Data Engineering and Data Science: A Case Study on Opinion Mining

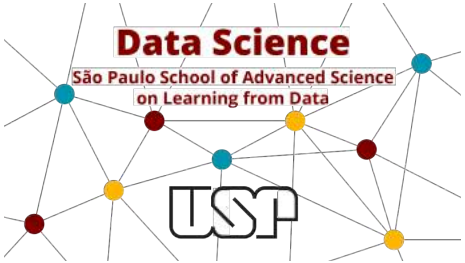
August/2019

Altigran Soares da Silva (alti)

alti@icomp.ufam.edu.br

Instituto de Computação

Universidade Federal do Amazonas



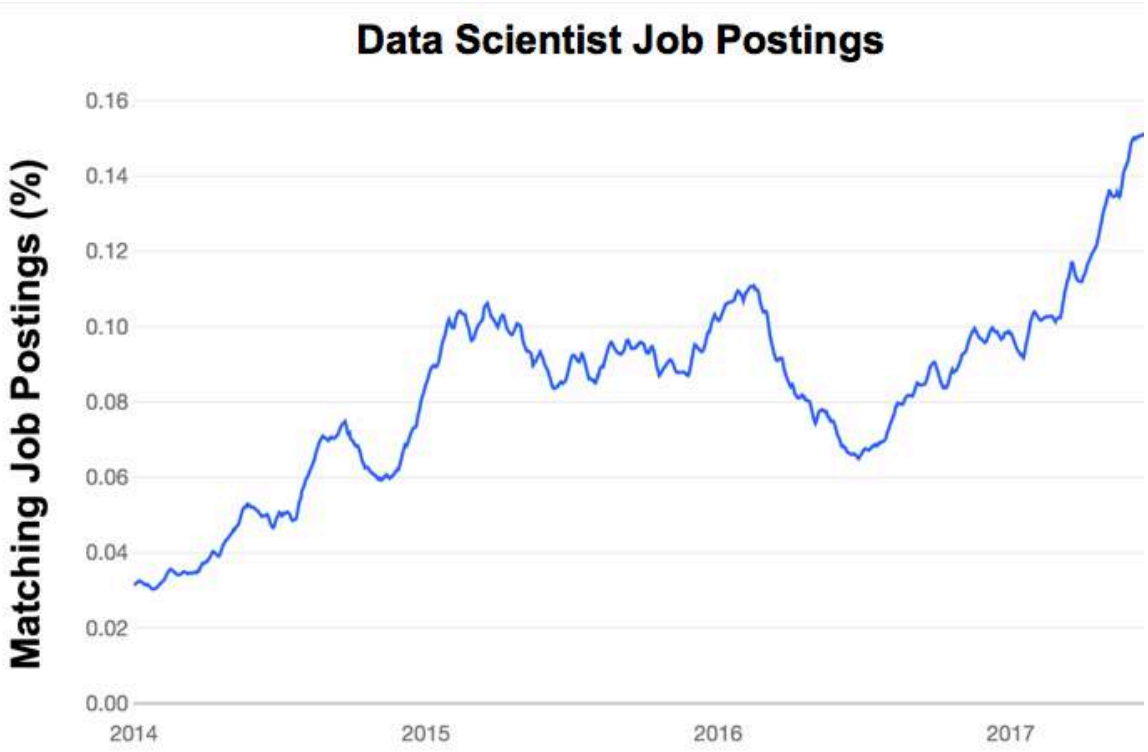
Summary

- Data Science : Huge interest from the industry and the academia
- However, no less than 80% of the time and efforts are spent with tasks related to the preparation of the data to be analyzed
  - Acquisition, extraction, deduplication, integration, cleaning, **protection**
- Huge demand, few people
- Heavy work, error prone. Need to be carried out automatically
- Great challenge for the community of Data Engineering Researchers
- Concrete Examples
  - Recent results of my research on Data Engineering Methods for allowing sentiment analysis over user-written opinionated text
  - Other stuff

Data Science: a decade-old *buzzword*

- "Information Platforms and the Rise of the Data Scientist"
  - Jeff Hammerbacher, Beautiful Data: The Stories Behind Elegant Data Solutions (2009)
- The *"sexiest job of the 21st century" (Harvard Business Review)*
- Students want to be taught "data science"
- Common believe:
  - Data Science is about Machine Learning and Statistical Modeling

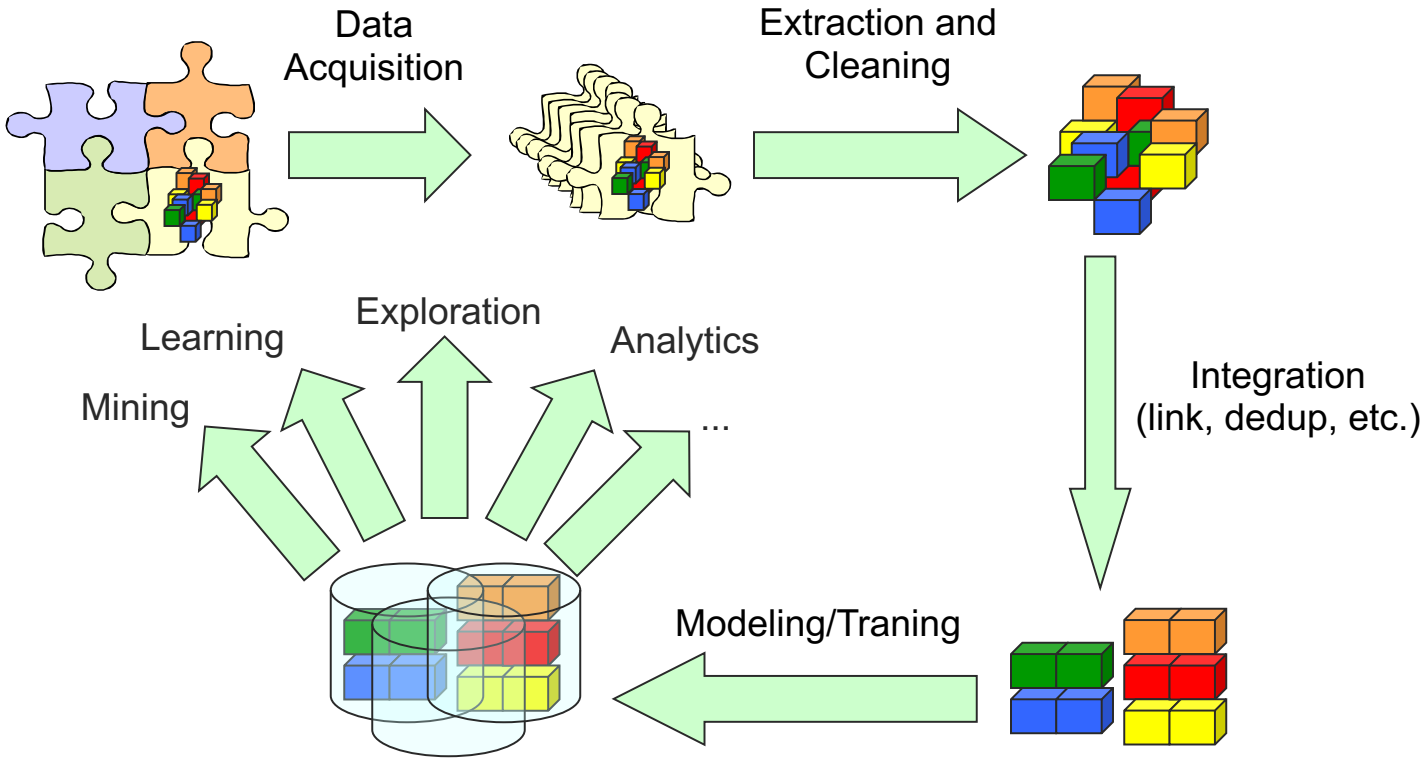
The *"sexiest job of the 21st century"*



Data Science: a definition (don't blame me!)

- Berman et al. CACM 61(4), April 2018
  - Processes and systems to extract knowledge or insight from data in various forms and translate it into action.
  - Interdisciplinary field that integrates approaches from statistics, data mining, predictive analytics
  - Incorporates advances in scalable computing and data management.

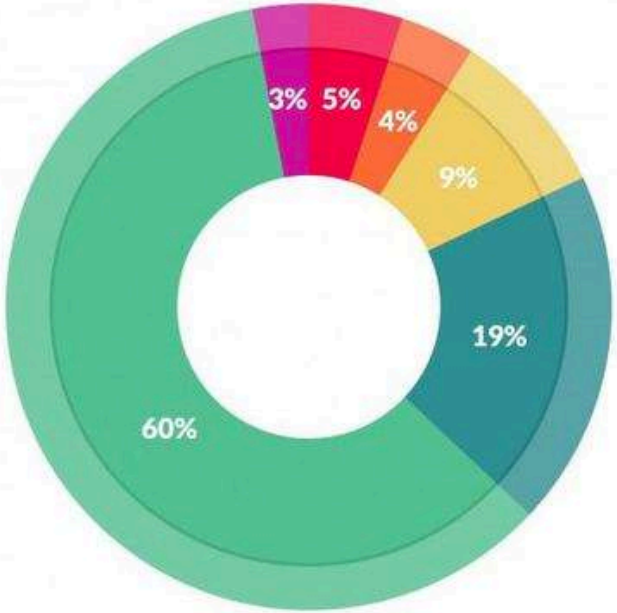
The Big Data Pipeline



H. V. Jagadish ACM SIGMOD Blog - 06/2012

Data Science: Reality (FORBES 2016)

- 80% of time of data scientists spent on data pre-processing, cleansing, etc.



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%*

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says

Too much work …

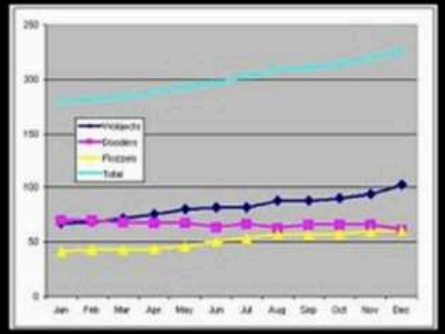
- The subset of interesting/useful data is not known in advance
- Thus, the preparation phase must include all owned data which is possibly relevant to data analysis
- Typically, no more than 10% to 12% of the full volume of data is really needed
- Too demanding in terms of time and resources
- Complex and error-prone
- Cost increases as a function of the exponential data growth.

Data Scientist

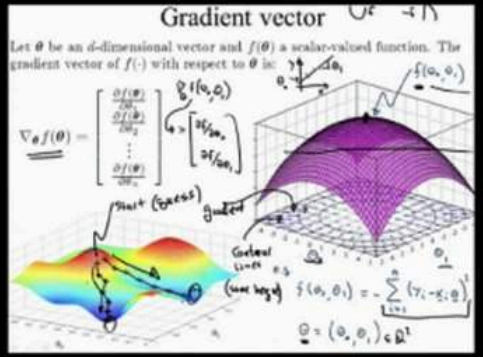
What my friends think I do

What my mom thinks I do

What society thinks I do



What my boss thinks I do



What I think I do

What I actually do

Economic Impact

- Revenue from big data hardware/software/services:
  - US$ 13 B in 2013
  - Annual growth rate of 60%.
- The UK Government's Information Economy Strategy states:
  - *"The overwhelming majority of information economy businesses – 95% of the 120,000 enterprises in the sector – employ fewer than 10 people"*

Department for Business, Innovation & Skills. Information economy strategy. http://bit.ly/1W4TPGU, 2013.

What is needed ?

- Organizations that can benefit from data-driven processes (i.e., data science) will not be able to spend substantial resources to carry out these processes
- Massive automation of Data Engineering processes is needed
- Manual intervention, if any, should be limited to high-level feedback and to the specification of exceptions

Department for Business, Innovation & Skills. Information economy strategy. http://bit.ly/1W4TPGU, 2013.

The Data Science Education panel @ ICDE 2017

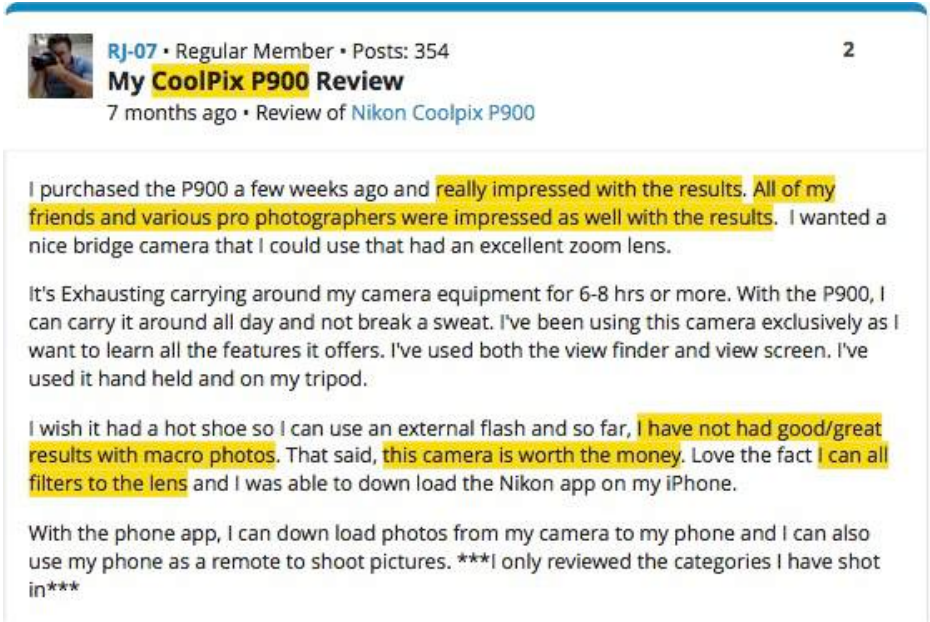
- **"Data Science Education: We're Missing the Boat, Again",**
- There is a black art to making our systems sing and dance at scale, even though we like to pretend everything happens automatically.
- How can we stop pretending and start teaching the black art in a principled way?
- A second wave of data science:
  - Ethics and Legal compliance,
  - Scientific reproducibility
  - Data quality
  - Algorithmic bias

The Data Civilizer Project

- **MIT/CSAIL**, Qatar CRI, TU/Berlin and the University of Waterloo
- Michael Stonebraker – Data Engineering Giant, Turing Award 2014
- Build an end-to-end data discovery system to assist data scientists
  - locating data of interest in an enterprise or on the public web,
  - transforming it to a common format
  - cleaning the data into usable information
  - performing schema matching to line up multiple data sets
  - performing entity consolidation.
- Papers: CIDR'17, SIGMOD'17 and ICDE'18
- Use case: Merck, "Big Pharma" +4000 databases, a data lake, uncountable numbers of files and data off the public web.

Opinion Mining

- Consumers love opinions
  - 92% of users put more trust on information about products and services published in social media by regular users, than on information published in other, more traditional sources, such as advertisement
  - 51% of the people who shop outside the Web, make decisions based on online comments and reviews



RJ-07 • Regular Member • Posts: 354 2

My CoolPix P900 Review

7 months ago • Review of Nikon Coolpix P900

I purchased the P900 a few weeks ago and really impressed with the results. All of my friends and various pro photographers were impressed as well with the results. I wanted a nice bridge camera that I could use that had an excellent zoom lens.

It's Exhausting carrying around my camera equipment for 6-8 hrs or more. With the P900, I can carry it around all day and not break a sweat. I've been using this camera exclusively as I want to learn all the features it offers. I've used both the view finder and view screen. I've used it hand held and on my tripod.

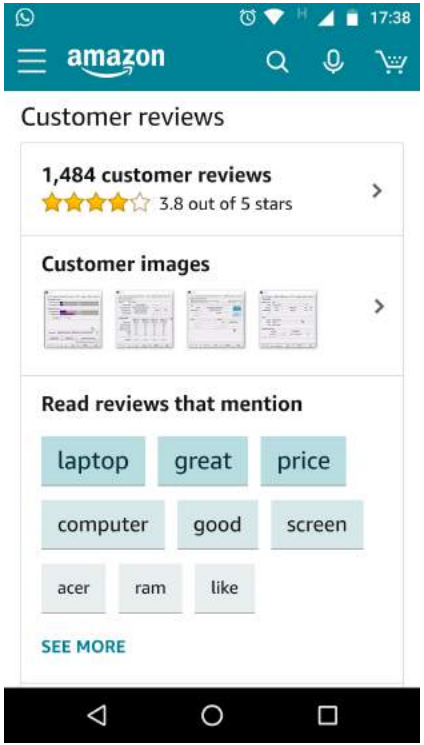
I wish it had a hot shoe so I can use an external flash and so far, I have not had good/great results with macro photos. That said, this camera is worth the money. Love the fact I can all filters to the lens and I was able to down load the Nikon app on my iPhone.

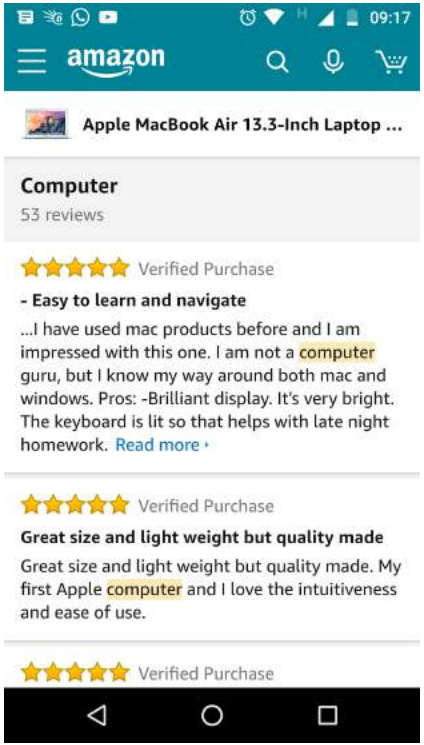
With the phone app, I can down load photos from my camera to my phone and I can also use my phone as a remote to shoot pictures. ***I only reviewed the categories I have shot in***

Opinion Mining

- Active research area in many communities
- Initial results already reaching the market
- A lot more to do
- Challenges
  - Too many reviews for humans to read and process
  - Reviews written informally
  - Useful opinions are sparse
  - Not all reviews are trustful

Amazon Themes





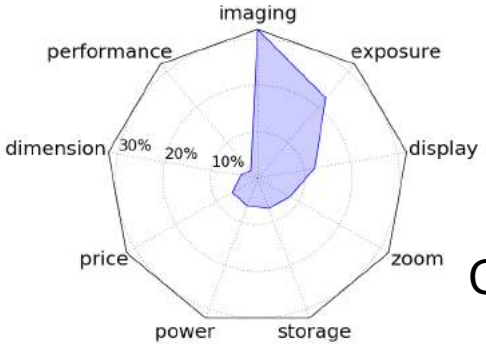
Possible Outcomes (Data Science “Products”)

- Predicting user group behavior
- Providing reliable, explainable and valuable recommendations
- Summarizing user group opinions
- Predicting users sentiment polarity towards products or their features
- Estimating the return of investment of advertisement content
- Which are the most important features on a given product?

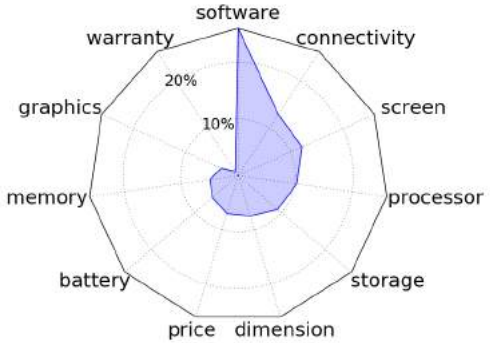
Opinion Mining @ IComp/UFAM

- Sources: Reviews on e-commerce sites and Internet Forums
- Problems:
  - Too many reviews to read for each product, specially for the most popular
  - Reviews are hard for customers to "process": read, understand, extract knowledge
- Overall Goal: organize opinions to make it easy processing them
  - In product reviews: link opinions on products to the attributes they refer to
  - In forums: link opinions to products they refer to
- Applications
  - Searching, recommendation, pricing, product design, etc.
- Projects: E-vox (FAPEAM), E-spot (CNPq), SocSens (PGCI/CAPES + NYU)

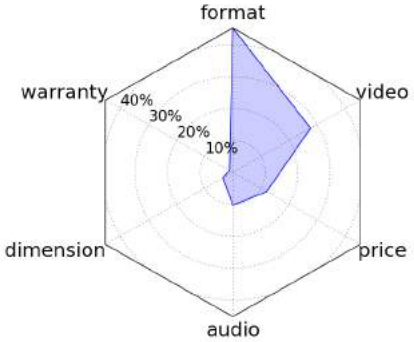
Which are the most important features?



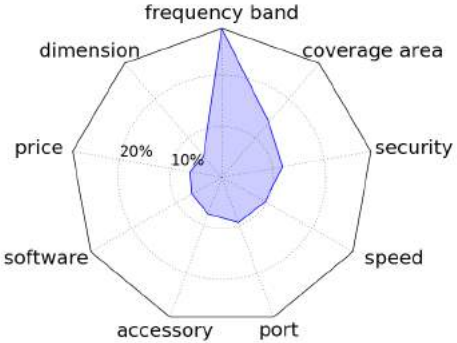
Cameras



Laptops



Midia Players



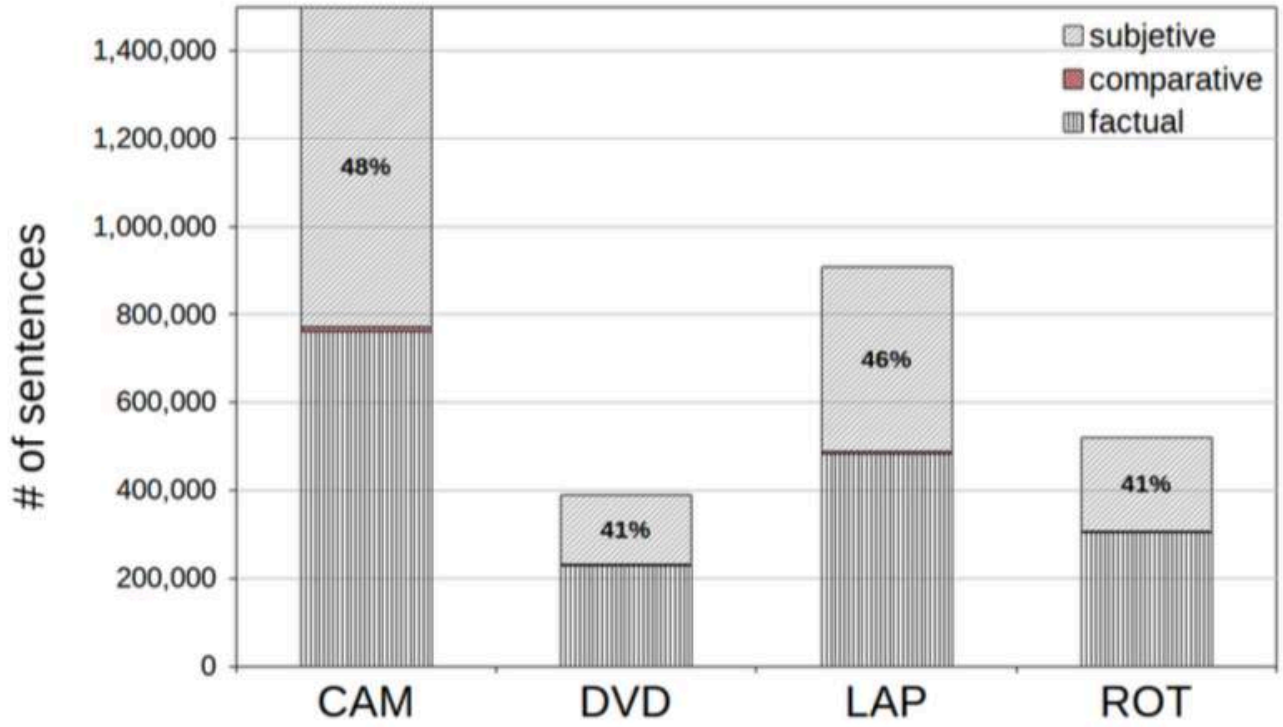
Routers

Experimental Dataset

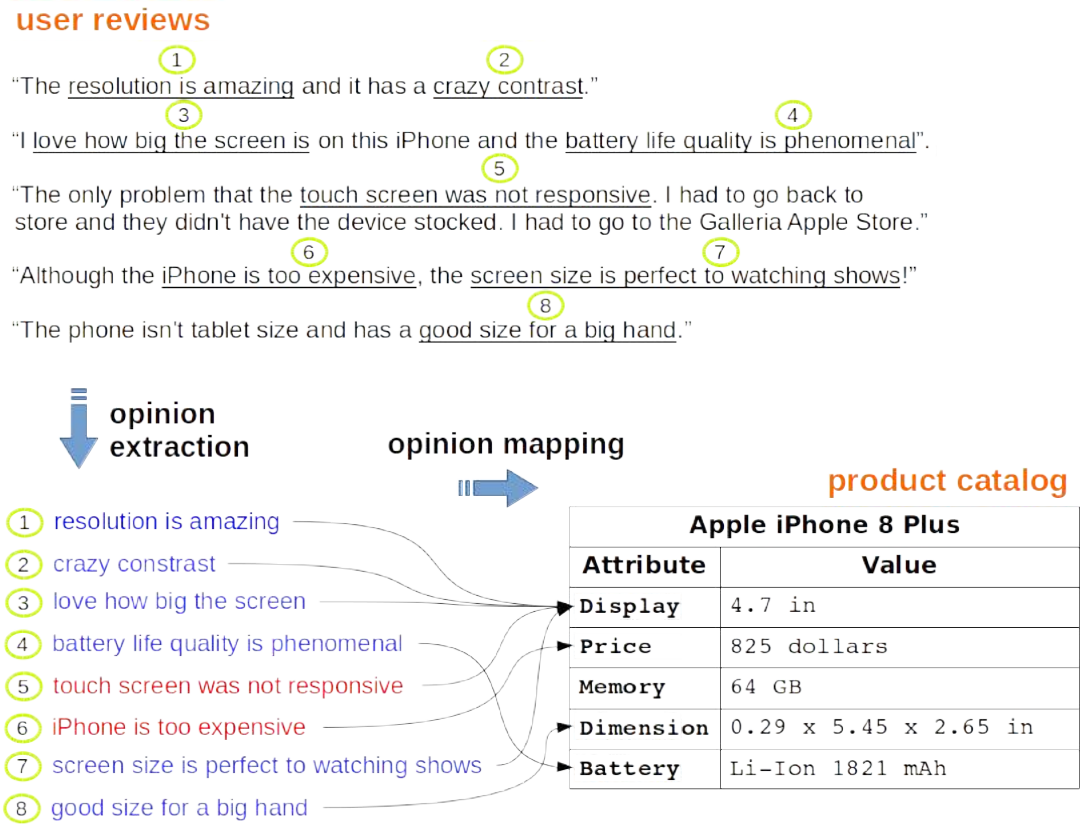
Category	Products	Reviews	Sentences	
			Total	Target
CAM	8,839	204,127	1,499,405	726,388
DVD	2,503	61,997	390,812	159,317
LAP	9,491	115,521	907,031	417,278
ROT	1,592	84,270	520,752	212,853
Total	**22,375**	**464,871**	**3,318,000**	**1,515,836**

Extracted from the "Amazon Product Dataset" Julian McAuley, UCSD

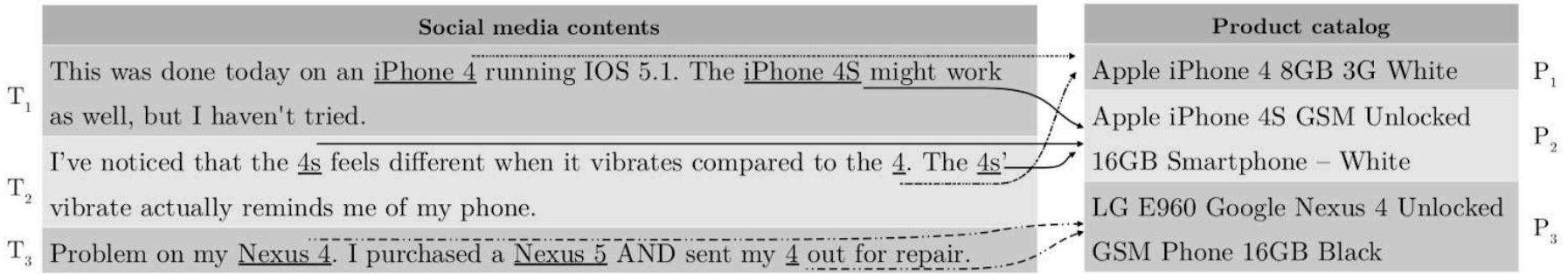
Sentences in Reviews



Problem 1: Linking Opinions to Product Attributes



Problem 2: Product Identification in Forum Posts



Problem 2: Product Identification in Forum Posts

- Strategy:
  - First: identify product mentions (surface forms) in the posts (ECIR'15)
  - Then: link mentions with the correct product from a catalog (CIKM'16)

- Overview
  - Distantly supervised approach.
  - Sample surface automatically extracted from product catalog
  - Then, training sentences are "synthesized" from real sentences from the corpus (forums) that match the surface forms.

JIT DBMSs for Semi-structured Text

- DIAS@EPFL and IComp@UFAM
- JIT DBMSs
  - Require no data preparation to answer queries on complex heterogeneous datasets;
  - Upon reception of a query, navigate related data sources making the necessary transformations and code-generating access paths on-the-fly.

Conversational Interfaces for RDBMSs

Retorne os nomes dos ***professores*** *que ministraram mais* ***turmas*** *que o* ***professor*** *João Marcos no semestre 2019/02*

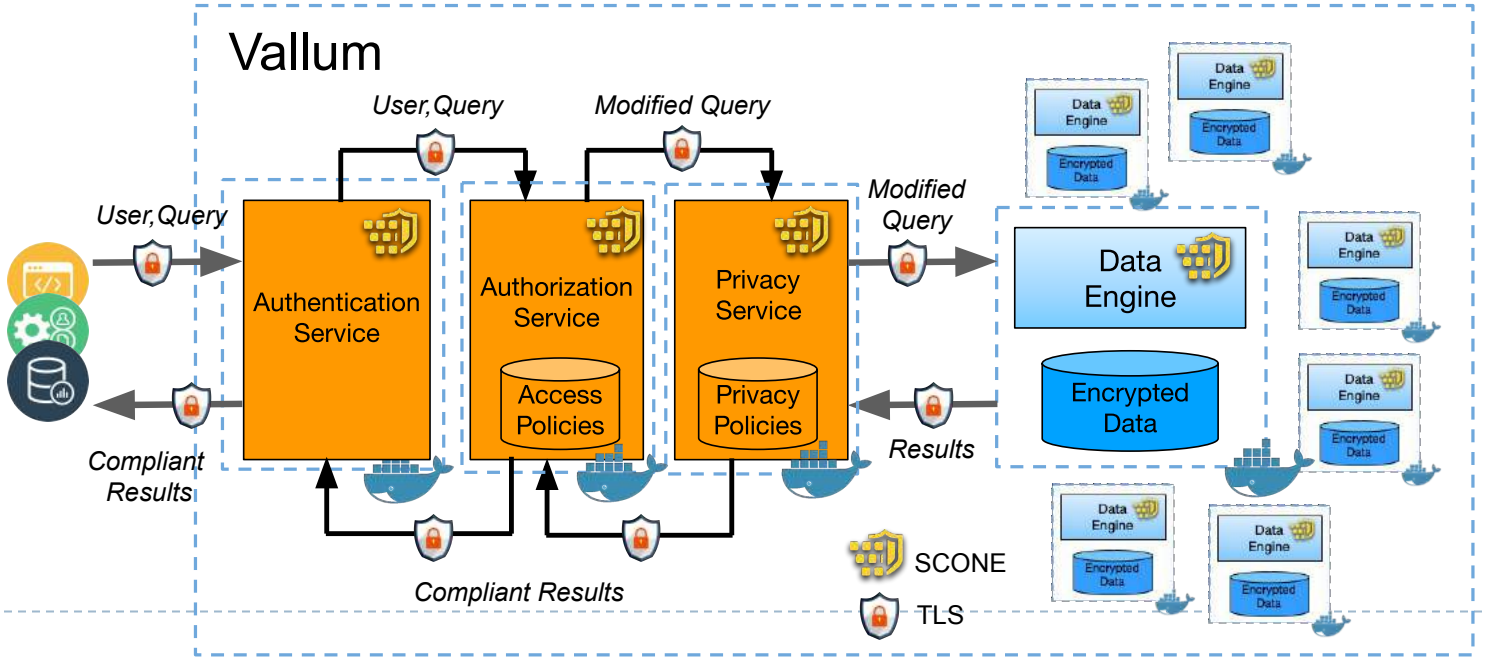
(a) Consulta formulada em linguagem natural

```
SELECT     nome, COUNT(*)
FROM       Turma T1, Docente D1
WHERE      T1.siape = D1.siape AND
           T1.semestre = "2019/02"
GROUP BY   nome
HAVING  COUNT(*) >  (SELECT  COUNT(*)
                     FROM    Turma T2, Docente D2
                     WHERE   T2.siape = D2.siape AND
                             D2.nome = "João Marcos"AND
                             T2.semestre = "2019/02")
```

(b) Consulta SQL a ser gerada

Vallum

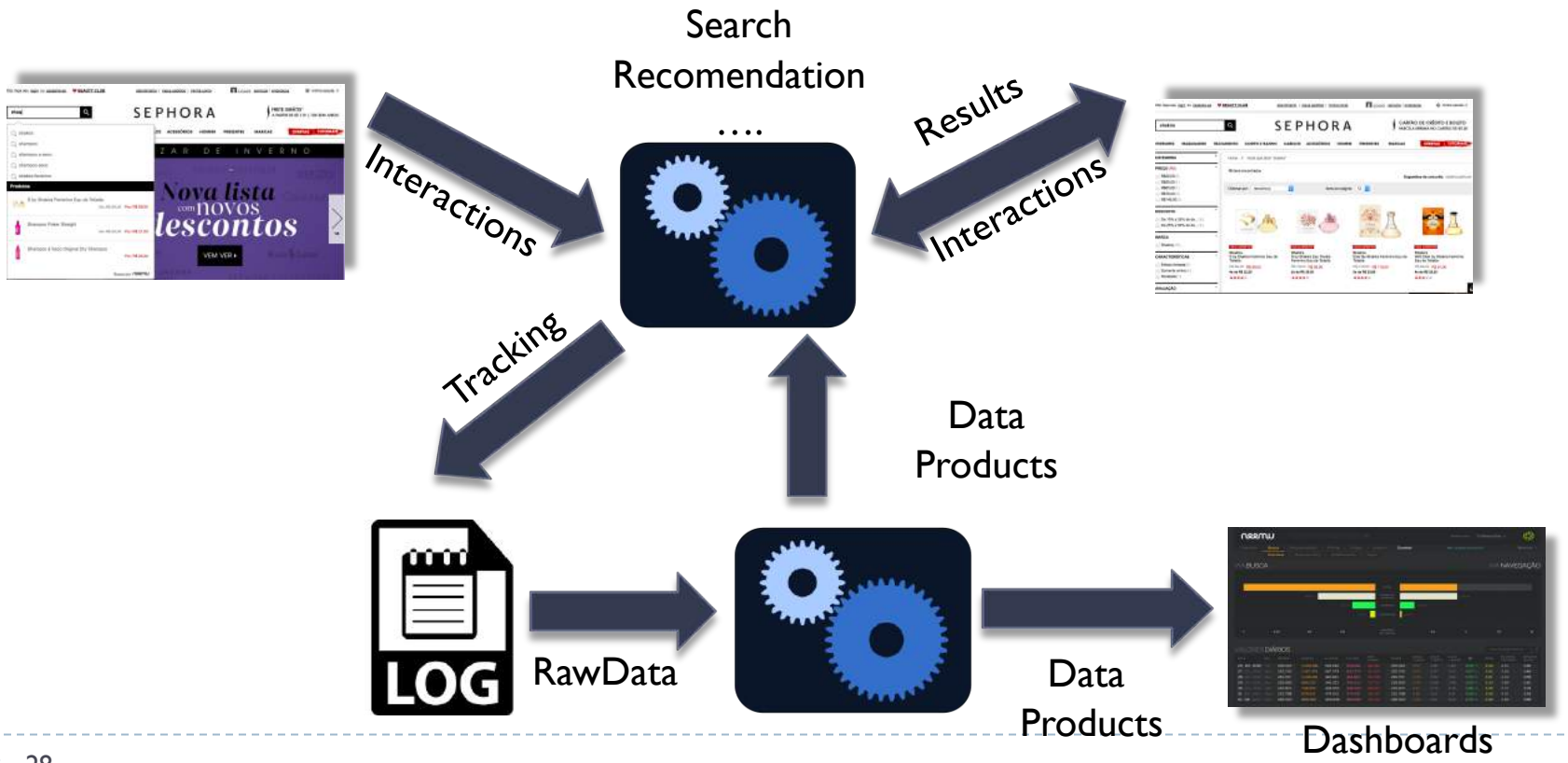
Vallum: Trusted layer for providing security and anonymity for existing data engines (e.g., Relational DBMS, Key/Value, etc.). ATMOSPHERE EU/BR.



- Created in 2010 by IComp/UFAM students and professors
- Big Data for e-Commerce
- Product-oriented Search and Recommendation Systems
  - Intensive use of data on user behavior
- 1/3 of Brazilian E-commerce transactions by 2014
  - +50 million monthly queries
  - +200 TB transaction data

- Acquired by Linx Systems in 2015: 70% market-share

Neemu Data Platform circa 2014



teewa

- Created in 2016 by students and professors from IComp/UFAM
- Chat-based platform to allow customers and local sellers to do business interactively
- O2O – Of-line to On-line
  - Some sellers cannot afford having a web site, but they carry a smartphone everywhere
- Shopping-oriented chatbots to handle repetitive tasks
- Owners/Clerks can handle special needs
- Funding from Monashees+ e ABCapital

Conclusions

- Data Science : Huge interest from the industry and academia
- However, no less than 80% of the time and efforts are spent with tasks related to the preparation of the data to be analyzed
  - Acquisition, extraction, deduplication, integration, cleaning, **protection**
- Huge demand, few people
- Heavy work, error prone. Need to be carried out automatically
- Great challenge for the community of Data Engineering Researchers