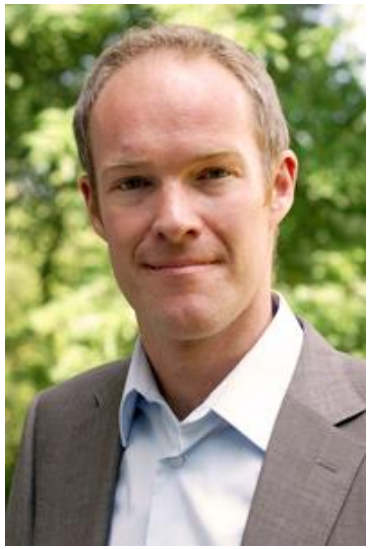
UNIVERSITY OF MANNHEIM

Team Project FSS 2018

Mining 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
2. Organization and Schedule
3. 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, the producer, consumers
- different websites describe different aspects of a product
  - technical spec vs consumer experience
- there are plenty of offers for a product online
  - we can collect price information on global scale
- many websites point us at similar products

Using information about products from the Web, we can

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

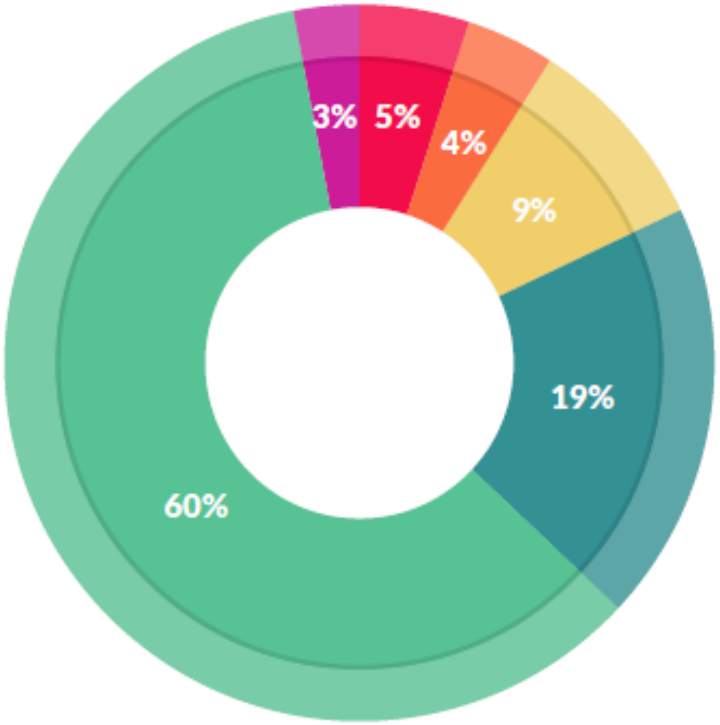
Project Goals

1. Gather and integrate product, price, and review data from multiple e-shops
2. Mine this data to discover
   - price/feature associations
   - feature/user perception associations
   - understand the market structure
   - understand consumer behaviour

Questions and Subtasks

1. Which e-shops to consider? → **Data Selection and Crawling**
2. Which data to extract? → **Feature Extraction**
3. How to recognize identical products? → **Identity Resolution**
4. How to group similar products? → **Categorization / Cluster Analysis**
5. How to understand user perception? → **Sentiment Analysis**
6. How to combine extracted information? → **Data Fusion**
7. **What patterns can be found in the data? → Data Mining**

How Do Data Scientists Spend Their Days?



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: CrowdFlower Data Science Report 2016: http://visit.crowdflower.com/data-science-report.html

Project Organization

Duration: 6 months (02.03.2018 – 02.09.2018)

ECTS: 12

Participants: 8 people

Type of work: Team and subgroup based

Milestones: 4 project phases

Evaluation:

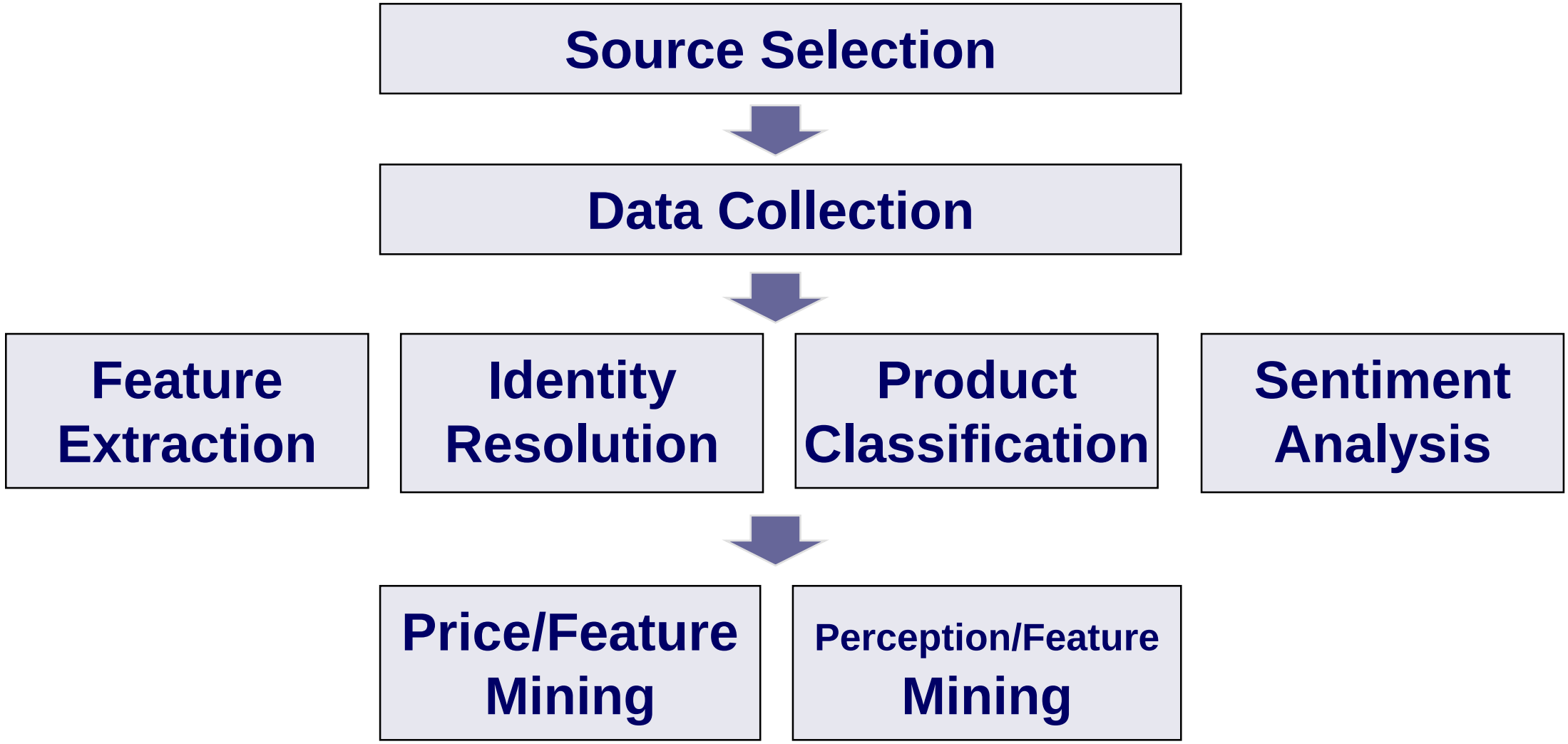
- Individual contribution to the deliverables
- Deliverables: Presentations, reports, code, data
- Every project phase determins 25% of your final grade

The Project Team

1. Heißler, Larissa
2. Bertsch, Matthias Helmuth
3. Demirxhiu, Ersejda
4. Leung, Chung Chi
5. Chowdhury, Abdullah Al Murad
6. Koseoglu, Bengi
7. Aghazada, Adila
8. Gjuzi, Anjeza

– 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?

Main Steps of the Project



Detailed Schedule

Date	Session
Friday, 02.03.2018	Kickoff meeting (today)
	Phase 1 (all members): Decide on a set of products and sources, crawling, basic feature extraction, product catalog construction
Friday, 09.03.2018	Meet Anna and report plan, division of work
Friday, 23.03.2018	Drop-out deadline: Dropping out after this date will result in failing the team project
Friday, 13.04.2018	**1st Deliverable:** 15 minutes presentation, code & data - Subgroup formation
	Phase 2 (in 4 subgroups): feature extraction, identity resolution, categorization, sentiment analysis
Friday, 20.04.2018	Meet Anna and report plans
Friday, 18.05.2018	**2nd Deliverable:** 10 minutes presentation from each subgroup, code & data
	Phase 3 (in 4 subgroups): Refinement of phase 2
Tuesday, 29.05.2018	Meet Anna and report plan
Sunday, 01.07.2018	**3rd Deliverable:** 8-12 pages report from each subgroup, code & data
	Phase 4 (in 2 subgroups): Mining of integrated product data and reviews
Friday, 20.07.2018	Meet Anna and report plan
Sunday, 02.09.2018	**4th Deliverable:** 8-12 pages report from each subgroup, code & data
Friday, 07.09.2018	Overall presentation 30 min + Feedback

Phase 1: Source Selection, Crawling, Basic Feature Extraction, Catalog Construction

Participants: All team members

Duration: 02.03.2018 – 03.04.2018

Deliverables: 15 minutes presentation, data & code, report who did what

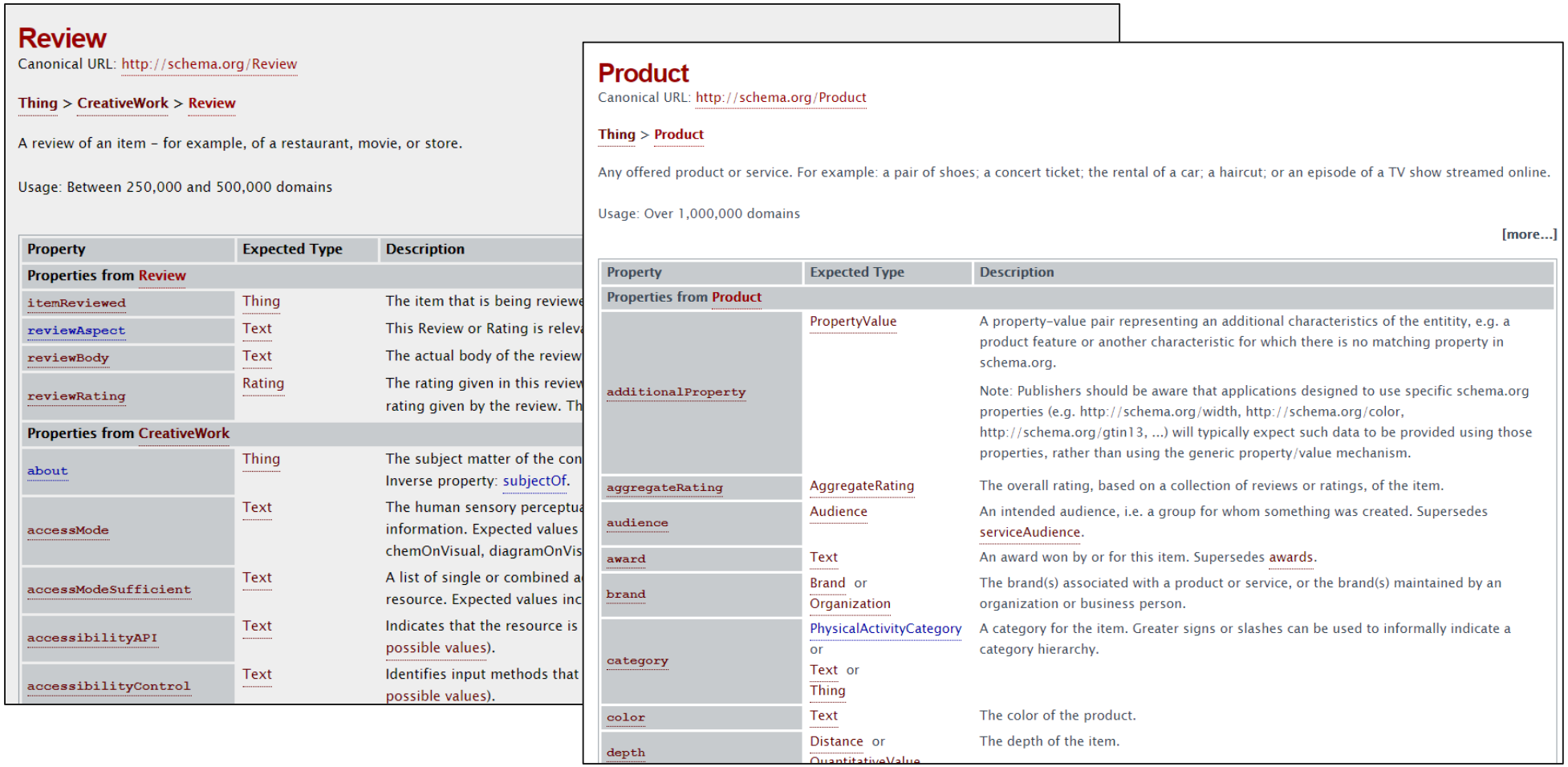
Tasks (2/5)

1. Decide on two main product categories
   - Select 2 non-similar main product categories, e.g. laptops and shoes (**NOT** phones, headphones, TVs)
   - Choose 3 – 4 subcategories for each main category, e.g. noise-cancelling, over-ear, on-ear, & sports headphones.
2. Decide on a set of e-shops
   - Analyse data for main players for each product category using sources such as Alexa and WDC
   - Select minimum 20 e-shops for each main product category
   - The selected e-shops should: Be located in 2 countries, be in English, **NOT** be marketplaces.
   - Partly (50%+) support the extraction of structured data by using schema.org and/or HTML tables

Tasks (3/5)

3. Crawl product pages
   - For each main category select 50 seed products. The selected products should not be too distinct from each other, e.g. i-phone 4s, samsung galaxy s8, nokia lumia 635.
   - Crawl products by following links from the seed product pages. Crawl 1000+ products per website
   - Your crawled results should include closely related products, e.g. i-phone 4 and i-phone 4s.
4. Extract product specifications, prices, category information, product IDs, and reviews
   - consider the schema.org annotations s:Product, s:Review, and s:Offer
   - use simple heuristics for locating relevant data: use annotations and identify web tables
5. Construct a product catalog
   - Consider google shopping to create a product catalog the covering 50 seed products per category
     - The catalog defines a central schema for describing and a single product hirarchie
   - Perform basic schema matching of product specifications to product attributes in catalog

schema.org Terms



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 being reviewe
reviewAspect	Text	This Review or Rating is releva
reviewBody	Text	The actual body of the review
reviewRating	Rating	The rating given in this review rating given by the review. Th
Properties from CreativeWork		
about	Thing	The subject matter of the con Inverse property: subjectOf.
accessMode	Text	The human sensory perceptua information. Expected values chemOnVisual, diagramOnVis
accessModeSufficient	Text	A list of single or combined a resource. Expected values inc
accessibilityAPI	Text	Indicates that the resource is possible values).
accessibilityControl	Text	Identifies input methods that 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.

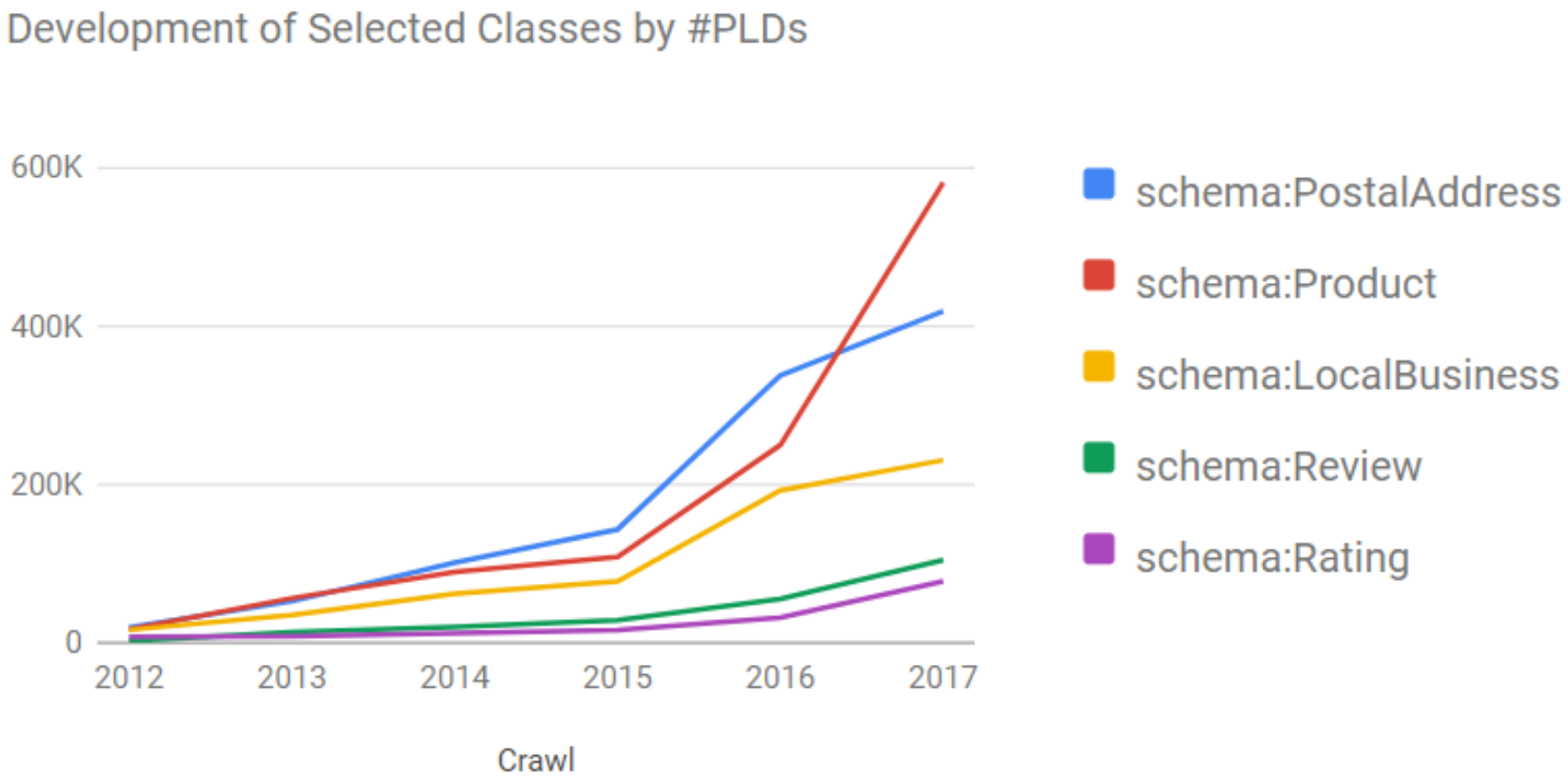
Usage: Over 1,000,000 domains

[more...]

Property	Expected Type	Description
Properties from Product		
additionalProperty	PropertyValue	A property-value pair representing an additional characteristics of the entitity, e.g. a product feature or another characteristic for which there is no matching property in schema.org. 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.

schema.org Annotations: Most Popular Classes



http://webdatacommons.org/structureddata/

Adoption by E-Commerce Websites 2014

Distribution by Top-Level Domain

TLD	#PLDs
com	38344
co.uk	3605
net	1813
de	1333
pl	1273
com.br	1194
ru	1165
com.au	1062
nl	1002

Adoption by Top-15:
60 %

Alexa Top-15 Shopping Sites

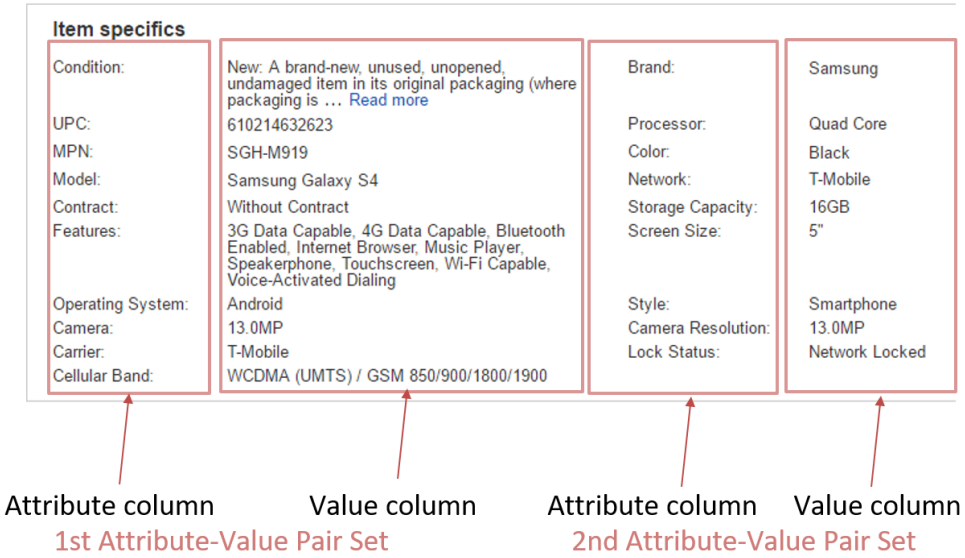
Website	schema:Product
Amazon.com	☒
Ebay.com	✓
NetFlix.com	☒
Amazon.co.uk	☒
Walmart.com	✓
etsy.com	☒
Ikea.com	✓
Bestbuy.com	✓
Homedepot.com	✓
Target.com	✓
Groupon.com	☒
Newegg.com	✓
Lowes.com	☒
Macys.com	✓
Nordstrom.com	✓

Properties used to Describe Products 2014

Top 15 Properties	PLDs	
	#	%
schema:Product/name	78,292	87 %
schema:Product/image	59,445	66 %
schema:Product/description	58,228	65 %
schema:Product/offers	57,633	64 %
schema:Offer/price	54,290	61 %
schema:Offer/availability	36,789	41 %
schema:Offer/priceCurrency	30,610	34 %
schema:Product/url	23,723	26 %
schema:Product/aggregateRating	21,166	24 %
schema:AggregateRating/ratingValue	20,513	23 %
schema:AggregateRating/reviewCount	14,930	17 %
schema:Product/manufacturer	10,150	11 %
schema:Product/brand	9,739	11 %
schema:Product/productID	9,221	10 %
schema:Product/sku	7955	9 %

Stuff that will help you later …

1. Product IDs
   - GTINs, MPNs, SKUs, ISBNs
   - solve the identity resolution problem
2. HTML Tables
   - contain key/value pairs
   - main souce of structured product specifications
3. Category Information
   - (Schema.org) bread crumps
   - Schema.org category
   - URLs fragments

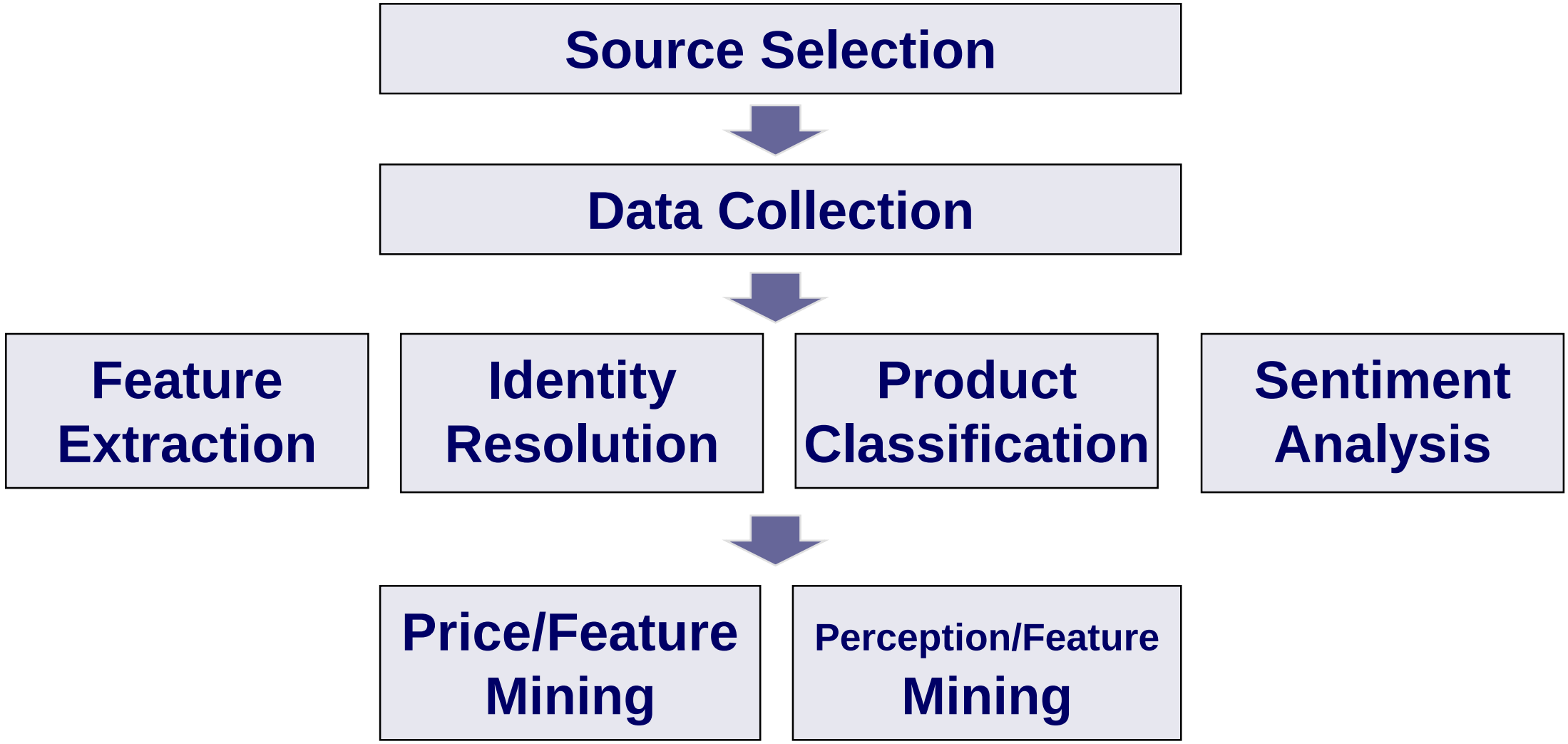


Telefon + Navi / Wearables / Smartwatches / SAMSUNG Gear S3

https://www.galeria-kaufhof.de/schuhe/damenschuhe/mokassins/

– Make sure that 50%+ of the selected shops provide these things

Main Steps of the Project



Phases 2 & 3: Feature Extraction, Identity Resolution, Categorization, Sentiment Analysis

Duration: 04.04.2018 – 15.05.2018 (Phase 2), 16.05.2018 – 01.07.2018 (Phase 3)

Participants: 4 subgroups of 2 persons each

Deliverables:
- 10 minutes presentation from each subgroup after phase 2
- 8 -12 pages report, data & code from each subgroup after phase 3

Tasks (2/4)

Subgroup 1: Feature Extraction

- Perform more sophisticated feature extraction techniques such as application of regex expressions, further consideration of table structure
- perform more sophisticated schema matching with product catalog

Subgroup 2: Identity Resolution

- Apply basic IR techniques and machine learning, exploit features produced by subgroup 1 in phase 2
- Apply transfer learning: use product identifiers as labeled data to learn a classification model and evaluate the model on unseen data

Tasks (4/4)

Subgroup 3: Product Categorization

- Apply multi-level classification techniques
- Consider product features for categorization
- Exploit existing categorization information / integrate category trees
- Exploit identity resolution results from team 2 in second phase.
- Cluster products into additional more fine grained cateogies (premium products within subcategory?)

Subgroup 4: Sentiment Analysis

- Gather and preprocess the reviews of the crawled websites
- Extract sentiment information for the specific product features and overall for the product
- Extract background information about raters (crawl additional review portars if necessary)

Phase 3 will be a refinement of Phase 2!

Phase 4: Data Mining

Participants: 2 subgroups of 4 persons each

Duration: 02.07.2018 – 01.09.2018

Deliverables: report, code & data from each subgroup, final overall presentation

Tasks

Bring your results together and extract interesting facts

- Combine the information extracted from the subtasks of phase 2 and 3
- Mine your data by performing correlation analysis
- Discover and report interesting facts concerning

  Subgroup 1 : **Price**

  e.g. Which features determine the price? How are prices distributed? By location? Product type?

  Subgroup 2 : **Reviews**

  e.g. Which specific customer groups prefer which sub-categories/products?, which features matter most for specific customer groups?, how do certain product features affect customer satisfaction?

Formal Requirements

Formal Requirements & Consultation

- Deliverables
  - Reports should be 8-12 pages single column
    - including appendixes
    - not including the bibliography
    - every additional page reduces your grade by 0.3
    - Created with Latex template of the Data and Web Science group (http://dws.informatik.uni-mannheim.de/en/thesis/masterthesis/)
  - Every deliverable (presentation/ report) should be accompanied with an excel sheet stating which team member conducted which subtask.
- Final grade
  - 25% for every phase, individual grade / not per team
  - Late submission: reduction of grade by 0.3 per day
- Consultation
  - Send one e-mail per team stating your questions to Anna, she answers questions or meets with you
  - Chris does second level support and gives feedback at presentations

How to structure your deliverables?

1. Problem definition
2. Profiling of selected data
3. Methodology
4. Evaluation (for phases 2 & 3)
5. Error Analysis (for phases 2 & 3)
6. Conclusion

Accompany your deliverables with the code and data you used

- ! The phase deadlines apply for the submission of your code and data as well

Submission of Deliverables

Presentations

The presentations will take place according to the schedule. For the exact time you will be informed via e-mail. The presentation slides should be provided by the end of the meeting.

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!

Potentially Useful Software

- Crawling
  - Scrapy : https://scrapy.org/
  - Any23
- Data Integration
  - Winte.r Framework : https://github.com/olehmberg/winter
  - Silk Framework : https://github.com/silk-framework/silk
- Data Mining, Machine Learning
  - RapidMiner : https://rapidminer.com/
- Natural Language Processing
  - Stanford NLP: https://nlp.stanford.edu/software/

Related Work (1/3)

- 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.
- 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.
- 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
- 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.
- Ristoski, Petar, and Peter Mika. "Enriching product ads with metadata from HTML annotations." *International Semantic Web Conference*. Springer, Cham, 2016
- Petrovski, Petar, Volha Bryl, and Christian Bizer. "Learning regular expressions for the extraction of product attributes from e-commerce microdata." Proceedings of the Second International Conference on Linked Data for Information Extraction-Volume 1267. CEUR-WS. org, 2014

Related Work (2/3)

- 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
- Robert Meusel, Christian Bizer, Heiko Paulheim: A Web-scale Study of the Adoption and Evolution of the schema.org Vocabulary over Time. 5th International Conference on Web Intelligence, Mining and Semantics (WIMS2015), Limassol, Cyprus, July 2015.
- Meusel, Robert, et al. "Exploiting microdata annotations to consistently categorize product offers at web scale." *International Conference on Electronic Commerce and Web Technologies*. Springer International Publishing, 2015
- 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
- Christen, Peter. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012
- Isele, Robert, and Christian Bizer. "Learning linkage rules using genetic programming." *Proceedings of the 6th International Conference on Ontology Matching-Volume 814*. CEUR-WS. org, 2011

Related Work (3/3)

- Petar Petrovski, Christian Bizer: Learning Expressive Linkage Rules from Sparse Data. Under review at the Semantic Web Journal, 2018.
- Poon, Hoifung, and Pedro Domingos. "Unsupervised ontology induction from text." Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.
- Navigli, Roberto, Paola Velardi, and Stefano Faralli. "A graph-based algorithm for inducing lexical taxonomies from scratch." IJCAI. Vol. 11. 2011.
- Ristoski, Petar, et al. "Large-scale taxonomy induction using entity and word embeddings." Proceedings of the International Conference on Web Intelligence. ACM, 2017.
- Silla, Carlos N., and Alex A. Freitas. "A survey of hierarchical classification across different application domains." Data Mining and Knowledge Discovery 22.1-2 (2011): 31-72.
- Melo, André, Heiko Paulheim, and Johanna Völker. "Type prediction in rdf knowledge bases using hierarchical multilabel classification." Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics. ACM, 2016.

Learning Targets

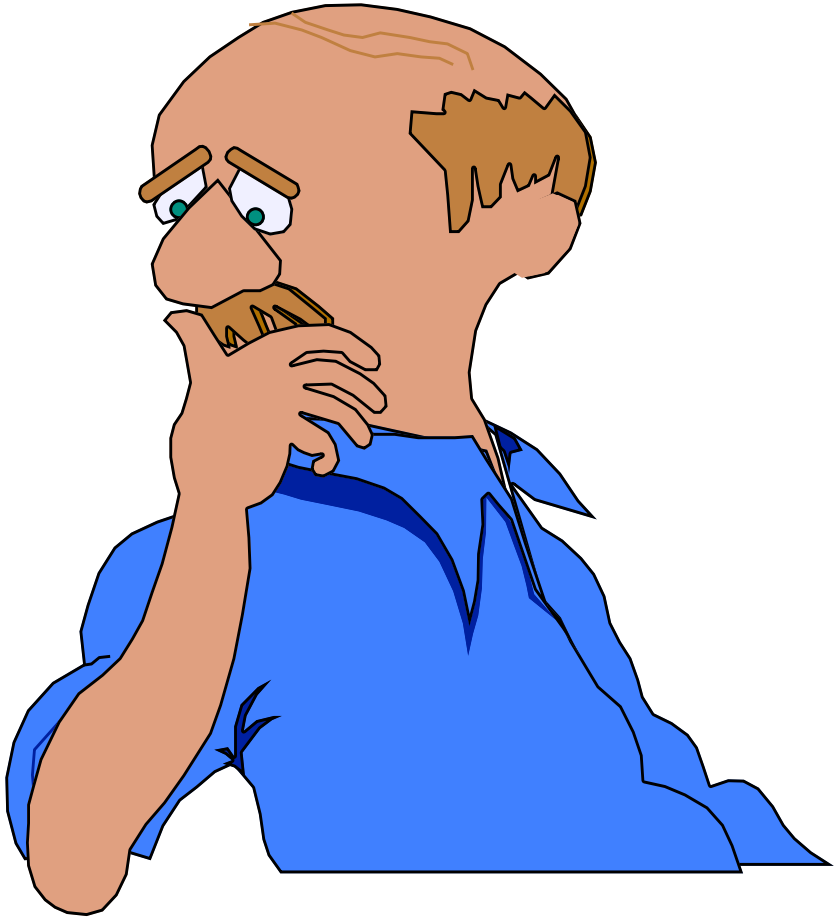
Improve your technical skills

- Work as a **Data Scientist**: gather, clean, profile, integrate, classify data in order to extract knowledge
- 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

Questions?



Project Infrastructure?

- Shared Document Space
  - for todo lists, brainstorming documents
  - Google Docs? Wiki?
- ILIAS Group
  - mailing to all participants
  - for sharing files
- Code Repository
  - GitHub?
- Data Repository
  - Google Drive? Dropbox?
- Anything else?