solute

# A Brief History
# of
# Category Classification

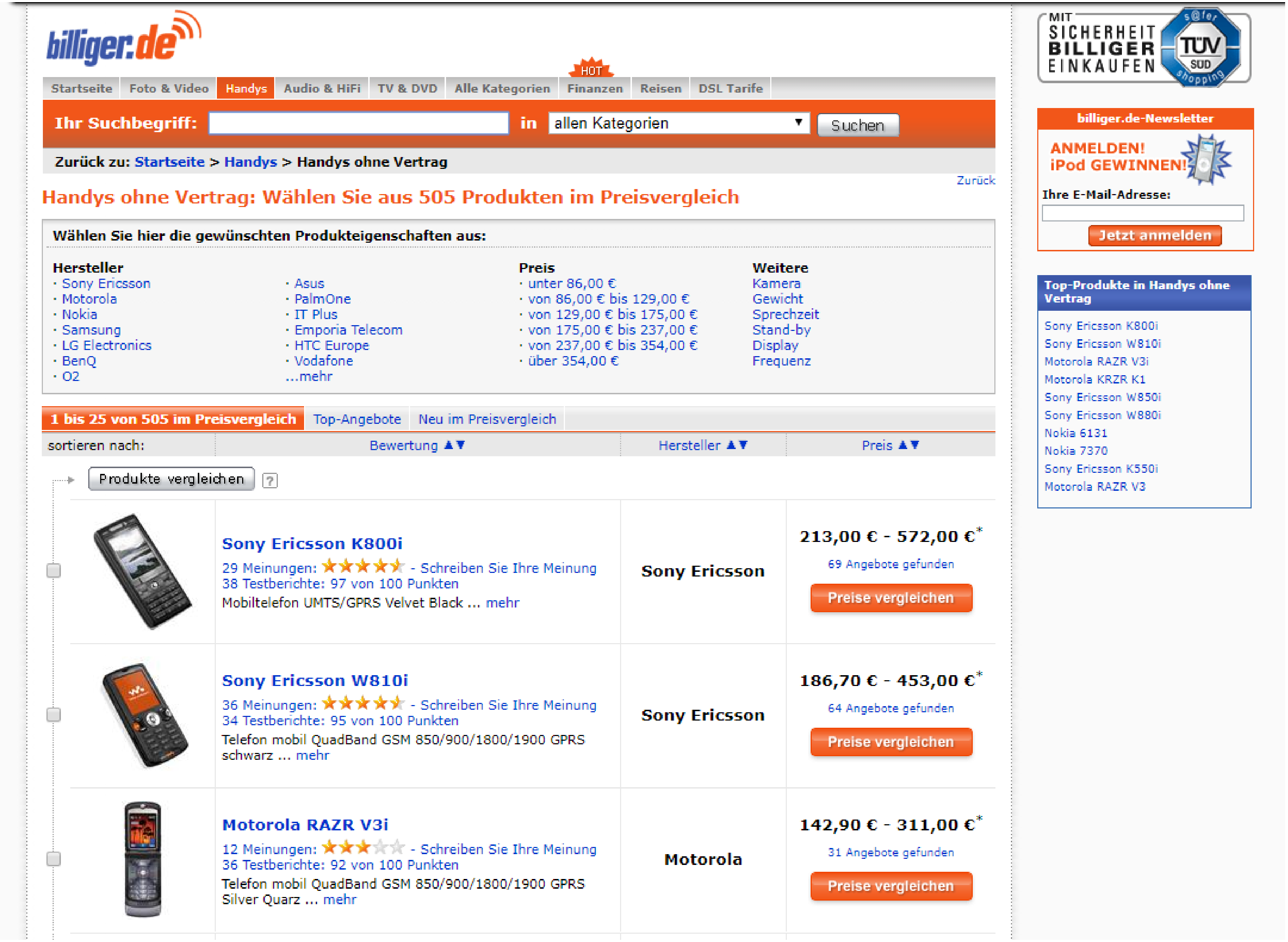
Patrick Schemitz
Senior Scientist
solute GmbH – billiger.de

billiger.de Shopping.de kaching. friends communication solute

# The Price Comparison Universe

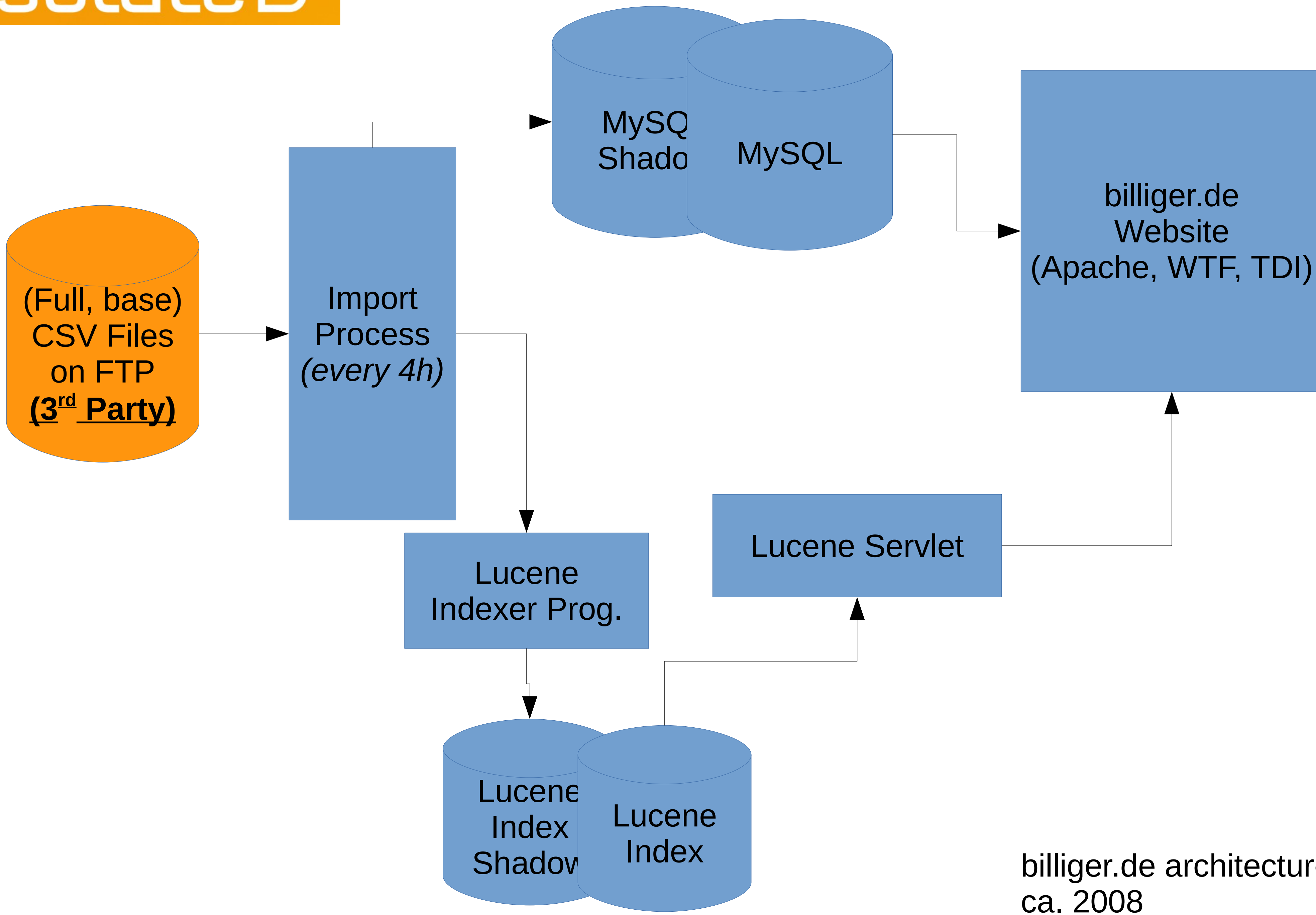
- Shops:
  provide files with offers (CSV mostly)
- Offers:
  An article, for a price, by a shop.
- **Products**:
  Group of offers for the same article.
  Created implicitly.
- **Categories**:
  Group of offers/products of same article type.
  Created explicitly.

# Status Quo Ante



# Status Quo Ante

- ***billiger.de*** was just a website, no “backend”
- Exports from a third party vendor (base CSVs) mentasys → pangora → become → connexity
- Categorization and product matching by third party




billiger.de architecture,
ca. 2008

# Drawbacks

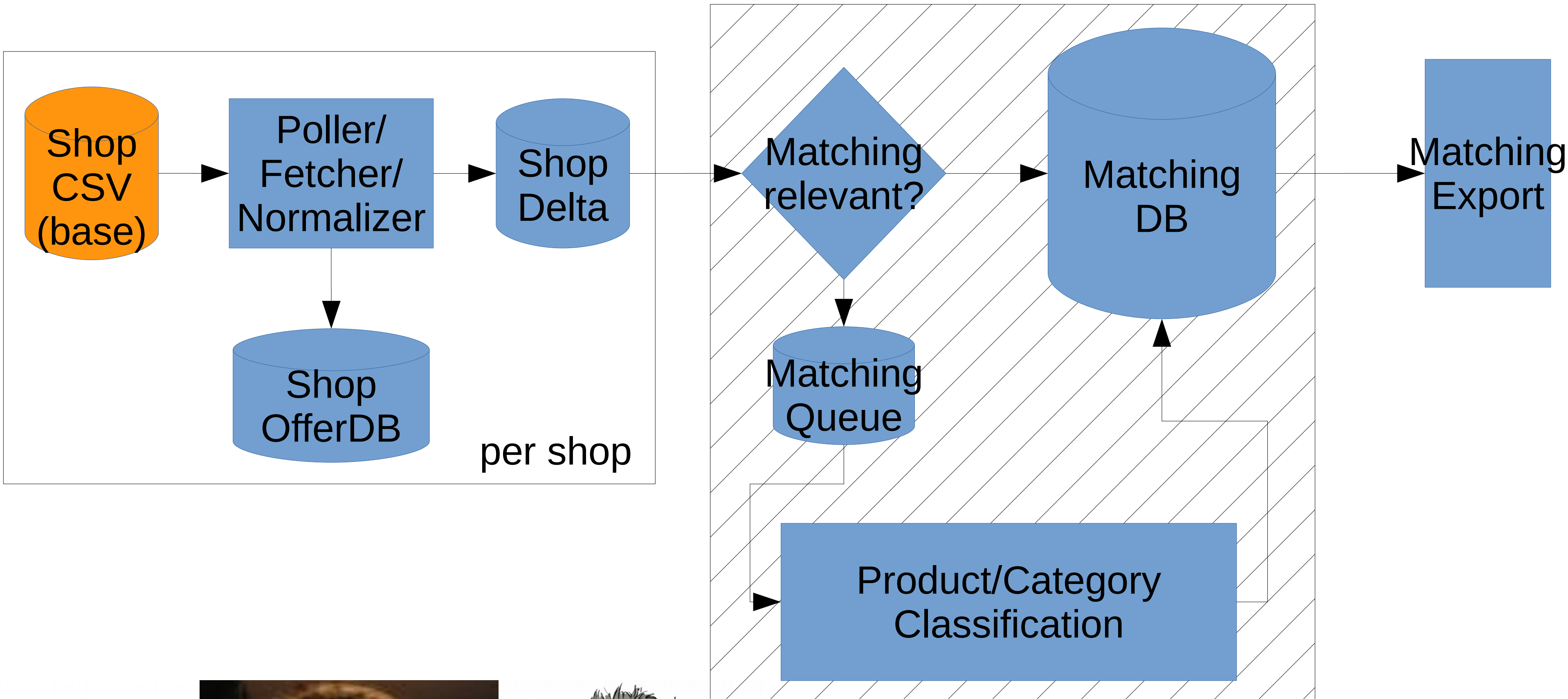
- Huge **lag**:
  imports only every 4h
- **Errors** hard to correct:
  no feedback API
- **Expensive**:
  revenue share with third party

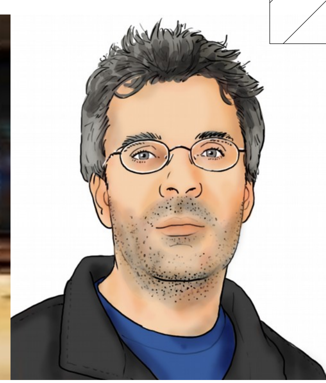
# “EMP”

- “Eigenes Matching Projekt”
- Matching := product and category classification
- Replace the orange box…

# “EMP”

- … a *lot* of work:

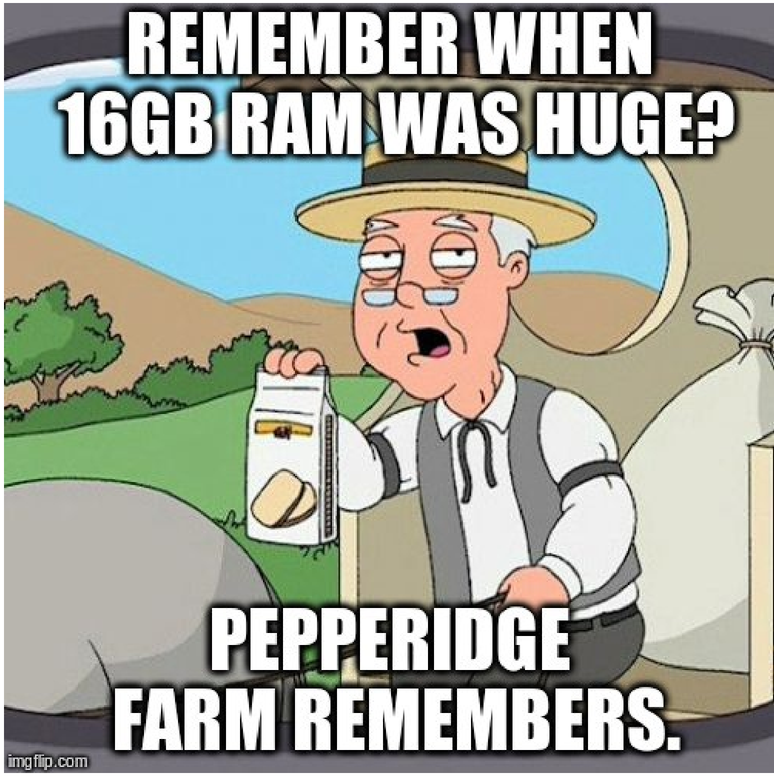


# Classification from Scratch

Back then...

- 16 GB RAM was *huge*
- scikit-learn, NumPy not much help (yet)
- Liblinear, libsvm (C/C++)
- Training: existing offers and categories
- Our first real ML project!
- First thing you write:




**Precision/Recall Script!**

# First Experiments

- All data is labeled
- Experiments with simple **search** in labeled data (tf/idf score, allowing missing words)
  → false positives, poor precision
  "Dunlop Winter Sport 205/55R16"
- Need "repellent" word scores for nearby categories!
  → **Support Vector Machine!**

# First SVM

- One model per category (one vs. many)
- Classify offer 2000×, pick highest score
- Acceptance threshold for each category
- Training based on liblinear code
- Features: just the words/tokens of the offers (cf. search ansatz)

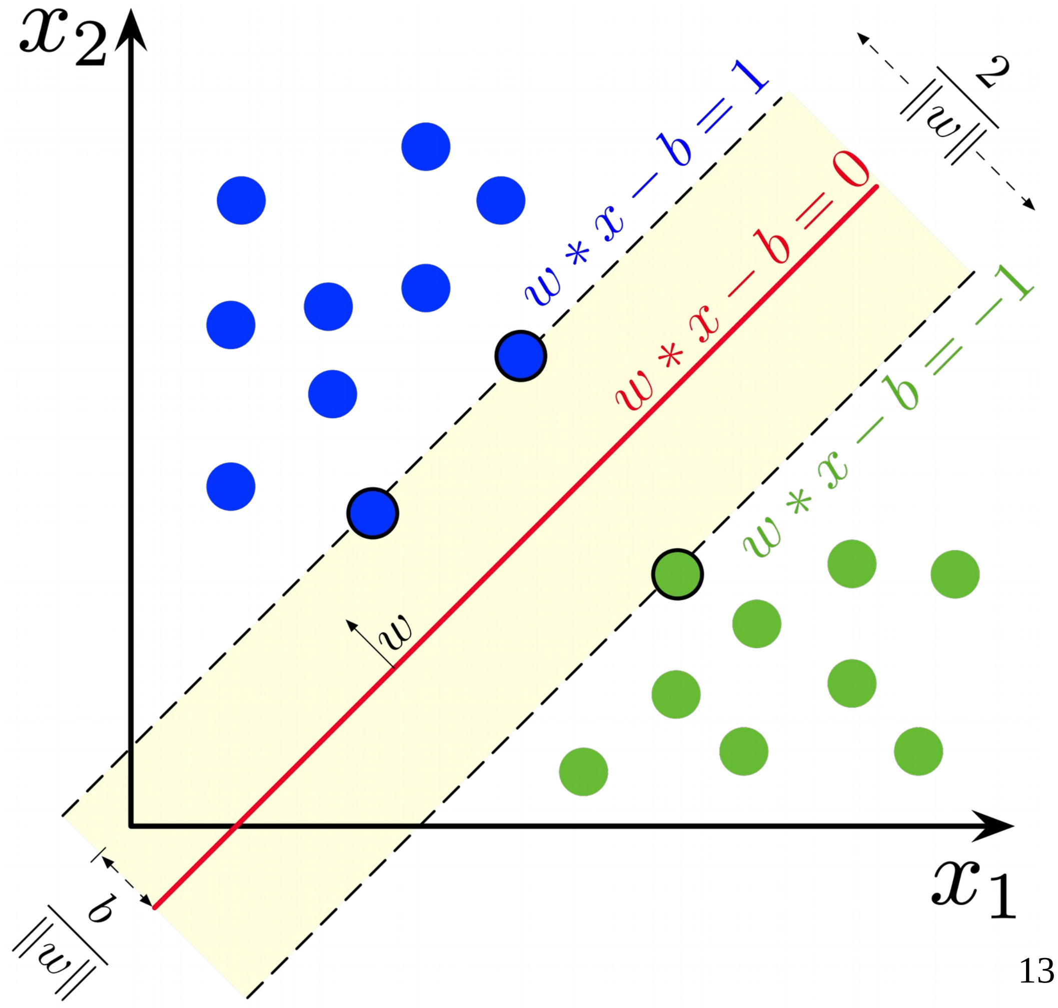
soluteQ

# Infrastructure Challenges

- 2000× classification is costly → “transposed” classification: each feature brings in its possible categories (typical: few hundred)
- Holding 2000+ models in memory (for classification): strip features with low weight; watch for Python ref counting vs. COW in multiprocessing
- Building 2000+ models: C++ code, mmap(2)ed binary files, multiple processes via fork().
- Cross validation for each category (precision/recall)
- Adjusting thresholds for each category (which is worse, fp or fn?)
- **LIVE**!

# SVM and Insights

- Hyperplane separating the points of the two classes
- Score := w · x – b
- Weights *mean* something



# SVM Models are Intelligible

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

| | |
|---|---|
| -3.04 | case |
| -3.06 | tasche |
| -3.24 | akkus |
| -3.61 | cd |
| -3.64 | dockingstation |
| -3.99 | taschen |
| -4.87 | zubehör |
| -4.99 | displayschutzfolien |
| -6.55 | für |

**MP3 Players 2012**

| | |
|---|---|
| 0.00403229 | schokobraun |
| 0.00323284 | beautiful |
| 0.00240695 | lieferumfang |

# Interpreting the Model

- Positive weights → Attraction
- Negative weights → Rejection
- Near-zero weights → neutral
- Negative features indicate *neighboring* categories
- Category taxonomy influences models

# Example: “Erotik”

- (TÜV Certificate → no sexy stuff on billiger.de)
- Some shops also carry adult toys
- Some syndication partners wanted the E
- Research: model adult vs. non-adult
- (We actually have several distinct adult categories)

# Erotik Model

| | |
|---|---|
| 5.64 | erotik |
| 5.44 | vibratoren |
| 3.55 | fetish |
| 3.45 | sextoys |
| 3.39 | sexspielzeuge |
| 3.13 | obsessive |
| 2.98 | fetisch |
| 2.94 | dildo |
| 2.85 | dorcel |

| | |
|---|---|
| -4.31 | oboy |
| -5.31 | produkte |
| -8.33 | bekleidung |

**Erotik 2012**

oboy, bekleidung
→ neighbors...

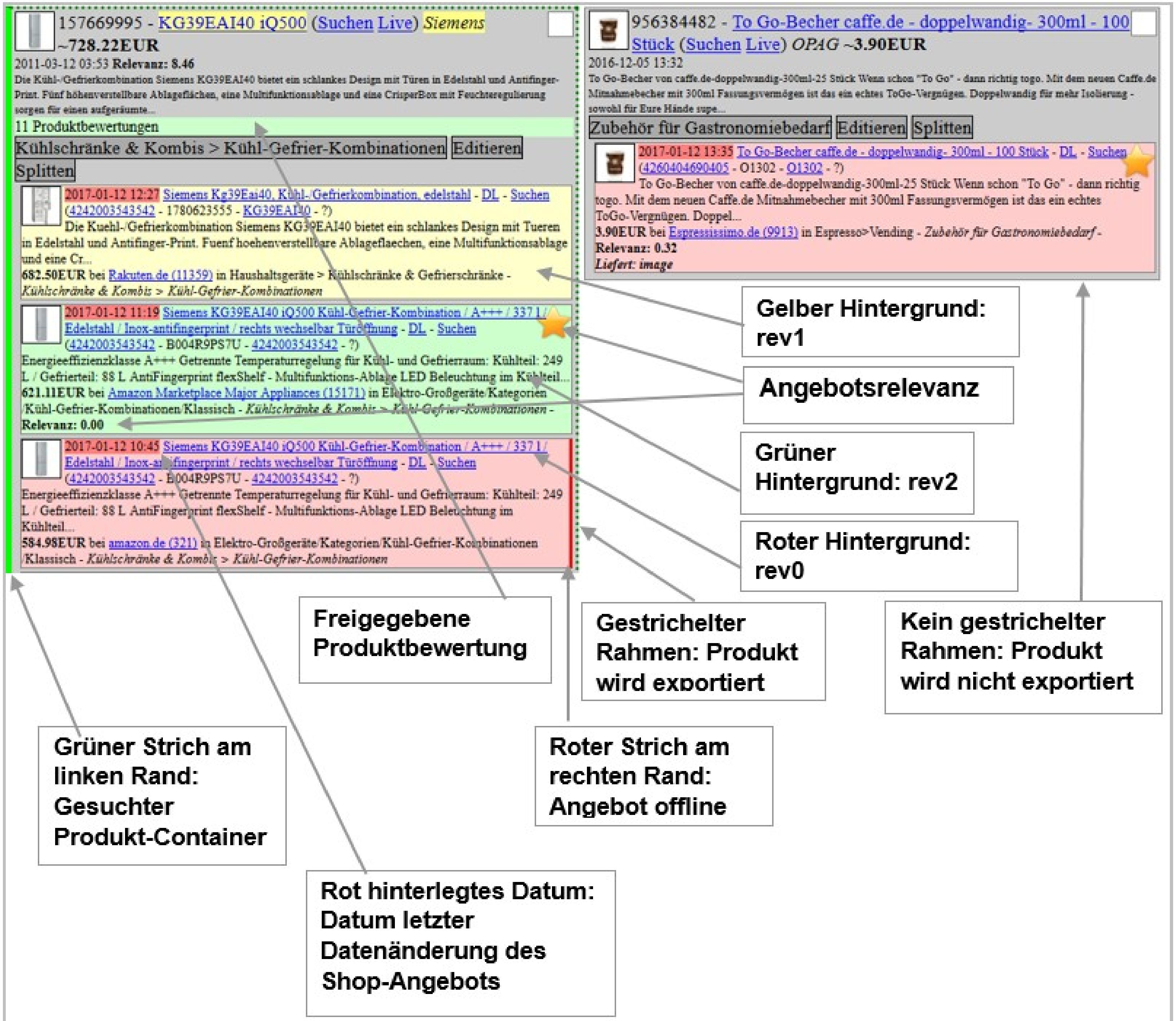
# (Un-) Sexiest Names

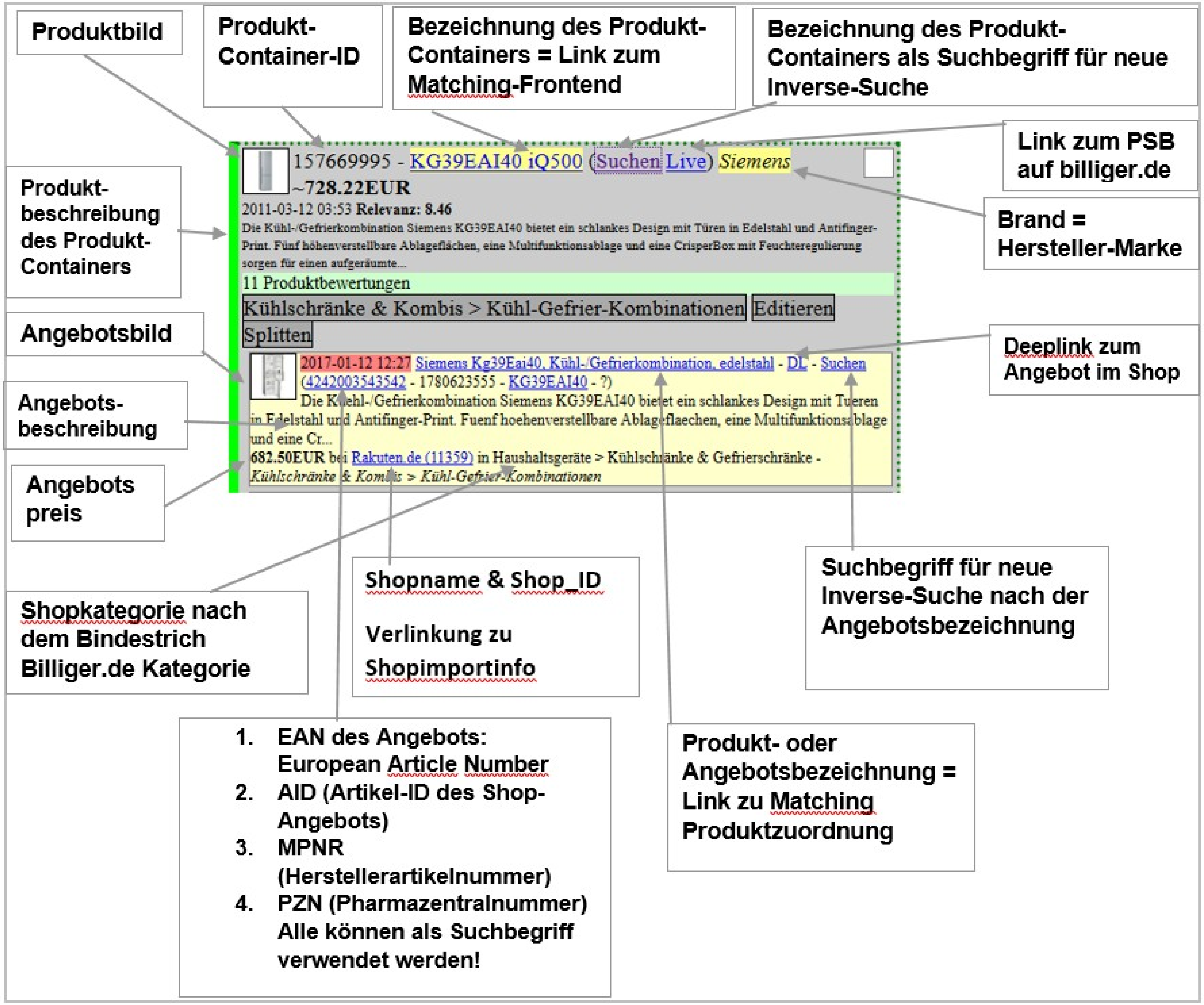
| | | | |
|---|---|---|---|
| 2.38 | heida | -0.72 | david |
| 2.13 | mandy | -0.73 | oliver |
| 1.97 | eugenie | -0.75 | sylvester |
| 1.83 | anetta | -0.88 | tom |
| 1.63 | molly | -0.90 | erika |
| 1.53 | vanessa | -1.07 | lara |
| 1.48 | anderson | -1.08 | lucia |
| 1.47 | nadine | -1.26 | max |

- Model vs. name list
- First sexy male name: **Sven** (pos. 17)!
- Jenna: pos. 25, Ron: pos. 28

# Living Data

- Source of labeled data dried up
  → need new source for training data
- New categories → need training data
- Imperfect results (also in product matching)
  → need cleanup imMeDiAtELy!!!1elf
- **InverseFE** to move offers between products (and categories)
- Re-purposed cleaned up offers as training data


157669995 - KG39EAI40 iQ500 (Suchen Live) Siemens ~728.22EUR
2011-03-12 03:53 Relevanz: 8.46
11 Produktbewertungen
Kühlschränke & Kombis > Kühl-Gefrier-Kombinationen Editieren Splitten
2017-01-12 12:27 Siemens Kg39Eai40, Kühl-/Gefrierkombination, edelstahl - DL - Suchen
682.50EUR bei Rakuten.de (11359) in Haushaltsgeräte > Kühlschränke & Gefrierschränke - Kühlschränke & Kombis > Kühl-Gefrier-Kombinationen
2017-01-12 11:19 Siemens KG39EAI40 iQ500 Kühl-Gefrier-Kombination / A+++ / 337 l / Edelstahl / Inox-antifingerprint / rechts wechselbar Türöffnung - DL - Suchen
621.11EUR bei Amazon Marketplace Major Appliances (15171)
Relevanz: 0.00
2017-01-12 10:45 Siemens KG39EAI40 iQ500 Kühl-Gefrier-Kombination / A+++ / 337 l / Edelstahl / Inox-antifingerprint / rechts wechselbar Türöffnung - DL - Suchen
584.98EUR bei amazon.de (321) in Elektro-Großgeräte/Kategorien/Kühl-Gefrier-Kombinationen/Klassisch - Kühlschränke & Kombis > Kühl-Gefrier-Kombinationen
956384482 - To Go-Becher caffe.de - doppelwandig- 300ml - 100 Stück (Suchen Live) OPAG ~3.90EUR
2016-12-05 13:32
Zubehör für Gastronomiebedarf Editieren Splitten
2017-01-12 13:35 To Go-Becher caffe.de - doppelwandig- 300ml - 100 Stück - DL - Suchen
3.90EUR bei Espressissimo.de (9913) in Espresso>Vending - Zubehör für Gastronomiebedarf -
Relevanz: 0.32
Liefert: image
Gelber Hintergrund: rev1
Angebotsrelevanz
Grüner Hintergrund: rev2
Roter Hintergrund: rev0
Freigegebene Produktbewertung
Gestrichelter Rahmen: Produkt wird exportiert
Kein gestrichelter Rahmen: Produkt wird nicht exportiert
Grüner Strich am linken Rand: Gesuchter Produkt-Container
Roter Strich am rechten Rand: Angebot offline
Rot hinterlegtes Datum: Datum letzter Datenänderung des Shop-Angebots



Produktbild
Produkt-Container-ID
Bezeichnung des Produkt-Containers = Link zum Matching-Frontend
Bezeichnung des Produkt-Containers als Suchbegriff für neue Inverse-Suche
Link zum PSB auf billiger.de
Brand = Hersteller-Marke
Produkt-beschreibung des Produkt-Containers
Angebotsbild
Angebots-beschreibung
Angebots preis
157669995 - KG39EAI40 iQ500 (Suchen Live) Siemens
~728.22EUR
2011-03-12 03:53 Relevanz: 8.46
Die Kühl-/Gefrierkombination Siemens KG39EAI40 bietet ein schlankes Design mit Türen in Edelstahl und Antifinger-Print. Fünf höhenverstellbare Ablageflächen, eine Multifunktionsablage und eine CrisperBox mit Feuchteregulierung sorgen für einen aufgeräumte...
11 Produktbewertungen
Kühlschränke & Kombis > Kühl-Gefrier-Kombinationen
Editieren
Splitten
2017-01-12 12:27 Siemens Kg39Eai40, Kühl-/Gefrierkombination, edelstahl - DL - Suchen
(4242003543542 - 1780623555 - KG39EAI40 - ?)
Die Kuehl-/Gefrierkombination Siemens KG39EAI40 bietet ein schlankes Design mit Tueren in Edelstahl und Antifinger-Print. Fuenf hoehenverstellbare Ablageflaechen, eine Multifunktionsablage und eine Cr...
682.50EUR bei Rakuten.de (11359) in Haushaltsgeräte > Kühlschränke & Gefrierschränke - Kühlschränke & Kombis > Kühl-Gefrier-Kombinationen
Deeplink zum Angebot im Shop
Suchbegriff für neue Inverse-Suche nach der Angebotsbezeichnung
Shopname & Shop_ID
Verlinkung zu Shopimportinfo
Shopkategorie nach dem Bindestrich Billiger.de Kategorie
1. EAN des Angebots: European Article Number
2. AID (Artikel-ID des Shop-Angebots)
3. MPNR (Herstellerartikelnummer)
4. PZN (Pharmazentralnummer)
Alle können als Suchbegriff verwendet werden!
Produkt- oder Angebotsbezeichnung = Link zu Matching Produktzuordnung

# New Feature: Price

- Price as token is too sensitive: 9.99€ != 10.00€
- Token cardinality!
- Price as float still too sensitive
- Sigmoid didn't work (forgot why)
- Solution: Price Bin (0..5€, 5-10€, …) as "token"
- Works well, often high weights

# New Feature: Shop Cat. Mk. 1

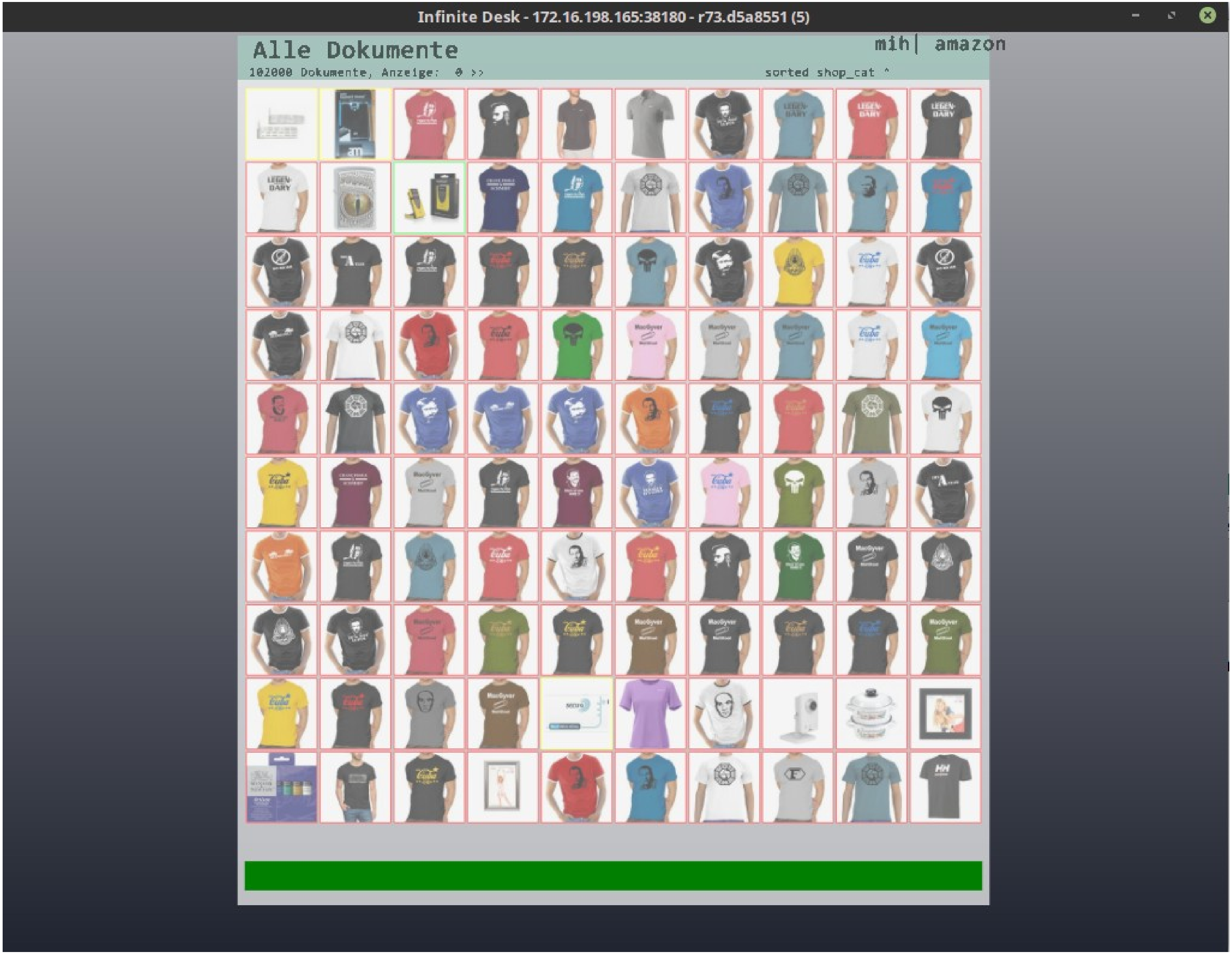
- Generate token from Shop ID and Shop Category → unique shop category
- MD5 over Shop ID and Shop Cat.
- Worked well (precision *and* recall)…
- ... until we
  - … got new, unlabelled shops
  - … had categories finer than shop cats
- Generalizes poorly to new shop cats

# New Feature: Shop Cat Mk. 2

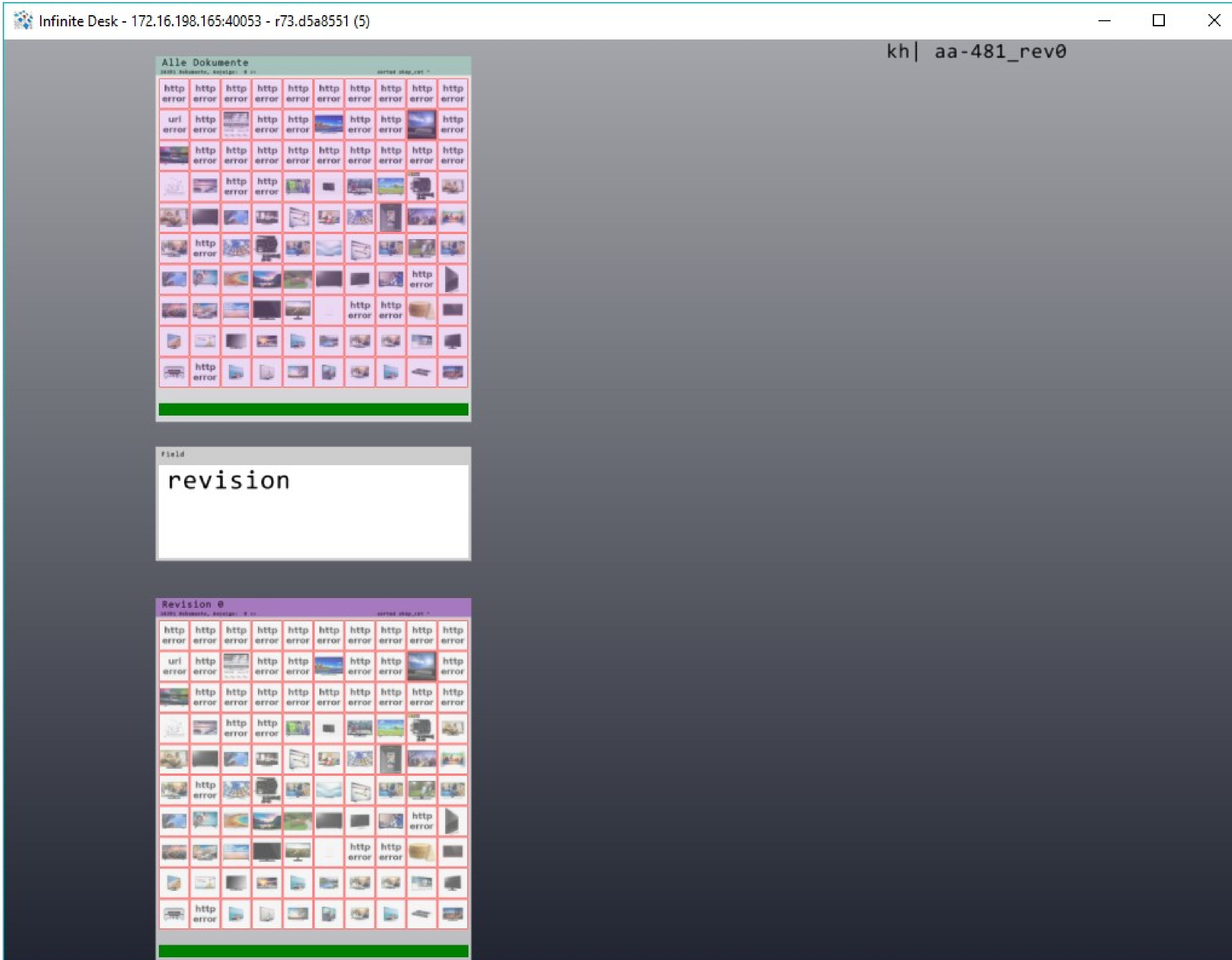
- Better generalizing solution:
- Normal tokenization
- … with “SC_” prefix for shop cat tokens
- No more Shop ID → generalizes over shops

# Fast Shop Cleanup

- Shops are new or complain about category classification → need tool to clean up offers of one (potentially large) shop, quick.
- Remapping Tool: cut (via search & filter) and assign category manually → labeled data.



Infinite Desk - 172.16.198.165:38180 - r73.d5a8551 (5)
Alle Dokumente
mih| amazon
102000 Dokumente, Anzeige: 0 >>
sorted shop_cat ^


Infinite Desk - 172.16.198.165:40053 - r73.d5a8551 (5)
kh| aa-481_rev0
Alle Dokumente
Field
revision
Revision 0

solute


Infinite Desk - 172.16.198.165:59401 - r73.d5a8551 (5)
pk| kinder-spiel
Live!
strazak gefunden
1 Dokumente, Anzeige: 0
sorted shop_cat ^
http error
strazak nicht gefunden
3977 Dokumente, Anzeige: 0 >>
http error
55.10 EUR
DOC ID: 3506720279
ID: 710550445
Feed ID: 0
Brand: smoby
Name: Strazak Sam Array aus Kunststoff (7600410604)
ShopKat: Baby & Spielwaren > Spiel & Spielwaren > Draußen spielen > Spielplätze
EAN: 3032164106042
Shop: 18026 - my.shopping.de
Kategorie: Freizeit & Musik > Basteln & Handarbeit > Malen
Beschreibung
Smoby Strazak Sam Array aus Kunststoff (7600410604)

# Growth (Pains)

- Matching DB: 20 mio offers → 200 mio offers, MySQL → Postgres
- Rebuilding the training binary: cur.fetchall("SELECT *") → own map/reduce + MogileFS → Nokia Disco → Deltas
- Updating training docs → compute new + more (b/c more categories) models: parallelize over multiple machines ("mapping training service")
- Processes around Category Classification: decisions by SEO vs. taxonomy; high turnover → loss of best practices; nearshoring vs. Inverse vs. Remapping Tool


THIS IS FINE.
Rothaus PILS