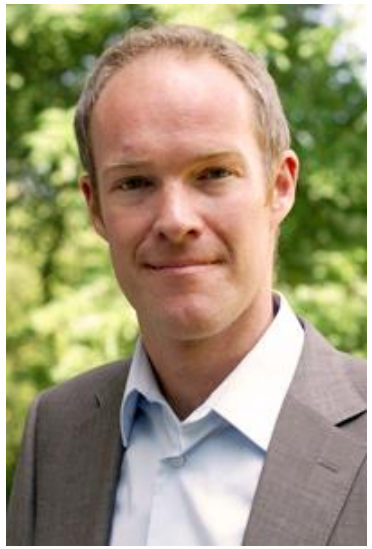
UNIVERSITY OF MANNHEIM

# Team Project HWS 2018

# Integrating Product Data from the Web

# Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Web Data Integration
  - Data and Web Mining
  - Linked Data Technologies
- Room: B6 - B1.15
- eMail: chris@informatik.uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30



# Hallo

- **Anna Primpeli**
- Graduate Research Associate
- Research Interests:
  - Data Extraction
  - Web Data Integration
  - Active Learning
  - Structured Data on the Web

- Room: B6, 26, C 1.04
- eMail: anna@informatik.uni-mannheim.de

# Agenda of Today‘s Kickoff Meeting

1. Introduction and Project Goals
2. The Product Data Corpus
3. Organization and Schedule
4. Specific Subtasks

# Motivation of the Team Project

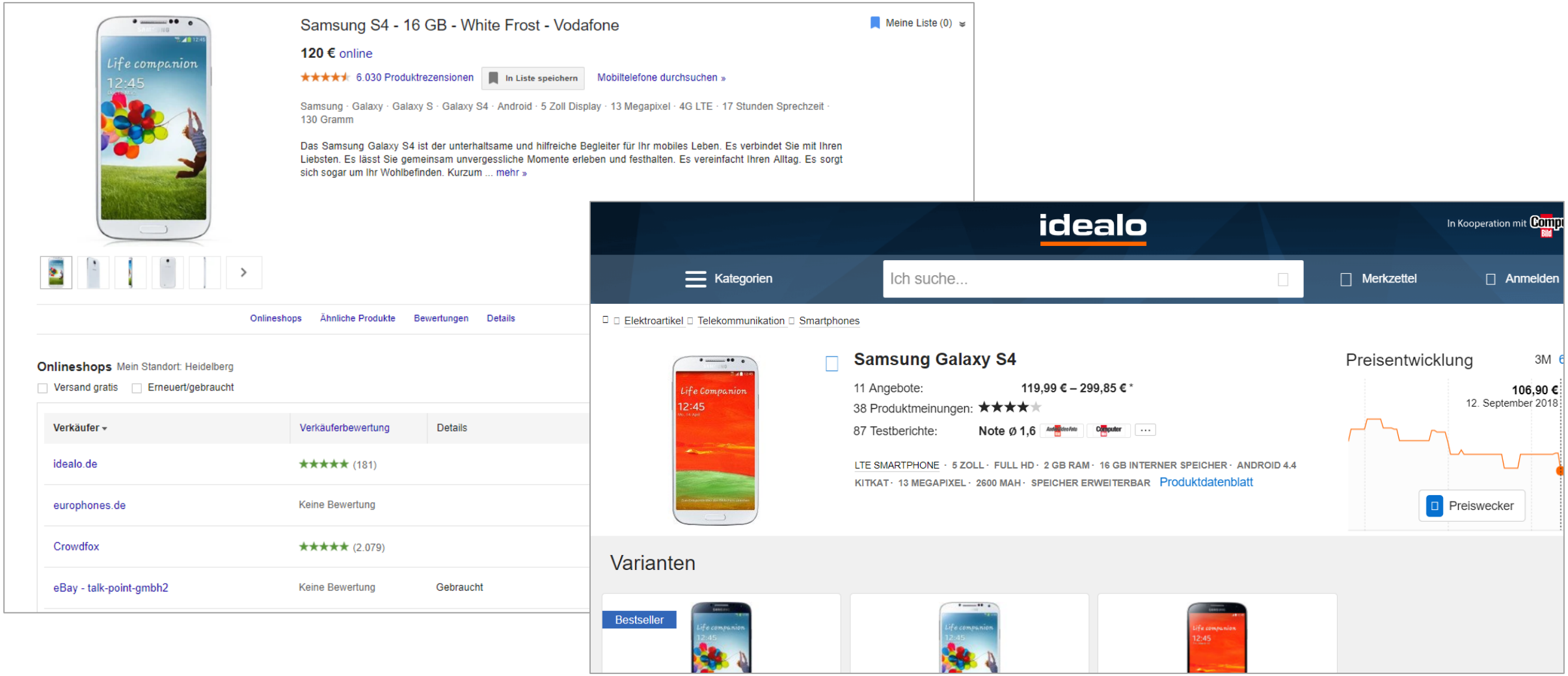
## The Web is a rich source of product information

- the same product is described by 100s of websites
  - by merchants (offers)
  - the producer (product specs)
  - by consumers (reviews and ratings)


## If we can determine which pages describe a specific product (identity resolution), we are able to

- build comprehensive product catalogues and search engines
- conduct global price comparison engines
- understand market structure and consumer preferences

# Identity Resolution is the Key Task for Downstream Applications



# Features that help us to Distinguish Products on the Web

- Product Identifiers
  - GTINs, UPCs, ISBNs, MPN, ….
- Product Titles
  - Product name plus selected features
- Product Descriptions
  - long free texts
- Specification Tables and Lists
  - Detailed features as key/value pairs
- Product Pictures

Samsung Galaxy S4 Verizon AT&T T-Mobile GSM Unlocked Smartphone SRF


| | |
|---|---|
| Produkttyp | Smartphone |
| Formfaktor | Touch |
| Integrierte Komponenten | Rückwärtige Kamera, Frontkamera, |
| Breite | 70 mm |
| Tiefe | 7.9 mm |
| Höhe | 137 mm |
| Gewicht | 130 g |
| Gehäusefarbe | White Frost |

Das Samsung Galaxy S4 ist der unterhaltsame und hilfreiche Begleiter für Ihr mobiles Leben. Es verbindet Sie mit Ihren Liebsten. Es lässt Sie gemeinsam unvergessliche Momente erleben und festhalten. Es vereinfacht Ihren Alltag.

Before we can use these features:

- values need to be cleansed and normalized
- we might want to apply information extraction in order to increase the structuredness of data

# Difficulty of the Task depends on the Product Category

- Books
  - wide adoption of identification schema (ISBNs)
  - problem mostly solved ☺
  - other features like title and author often only used for sanity checks

- Phones / Computers / Cameras
  - rather structured descriptions, often including tables/lists
  - different sites often describe same features

  ➔ identity resolution methods for structured data can be applied


- Cloths / Bags / ….
  - rather unstructured descriptions, not too many tables/lists
  - only weak agreement of attributes

  ➔ identity resolution / disambiguation methods for texts need to be applied (bag of words methods)

# Identity Resolution Methods

| | |
|---|---|
| Text-oriented<br>(bag of words) | Attribute-oriented<br>(weighted matching rules) |
| Unsupervised<br>(TF/IDF+Cosine) | Supervised<br>(random forest) |
| Symbolic<br>(matching rules) | Sub-symbolic<br>(embeddings) |

# Project Goals

- Integrate product data from a large number of websites, using:
  - different identity resolution methods
  - different information extraction methods

- Compare performance of methods w.r.t.:
  - product categories (structured vs. semi-structured input)
  - product popularity (head vs. tail products)

- Evaluate the usefulness of weak supervision that is found on the Web for product matching
  - weak supervision = product identifiers such as GTINs

# Learning Targets

**Improve your technical skills**

- Work as a **Data Scientist**: clean, profile, integrate, classify data, classify record pairs
- Understand the nature of **Web Data**
- Improve your technical expertise / programming skills

**Improve your soft skills**

- Work as part of a bigger team on a more complex project
- Organize yourself and assign tasks based on your skills
- Communicate and coordinate your work

# 2. The Product Data Corpus

1. Semantic Annotations in HTML Pages
2. Web Data Commons Project
3. Web Data Commons – Silver Standard for Large-Scale Product Matching

# Semantic Annotation of HTML Pages: Schema.org

schema.org

Search

Home Schemas Documentation

## Thing > Organization > LocalBusiness

A particular physical business or branch of an organization. Examples of LocalBusiness include a restaurant, a particular branch of a restaurant chain, a branch of a bank, a medical practice, a club, a bowling alley, etc.

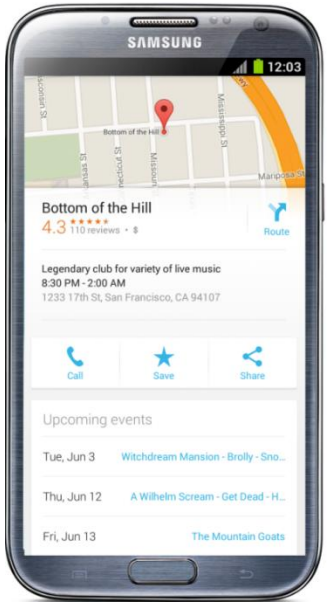
| Property | Expected Type | Description |
|---|---|---|
| **Properties from Thing** | | |
| description | Text | A short description of the item. |
| image | URL | URL of an image of the item. |
| name | Text | The name of the item. |
| url | URL | URL of the item. |
| **Properties from Place** | | |
| address | PostalAddress | Physical address of the item. |
| aggregateRating | AggregateRating | The overall rating, based on a collection of reviews or ratings, of the item. |
| containedIn | Place | The basic containment relation between places. |

- ask site owners since 2011 to annotate data for enriching search results
- 675 Types: Event, Place, Local Business, Product, Review, Person
- Encoding: Microdata, RDFa, JSON-LD

# Usage of Schema.org Data @ Google

**Gramercy Tavern - Flatiron - New York, NY | Yelp**
www.yelp.com › Restaurants › American (New)
★★★★½ Rating: 4.5 - 1,288 reviews - Price range: $$$$
Jeff C and I were in **New York** for vacation, and I wanted to treat him to a nice dinner for ..... **Gramercy Tavern** is certainly a legendary **NY** dining establishment.

**Gramercy Tavern Restaurant - New York, NY | OpenTable**
www.opentable.com › ... › Gramercy restaurants
★★★★½ Rating: 4.7 - 508 reviews - Price range: $50 and over
Book now at **Gramercy Tavern** in **New York**, explore menu, see photos and read 508 reviews: "The menu was so limited but it was worth trying, food was deli..."




Data snippets within search results

Data snippets within info boxes


## The Black Keys

Band

The Black Keys is an American rock duo formed in Akron, Ohio in 2001. The group consists of Dan Auerbach and Patrick Carney. Wikipedia

**Origin:** Akron, Ohio, United States

**Members:** Dan Auerbach, Patrick Carney

**Record labels:** Fat Possum Records, Nonesuch Records, V2 Records, Alive Naturalsound Records

**Awards:** Grammy Award for Best Rock Album, more

### Upcoming events

| | |
|---|---|
| Jun 20 Fri | The Black Keys<br>Neuhausen ob Eck (near you) |
| May 16 Fri | The Black Keys<br>Gulf Shores, AL |
| Jun 22 Sun | The Black Keys<br>Scheeßel |

# Example: Microdata Annotations in HTML

```
<div  itemtype="http://schema.org/Hotel">
  <span itemprop="name">Vienna Marriott Hotel</span>
  <span itemprop="address" itemscope="" itemtype="http://schema.org/PostalAddress">
        <span itemprop="streetAddress">Parkring 12a</span>
        <span itemprop="addressLocality">Vienna</span>
 </span>
   <div itemprop="aggregateRating" itemscope itemtype="http://schema.org/AggregateRating">
         <span itemprop="ratingValue"> 4 </span> stars-based on
         <span itemprop="reviewCount"> 250 </span> reviews.
   </div>
```

# Product-related schema.org Classes

## Review

Canonical URL: http://schema.org/Review

Thing > CreativeWork > Review

A review of an item – for example, of a restaurant, movie, or store.

Usage: Between 250,000 and 500,000 domains

| Property | Expected Type | Description |
|---|---|---|
| **Properties from Review** | | |
| itemReviewed | Thing | The item that is bei |
| reviewAspect | Text | This Review or Ratin |
| reviewBody | Text | The actual body of t |
| reviewRating | Rating | The rating given in t rating given by the r |
| **Properties from CreativeWork** | | |
| about | Thing | The subject matter Inverse property: su |
| accessMode | Text | The human sensory information. Expect chemOnVisual, diag |
| accessModeSufficient | Text | A list of single or co resource. Expected |
| accessibilityAPI | Text | Indicates that the re possible values). |
| accessibilityControl | Text | Identifies input met possible values). |

## Product

Canonical URL: http://schema.org/Product

Thing > Product

Any offered product or service. For example: a pair of shoes; a concert ticket; the rental of a car; a haircut; or an episode of a TV show streamed online.

Usage: Over 1,000,000 domains

[more...]

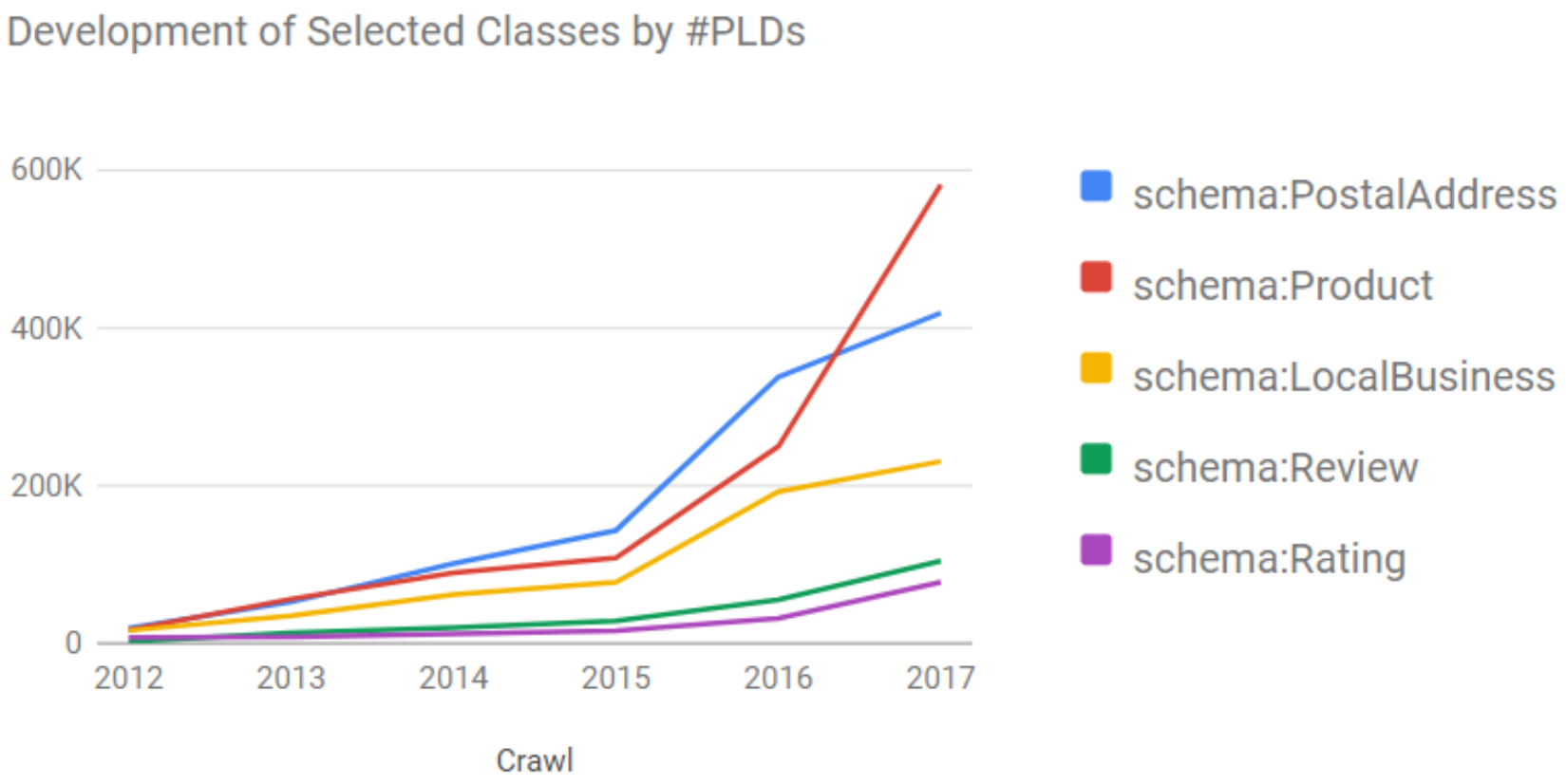
| Property | Expected Type | Description |
|---|---|---|
| **Properties from Product** | | |
| additionalProperty | PropertyValue | A property-value pair representing an additional characteristics of the entity, e.g. a product feature or another characteristic for which there is no matching property in schema.org.<br>Note: Publishers should be aware that applications designed to use specific schema.org properties (e.g. http://schema.org/width, http://schema.org/color, http://schema.org/gtin13, ...) will typically expect such data to be provided using those properties, rather than using the generic property/value mechanism. |
| aggregateRating | AggregateRating | The overall rating, based on a collection of reviews or ratings, of the item. |
| audience | Audience | An intended audience, i.e. a group for whom something was created. Supersedes serviceAudience. |
| award | Text | An award won by or for this item. Supersedes awards. |
| brand | Brand or Organization | The brand(s) associated with a product or service, or the brand(s) maintained by an organization or business person. |
| category | PhysicalActivityCategory or Text or Thing | A category for the item. Greater signs or slashes can be used to informally indicate a category hierarchy. |
| color | Text | The color of the product. |
| depth | Distance or QuantitativeValue | The depth of the item. |

Meusel, Robert, Petar Petrovski, and Christian Bizer. "The webdatacommons microdata, rdfa and microformat dataset series." *International Semantic Web Conference*. Springer, Cham, 2014.

# The Web Data Commons Project

- extracts all Microformat, Microdata, RDFa, JSON-LD data from the Common Crawl
- analyzes and provides the extracted data for download
- statistics about some extraction runs
  - 2017 CC Corpus: 3.1 billion HTML pages → 38.2 billion RDF triples
  - 2016 CC Corpus: 3.1 billion HTML pages → 44.2 billion RDF triples
  - 2014 CC Corpus: 2.0 billion HTML pages → 20.4 billion RDF triples
  - 2012 CC Corpus: 3.0 billion HTML pages → 7.3 billion RDF triples
- uses 100 machines on Amazon EC2
  - approx. 2000 machine/hours
    (100 spot instances of type c3.xlarge) ➔ 350 Euro
- http://www.webdatacommons.org/structureddata/

# schema.org Annotations: Most Popular Classes




http://webdatacommons.org/structureddata/

# Properties used to Describe Products 2017

| Top 15 Properties | PLDs | |
|---|---|---|
| | # | % |
| schema:Product/name | 535,625 | 92% |
| schema:Offer/price | 462,444 | 80% |
| schema:Product/offers | 462,233 | 79% |
| schema:Offer/priceCurrency | 430,556 | 74% |
| schema:Product/image | 419,391 | 72% |
| schema:Product/description | 377,639 | 65% |
| schema:Offer/availability | 337,876 | 58% |
| schema:Product/url | 263,720 | 45% |
| schema:AggregateRating/ratingValue | 184,004 | 32% |
| schema:Product/sku | 126,696 | 22% |
| schema:AggregateRating/reviewCount | 112,408 | 19% |
| schema:Product/aggregateRating | 101,434 | 17% |
| schema:Product/brand | 73,934 | 13% |
| schema:Product/productID | 35,211 | 6% |
| schema:Product/manufacturer | 21,967 | 4% |

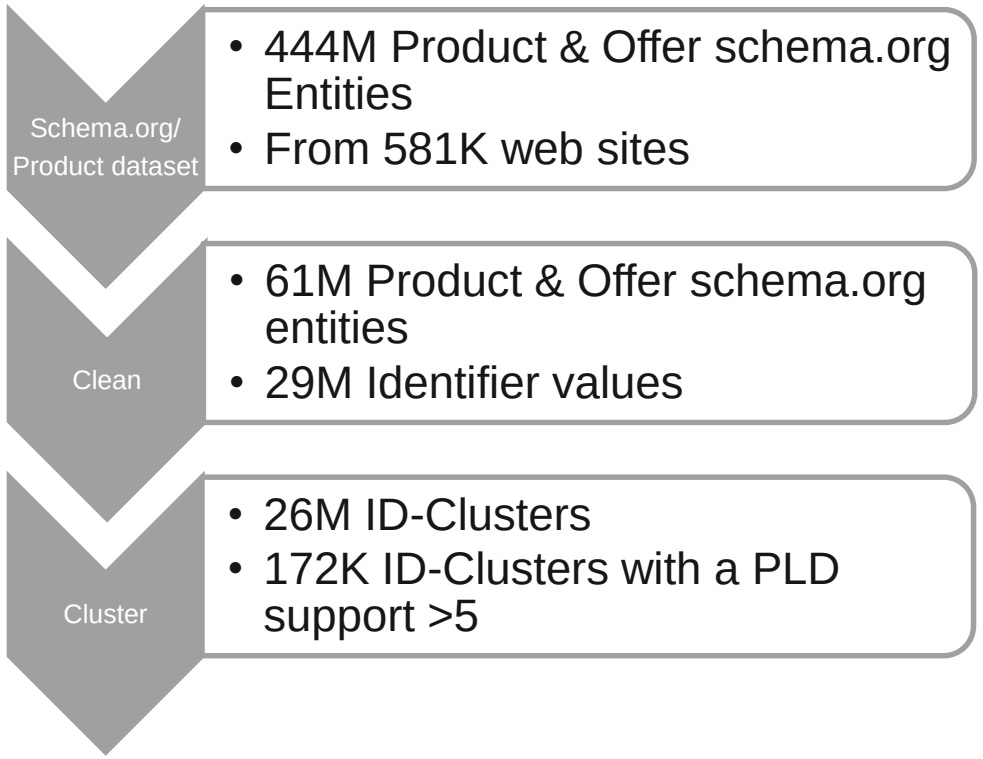
Samsung Galaxy S4 Verizon AT&T T-Mobile GSM Unlocked Smartphone SRF

Das Samsung Galaxy S4 ist der unterhaltsame und hilfreiche Begleiter für Ihr mobiles Leben. Es verbindet Sie mit Ihren Liebsten. Es lässt Sie gemeinsam unvergessliche Momente erleben und festhalten. Es vereinfacht Ihren Alltag.

UPC 610214632623

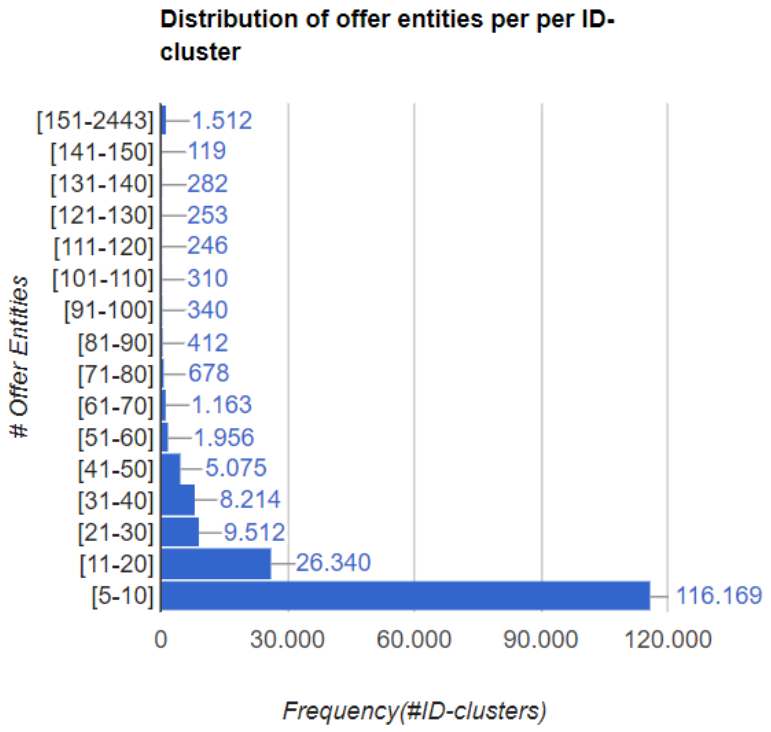
# The WDC Silver Standard for Large-Scale Product Matching

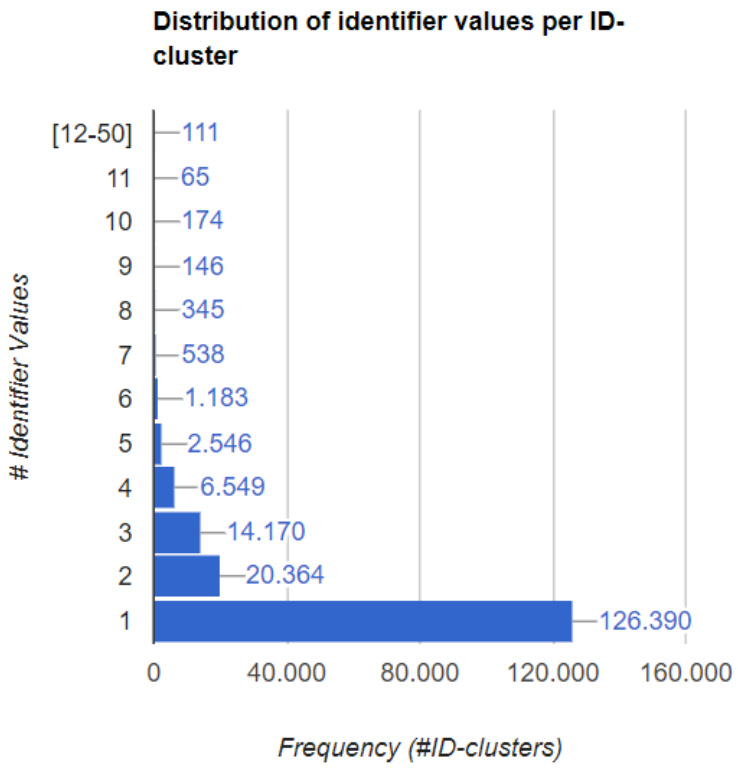
- Silver Standard grouping schema.org product annotations by identifier value.
  - all WDC product data is included that
  - provides come sort of product ID
- Initial cleaning steps are performed
- Results clusters of product descriptions from different websites that share identifier values.

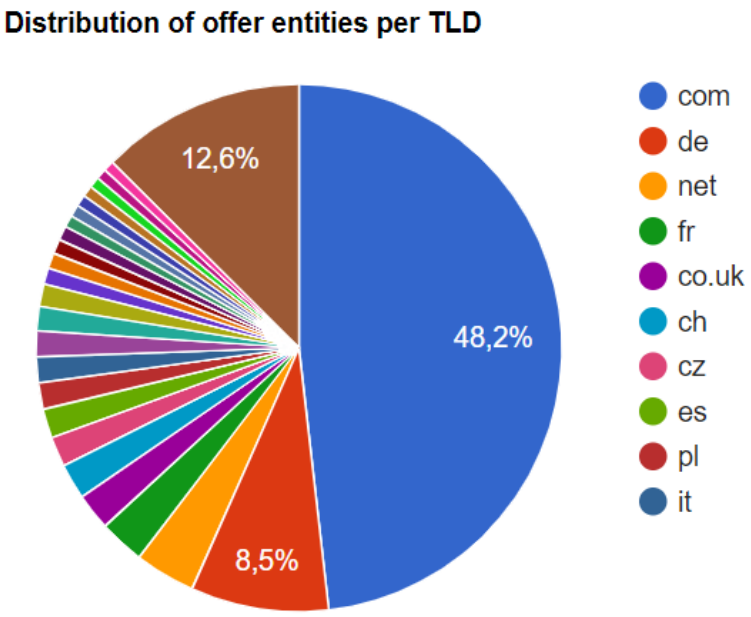



Details and Download
http://webdatacommons.org/largescaleproductcorpus/index.html

# The WDC Silver Standard for Large-Scale Product Matching



# 3. Organization and Schedule

**Duration:** 6 months (28.09.2018 – 28.03.2018)

**ECTS:** 12

**Participants:** 8 people

**Type of work:** Team and subgroup based

**Milestones:** 4 project phases

**Evaluation:**

- Individual contribution to the deliverables
- Deliverables: Presentations, final report, code, data
- Every project phase determines 25% of your final grade

# Questions and Subtasks

1. Which two product categories should we use? → **Corpus Profiling, Data Selection**
2. Is the corpus for these categories dense and clean enough? → **Corpus Profiling, Additional Crawling**
3. How do supervised and unsupervised IR methods perform?→ **Basic IR**
4. Which features should be extracted to help IR? → **Feature Extraction**
5. How do IR methods performs given cleaner features? → **IR with Feature Extraction**
6. How do your different methods compare? → **Comparision of Results**

# Main Steps of the Project

**Data Collection**

↓

**IR with BoW**

↓

**Feature Extraction**

↓

**IR with Enhanced Features**

↓

**Comparison of Results**

# Detailed Schedule

| Date | Session |
|---|---|
| **Friday, 28.09.2018** | Kickoff meeting (today) |
| | **Phase 1 (all members): Data Collection** |
| **Friday, 12.10.2018** | Meet Anna and report current results |
| **Friday, 12.10.2018** | Drop-out deadline: Dropping out after this date will result in failing the team project |
| **Friday, 26.10.2018** | **1st Deliverable:** 20 minutes presentation, data |
| | **Phase 2 (in subgroups): IR with BoW + Gold Standard Creation** |
| **Friday, 16.11.2018** | Meet Anna and report current results |
| **Friday, 14.12.2018** | **2nd Deliverable:** 20 minutes presentation from each subgroup, code & data |
| | **Phase 3 (in subgroups): Feature Extraction** |
| **Friday, 11.01.2019** | Meet Anna and report current results |
| **Friday, 25.01.2019** | **3rd Deliverable:** 20 minutes presentation from each subgroup, code & data |
| | **Phase 4 (in subgroups): IR with Features** |
| **Friday, 01.03.2019** | Meet Anna and report current results |
| **Monday, 25.03.2019** | **4th Deliverable:**15-20 pages overall report, code & data |
| **Friday, 29.03.2019** | Overall presentation 30 min + Feedback |

# Phase 1: Data Collection

**Participants:** All team members

**Duration:** 28.09.2018 – 26.10.2018

**Deliverables:** 20 minutes presentation, data, report who did what

**Tasks**

1. Decide on two product categories
   - Select 2 non-similar product categories, one structured, one less structured e.g. laptops and shoes (**NOT** phones, headphones, TVs)
2. Decide on a set of products
   - Collect a set of products from each category together with respective IDs (>100 products/category)
3. Create your subcorpus
   - Identify the relevant ID-clusters from WDC Large-Scale GS using the product identifiers
   - Profile the data / perform additional cleansing steps / maybe crawl additional data
   - Report detailed statistics about the initial and final subcorpus (cluster sizes, feature frequency)

# Expected Result of Phase 1

- Product offer dataset as basis for evaluating different identity resolution methods in the following phases.
- Expected profile of your data set
  - >=2 categories
  - >100 products per category
  - >10 and median 20 pages from different PLDs per product
    - majority of PLDs should be .com/ co.uk
  - All pages should contain schema:title, schema:description, and a product ID (not necessarily annotated).
  - One category: Rather structured product descriptions containing detailed specification tables/lists
  - Other category: Less structured descriptions, not necessarily containing tables/lists.
- Expected format of the dataset
  - Same format as WDC Gold Standard for Product Matching and Product Feature Extraction
  - http://www.webdatacommons.org/productcorpus/

# Phase 2 : Identity Resolution with BoW Models + Gold Standard Creation

**Duration:** 26.10.2018 – 30.11.2018

**Participants:** 2 subgroups of 4 persons each

**Deliverables:**
– 20 minutes presentation from each subgroup, data & code, report who did what

**Tasks**

Use BoW models with different input data and apply:

**Subgroup 1: Unsupervised IR methods**

- TF/IDF+cosine, embeddings, domain-specific heuristics

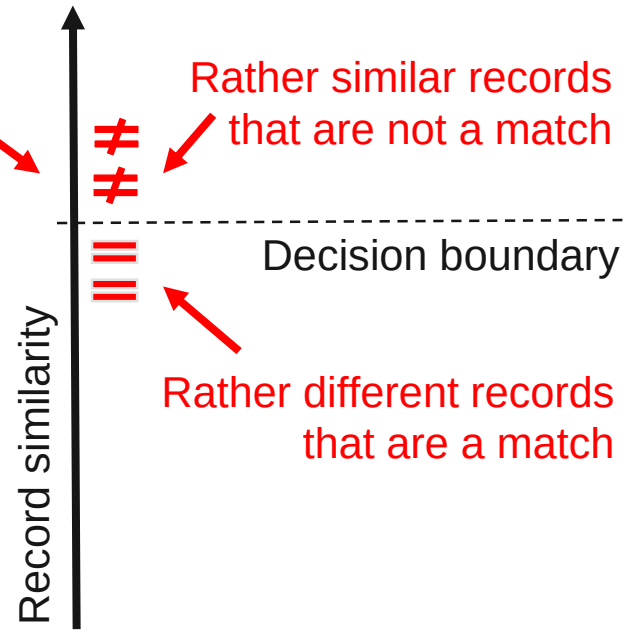
**Subgroup 2: Supervised IR methods**

- word weights, decision trees, random forrests, deep learning

Evaluate on the WDC Silver Standard using a category-specific gold standard.

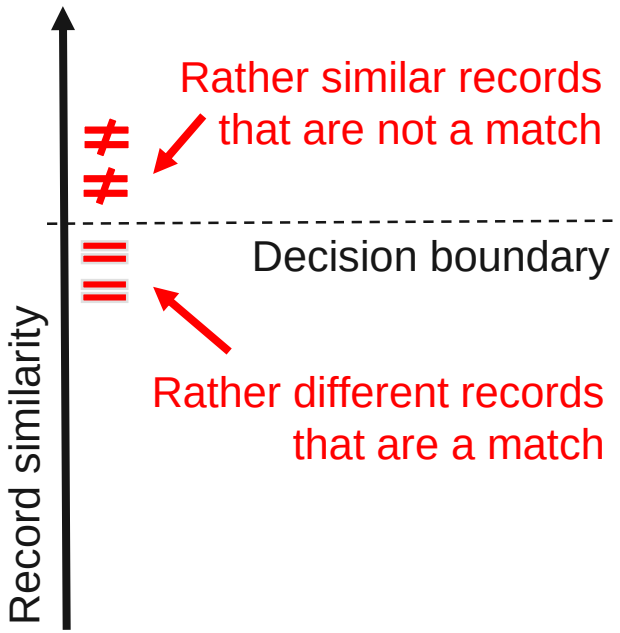
# How to Build an Good Gold Standard?

- You need ground truth (gold standard) for your evaluation.
- To create a gold standard, manually label a set of record pairs as matches or non-matches including corner cases
- Rule of thumb for creating an *interesting* gold standard with *acceptable* manual effort:
  1. match records using several simple matching techniques (similar to multi-pass blocking) and
  2. reuse existing information about matches (e.g. ISBN or GTIN numbers that exist in multiple sources)
  3. <u>manually</u> verify a fair amount of the resulting pairs (e.g. 500 pairs) including
     1. matching record pairs (randomly chosen, 20% of GS)
     2. corner cases (30% of GS)
     3. non-matching record pairs (randomly chosen, 50% of GS)



# How Select Pairs of Offers for the Gold Standard?

1. Sort ID-Clusters within each product category by amount of offers.
2. Randomly choose 100 large clusters and 50 mid-size clusters.
3. Calculate the Jaccard-Similarity of the product titles within each cluster and sort pairs by similarity.
4. Calculate the Jaccard-Similarity of a product title from each cluster and the product titles of all offers within oter clusters of the same category.
5. Manually verify the following pairs for your gold standard:
   1. Normal positives: Randomly choose 1 matching pair from the top-3 similarity quantiles of each cluster.
   2. Corner cases: Randomly choose 1 pair from the fourth similarity quantile of each cluster.
   3. Negatives: Randomly choose 3 pairs top-2 similarity quantiles of the out of cluster pairs.



# Phase 1 (update) and 2: Checklist

Phase 1

- A training corpus consisting of groups of offers with the same identifiers (ID-Clusters)
- Profiling results about the training corpus

Phase 2

- One gold standard per product category including corner cases
- Unsupervised IR methods on your gold standards
- Supervised IR methods which use your training corpus for learning
- Results and error analysis for unsupervised and supervised IR

# Phases 3: Feature Extraction

**Duration:** 14.12.2018 – 25.01.2018

**Participants:** 2 subgroups of 4 persons each

**Deliverables:**
– 20 minutes presentation from each subgroup, data & code, report who did what

**Tasks**

Apply advanced feature extraction methods

**Subgroup 1: Closed Feature Extraction methods**

- dictionary based using auxilary data for product properties and values

**Subgroup 2: Open Feature Extraction methods**

- Exploit HTML tables and HTML lists
- Perform schema matching on extracted data

**DO NOT create a wrapper for every page!**

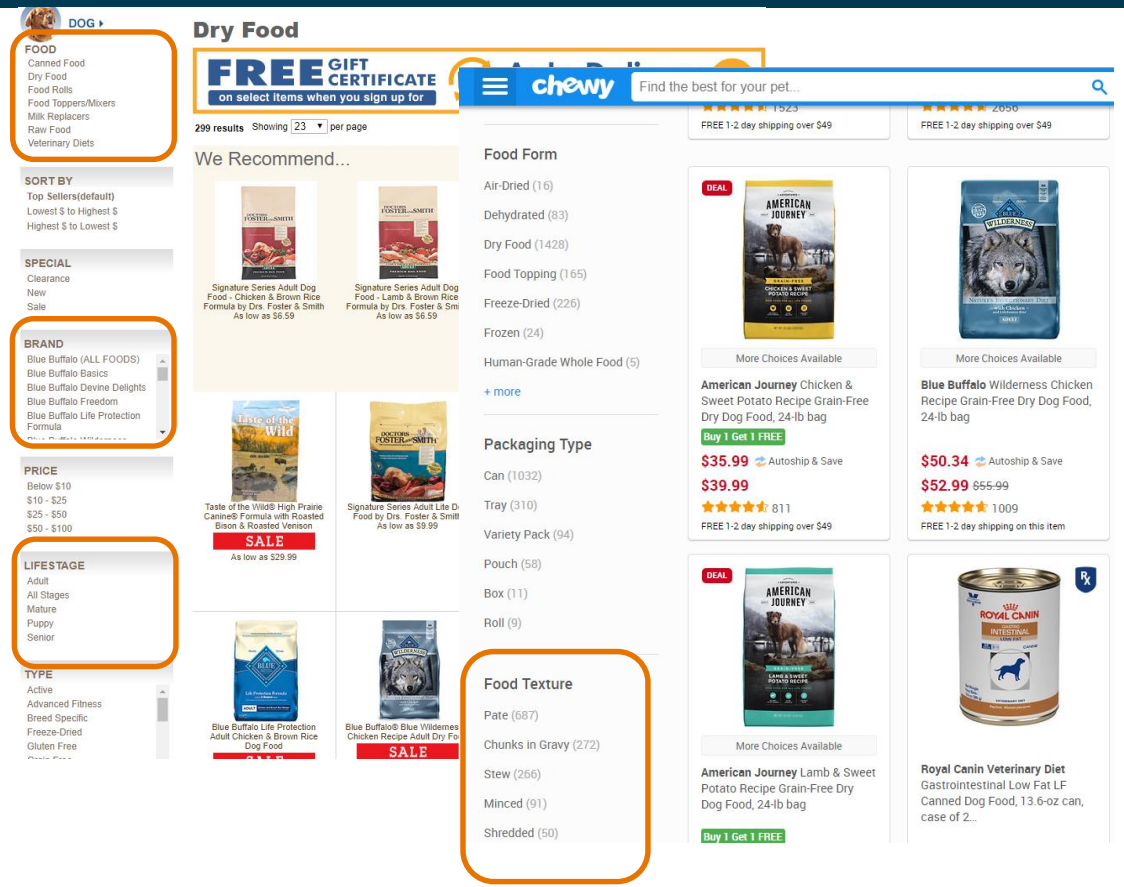
Evaluate your feature extraction techniques

# Closed Feature Extraction

1. Create a dictionary for every product category

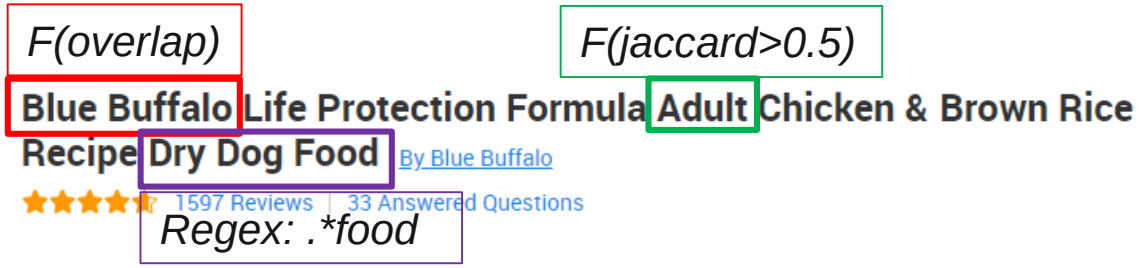
**HOW?** You can exploit the filters of related webpages and create an inverted index

| Attribute | Values |
|---|---|
| Food type | Canned food, dry food, raw food ... |
| Brand | Blue Buffalo, WholeHearted... |
| Lifestage | adulthood, all stages ... |
| Packaging | Can, tray... |




2. Use the dictionary to extract features from your offers

- Directly: overlap, window + sim.metric
- With regexes: .*food →Food type



# Open Feature Extraction

1. Consider HTML tables and extract key value pairs

2. Classify tables as specification and non-specification ones.

- Using simple heuristics: e.g. any HTML table with two columns is a specification one
- Learn a classification model using HTML table features
- Apply existing learned models (see [1,2,3] )

3. Perform N→1 schema matching using your dictionary (see previous slide)

- label based
- instance based
- duplicate-based schema matching (using the ID-Clusters)

| | |
|---|---|
| ITEM NUMBER | 46861 |
| WEIGHT | 30.0 pounds |
| BRAND | Blue Buffalo |
| BRAND | Blue Life Protection |
| BREED SIZE | Small Breeds, Medium Breeds, Large Breeds |
| FOOD FORM | Dry Food |
| LIFESTAGE | Adult |
| MADE IN | United States |
| SPECIAL DIET | No Corn No Wheat No Soy |

| Attribute | Values |
|---|---|
| Food type | Canned food, dry food, raw food ... |
| Brand | Blue Buffalo, WholeHearted... |
| Lifestage | adulthood, all stages ... |
| Packaging | Can, tray... |

*Example specification table from*

Rating Snapshot

Select a row below to filter reviews.

| | |
|---|---|
| 5★ | 604 |
| 4★ | 89 |
| 3★ | 14 |
| 2★ | 5 |
| 1★ | 21 |

Most Helpful Favorable Review

```
▼<table class="bv-section-summary-table bv-table-full-width" role=
"presentation"> == $0
  ▼<tbody>
    ▼<tr class="bv-section-summary-inline">
      ▼<td data-bv-v="inlineHistogram:9" class="bv-inline-histogram
      bv-section-summary-block">
        ▼<div class="bv-inline-histogram-ratings">
          ▶<div class="bv-inline-histogram-header">…</div>
           <span class="bv-histogram-filter-helper">Select a row
           below to filter reviews.</span>
```

*Example of a non-specification table from petco.com*

# Evaluate

Create a gold standard for feature extraction

1. Select 20 HTML pages from different websites
2. Annotate **all** product feature-value pairs found in each page
3. Map your features to your dictionary

On Feature Extraction GS: http://webdatacommons.org/productcorpus/index.html#toc3

Evaluate and compare your approaches

- Which approach worked better for which category?
- What went wrong?

[1] Qiu, Disheng, et al. "Dexter: large-scale discovery and extraction of product specifications on the web." *Proceedings of the VLDB Endowment* 8.13 (2015)
[2] Petar Petrovski, Christian Bizer: Extracting Attribute-Value Pairs from Product Specifications on the Web. International Conference on Web Intelligence (WI2017).
[4] Petrovski, et al.: Integrating Product Data from Websites offering Microdata Markup. WWW2014 Workshops.
[5] Zheng, Mukherjee, Dong: OpenTag: Open Attribute Value Extraction from Product Profiles. KDD, 2018.
[6] svn://svn.fmi.uni-mannheim.de/aprimpel/spectables-extractor/wstl-extractor

# Phases 1,2,3 – Checklist (1/2)

Phase 1

- A training corpus consisting of groups of offers with the same identifiers (ID-Clusters)
- Profiling results about the training corpus

Phase 2

- One gold standard per product category including corner cases
- Unsupervised IR methods on your gold standards
- Supervised IR methods which use your training corpus for learning
- Results and error analysis for unsupervised and supervised IR

# Phases 1,2,3 – Checklist (2/2)

Phase 3

- One dictionary per product category
- One gold standard for feature extraction per product category
- Evaluation of your closed feature extraction method
- A method for classifying HTML tables as specification/non-specification
- Evaluation of your open feature extraction method
- Error analysis of your implemented feature extraction methods
- Extracted feature-value pairs from your corpus

# Phase 4: Identity Resolution with Enhanced Features

**Duration:** 25.01.2018 – 25.03.2018

**Participants:** 2 subgroups of 4 persons each

**Deliverables:**
- 30 minutes presentation
- 15-20 pages overall report, data & code, report who did what

**Tasks**

Apply IR methods that exploit the enhanced features (extracted in phase 3)

Compare your results

# Phase 4: Input for IR

## Input for IR Phase 2

| OfferName1 | OfferName2 |
|---|---|
| lenovo - thinkpad x1 carbon 14in laptop - intel core i5 - 8gb memory - 256gb solid state drive - black | lenovo - thinkpad x1 carbon 14in laptop - intel core i7 - 16gb memory - 512gb solid state drive - black |

- BoW model based **on the whole schema.org name/description of the offer**
- Supervised methods based on word co-occurrence or different sim. scores **on the whole name**

| Entity Pair | lenovo | 512gb | token_x |
|---|---|---|---|
| $o_1 - o_2$ | 1 | 0 | 1/0 |
| $o_1 - o_3$ | | | |
| $o_x - o_y$ | | | |

Feature generation based on word co-occurrence

| Entity Pair | Lev(Name) | Monge(Name) | Sim (Name) |
|---|---|---|---|
| $o_1 - o_2$ | 0.9 | 0.7 | $[0-1]$ |
| $o_1 - o_3$ | | | |
| $o_x - o_y$ | | | |

Feature generation sim. score of the complete name

## Input for IR Phase 4

- Calculate sim. scores between the values of the extracted features from phase 3

| Entity Pair | brand | memory | color | feature_x |
|---|---|---|---|---|
| $o_1 - o_2$ | $f_{\text{sim}}(brand)$ | $f_{\text{sim}}(memory)$ | $f_{\text{sim}}(color)$ | $f_{\text{sim}}(feature\_x)$ |
| $o_1 - o_3$ | $f_{\text{sim}}(brand)$ | $f_{\text{sim}}(memory)$ | $f_{\text{sim}}(color)$ | $f_{\text{sim}}(feature\_x)$ |
| $o_x - o_y$ | $f_{\text{sim}}(brand)$ | $f_{\text{sim}}(memory)$ | $f_{\text{sim}}(color)$ | $f_{\text{sim}}(feature\_x)$ |

$f_{\text{sim}}$(brand) : jaccard sim. score

$f_{\text{sim}}$(memory): exact sim. score

$f_{\text{sim}}$(color): levenshtein sim. score

$f_{\text{sim}}$(feature_x): x sim. score

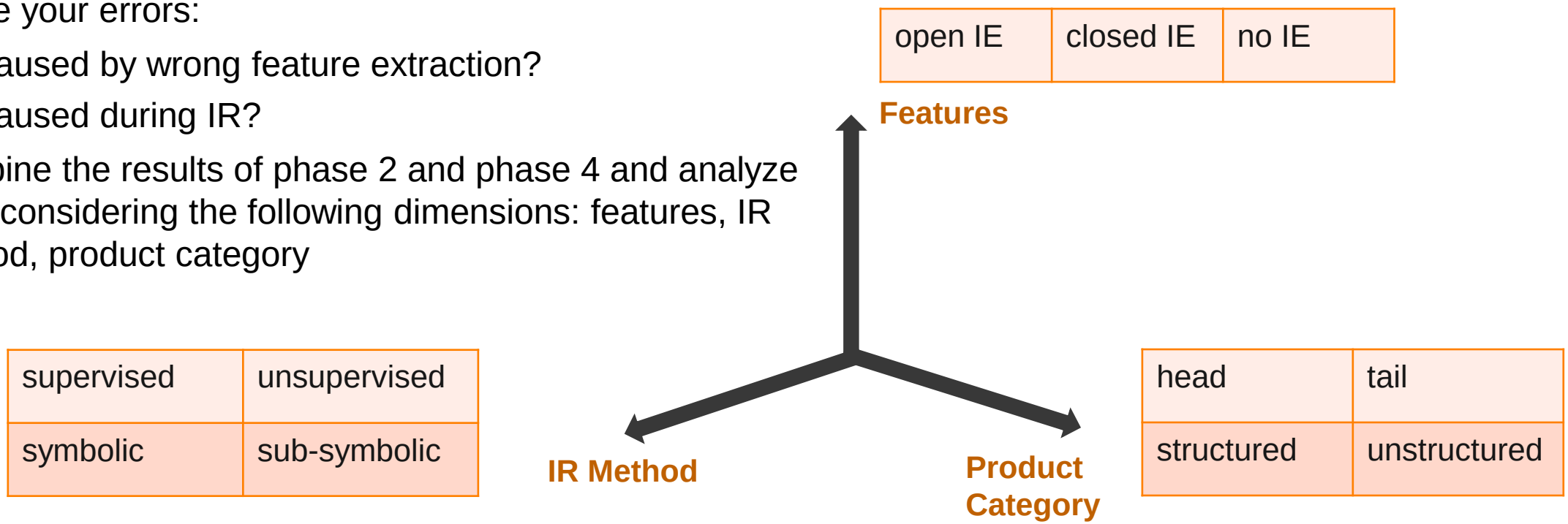
# Phase 4: IR Methods

- Unsupervised: Handwritten rules
  - Try different combinations of features
  - Assign manually weights and thresholds
  - Example: 0.3*$f_{\text{sim}}$(memory) +0.2*$f_{\text{sim}}$(brand) +0.5*$f_{\text{sim}}$(model) > 0.7

- Supervised: Similar to phase 2
  - Learn feature weights (linear, logistic regression)
  - Decision trees
  - Random forests

# Phase 4: Evaluation and Comparison to Phase 2

- Evaluate your results using the gold standard of phase 2
- Analyse your results
  - Compare to the results of phase 2 (make sure the pairs in the training subsets and gs are the same!)
  - Profile your errors:
    - caused by wrong feature extraction?
    - caused during IR?
  - Combine the results of phase 2 and phase 4 and analyze them considering the following dimensions: features, IR method, product category



# Phase 4: WInte.r Framework (1/2)

- You can run unsupervised and supervised IR methods on feature-based input using the WInte.r Framework
- Translate your data to be compatible with WInte.r

Input data

```
<movie>
          <id>1-9311</id>
          <title>Winter's Bone</title>
</movie>
<movie>
          <id>1-9312</id>
          <title>Black Swan</title>
          <date>2011-01-01</date>
          <globe>yes</globe>
</movie>
<movie>
          <id>1-9313</id>
          <title>Blue Valentine</title>
          …
</movie>
```

Gold Standard

```
gold.csv:

1-9309,2-9309,true
1-9310,2-9310,true
1-9311,2-9311,false
1-9312,2-9312,true
1-9313,2-9313,false
1-9314,2-9314,false
1-9315,2-9315,false
1-9316,2-9316,true
```

- Define your data model

# Phase 4: WInte.r Framework (2/2)

– Create handwritten rules (unsupervised IR)

```
LinearCombinationMatchingRule<Movie, Attribute> rule =
  new LinearCombinationMatchingRule<>(0.5);          // final threshold

rule.addComparator(new MovieTitleComparator(), 0.6);  // comparator & weight

rule.addComparator(new MovieDateComparator(), 0.4);   // comparator & weight
```

– Or use the WekaMatchingRule class to learn feature-based rules

```
// create the matching rule
String options[] = new String[] { "-S" };
String modelType = "SimpleLogistic"; // use a logistic regression
WekaMatchingRule<Movie, Attribute> matchingRule
 = new WekaMatchingRule<>(0.5, modelType, options);

// add comparators
matchingRule.addComparator(new MovieDirectorComparatorLevenshtein());
matchingRule.addComparator(new MovieTitleComparatorLevenshtein());

// load the training set
MatchingGoldStandard gsTraining = new MatchingGoldStandard();
gsTraining.loadFromCSVFile(new File("training.csv"));

// train the matching rule's model
RuleLearner<Movie, Attribute> learner = new RuleLearner<>();
learner.learnMatchingRule(dataAcademyAwards, dataActors, null, matchingRule,
gsTraining);
```

– **Detailed docu:** *https://github.com/olehmberg/winter/blob/master/README.md*

# Phase 4: Final Report

- Should include description and results of all phases
- Report structure
  - Introduction - Motivation
  - Phase 1: data collection, dataset profiling, cluster creation, cluster profiling
  - Phase 2: IR gold standard, IR methods, results, error analysis
  - Phase 3: FE gold standard, open FE methods, closed FE methods, results, error analysis
  - Phase 4: IR with enhanced features, results, error analysis, comparison to the results of phase 2, final comparison based on the three dimensions (FE, IR method)
  - Conclusion

# Submission of Deliverables

**Presentation Slides**

- Send slide until the submission deadline.
- The exact time of the presentation will be determined case by case.

**Team and Subgroup Reports**

- Send one e-mail per team or subteam until the deadline date according to the schedule

**Data and Code**

- Add your data and code in a zipped folder and send (URL) via e-mail

**Member to subtask report**

- Send one excel sheet per team explaining who did what together with the deliverables.

All deliverables should be sent to Chris & Anna!

# References: Identity Resolution in General


**Lecture Slides**

- Bizer: **Web Data Integration – Chapter: Identity Resolution**, 2017 (see lecture archive)

**Book Chapters**

- Doan, Halevy, Ives: **Principles of Data Integration**. Chapter 7, Morgan Kaufmann, 2012.
- Peter Christen: **Data Matching**. Springer 2012.
- Naumann, Herschel: **An Introduction to Duplicate Detection**. Morgan & Claypool Publishers, 2010.
- Christophides, et al: **Entity Resolution in the Web of Data**. Morgan & Claypool Publishers, 2015.

**Papers**

- Köpcke, Hanna, Andreas Thor, and Erhard Rahm.: **Evaluation of entity resolution approaches on real-world match problems**. *Proceedings of the VLDB Endowment* 3.1-2 (2010): 484-493
- Isele, Robert, and Christian Bizer.: **Learning linkage rules using genetic programming.** *Proc. VLDB Endow. 5, 11, 2012.*

# References: Identity Resolution for Product Data


- Kannan, Anitha, et al. **Matching unstructured product offers to structured product specifications**. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011
- Akritidis, Bozanis: **Effective Unsupervised Matching of Product Titles with k-Combinations and Permutations.** Inista, 2018.
- Petrovski, Petar, Volha Bryl, and Christian Bizer.: **Integrating product data from websites offering microdata markup.** *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014.
- Petar Petrovski, Anna Primpeli, Robert Meusel, Christian Bizer: **The WDC Gold Standards for Product Feature Extraction and Product Matching**. 17th International Conference on Electronic Commerce and Web Technologies (EC-Web 2016), Porto, Portugal, September, 2016.
- Petar Petrovski, Christian Bizer: **Learning Expressive Linkage Rules from Sparse Data**. Under review at the Semantic Web Journal, 2018.

# References: Deep Learning for Identity Resolution


- Mudgal, Sidharth, et al.: **Deep Learning for Entity Matching: A Design Space Exploration**. *Proceedings of the 2018 International Conference on Management of Data*. ACM, 2018.
- Shah, Kashif, Selcuk Kopru, and Jean David Ruvini.: **Neural Network based Extreme Classification and Similarity Models for Product Matching**. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. Vol. 3. 2018.
- Ajinkya More (WalmartLabs): **Product Matching in eCommerce using Deep Learning.** https://medium.com/walmartlabs/product-matching-in-ecommerce-4f19b6aebaca
- Ristoski, Petar, et al.: **A machine learning approach for product matching and categorization."** *Semantic Web Journal, 9(4):1-22, 2018*.
- Fernandez, Raul Castro, et al. "**Seeping semantics: Linking datasets using word embeddings for data discovery**." *34th IEEE International Conference on Data Engineering, ICDE, Paris, France*. 2018.

# Related Work for Feature Extraction


- Qiu, Disheng, et al.: **Dexter: Large-scale Discovery and Extraction of Product Specifications on the Web.** *Proceedings of the VLDB Endowment* 8.13 (2015): 2194-2205
- Petar Petrovski, Christian Bizer: **Extracting Attribute-Value Pairs from Product Specifications on the Web.** International Conference on Web Intelligence (WI2017), pp. 558-565, Leipzig, Germany, August 2017.
- Zheng, Mukherjee, Dong: **OpenTag: Open Attribute Value Extraction from Product Profiles**. KDD, 2018.
- Dalvi, Nilesh, Ravi Kumar, and Mohamed Soliman.: **Automatic wrappers for large scale web extraction.** *Proceedings of the VLDB Endowment* 4.4 (2011): 219-230
- Probst, Katharina, et al.: **Semi-Supervised Learning of Attribute-Value Pairs from Product Descriptions**. *IJCAI*. Vol. 7. 2007
- Ristoski, Petar, et al.: **A machine learning approach for product matching and categorization."** *Semantic Web Journal, 9(4):1-22, 2018*.
- Petar Petrovski, Anna Primpeli, Robert Meusel, Christian Bizer: **The WDC Gold Standards for Product Feature Extraction and Product Matching**. 17th International Conference on Electronic Commerce and Web Technologies (EC-Web 2016), Porto, Portugal, September, 2016.

# Potentially Useful Software

- Identity Resolution
  - Winte.r Framework : https://github.com/olehmberg/winter
  - Silk Framework : https://github.com/silk-framework/silk
  - DeepMatcher : https://github.com/anhaidgroup/deepmatcher
- Information Extraction
  - Specification table classifier (template project): https://github.com/petrovskip/wstl-extractor
  - Feature extraction gold standard tool: https://github.com/aprimpeli/LabellingTool
  - Stanford NLP: https://nlp.stanford.edu/software/
- Crawling
  - Scrapy : https://scrapy.org/

# The Project Team

1. Michael, Anne Katrin
2. Ly, Duc Tai
3. Le, Phuong Anh
4. Zhang, Shenghan
5. Yeu, Se Won
6. Amedani, Jurgen
7. Shkrepa, Lerida
8. Erazo Guevara, Maria Alejandra

– A Short Round of Introductions
  - What are you studying? Which semester?
  - Which DWS courses did you already attend?
  - What are your programming and data wrangling skills?

# Questions?