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

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

<table>
<thead>
<tr>
<th>Text-oriented</th>
<th>Attribute-oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td>(bag of words)</td>
<td>(weighted matching rules)</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Supervised</td>
</tr>
<tr>
<td>(TF-IDF+Cosine)</td>
<td>(random forest)</td>
</tr>
<tr>
<td>Symbolic</td>
<td>Sub-symbolic</td>
</tr>
<tr>
<td>(matching rules)</td>
<td>(embeddings)</td>
</tr>
</tbody>
</table>
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

- 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) ▼

⭐⭐⭐⭐⭐ Rating: 4.5 - 1,286 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 ▼
⭐⭐⭐⭐⭐ Rating: 4.7 - 500 reviews - Price range: $$$ and over
Book now at Gramercy Tavern in New York, explore menu, see photos and read 500 reviews: “The menu was so limited but it was worth trying, food was deli...”

**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](https://en.wikipedia.org/wiki/The_Black_Keys)

**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: The Black Keys - Neuhausen ob Eck (near you)
- May 16: The Black Keys - Gulf Shores, AL
- Jun 22: The Black Keys - Schlesgel
Example: Microdata Annotations in HTML

```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>
</div>
```
Product-related schema.org Classes

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 $\rightarrow$ 38.2 billion RDF triples
  - 2016 CC Corpus: 3.1 billion HTML pages $\rightarrow$ 44.2 billion RDF triples
  - 2014 CC Corpus: 2.0 billion HTML pages $\rightarrow$ 20.4 billion RDF triples
  - 2012 CC Corpus: 3.0 billion HTML pages $\rightarrow$ 7.3 billion RDF triples
- uses 100 machines on Amazon EC2
  - approx. 2000 machine/hours
    (100 spot instances of type c3.xlarge) $\rightarrow$ 350 Euro
- http://www.webdatacommons.org/structureddata/
schema.org Annotations: Most Popular Classes

http://webdatacommons.org/structureddata/
## Properties used to Describe Products 2017

<table>
<thead>
<tr>
<th>Top 15 Properties</th>
<th>PLDs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
</tr>
<tr>
<td>schema:Product/name</td>
<td>535,625</td>
</tr>
<tr>
<td>schema:Offer/price</td>
<td>462,444</td>
</tr>
<tr>
<td>schema:Product/offers</td>
<td>462,233</td>
</tr>
<tr>
<td>schema:Offer/priceCurrency</td>
<td>430,556</td>
</tr>
<tr>
<td>schema:Product/image</td>
<td>419,391</td>
</tr>
<tr>
<td>schema:Product/description</td>
<td>377,639</td>
</tr>
<tr>
<td>schema:Offer/availability</td>
<td>337,876</td>
</tr>
<tr>
<td>schema:Product/url</td>
<td>263,720</td>
</tr>
<tr>
<td>schema:AggregateRating/ratingValue</td>
<td>184,004</td>
</tr>
<tr>
<td>schema:Product/sku</td>
<td>126,696</td>
</tr>
<tr>
<td>schema:AggregateRating/reviewCount</td>
<td>112,408</td>
</tr>
<tr>
<td>schema:Product/aggregateRating</td>
<td>101,434</td>
</tr>
<tr>
<td>schema:Product/brand</td>
<td>73,934</td>
</tr>
<tr>
<td>schema:Product/productID</td>
<td>35,211</td>
</tr>
<tr>
<td>schema:Product/manufacturer</td>
<td>21,967</td>
</tr>
</tbody>
</table>

The Samsung Galaxy S4 is the entertaining and helpful companion for your mobile life. It connects you with your loved ones. It allows you to experience and preserve unforgettable moments together. It simplifies your everyday life.

UPC 610214632623

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.
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 some 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? \(\rightarrow\) Corpus Profiling, Data Selection
2. Is the corpus for these categories dense and clean enough? \(\rightarrow\) Corpus Profiling, Additional Crawling
3. How do supervised and unsupervised IR methods perform? \(\rightarrow\) Basic IR
4. Which features should be extracted to help IR? \(\rightarrow\) Feature Extraction
5. How do IR methods perform given cleaner features? \(\rightarrow\) IR with Feature Extraction
6. How do your different methods compare? \(\rightarrow\) Comparison of Results
Main Steps of the Project

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

Participants: All team members
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/](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 forests, 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. manually 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 other 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 auxiliary 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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food type</td>
<td>Canned food, dry food, raw food ...</td>
</tr>
<tr>
<td>Brand</td>
<td>Blue Buffalo, WholeHearted...</td>
</tr>
<tr>
<td>Lifestage</td>
<td>adulthood, all stages ...</td>
</tr>
<tr>
<td>Packaging</td>
<td>Can, tray...</td>
</tr>
</tbody>
</table>

2. Use the dictionary to extract features from your offers

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

\[
F(\text{overlap}) \quad \text{and} \quad F(\text{jaccard}>0.5)
\]

Regex: .*food
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)

Example specification table from chewy.com

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food type</td>
<td>Canned food, dry food, raw food...</td>
</tr>
<tr>
<td>Brand</td>
<td>Blue Buffalo, WholeHearted...</td>
</tr>
<tr>
<td>Lifestage</td>
<td>adulthood, all stages...</td>
</tr>
<tr>
<td>Packaging</td>
<td>Can, tray...</td>
</tr>
</tbody>
</table>

Example of a non-specification table from petco.com

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>Values</td>
</tr>
<tr>
<td>Attribute</td>
<td>Values</td>
</tr>
<tr>
<td>Attribute</td>
<td>Values</td>
</tr>
</tbody>
</table>
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?

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
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

<table>
<thead>
<tr>
<th>OfferName1</th>
<th>OfferName2</th>
</tr>
</thead>
<tbody>
<tr>
<td>lenovo - thinkpad x1 carbon 14in laptop - intel core i5 - 8gb memory - 256gb solid state drive - black</td>
<td>lenovo - thinkpad x1 carbon 14in laptop - intel core i7 - 16gb memory - 512gb solid state drive - black</td>
</tr>
</tbody>
</table>

- 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

Input for IR Phase 4

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

<table>
<thead>
<tr>
<th>Entity Pair</th>
<th>brand</th>
<th>memory</th>
<th>color</th>
<th>feature_x</th>
</tr>
</thead>
<tbody>
<tr>
<td>$o_1 - o_2$</td>
<td>$f_{sim}(brand)$</td>
<td>$f_{sim}(memory)$</td>
<td>$f_{sim}(color)$</td>
<td>$f_{sim}(feature_x)$</td>
</tr>
<tr>
<td>$o_1 - o_3$</td>
<td>$f_{sim}(brand)$</td>
<td>$f_{sim}(memory)$</td>
<td>$f_{sim}(color)$</td>
<td>$f_{sim}(feature_x)$</td>
</tr>
<tr>
<td>$o_x - o_y$</td>
<td>$f_{sim}(brand)$</td>
<td>$f_{sim}(memory)$</td>
<td>$f_{sim}(color)$</td>
<td>$f_{sim}(feature_x)$</td>
</tr>
</tbody>
</table>

$\begin{align*}
        f_{sim}(brand) : & \text{ jaccard sim. score} \\
        f_{sim}(memory): & \text{ exact sim. score} \\
        f_{sim}(color): & \text{ levenshtein sim. score} \\
        f_{sim}(feature_x): & \text{ x sim. score}
\end{align*}$
Phase 4: IR Methods

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

- 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

<table>
<thead>
<tr>
<th>Features</th>
<th>open IE</th>
<th>closed IE</th>
<th>no IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR Method</td>
<td>supervised</td>
<td>unsupervised</td>
<td>symbolic</td>
</tr>
<tr>
<td>Product Category</td>
<td>head</td>
<td>tail</td>
<td>structured</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Input data</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;movie&gt;</code></td>
</tr>
<tr>
<td><code> &lt;id&gt;1-9311&lt;/id&gt;</code></td>
</tr>
<tr>
<td><code> &lt;title&gt;Winter's Bone&lt;/title&gt;</code></td>
</tr>
<tr>
<td><code>&lt;/movie&gt;</code></td>
</tr>
<tr>
<td><code>&lt;movie&gt;</code></td>
</tr>
<tr>
<td><code> &lt;id&gt;1-9312&lt;/id&gt;</code></td>
</tr>
<tr>
<td><code> &lt;title&gt;Black Swan&lt;/title&gt;</code></td>
</tr>
<tr>
<td><code> &lt;date&gt;2011-01-01&lt;/date&gt;</code></td>
</tr>
<tr>
<td><code> &lt;globe&gt;yes&lt;/globe&gt;</code></td>
</tr>
<tr>
<td><code>&lt;/movie&gt;</code></td>
</tr>
<tr>
<td><code>&lt;movie&gt;</code></td>
</tr>
<tr>
<td><code> &lt;id&gt;1-9313&lt;/id&gt;</code></td>
</tr>
<tr>
<td><code> &lt;title&gt;Blue Valentine&lt;/title&gt;</code></td>
</tr>
<tr>
<td><code>...</code></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gold Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>gold.csv:</strong></td>
</tr>
<tr>
<td>1-9309,2-9309,true</td>
</tr>
<tr>
<td>1-9310,2-9310,true</td>
</tr>
<tr>
<td>1-9311,2-9311,false</td>
</tr>
<tr>
<td>1-9312,2-9312,true</td>
</tr>
<tr>
<td>1-9313,2-9313,false</td>
</tr>
<tr>
<td>1-9314,2-9314,false</td>
</tr>
<tr>
<td>1-9315,2-9315,false</td>
</tr>
<tr>
<td>1-9316,2-9316,true</td>
</tr>
</tbody>
</table>

- Define your data model
Phase 4: WInte.r Framework (2/2)

- Create handwritten rules (unsupervised IR)

```java
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

```java
// 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

Papers
References: Identity Resolution for Product Data


References: Deep Learning for Identity Resolution


- Ajinkya More (WalmartLabs): **Product Matching in eCommerce using Deep Learning.**


Related Work for Feature Extraction


- Zheng, Mukherjee, Dong: OpenTag: Open Attribute Value Extraction from Product Profiles. KDD, 2018.


Potentially Useful Software

- Identity Resolution
  - Winte.r Framework: [https://github.com/olehmberg/winter](https://github.com/olehmberg/winter)
  - Silk Framework: [https://github.com/silk-framework/silk](https://github.com/silk-framework/silk)
  - DeepMatcher: [https://github.com/anhaidgroup/deepmatcher](https://github.com/anhaidgroup/deepmatcher)

- Information Extraction
  - Specification table classifier (template project): [https://github.com/petrovskip/wstl-extractor](https://github.com/petrovskip/wstl-extractor)
  - Feature extraction gold standard tool: [https://github.com/aprimpeli/LabellingTool](https://github.com/aprimpeli/LabellingTool)
  - Stanford NLP: [https://nlp.stanford.edu/software/](https://nlp.stanford.edu/software/)

- Crawling
  - Scrapy: [https://scrapy.org/](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?