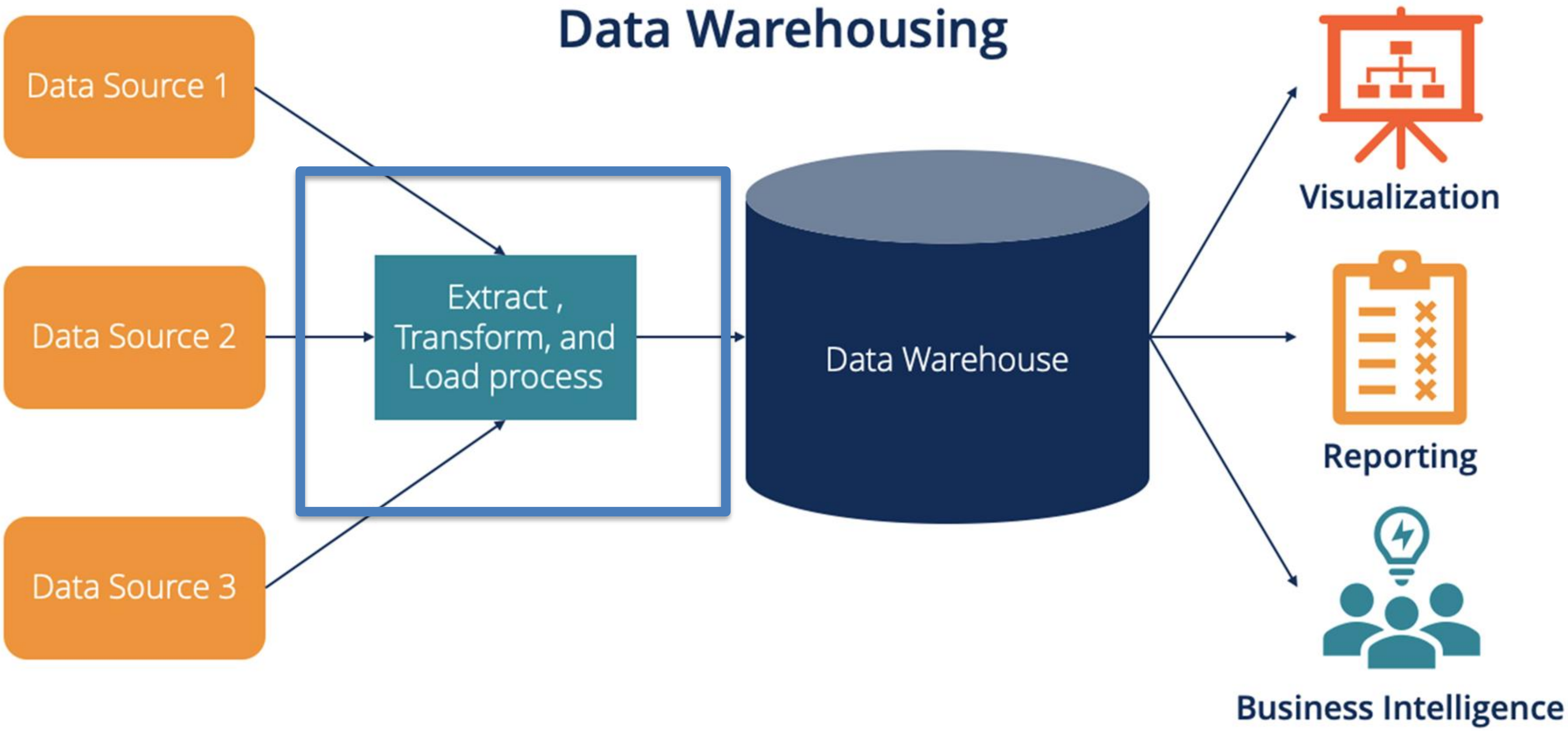
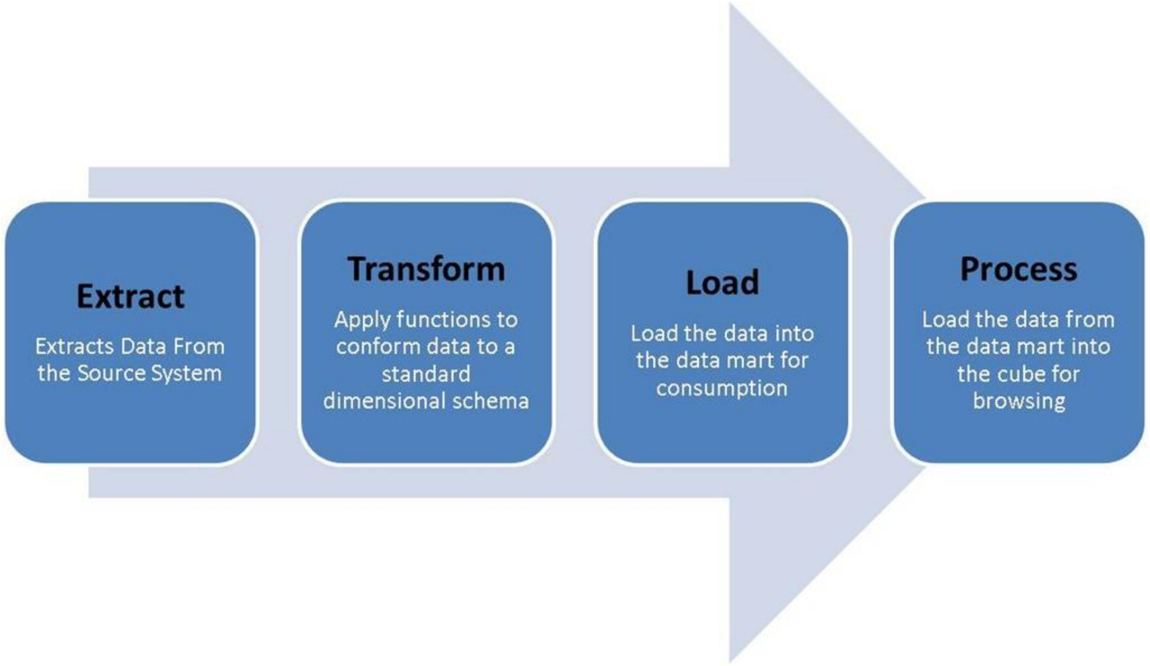
Part b Design aspects of a DW

1. **Select:** Which data and what for
2. **Transform:** so-called **ETL**: Extraction, cleaning, Transform and Load data
3. **Store and process data:** data Mars, metadata, aggregations

Step 2: ETL: extraction, cleaning, transform and load data





Extract
Extracts Data From the Source System

Transform
Apply functions to conform data to a standard dimensional schema

Load
Load the data into the data mart for consumption

Process
Load the data from the data mart into the cube for browsing

- It is important to understand that a data warehouse has the purpose of integrating **different sources** of data, not just COLLECTING **new data.**
- So, new data are added, deleted, and updated in the ORIGINAL sources (e.g. an OLTP, or in the original source).
- The data warehouse must **extract** new data as they are generated, detect and handle ***changes*** in old data, and **integrate** data from the different sources.

What is ETL

- Extraction–transformation–loading (ETL) tools are pieces of software responsible for
  - the extraction of data from several sources,
  - its cleansing, customization, reformatting, integration, and
  - storage into a data warehouse.
- Building the ETL process is potentially **one of the biggest tasks of creating a warehouse**; it is complex, time consuming, and consumes most of data warehouse project’s implementation efforts, costs, and resources.

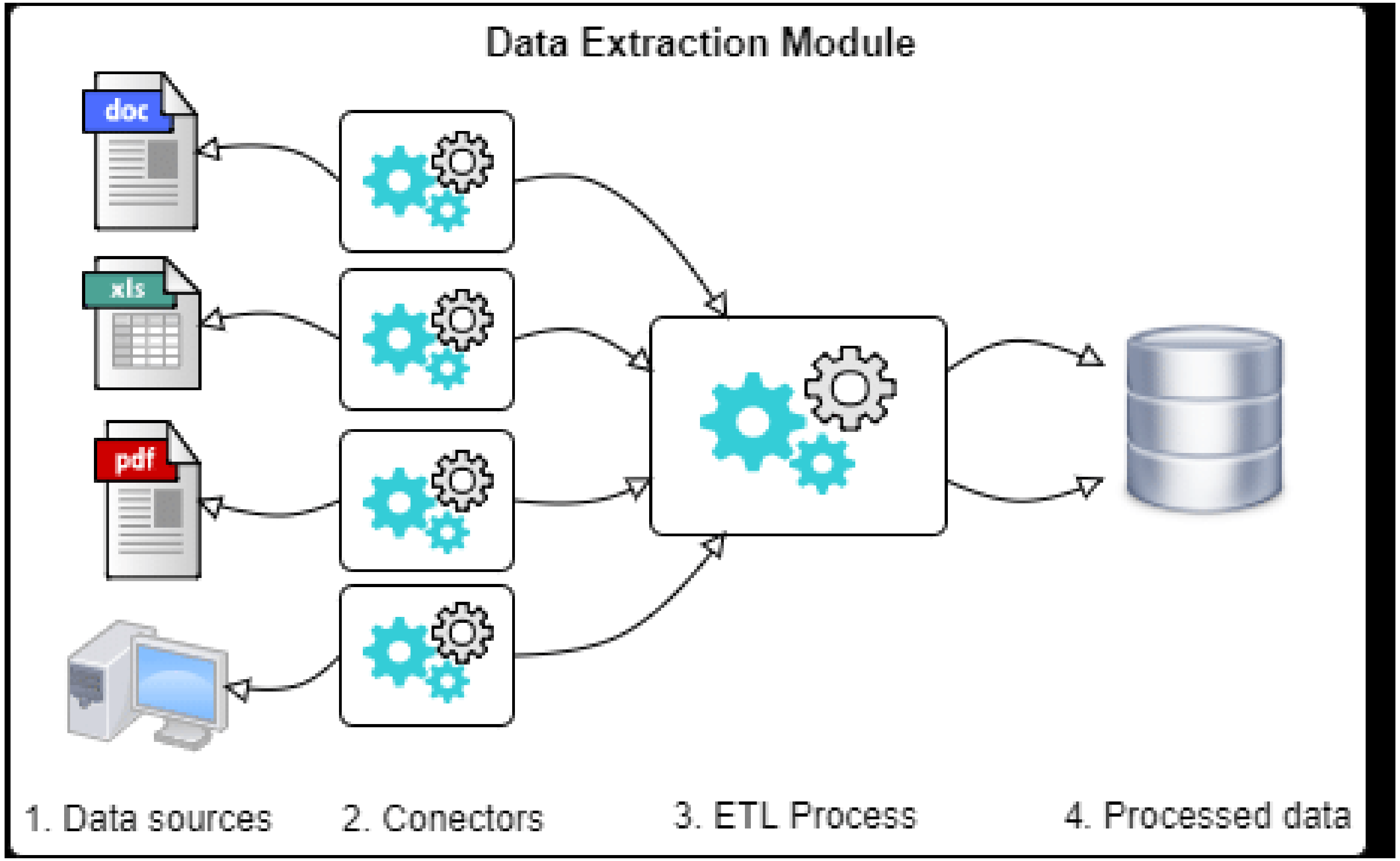
ETL Functional Elements

- ETL systems have a common purpose: **they move data from one database to another.**
- Generally, ETL systems move data from OLTP systems (or from external sources) to a data warehouse.
- An ETL system consists of **four distinct functional elements**:
  - Extraction
  - Transformation (cleaning, alignment of data, ecc. ..will see)
  - Loading (the result of Extraction and Transformation on the DW)
  - Adding Metadata to the DW

ETL 1. Extraction

- The first step in any ETL scenario is data extraction.
- The ETL extraction step is responsible for extracting data from the source systems.
- Each data source has its distinct set of characteristics that need to be managed in order to effectively extract data for the ETL process.
- The process needs to integrate systems that have different platforms, such as different database management systems, different operating systems, and different communications protocols.

Extraction


Data Extraction Module
doc
xls
pdf
1. Data sources
2. Conectors
3. ETL Process
4. Processed data

Issues: Extraction frequency

- There are several ways to perform the extract:
  - **Update notification** - if the **source system** is able to provide a notification that a record has been changed in the original data source and describe the change (e.g. a new shipment has been completed, and order has been filed..), this is the easiest way to get the data.
  - **Incremental extract** – No notifications, so in given **time intervals** the extraction process start, source system should be able to identify which records have been modified and provide an extract of such records. During further ETL steps, the system needs to identify changes and propagate it down.
  - **Full extract** - some systems are not able to identify which data has been changed at all, so a full extract is the only way one can get the data out of the system. The full extract requires keeping a copy of the last extract in the same format in order to compare and be identify changes. Full extract handles deletions as well.
  - **Extract from unstructured resources** – If data are not structured (not a database) system extracts *either in real time (in streaming) or incrementally*, but new data are simply added to old data (e.g. new tweets discussing about a given product).

Take away message

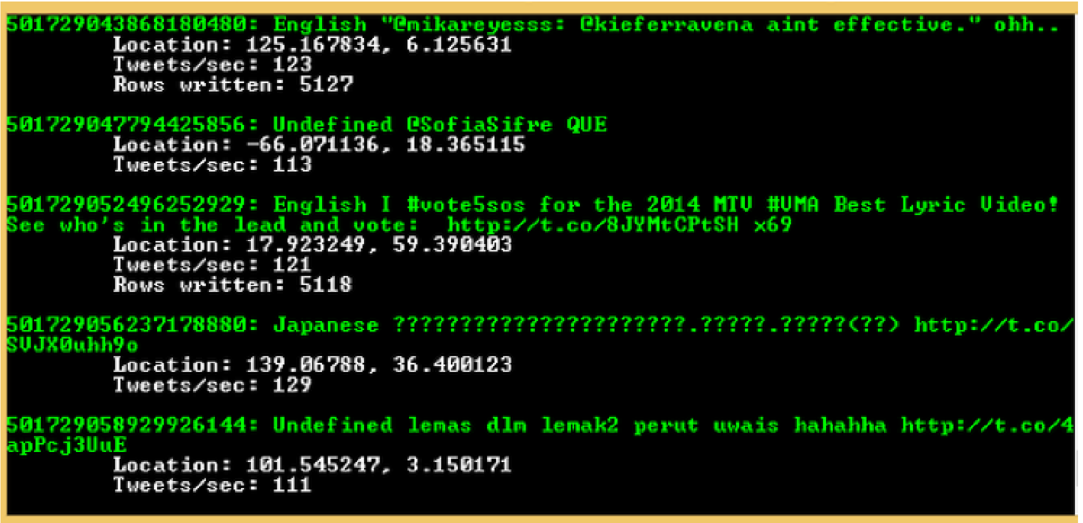
- You are not responsible for the extraction process, **IT people will be**
- Your responsibility is to help deciding – having in mind objectives of the analysis and timing constraints – **which data should be extracted, and (about) what frequency of extraction**.
- E.g., if the objective is to predict credit card frauds, need *real-time updating*. If objective is to analyze and compare point-of-sales, weekly or monthly extraction can be enough

In class exercise: what updating policy and which sources would you use for these applications? (real-time vrs incremental)

- Telephony (Churn prediction)
- Transportation (traffic management)
- Energy and utilities (energy savings)
- Health (remote healthcare; epidemic warning systems)
- Natural systems (water management)
- Law, defense, cybersecurity (surveillance systems, cybersecurity detection)
- Stock market (marked data analysis)
- Fraud detection (credit card, insurance, taxes)
- eScience (weather prediction)

What about unstructured data?

- Need specialized software to download data streams (e.g. Twitter API)
- But most of all, it needs complex **transformation**
- We first consider transformation methods only for structured data (since manipulating unstructured data such as text and images requires complex artificial intelligence based algorithms – will shortly describe this later on)



```
501729043868180480: English "@mikareyesss: @kieferravena aint effective." ohh..
        Location: 125.167834, 6.125631
        Tweets/sec: 123
        Rows written: 5127

501729047794425856: Undefined @SofiaSifre QUE
        Location: -66.071136, 18.365115
        Tweets/sec: 113

501729052496252929: English I #vote5sos for the 2014 MTV #VMA Best Lyric Video!
See who's in the lead and vote:  http://t.co/8JYMtCPtSH x69
        Location: 17.923249, 59.390403
        Tweets/sec: 121
        Rows written: 5118

501729056237178880: Japanese ??????????????????????.?????.?????(??) http://t.co/
SVJX0uhh9o
        Location: 139.06788, 36.400123
        Tweets/sec: 129

501729058929926144: Undefined lemas dlm lemak2 perut uwais hahahha http://t.co/4
apPcj3UuE
        Location: 101.545247, 3.150171
        Tweets/sec: 111
```

ETL

2. Transformation

- The second step in any ETL scenario is data transformation.
- Objective: make some cleaning and conforming on the incoming data to gain accurate data which is **correct, complete, consistent, and unambiguous.**
- For all datatypes, this process includes data cleaning, transformation, and integration. It defines the granularity of fact tables, the dimension tables, data structures, etc.
- **Note:** if source data is unstructured (images, text, signals), transformation also imply converting from unstructured to structured (tables)!! We now consider only transformation of structured data, text image and signal processing will be introduced later!
- **All transformation rules and the resulting schemas must be described in the metadata repository**.
- Will see later, but **your responsibility** (as business experts in a BI project) is that a **comprehensible** (by business people) description of what kind of transformations are performed on the data is maintained!

Data transformation and cleaning

Data in the real world is **dirty**

- **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=“”
- **noisy**: containing errors or outliers (spelling, phonetic and typing errors, word transpositions, multiple values in a single free-form field)
  - e.g., Salary=“-10”
- **inconsistent**: containing discrepancies in codes or names (synonyms and nicknames, prefix and suffix variations, abbreviations, truncation and initials)
  - e.g., Age=“42” Birthday=“03/07/1997”
  - e.g., Was rating “1,2,3”, now rating “A, B, C”
  - e.g., discrepancy between duplicate records

Why data is dirty?

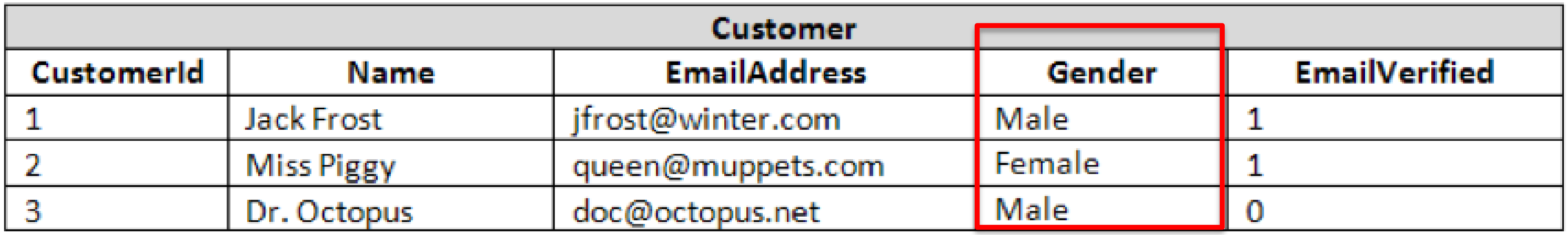
- Incomplete data comes from:
  - Not available data when collected
  - Criteria changed (e.g. collect twitter messages with user ID, then GDR rules e no more user ID collected)
  - Human/hardware/software problems
- Noisy data come from:
  - Faulty intruments, Human errors, transmission errors
- Inconsistent/redundant data comes from:
  - Different data sources with different data description models

What can we do with «dirty» data?

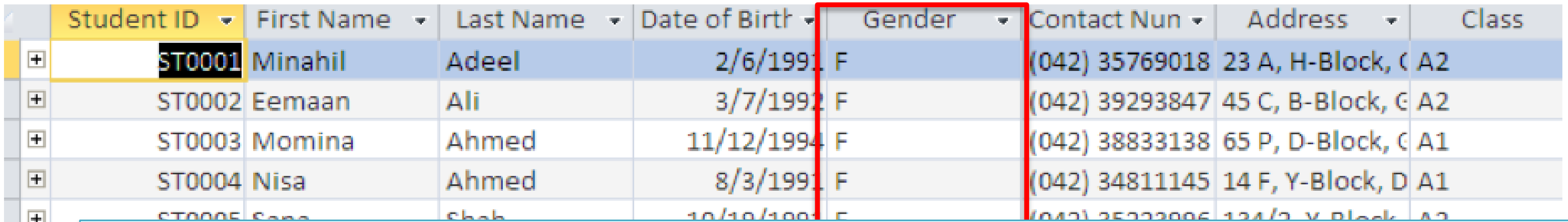
- You work on data cleaning when using Watson Studio (the process is called data REFINERY)
- Incomplete, noisy data: if an attribute in a table has mostly noisy of empty values, better to cancel the entire column!
- If a column has a *missing* or unclear attribute name, you can change it
- There are, however, many machine learning algorithms to cope with incomplete data (automated «imputation»)
- Inconsistent data are a more complex problem

Example of inconsistent data

- As a small example, assume you have data coming from two different source systems which you want to merge in the data warehouse: there might be some differences between the two.
- For example, one source may denote the *gender* as Male and Female while other may denote as F and M.



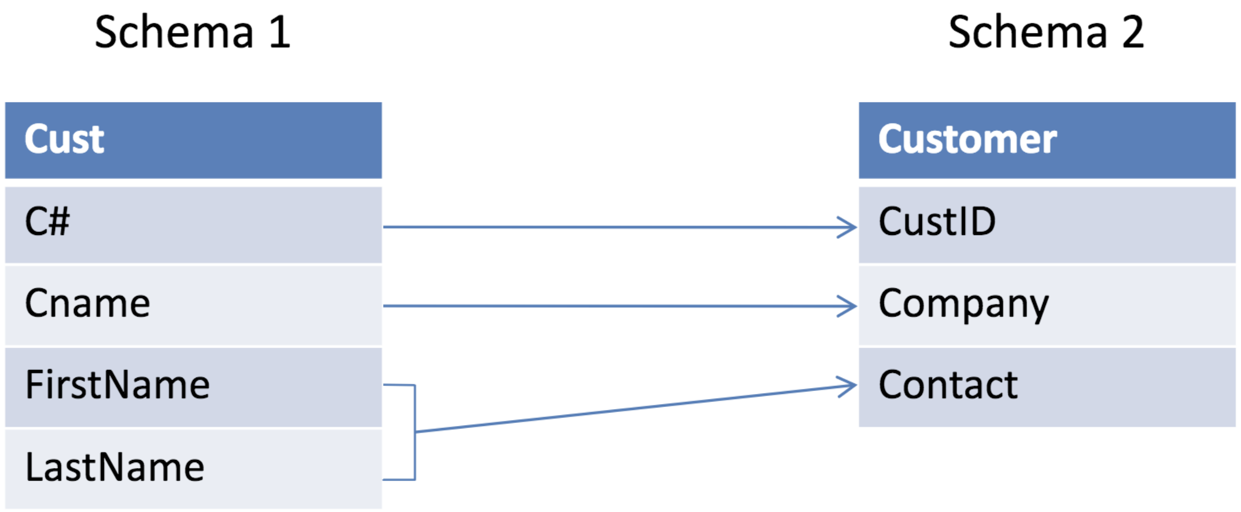
Customer				
CustomerId	Name	EmailAddress	Gender	EmailVerified
1	Jack Frost	jfrost@winter.com	Male	1
2	Miss Piggy	queen@muppets.com	Female	1
3	Dr. Octopus	doc@octopus.net	Male	0



Student ID	First Name	Last Name	Date of Birtł	Gender	Contact Nun	Address	Class
ST0001	Minahil	Adeel	2/6/1991	F	(042) 35769018	23 A, H-Block, (	A2
ST0002	Eemaan	Ali	3/7/1992	F	(042) 39293847	45 C, B-Block, G	A2
ST0003	Momina	Ahmed	11/12/1994	F	(042) 38833138	65 P, D-Block, (	A1
ST0004	Nisa	Ahmed	8/3/1991	F	(042) 34811145	14 F, Y-Block, D	A1

Comparing these two Tables there is another mismatch in the way the same information is encoded. Which one?

Mismatches in the schemas require CONFORMATION (also called reconciliation)



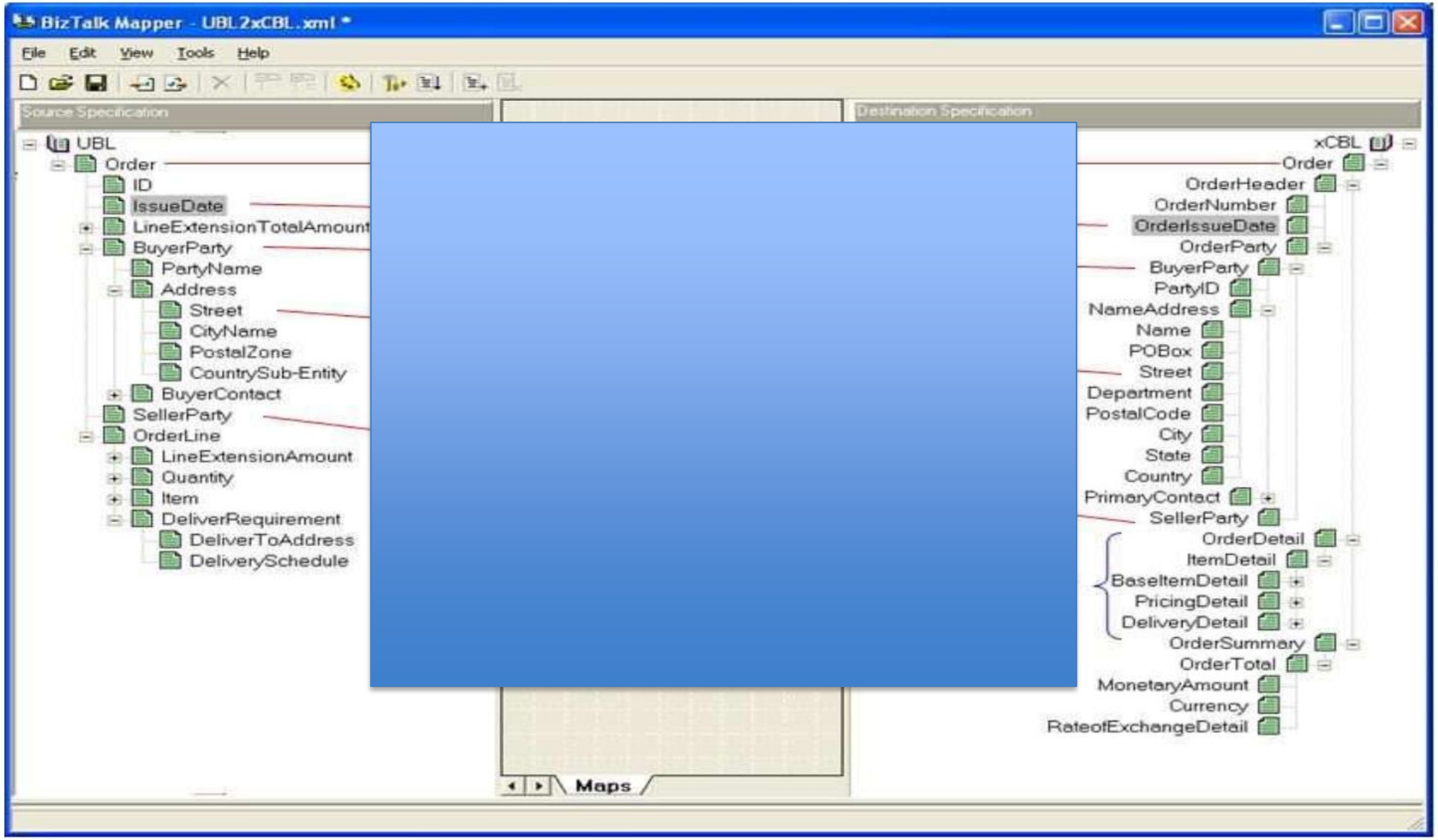
Schema 1

Cust
C#
Cname
FirstName
LastName

Schema 2

Customer
CustID
Company
Contact

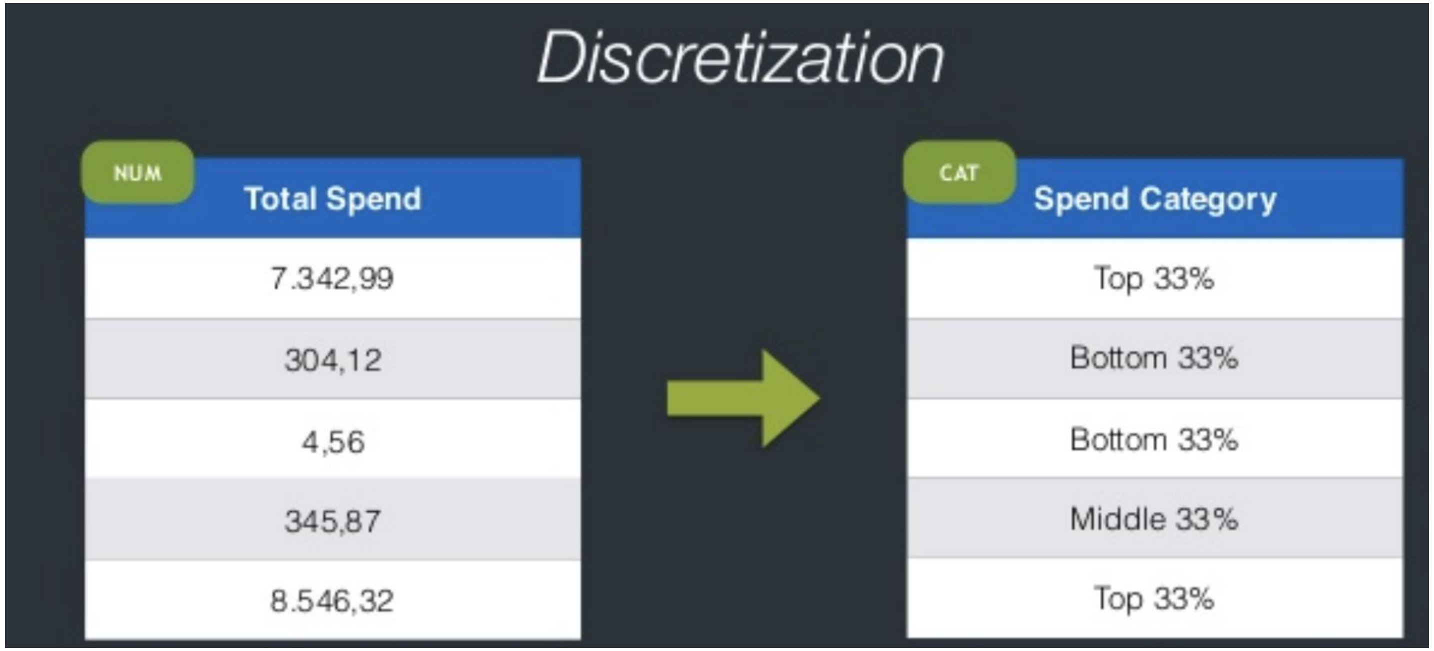
Example: aligning attribute names ("reconciling" data)



Summary of types of transformations: MODIFICATION, CONFORMATION, ADDITION

- MODIFICATION (changing the name of an attribute or value):
  - Example of **attribute** modification: If you are storing the gender in target as M and F, you may need to "transform" Male and Female to M and F (or viceversa). You may write a simple CASE statement (a RULE), or you may just write code which translates Male --> M and Female --> F.
  - Example of **values** modification: **Discretizing** attribute values: e.g., if you have «Age» as an attribute, you can define rules to change all values according. e.g., to the rule: IF age<=20 then change value to YOUNG; IF 20 < age <55 then change value to ADULT; ELSE change value to ELDER
  - You can also **define a hierarchy** of values for subsequent AGGREGATION operations. For example , if you have dates in your dataset (D1: 08/12/2020) you can define a time hierarchy day→week→month→semester→year. Now D1 can be replaced by different values according to the hierarchy:
    D1→week2→december→semester2→year2020

Example of MODIFICATION (discretization)



Discretization

NUM

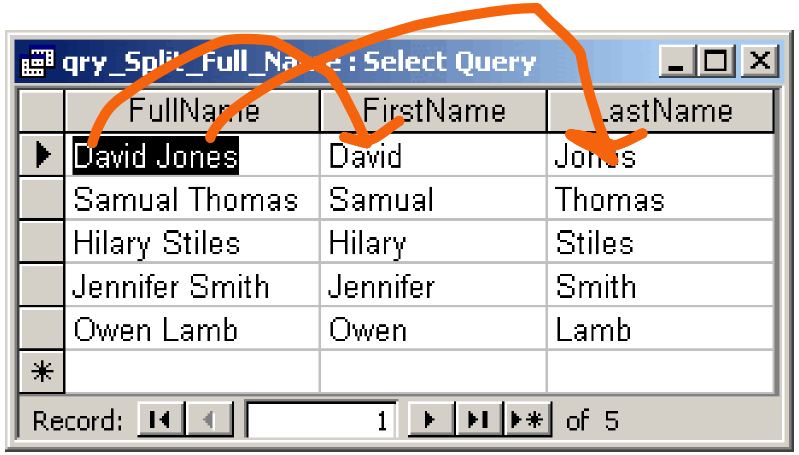
Total Spend
7.342,99
304,12
4,56
345,87
8.546,32

CAT

Spend Category
Top 33%
Bottom 33%
Bottom 33%
Middle 33%
Top 33%

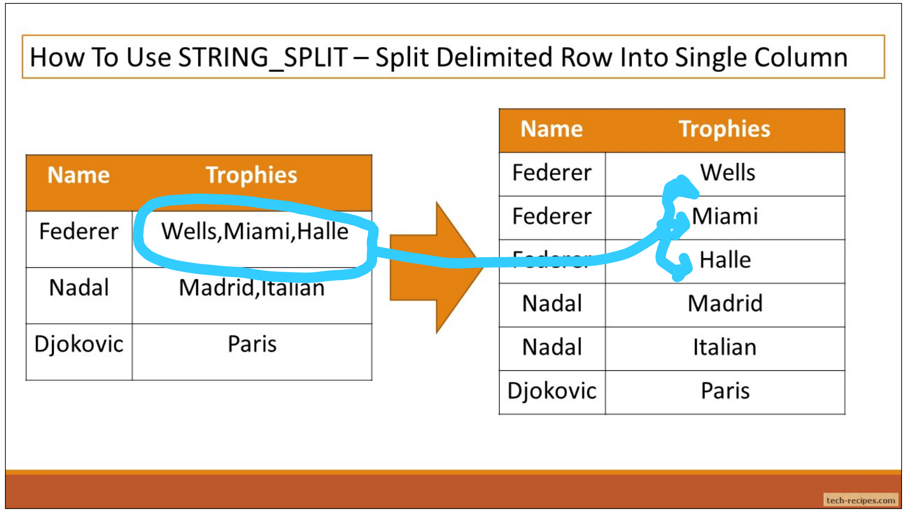
Summary of types of transformations: MODIFICATION, CONFORMATION, ADDITION (2)

- **CONFORMATION** (making two attribute compatible) : If you want to encode the Name attribute in two attributes: *First Name, Family Name*, then you must **split** the values in each record of Table 1 and record the data separately in the Target Table. Again, you do this writing some code and documenting it with a RULE.



qry_Split_Full_Name : Select Query

FullName	FirstName	LastName
David Jones	David	Jones
Samual Thomas	Samual	Thomas
Hilary Stiles	Hilary	Stiles
Jennifer Smith	Jennifer	Smith
Owen Lamb	Owen	Lamb



How To Use STRING_SPLIT – Split Delimited Row Into Single Column

Name	Trophies
Federer	Wells,Miami,Halle
Nadal	Madrid,Italian
Djokovic	Paris

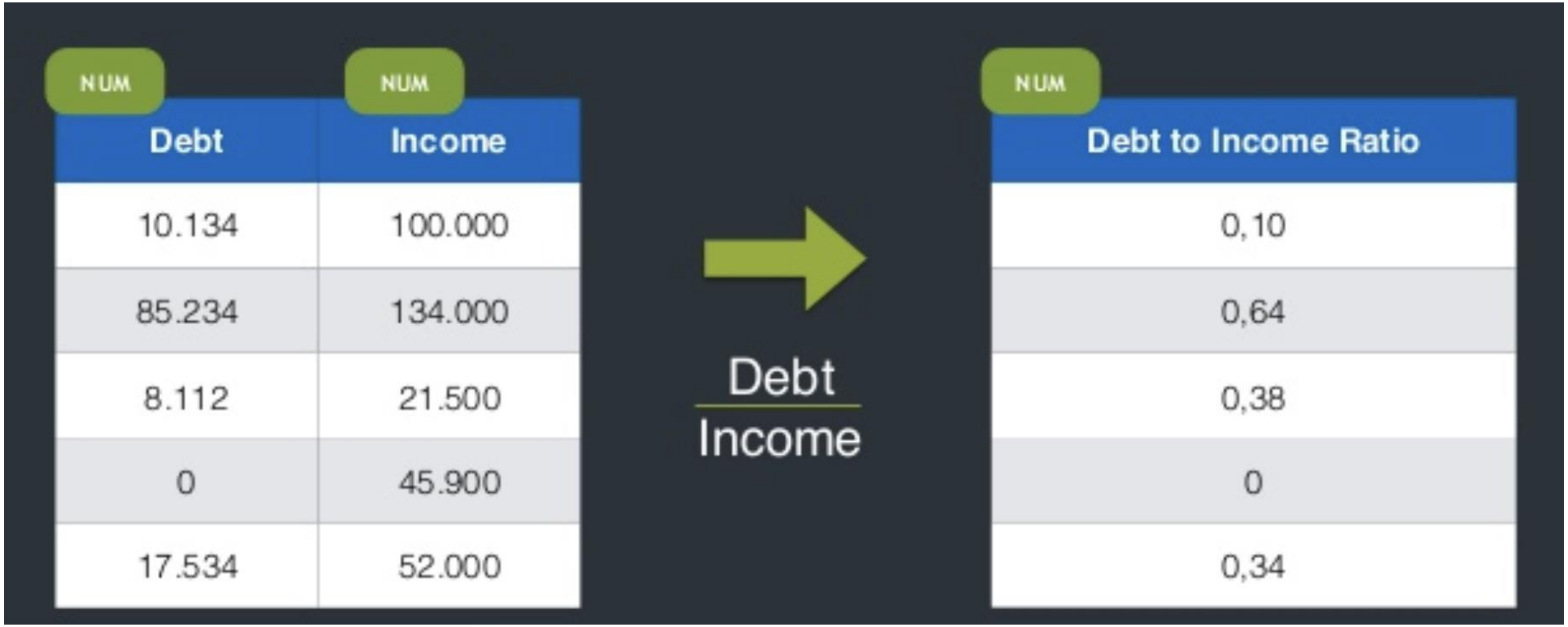
Name	Trophies
Federer	Wells
Federer	Miami
Federer	Halle
Nadal	Madrid
Nadal	Italian
Djokovic	Paris

tech-recipes.com

Summary of types of transformations: MODIFICATION, CONFORMATION, ADDITION (3)

- **ADDITION** (adding a new attribute, also called *augmentation*): In the same way, if you have a *Revenue* field in a Table maintained in Italy and another *Revenue* Field from Germany, and you need a Total Revenue in your target warehouse, you will write a function which **calculates** the sum and stores it in another column. All these modifications, additions, conformation are part of the Transform stage. These transformations must be encoded in **RULES** readable by non-ICT users.
- **IMPORTANT: the SYNTAX and SEMANTICS of the data you combine and store is a CRITICAL FACTOR. Syntactic and semantic mismatches are a major source of problems when aggregating data!**
- You will practice on these transformations during Labs

Example of ADDITION: computing a new attribute



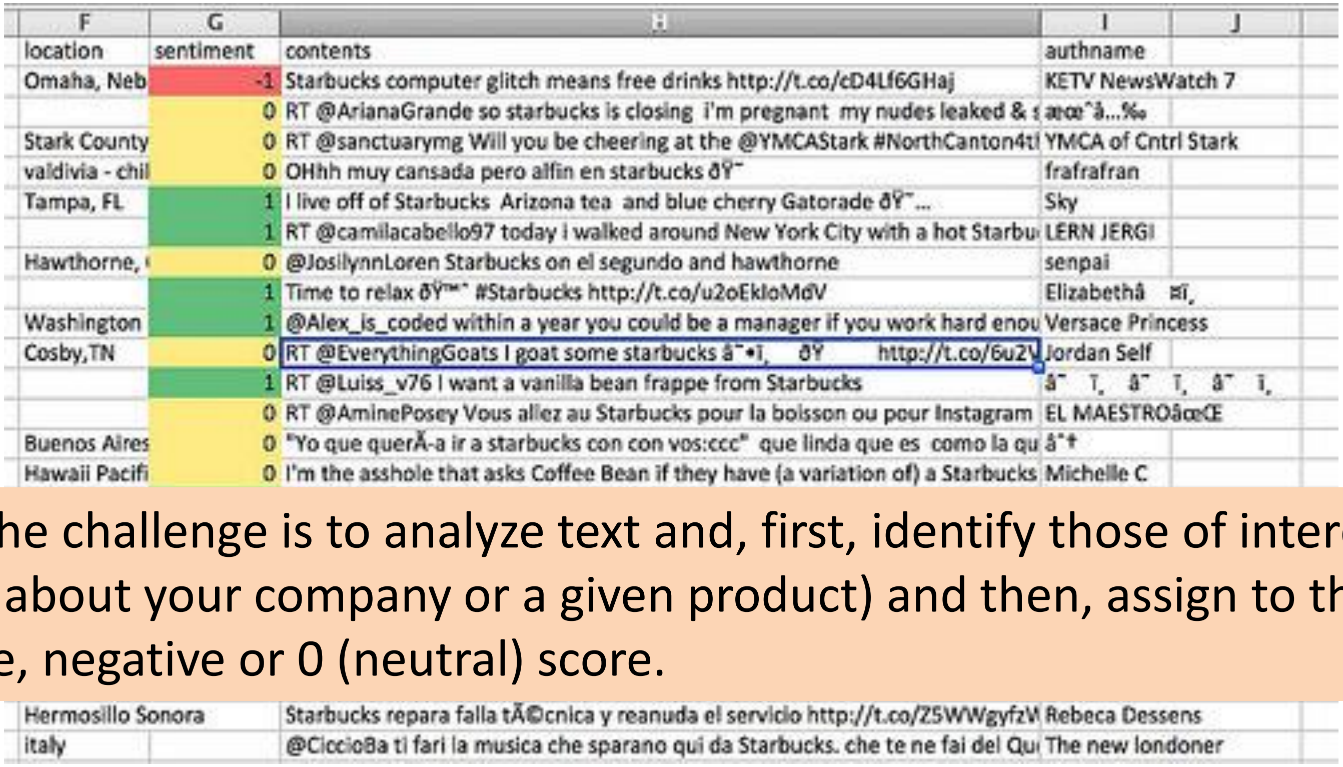
Debt	Income
10.134	100.000
85.234	134.000
8.112	21.500
0	45.900
17.534	52.000

$$\frac{\text{Debt}}{\text{Income}}$$

Debt to Income Ratio
0,10
0,64
0,38
0
0,34

Transforming unstructured data

- Way more complex! First, we need to transform from unstructured to stuctured
- Example: sentiment analysis in Twitter



F	G	H	I	J
location	sentiment	contents	authname	
Omaha, Neb	-1	Starbucks computer glitch means free drinks http://t.co/cD4Lf6GHaj	KETV NewsWatch 7	
	0	RT @ArianaGrande so starbucks is closing i'm pregnant my nudes leaked & s	ãœ˜å…‰	
Stark County	0	RT @sanctuarymg Will you be cheering at the @YMCAStark #NorthCanton4tl	YMCA of Cntrl Stark	
valdivia - chil	0	OHhh muy cansada pero alfin en starbucks ðŸ˜	frafrafran	
Tampa, FL	1	I live off of Starbucks Arizona tea and blue cherry Gatorade ðŸ˜…	Sky	
	1	RT @camilacabello97 today i walked around New York City with a hot Starbu	LERN JERGI	
Hawthorne, (	0	@JosilynnLoren Starbucks on el segundo and hawthorne	senpai	
	1	Time to relax ðŸ™ˆ #Starbucks http://t.co/u2oEkloMdV	Elizabethâ ¤ï¸	
Washington	1	@Alex_is_coded within a year you could be a manager if you work hard enou	Versace Princess	
Cosby,TN	0	RT @EverythingGoats I goat some starbucks â˜•ï¸ ðŸ http://t.co/6u2V	Jordan Self	
	1	RT @Luiss_v76 I want a vanilla bean frappe from Starbucks	â˜ ï¸ â˜ ï¸ â˜ ï¸	
	0	RT @AminePosey Vous allez au Starbucks pour la boisson ou pour Instagram	EL MAESTROâœŒ	
Buenos Aires	0	"Yo que querÃa ir a starbucks con con vos:ccc" que linda que es como la qu	â˜†	
Hawaii Pacifi	0	I'm the asshole that asks Coffee Bean if they have (a variation of) a Starbucks	Michelle C	
Hermosillo Sonora		Starbucks repara falla tÃ©cnica y reanuda el servicio http://t.co/Z5WWgyfzW	Rebeca Dessens	
italy		@CiccioBa ti fari la musica che sparano qui da Starbucks. che te ne fai del Qu	The new londoner	

Here, the challenge is to analyze text and, first, identify those of interest (e.g. talking about your company or a given product) and then, assign to the text a positive, negative or 0 (neutral) score.

Transforming unstructured data (2)

- What you get from this transformation (let’s ignore HOW for now)?

Table: Starbucks Twitter Sentiment

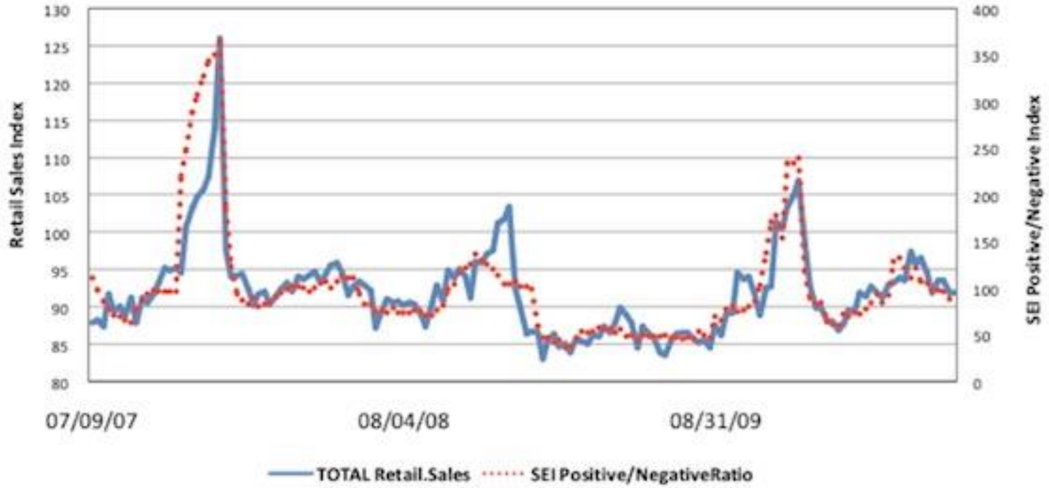
date	positive	negative	neutral
1/04/2016	500	237	1715
2/04/2016	451	277	2015
3/04/2016	816	300	3016

Transformation: aggregating heterogeneous data

- We already mentioned a simple example of aggregation (summing revenues data from different DBs in maintained in different departments)
- Aggregation on heterogeneous data may be far more complex
- E.g. we may want to aggregate *sentiment data* with *sales* to discover what went wrong (or what was the winning move users appreciated best)

Example (Social Engagement Index)

- http://www.brandamplitude.com/blog/innovation/item
- /announcing-breakthrough-in-measuring-the-impact-of-social-media-on-sales



Summary of data transformation

- It may be relatively simple if data are homogeneous (come from the same source and are structured)
- But this is the dream.. Usually data transformation is very complex and time-consuming and needs state-of-the-art software tools and also human supervision
- By far the most complex step in ETL

ETL

3.Loading

- Third step of ETL is Loading
- The ETL loading element is responsible for loading transformed data into the data warehouse database.
- The data warehouse is often taken *offline* during update operations so that data can be loaded faster
- If data are real-time streams (sensor data, social data..), or near-real time approach is used, then out-of-service is not perceived
- Loading basically implies to decide the destinantion and the updating frequency, that can be different for different sources (plus other security requirements)

DW ETL Tools

- **Some of the Well Known ETL Tools**
- The most well known commercial tools are Ab Initio, IBM InfoSphere DataStage, Informatica, Oracle Data Integrator and SAP Data Integrator.

Case Study (HW 4)

- Download the paper at http://bmjopen.bmj.com/content/bmjopen/6/8/e010962.full.pdf describing the use case of Dutch Red Cross data warehouse (also on course web site)
- Answer the following:
  - What type of data have been integrated, from which sources?
  - Can you draw the schema of all needed tables?
    - What are the objects? What are the attributes? What are the relationships? What is the “semantics” of relationships?
  - Can you list some of the TRANSFORM operations that were needed to harmonize data during the ETL process?
  - Which additional challenges are posed to the warehouse by the specific application domain?
  - Can you list the main categories of data which have been integrated?
  - Can you list and summarize the main data analytic tasks supported by the wharehouse?