Your Tasks

Business Intelligence

1st Semester Data Science Master

Prof. Dr. habil. Alexander Löser

Berlin University of Applied Sciences Berlin

Overview

In the first task you will collect and analyze data with a particular focus on Web searchers in the year 2006. You might consider this as an exercise to learn how to wrangle data (with ROLAP and SQL) and how to put insights into a meaningful presentation together in six weeks.

In the second task you will analyze the economy of a data product with a canvas. You are free in your choice of a company and a single data product. This task should help you to focus on a meaningful and doable idea for a data product that you and your team will work on the next two years. You can already apply your knowledge from this and other classes, such as programming with python or applying your machine learning basics. In the next semesters you can work on your “baby”, such in the 2nd semester in our text mining or deep learning classes, in the 3rd semester in enterprise data science or in the 4th semester in your master thesis.

During this class, we will introduce you to companies, you might also spot a topic which will help them and you might work out a data product with them. Some people might also be lucky and will be invited to apply for a student researcher position with our research group DATEXIS.COM

Task 1: Insights about Events from your Fellows in 2006 (50 points)

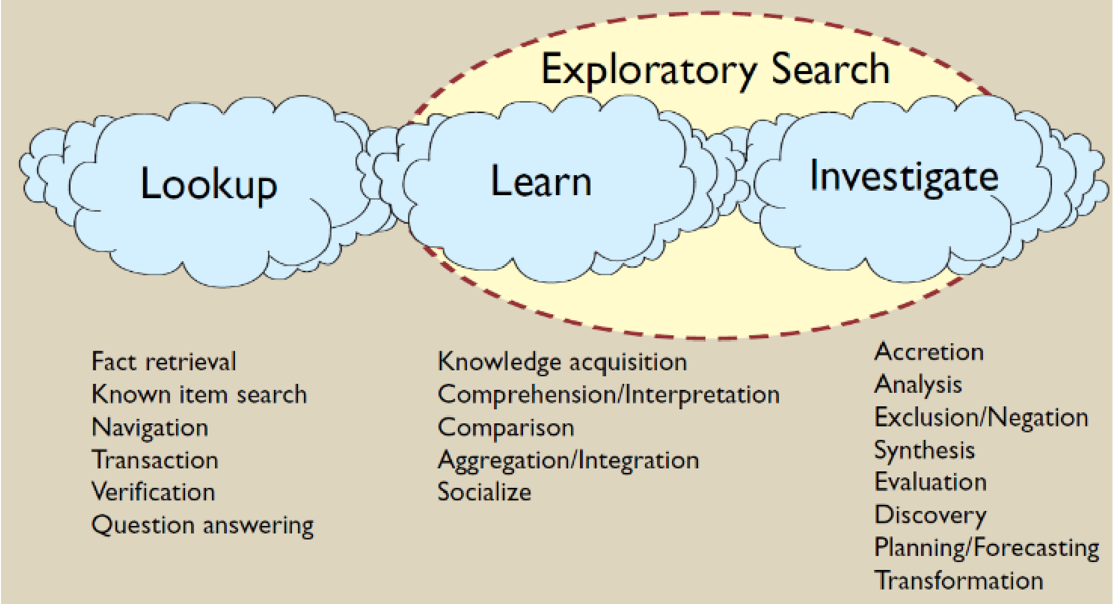
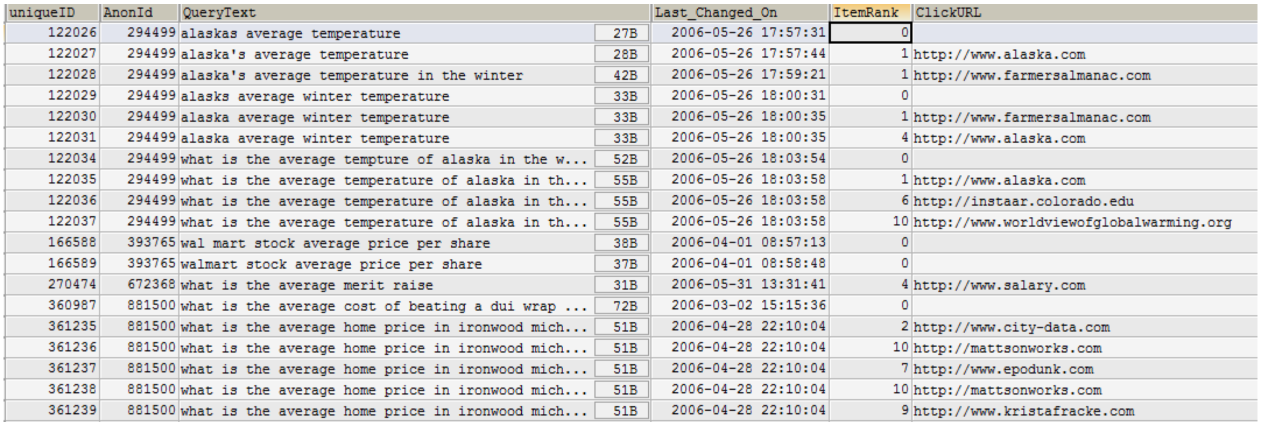


Figure 1: Searchers follow at a broad scale three stages ‘Lookup, Learn und Investigate’ in their iterative process. Each stage may contain several events, such as typing a query, browsing results, refining the query, clicking, staying on the page, closing the page etc. Source [Fehler! Verweisquelle konnte nicht gefunden werden.**, 5]**

Given is a snapshot of > 36 Mio. query logs from mostly US citizen and clicked results from March to June 2006 from the AOL search engine. During those times, Google search provided the technology. The following Figure shows an example snapshot:



uniqueID	AnonId	QueryText		Last_Changed_On	ItemRank	ClickURL
122026	294499	alaskas average temperature	27B	2006-05-26 17:57:31	0	
122027	294499	alaska's average temperature	28B	2006-05-26 17:57:44	1	http://www.alaska.com
122028	294499	alaska's average temperature in the winter	42B	2006-05-26 17:59:21	1	http://www.farmersalmanac.com
122029	294499	alasks average winter temperature	33B	2006-05-26 18:00:31	0	
122030	294499	alaska average winter temperature	33B	2006-05-26 18:00:35	1	http://www.farmersalmanac.com
122031	294499	alaska average winter temperature	33B	2006-05-26 18:00:35	4	http://www.alaska.com
122034	294499	what is the average tempture of alaska in the w...	52B	2006-05-26 18:03:54	0	
122035	294499	what is the average temperature of alaska in th...	55B	2006-05-26 18:03:58	1	http://www.alaska.com
122036	294499	what is the average temperature of alaska in th...	55B	2006-05-26 18:03:58	6	http://instaar.colorado.edu
122037	294499	what is the average temperature of alaska in th...	55B	2006-05-26 18:03:58	10	http://www.worldviewofglobalwarming.org
166588	393765	wal mart stock average price per share	38B	2006-04-01 08:57:13	0	
166589	393765	walmart stock average price per share	37B	2006-04-01 08:58:48	0	
270474	672368	what is the average merit raise	31B	2006-05-31 13:31:41	4	http://www.salary.com
360987	881500	what is the average cost of beating a dui wrap ...	72B	2006-03-02 15:15:36	0	
361235	881500	what is the average home price in ironwood mich...	51B	2006-04-28 22:10:04	2	http://www.city-data.com
361236	881500	what is the average home price in ironwood mich...	51B	2006-04-28 22:10:04	10	http://mattsonworks.com
361237	881500	what is the average home price in ironwood mich...	51B	2006-04-28 22:10:04	7	http://www.epodunk.com
361238	881500	what is the average home price in ironwood mich...	51B	2006-04-28 22:10:04	10	http://mattsonworks.com
361239	881500	what is the average home price in ironwood mich...	51B	2006-04-28 22:10:04	9	http://www.kristafracke.com

Figure 2: Example search queries containing the word "average" and clicked pages for userID 29499, 393765, 672368 and 881500. A value of "0" in column ItemRank indicates a user without a click on the particular search result page. Column ItemRank indicates the ranking position if a result was clicked.

You have access to these queries in your virtual image of the EXASOL main memory RDBMS. In addition, we provide you data several hundred thousand categories from the DMOZ open directory category project and several hundred thousand movies and actors from IMDB.

Your task: Analyze this data set as a team, enrich it with your data sources and give an insightful talk (15 minutes).

Subtasks

1. **Define five interesting analysis question on this data set**. You might pick up more, since not all chosen questions are answerable with the existing data and your additional data sources.
2. **Import missing data from one additional data source of your choice** for resolving your queries into the database. Use your knowledge on JDBC and Extract-Transform-Load. Please check legal issues when importing data from "The Web".
3. **Formalize at least five of your queries with ROLAP Statements on EXASOL**. Utilize operators such as SLICE, DICE, CUBE, ROLLUP, PARTITION BY, GROUPING SETS and other standard SQL statements such as joins, unions or intersections etc. (see the EXASOL manual for details on the syntax).You might also use PANDAS or Python functions to predict from the data.
4. **Display your results as charts**, for example with http://d3js.org or JFreeChart
5. **Create a presentation for about 15 minutes** and explain your analysis goal, your data sets, showcase selected "cool/surprising" queries and results/insights, explain why this is an important valuable finding, show your schema and explain your workload.
6. **Create an appendix in your presentation**, where you **show the ROLAP queries and results** as screenshots. Name on each slide, what this query should have done. Add to the appendix screenshots of the tables you created, including schema information.
7. **Upload this presentation to the Moodle-system** with a filename <your name> (PDF/PPT) and present it in front of your peers. Check if your peers liked it and considered it insightful. ☺

Just to inspire you (not to copy!) So far I have heard cool talks about:

- Impact of Oscars in 2006 on actors query traffic

- Popularity of online retailers/social media sites/car brands/soccer/baseball/basketball teams/ …
- Disasters and diseases: H5N1 what happened? Hurricanes in 2006? Airplane crashes?
- Iran war, do assaults and troop events correlate with news?

I would be highly interested in: Understanding sessions, how users search, when users break up searches, typical user intentions or user journeys etc.

There are many interesting data sources on the Web; some do come in an easily joinable format.

- https://query.wikidata.org
- www.programmableweb.com
- www.Internetarchive.com + Memeto Protocol [**Fehler! Verweisquelle konnte nicht gefunden werden.**]

Evaluation criteria (50 points).

- Novelty and wow-factor regarding your topic
- Variety and soundness of SQL/ROLAP queries, results and table schema
- Efforts into data importing and cleansing
- Analytical insights
- Talk and slides

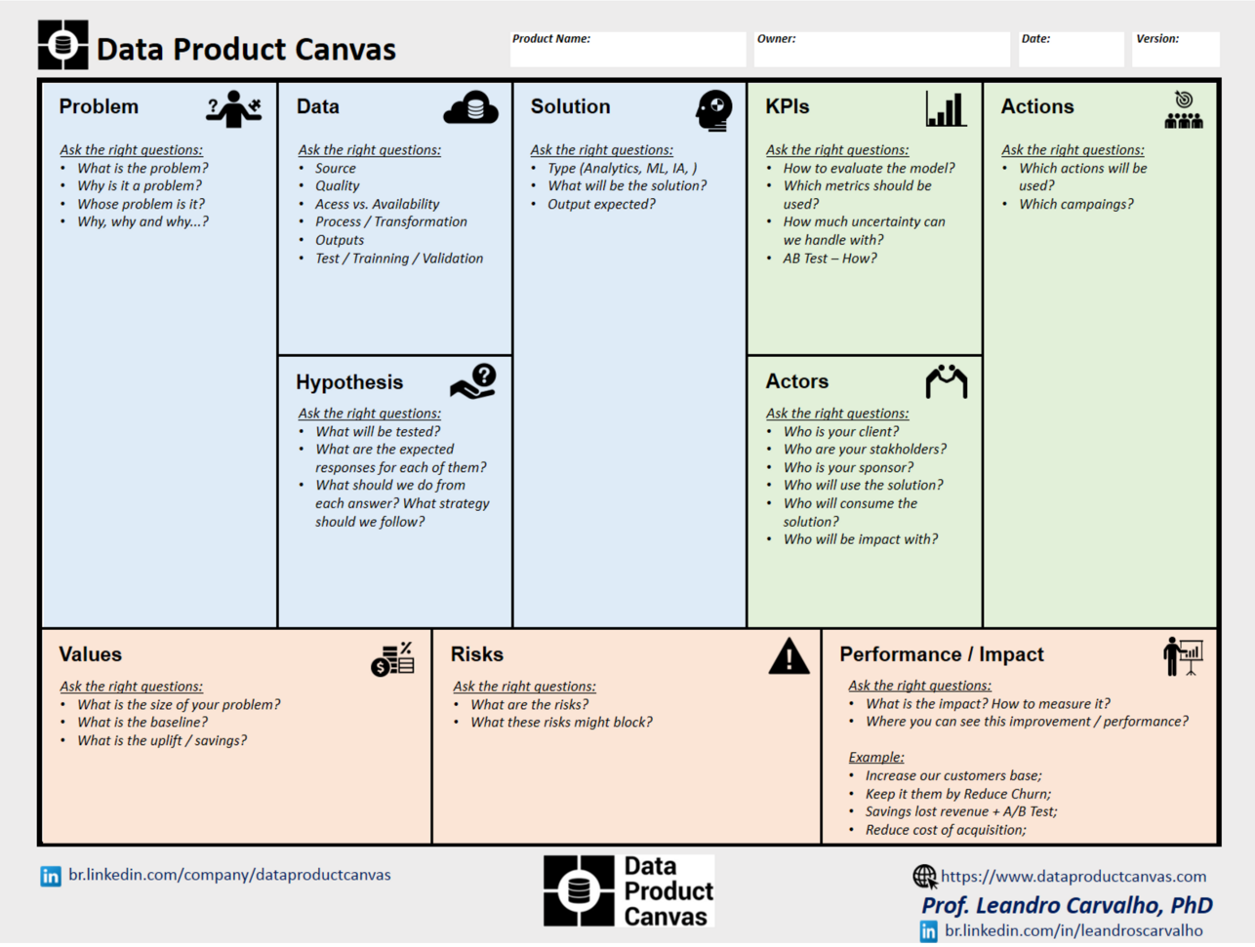
Important side note: This task is not only about analyzing the data and learning about ROLAP. It is more about team building and get things done, since the time is short to the first talks. Please start early to gather your team and to define your queries for the presentation.

Task 2: Analyze existing Data Products in the Platform Economy (50 Pts.)

Basically, the next two years you will learn a lot about data and how to get insights from. To prepare this step, you will pick a B2B, B2C or O2O platform and analyze one (!) existing data product. Please select one of these platform providers:

- Idealo.de
- Immoscout.de
- DuoLingo.com
- Mobile.de
- Babbel.com
- Zalando.de
- **Or a company of your choice, please send me an email before with your suggestion**

1. **Introduce the company**, show how large they are, how much turnover and revenue they make, since when they are on the market etc. and where they are located and who owns it.
2. **Select a data product/service, explain it and name and explain the business model behind.**
3. **Fill out this Data Product Canvas**. Prove your findings by references from articles or the web.



Data Product Canvas

Product Name:

Owner:

Date:

Version:

Problem

Ask the right questions:
- *What is the problem?*
- *Why is it a problem?*
- *Whose problem is it?*
- *Why, why and why...?*

Data

Ask the right questions:
- *Source*
- *Quality*
- *Acess vs. Availability*
- *Process / Transformation*
- *Outputs*
- *Test / Trainning / Validation*

Hypothesis

Ask the right questions:
- *What will be tested?*
- *What are the expected responses for each of them?*
- *What should we do from each answer? What strategy should we follow?*

Solution

Ask the right questions:
- *Type (Analytics, ML, IA,)*
- *What will be the solution?*
- *Output expected?*

KPIs

Ask the right questions:
- *How to evaluate the model?*
- *Which metrics should be used?*
- *How much uncertainty can we handle with?*
- *AB Test – How?*

Actors

Ask the right questions:
- *Who is your client?*
- *Who are your stakholders?*
- *Who is your sponsor?*
- *Who will use the solution?*
- *Who will consume the solution?*
- *Who will be impact with?*

Actions

Ask the right questions:
- *Which actions will be used?*
- *Which campaings?*

Values

Ask the right questions:
- *What is the size of your problem?*
- *What is the baseline?*
- *What is the uplift / savings?*

Risks

Ask the right questions:
- *What are the risks?*
- *What these risks might block?*

Performance / Impact

Ask the right questions:
- *What is the impact? How to measure it?*
- *Where you can see this improvement / performance?*

Example:
- *Increase our customers base;*
- *Keep it them by Reduce Churn;*
- *Savings lost revenue + A/B Test;*
- *Reduce cost of acquisition;*

br.linkedin.com/company/dataproductcanvas

Data Product Canvas

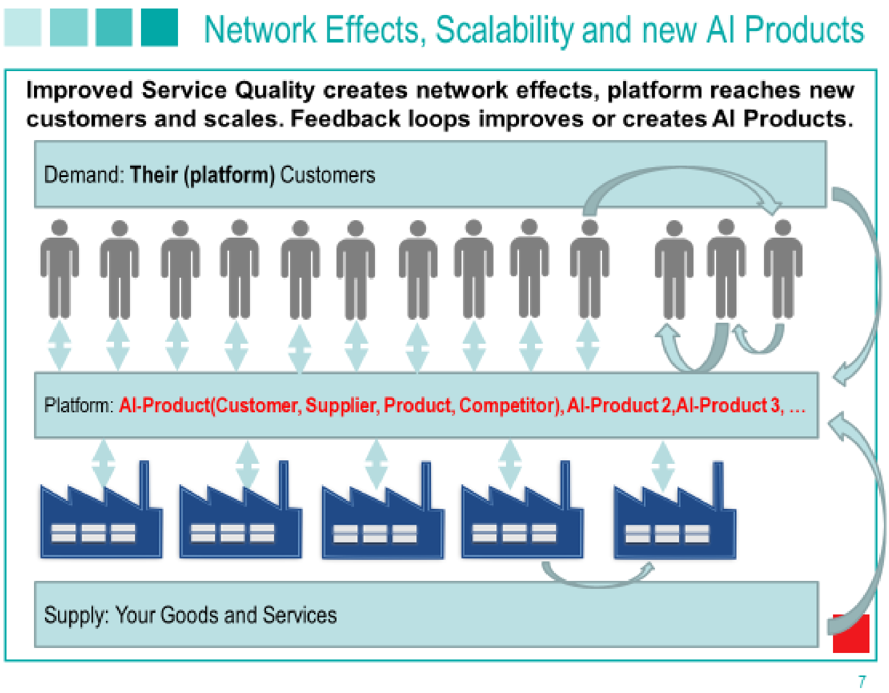
https://www.dataproductcanvas.com

Prof. Leandro Carvalho, PhD

br.linkedin.com/in/leandroscarvalho

Abbildung 3: https://medium.com/@leandroscarvalho/data-product-canvas-a-practical-framework-for-building-high-performance-data-products-7a1717f79f0

4. **Define Elements of the Platform Economy.** Name core entities, such as Customers, Supplier, Products, Competitors, etc. in the platform business of the customer of our choice. Abstract and discuss one core services the platform is running as a function that leverages these core entities as input variables.
5. **Estimate potential feedback-loops** and how they would enhance the data product.



Network Effects, Scalability and new AI Products

Improved Service Quality creates network effects, platform reaches new customers and scales. Feedback loops improves or creates AI Products.

Demand: **Their (platform)** Customers

Platform: **AI-Product(Customer, Supplier, Product, Competitor), AI-Product 2, AI-Product 3, …**

Supply: Your Goods and Services

7

6. **Attracting Customers/ Community for Growth.** Discuss on one slide what the company does for establishing growth with their customers or their community.

Your task: Please prepare a presentation for 15 minutes in a team.

Evaluation Criteria

1. Presentation of the Company
2. Presentation of one Data Product/Data Service
3. Presentation of the Business Modell for the selected data product
4. Data Product Canvas
5. Abstraction of Core Entities and Identification for one services as a function (see also lecture)
6. Discussion of Growth strategies
7. Your credible sources to justify your findings

Again, this is also about building a team. You can continue with the team from task one or refine your team.

Selected Literature

1. Greg Pass, Abdur Chowdhury, Cayley Torgeson: A picture of search. Infoscale 2006: 1

2. http://en.wikipedia.org/wiki/AOL_search_data_leak (letzter Besuch 26.9.2013)
3. Daniel E. Rose, Danny Levinson: Understanding user goals in web search. WWW 2004: 13-19
4. David J. Brenes et.al: Stratified analysis of AOL query log. Inf. Sci. 179(12): 1844-1858 (2009)
5. G. Marchionini. Exploratory search: from searching to understanding. CACM, 49:41-46, 2006
6. http://archive.org/web/web.php Internet Archive (last visit 29/9/2022).
7. Business Model Canvas. Alexander Osterwalder. https://en.wikipedia.org/wiki/Business_Model_Canvas (last visit 29/9/2022)
8. The Unreasonable Effectiveness of Data. Halevy, Norvig, Pereira. IEEE Intelligent Systems 2009. https://static.googleusercontent.com/media/research.google.com/de//pubs/archive/35179.pdf (last visit 29/9/2022)