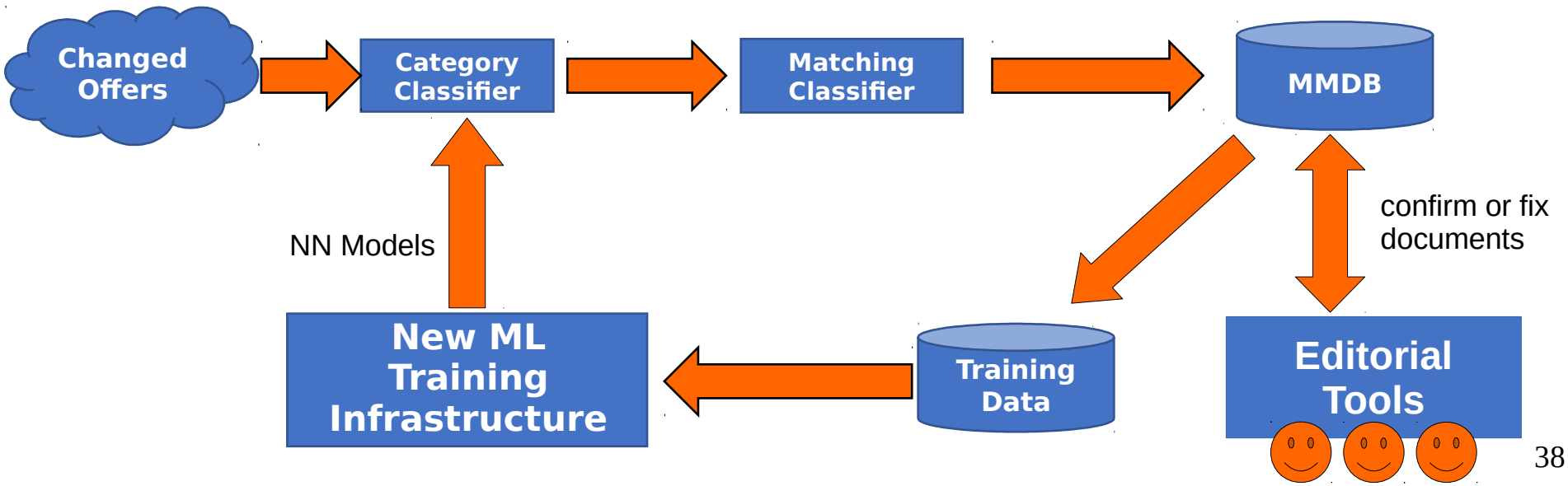
New Vision


Changed Offers
Category Classifier
MM Service
Matching Classifier
MMDB
NN Models
Annotation Tasks
New ML Training Infrastructure
Training Data
Editorial Tools
Editors Improve training data

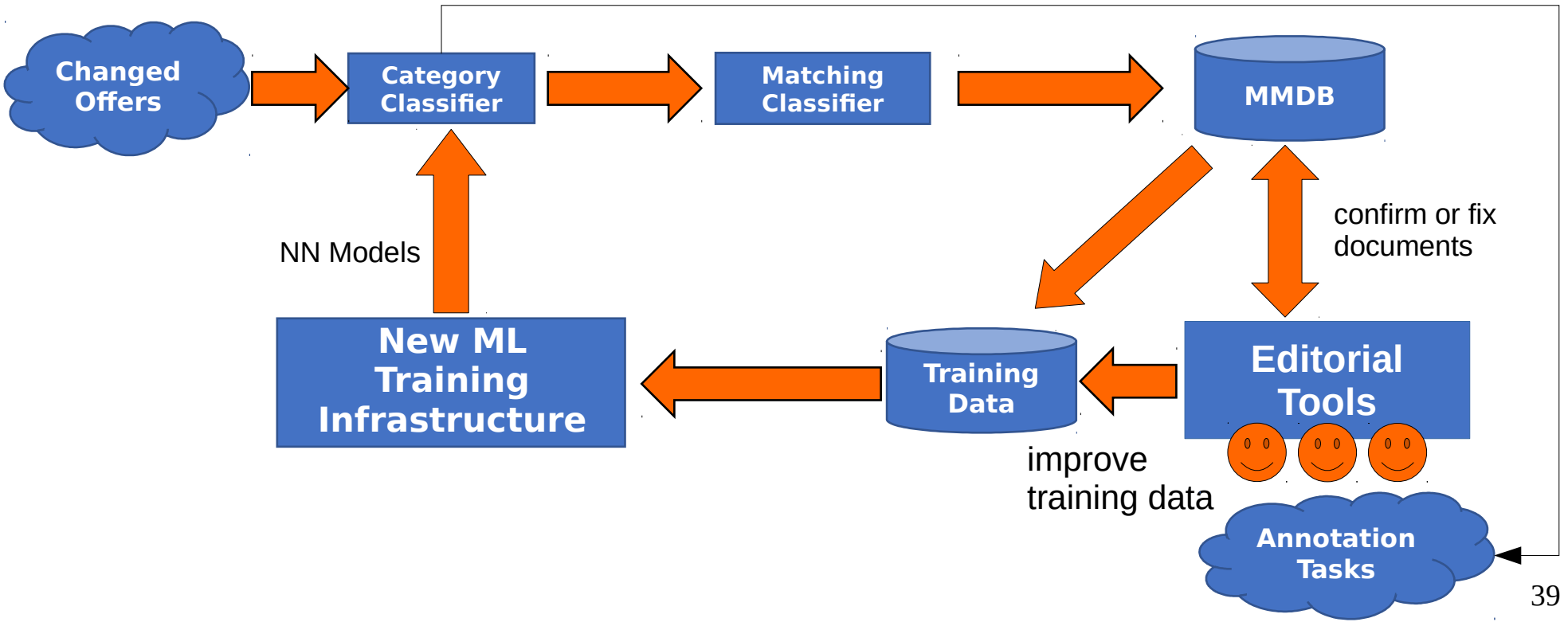
Transition: Operating both models February 2019



Go live NN Category Classifier March 2019



Add Annotation System to the mix Q4 2019



Some Takeaways

- Long time training data management is hard
  - What is your system’s life cycle?
  - Moving targets, aging data: If your domain is changing over time make sure your training data keeps up (forgetting data is harder than adding data!)
  - Technology transitions cost time but come with flexibility benefits afterwards which can revitalize creativity

Some Takeaways

- Well-founded decisions improve the status quo, but need to be challenged every now and then
  - If a fix solves just part of the problem, keep thinking about the whole solution
  - Remember this?

Does not fix:
- hard to correct errors: No feedback API
  - ✔ At least we have the data in our hands now!

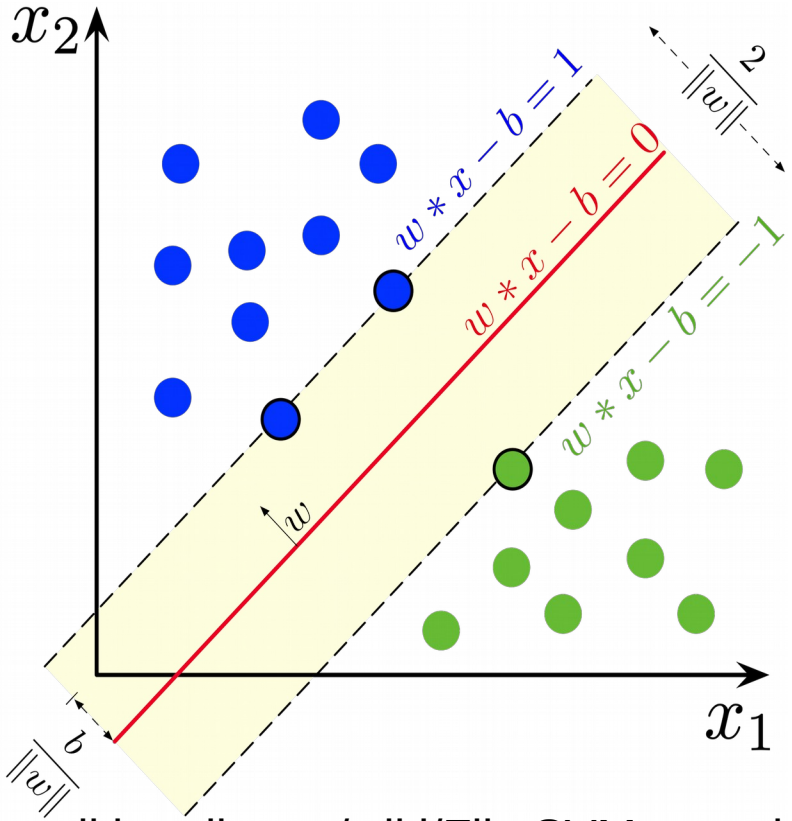
Thanks!

Special thanks to **Patrick Schemitz** who provided the information about the early architecture and evolution of the system and **Christian Schramm** for being a driving force of change and a source of motivation.

Backup Slides

SVM

- Hyperplane separating the points of the two classes
- Score: **w · x – b**
- Weights help interpreting the output

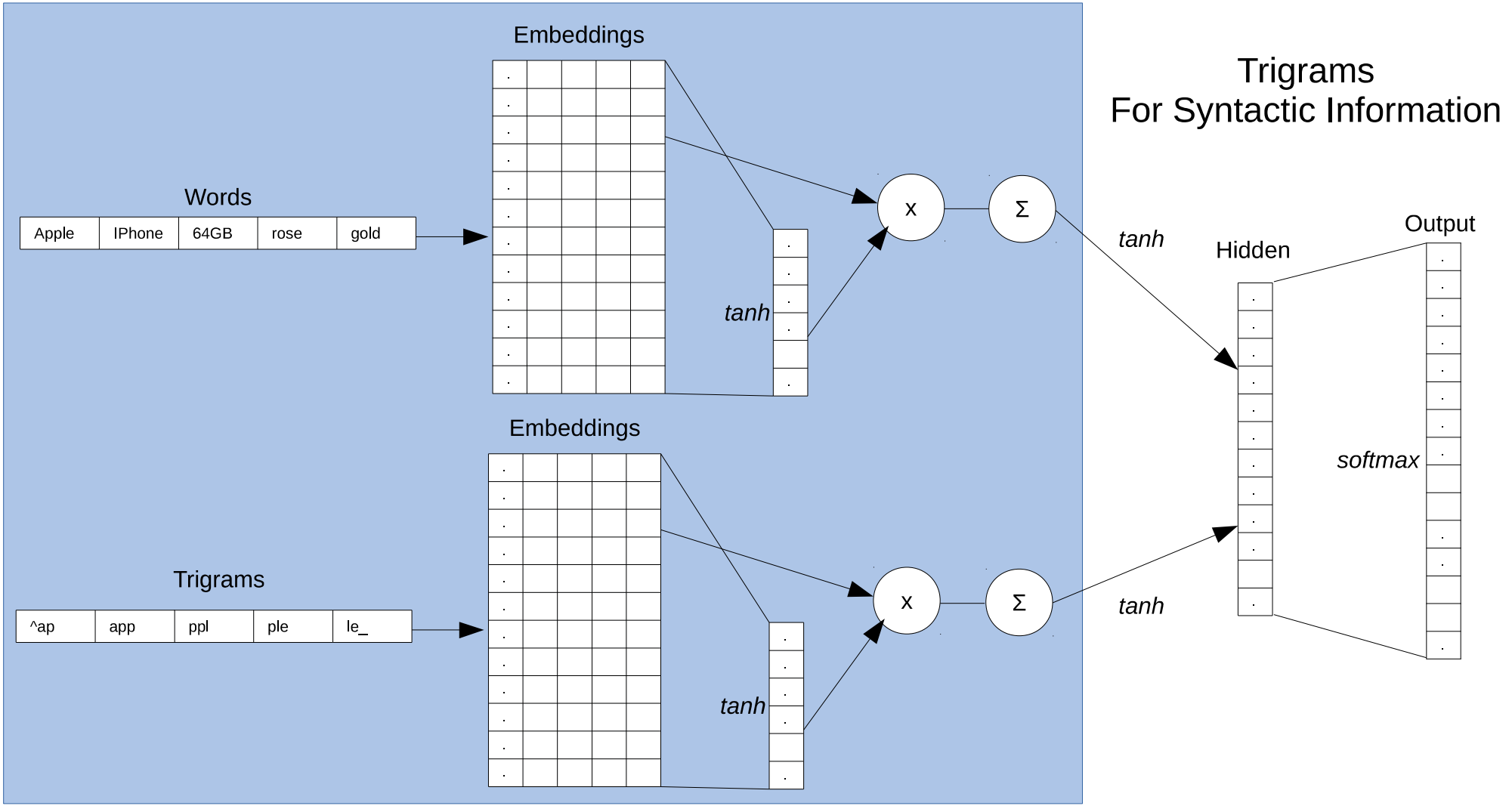


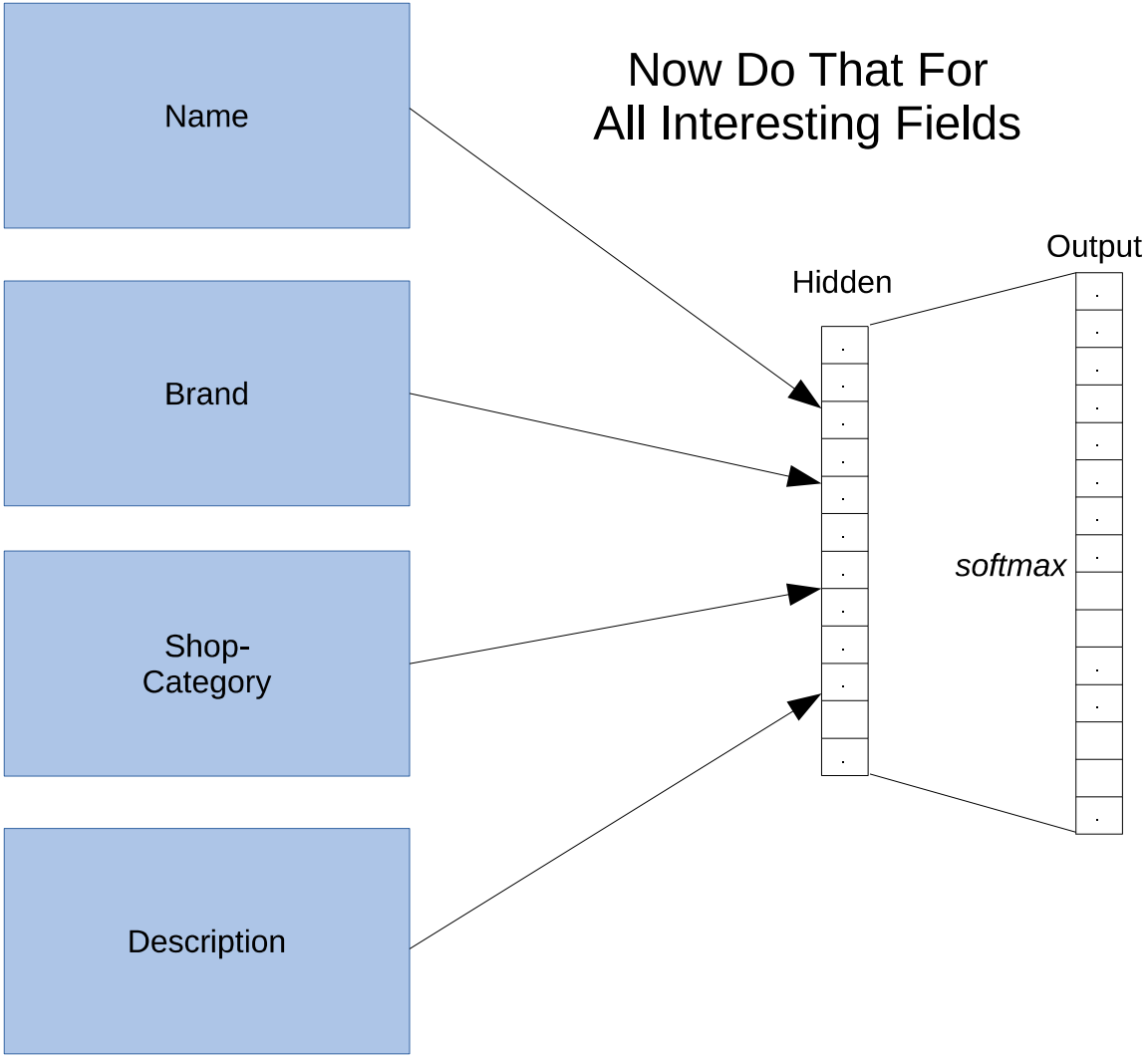
https://commons.wikimedia.org/wiki/File:SVM_margin.png

Category *MP3 Players* (2012)

Token	Weight
nwz	4.41
mpaxx	4.17
technimax	3.66
yp	3.59
4gb	3.44
techniplayer	3.28
xemio	3.18
mpixx	3.13
2gb	3.12
8gb	2.98
sansa	2.76
…	…

schokobraun	0.00403229
beautiful	0.00323284
lieferumfang	0.00240695
…	…
case	-3.04
tasche	-3.06
akkus	-3.24
cd	-3.61
dockingstation	-3.64
taschen	-3.99
zubehör	-4.87
displayschutzfolien	-4.99
für	-6.55


Trigrams
For Syntactic Information
Embeddings
Words
Apple
IPhone
64GB
rose
gold
tanh
x
Σ
tanh
Hidden
Output
softmax
Embeddings
Trigrams
^ap
app
ppl
ple
le_
tanh
x
Σ
tanh



Use Cross-Entropy Loss for a true multiclass model that outputs probabilities

Changes outside of mmservice

- Process shop offers in an event based manner and use ElasticSearch as a storage system
  - PyCon.DE 2017 Axel Arnold - And now to something ELSE: Real Time Data Processing @ billiger.de https://www.youtube.com/watch?v=en7XcpYxLU4