Data Science Opportunities

Claudio Delrieux

Data Science Opportunities, Part I

Business type/sector	Raw data examples	Business opportunities
1. **Bank, credit, and insurance**	Transaction history Registration forms External references such as the credit protection service Micro and macro economic indices Geographic and demographic data	Credit approval Interest rates charges Market analysis Prediction of default Fraud detection Identifying new niches Credit risk analysis
2. **Security**	Access history Registration form Texts of news and Web content	Pattern detection of physical or digital behaviors that offer any type of risk
3. **Health**	Medical records Geographic and demographic data Sequencing genomes	Predictive diagnosis (forecast) Analysis of genetic data Detection of diseases and treatments Map of health based on historical data Adverse effects of medication/treatments
4. **Oil, gas, and electricity**	Distributed sensor data	Optimization of production resources Prediction/fault and found detection

Data Science Opportunities, Part II

Business type/sector	Raw data examples	Business opportunities
5. Retail	Transaction history Registration form Purchase path in physical and/or virtual stores Geographic and demographic data Advertising data Customer complaints	Increasing sales by product mix optimization base on behavior patterns during purchase Billing analysis (as-is, trends), the high volume of customers and transactions, credit profile by region Increasing satisfaction/loyalty
6. Production	Data management systems/ERP production Market data	Optimization of production over sales Decreased time/amount of storage Quality control
7. Representative organizations	Customer's registration form Event data Business process management and CRM systems	Suggestions of optimal combinations of company profiles, customers, and business leverage to suppliers Synergy opportunities identification
8. Marketing	Micro and macroeconomic indices Market research Geographic and demographic data Content generated by users Data from competitors	Market segmentation Optimizing the allocation of advertising resources Finding niche markets Performance brand/product Identifying trends

Data Science Opportunities, Part III

Business type/sector	Raw data examples	Business opportunities
9. Education	Transcripts and frequencies Geographic and demographic data	Personalization of education Predictive analysis for school evasion
10. Financial/ economic	List of assets and their values Transaction history Micro and macroeconomics indexes	Identify the optimal value of buying complex assets with many analysis variables (vehicles, real estate, stocks, etc.) Determining trends in asset values Discovery of opportunities
11. Logistic	Data products Routes and delivery points	Optimization of good flows Inventory optimization
12. E-commerce	Customer registration Transaction history Users' generated content	Increase free users' conversion rate for paying users by detecting the heavier preferences of users
13. Games, social networks, and platforms	Access history Registration of users Geographic and demographic data	Increase free user conversion rate for paying users by detecting the behavior and preferences of users
14. Recruitment	Registration of prospects employees Professional history, CV Connections on social networks	The person's profile evaluation for a specific job role Criteria for hiring, promotions, and dismissal Better allocation of human resources

The End

Data Science Project Workflow

Claudio Delrieux

Data Science Project Workflow, Part I

- The development of a typical data science (DS) project involves a process workflow comprised of several stages
- First and foremost, we have to fully understand the purpose of the project, i.e., what is the business opportunity or corporate goal that is pursued?

Data Science Project Workflow, Part II

- Apart from legal and technical issues, a (somehow arbitrary) demarcation of these development stages may be the following:
  - Data procurement and munging
  - Data wrangling and representation
  - Data mining and analysis
  - Data products and analytics
  - Data visualization and visual analytics

Data Science Project Workflow, Part III

- Sometimes these stages, as activities, are intertwined (for example, munging and wrangling are performed in a single process)
- Also, other activities may require iteration procedures (for example, the results of a data analysis process may trigger yet another data analysis process)

The End

Data Procurement and Munging

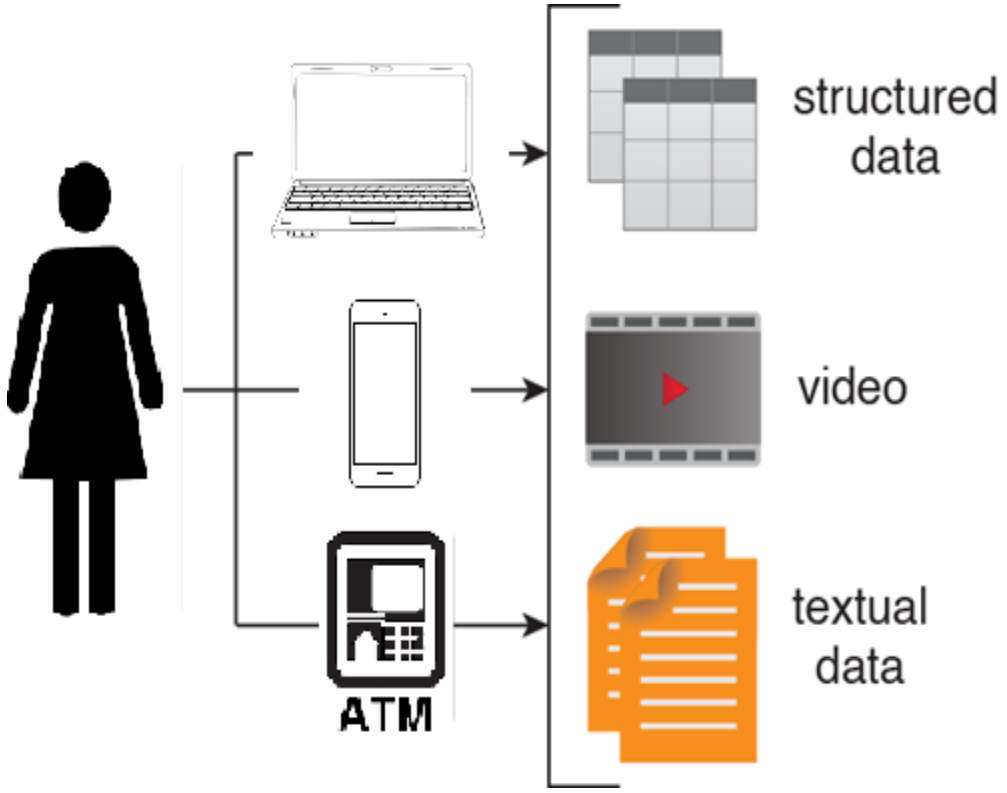
Claudio Delrieux

Data Procurement and Munging, Part I

- Data sources are diverse, abundantly available, and new ones arise by the day. As a coarse classification, we can mention:
  - Human vs. machine generated
  - Corporate, governmental, scientific, etc.
  - Free, purchased in data markets, or requiring explicit extraction

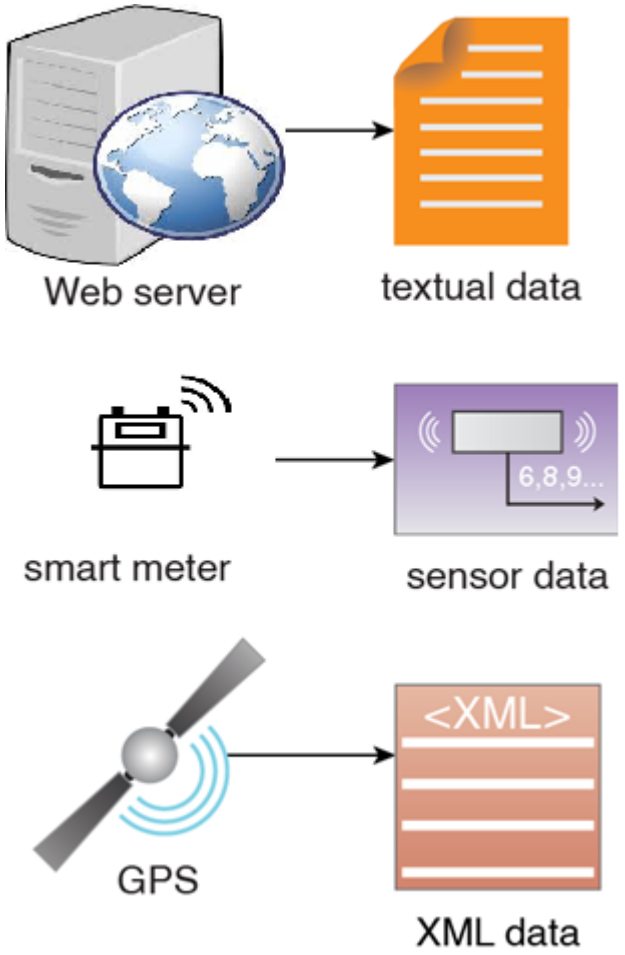
Data Procurement and Munging, Part II

- Human generated data ranges widely
  - Online services (for example, banking, retail)
  - Social networks
  - Email and web activity
  - Messaging
  - Many other…



Data Procurement and Munging, Part III

- Machine generated data is even more diverse
  - Data loggers
  - Smartphones
  - GPS
  - IoT
  - Sensor data
  - Scientific instrumentation



Data Procurement and Munging, Part IV

- Corporate datasets may be **public** (for example, datasets delivered by platforms like Netflix, Google, and many others) or **private**, when the enterprise uses its own data for specific purposes
- Public datasets may be freely available or can be purchased in data markets; may be static or streaming (for example, stock market data)

Data Procurement and Munging, Part V

- Government and institutional datasets are freely available, but typically require legacy systems or models (for example, census data)
- Scientific datasets abound, ranging from curated tables of specific experiments to raw huge datasets from large facilities like the ALMA radio telescope or the high energy hadron collider

The End

Data Wrangling and Representation

Claudio Delrieux

Data Wrangling and Representation, Part I

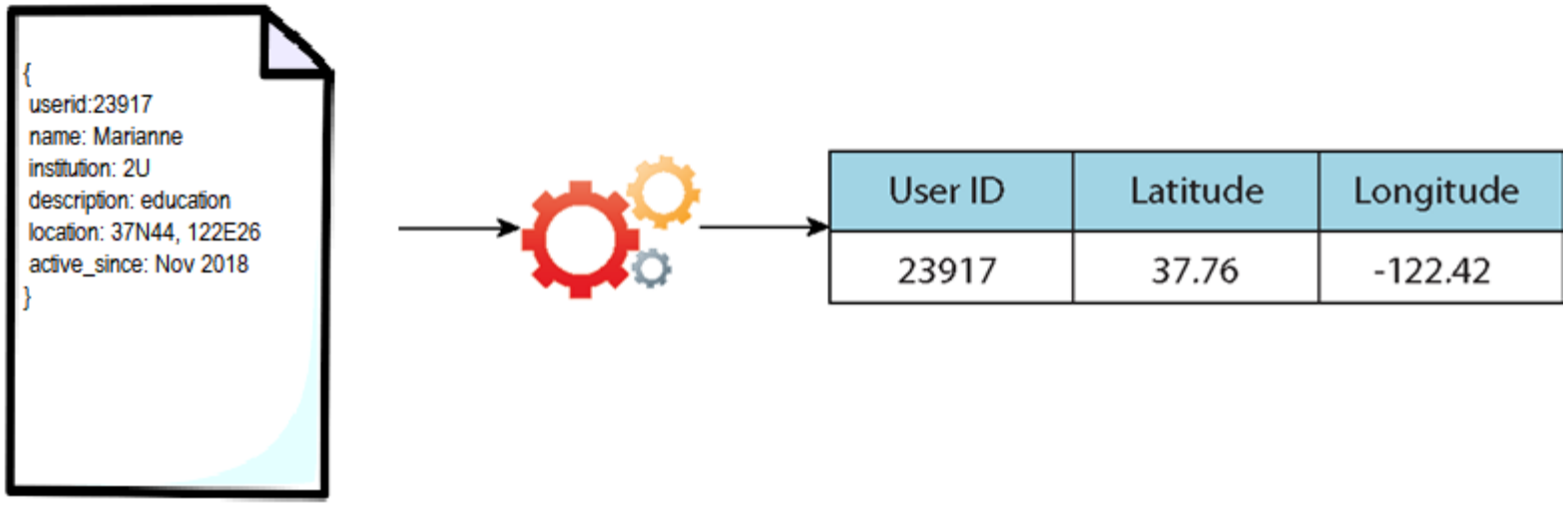
- All the prior data sources differ in nature, and thus have different representations.
  - Unstructured data (video, free text, voice)
  - Semi-structured (json, XML, schema-less files, and data-frames)
  - Structured data (tables, SQL, csv)

Data Wrangling and Representation, Part II

- Unstructured data is the most widespread and substantial (more than 90% of overall storage and streaming).
- It requires special purpose “analytics” for data extraction (typically on *ad-hoc* basis).
- It is neither robust nor scalable.
- Recent NLP products for free text and audio are stable, but images and video are still a challenge.

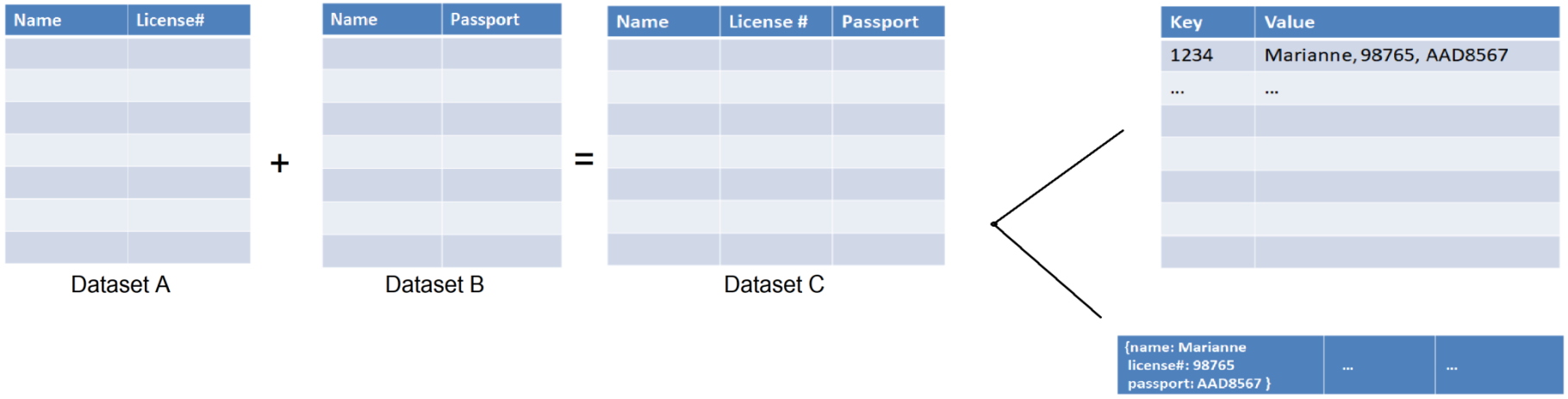
Data Wrangling and Representation, Part III

- Semi-structured has “dynamic” schemata, inferred from the file or data frame.
- *Readers* are required to convert this to structured formats, usually available or adaptable for most dataset types.



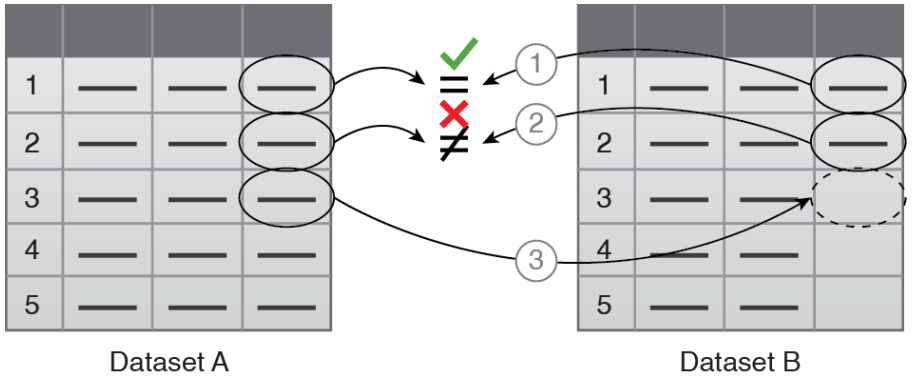
Data Wrangling and Representation, Part IV

- After data (typically “cells”) are curated, other preparation tasks may be required.
- Fusing together two or more datasets involves some kind of field reconciliation.
- Also, the expected analysis task at hand may entail different final representations.



Data Wrangling and Representation, Part V

- Once acquired the datasets and accessed the required fields, the next step is to transform them into clean and adequate formats.
- This may require fusion, validation, imputation, corrupt data filtering, consistent field name assignment, and much more.



The End

Data Mining and Analysis

Claudio Delrieux

Data Mining and Analysis, Part I

- This is our main interest (but be warned of the 80-20 rule, most of the time we will be performing the prior tasks).
- Our first aim is to explore the dataset in search of hidden patterns, rules, or trends that are aligned with the goals of the DS project.

Data Mining and Analysis, Part II

- This *exploratory* analysis (typically performed using data mining and statistical techniques) may uncover useful properties.
- We may analyze how the values of the different data variables are distributed, and if these distributions may be fit to a given known model.

Data Mining and Analysis, Part III

- Also, groups of variables may be analyzed together to see if any correlation or trend arises.
- Finally, records can be examined groupwise, trying to uncover hidden patterns or repeating tendencies.

Data Mining and Analysis, Part IV

- The second aim after this exploratory analysis is to *confirm* the prior findings with the whole dataset, testing the significance of the underlying model.
- After this confirmation, new insights may arise (in an iterative manner) until we arrive into a stable **data model.**

The End

Data Products and Analytics

Claudio Delrieux

Data Products and Analytics, Part I

- Once a stable data model is achieved, depending strongly on the DS project, we might need to implement it as a data product or as analytics.
- Data products are processing tools that accomplish the goal or serve as part of a data science project.

Data Products and Analytics, Part II

- Data products and analytics implement processing tasks that may range from simple math computer over isolated values, to very complex machine learning processes.
- This processing may be also assisted, autonomous asynchronous or periodic, continuous, or streamlined.

Data Products and Analytics, Part III

- When automatized and performed in systemic ways, data *analysis* leads to analytics processes, which may also be very diverse in nature.
- Analytics play a considerable role in contemporary business intelligence.
- Typically we consider descriptive, diagnostic, prescriptive, and predictive purposes.

Data Products and Analytics, Part IV

- Analytics' results, regarded as new data fields in datasets, or even as new datasets, may trigger new data analysis processes, also in an iterative manner.
- This may also inform and provide new perspectives on the original project purpose.

The End

Data Visualization

Claudio Delrieux

Data Visualization, Part I

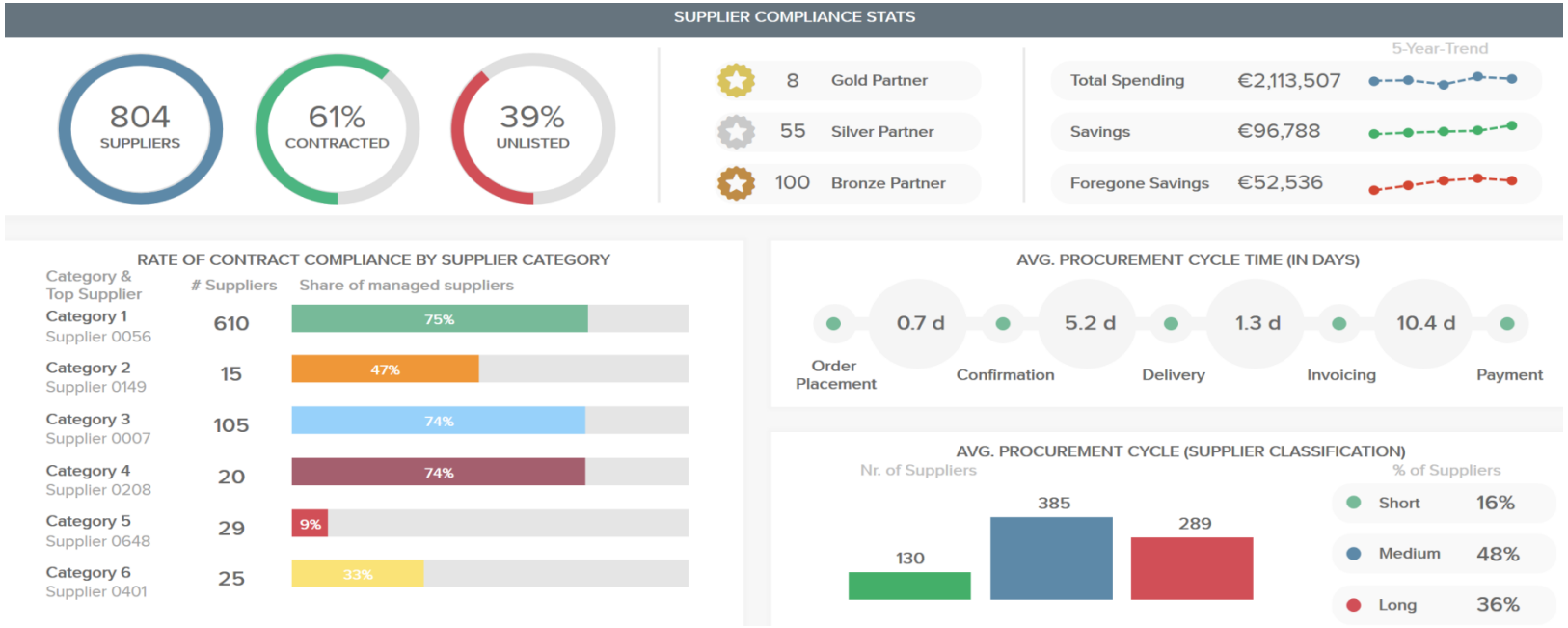
- Visualization takes advantage of the superior human capabilities to make sense of large amounts of information in a single view.

Data Visualization, Part II

- On the other hand, inadequate data visualization may be misleading or wrong.
- It is thus relevant to understand what are the values to be represented and to choose the right view (heatmap, scatterplot, bar chart, etc.).
- Contemporary visualization libraries provide adequate means for developing interactive and actionable data views.

Data Visualization, Part III

- Actionable dashboards are paramount in complex data science projects, especially to convey elaborate messages.



Data Visualization, Part IV

- However, dashboards are not mere “fireworks,” and thus careful understanding of their purpose and functioning is required.
- In this context, dashboards have to be designed to convey appropriate information about the key performance indicators (**KPI**s) in a meaningful way.

The End

Other Data Science Aspects

Claudio Delrieux

Other Data Science Aspects, Part I

Even though the following aspects may not be part of the processing pipeline, care must be taken to properly understand these issues:

- What are the goals of the data science (DS) project?
- What are the expected outcomes?
- What are the key indicators required?
- Are we conveying results adequately?

Other Data Science Aspects, Part II

In addition, the DS development may require embedding procedures for at least these two aspects:

1. Governance rules and actions once the project is deployed
2. Metadata management for traceability and auditing purposes

Other Data Science Aspects, Part III

Special care must be taken in these aspects:

- Establish adequate security policies
- Manage data in ways that no overt privacy breaches arise (and take all the legal provisos required)
- Understand the underlying hardware architecture and deploy issues

The End

Anscombe's Quartet
Project Statement and Dataset

Claudio Delrieux

Problem Statement: Dataset, Part I

- We are presented with the following four series of <x,y> values.
- In each, we have to predict the most likely value for a given new <x,y> point.

Problem Statement: Dataset, Part II

I	
x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

II	
x	y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

III	
x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

IV	
x	y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

Problem Statement: Dataset, Part III

- In all four cases, summary statistics are almost identical, thus giving little cue.

Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	y = 3.00 + 0.500x
Coefficient of determination R^2	0.67

The End

Text Classification
Project Statement and Dataset

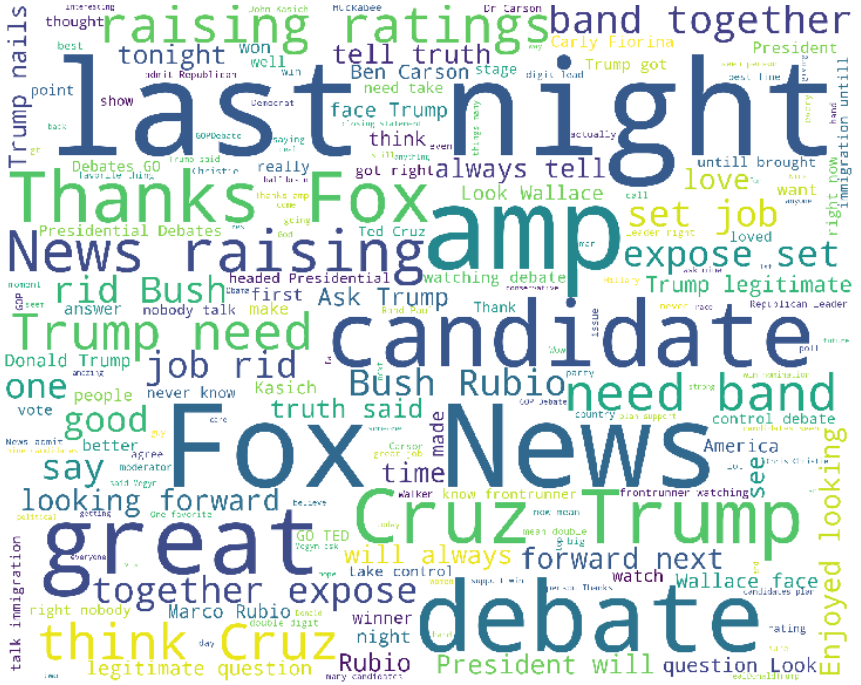
Claudio Delrieux

Problem Statement: Dataset, Part I

- Text classification for sentiment analysis is one of the leading information sources for enterprises to get a “360 view” about customers, citizens, patrons etc.
- It consists in computationally analyzing text messages and telling whether the underlying sentiment is positive, negative, or neutral.

Problem Statement: Dataset, Part II

- We will analyze data available from the first 2016 GOP Presidential Debate available online at [Python NLTK sentiment analysis](#).



Problem Statement: Dataset, Part III

- Take into account that filtering and class balance may be required to arrive into a trustable scoring.
- Also we will require an NLP library to remove stop words, emoticons, and to *tokenize* and *stemmize* words.

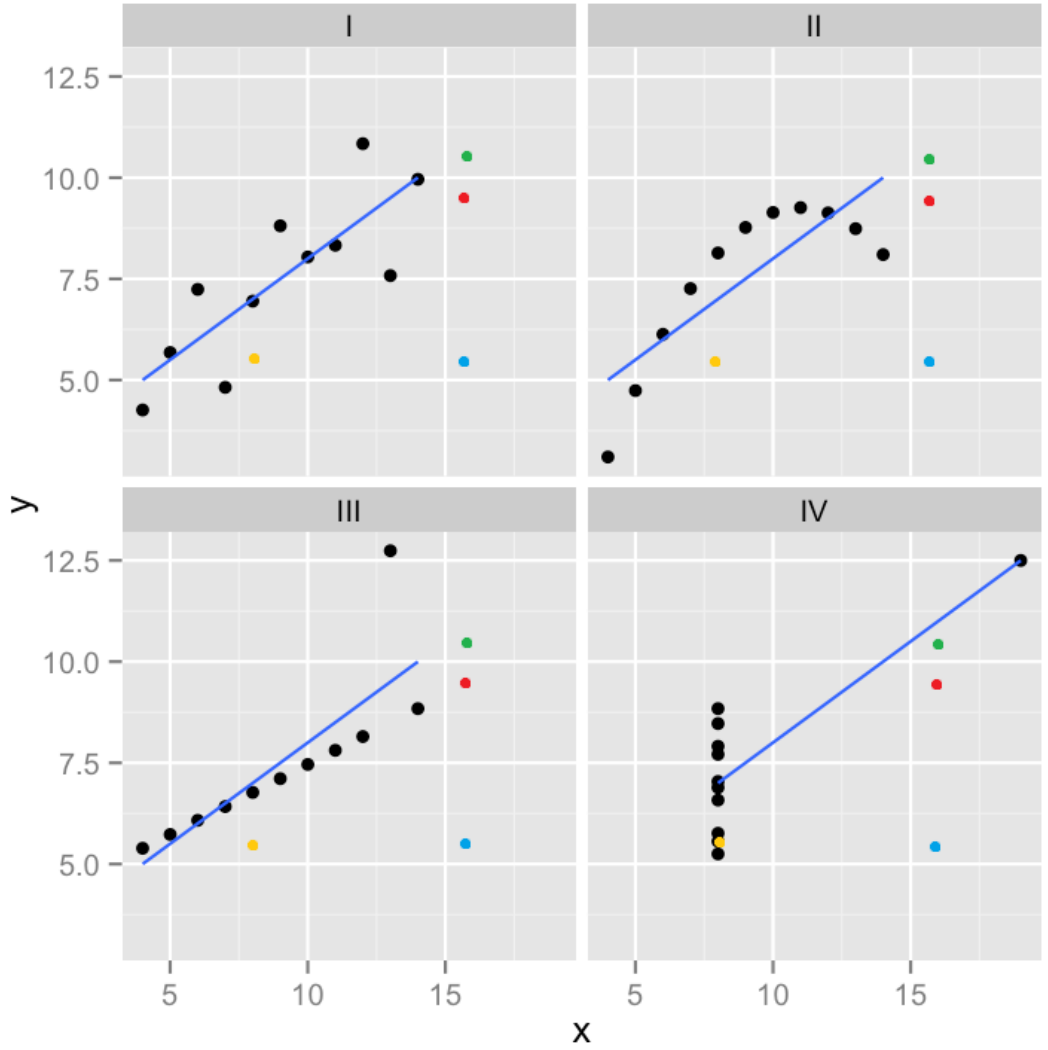
The End

Projects 1.1 and 1.2
Results and Overview

Claudio Delrieux

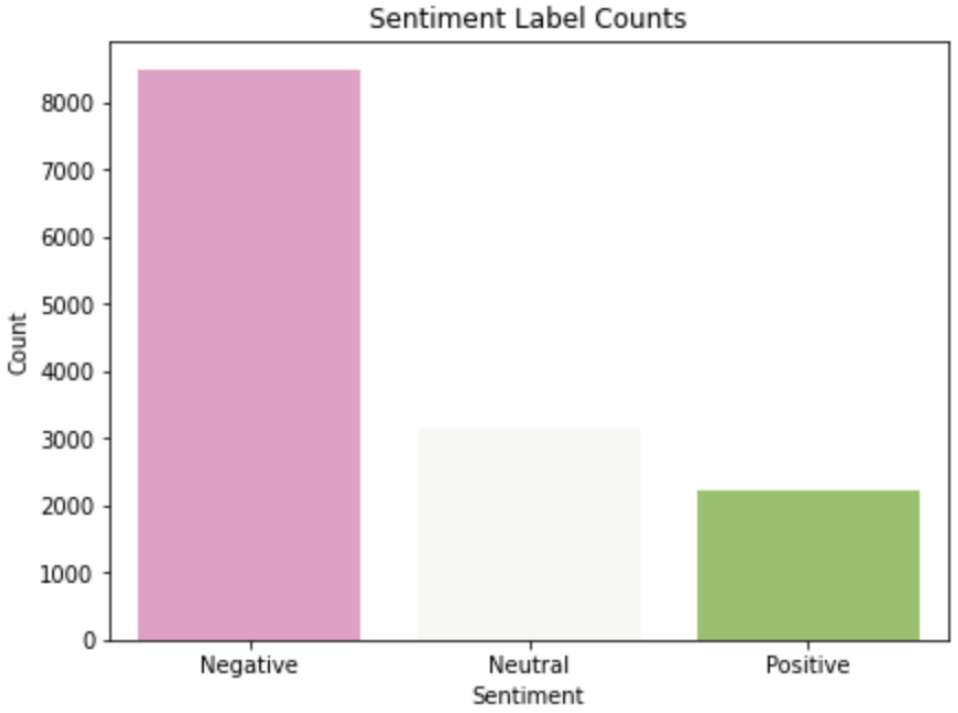
Projects 1.1 and 1.2: Results and Overview, Part I

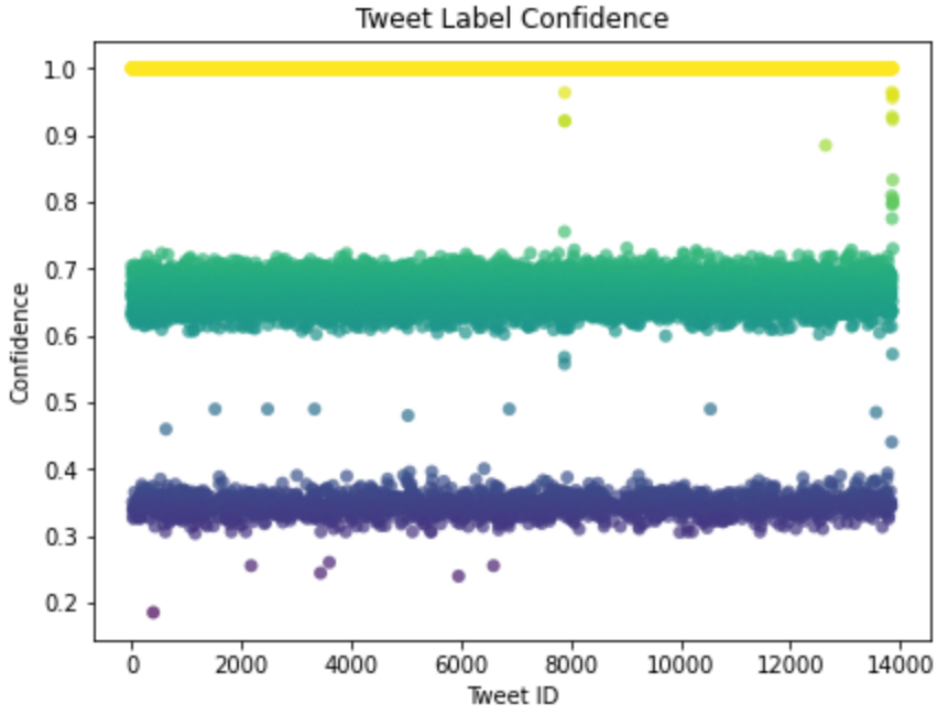
- In Anscombe's quartet, plotting the datasets helps make clear to which one the new points may belong



Projects 1.1 and 1.2: Results and Overview, Part II

- In the text classification, data curation was required given that the tweet label confidence was uneven





Projects 1.1 and 1.2: Results and Overview, Part III

- After filtering out and balancing the classes, we arrived into a reduced dataset (3,600 out of 13,800 tweets), used to train a classifier
- Using NLTK, we first remove *stop words* which have no significance in our analysis; same with URLs, hashtags, and usernames; then we convert all words to their *stems*, using the NLTK stemmizer

Projects 1.1 and 1.2: Results and Overview, Part IV

- In order to train a classifier, we need to create a *feature vector* identifying each tweet; this is called a feature vector
- We identify and create a dictionary of unique words using a *vectorizer*, which helps translate each word into a unique integer code, extracting the *vocabulary* in our dataset

Projects 1.1 and 1.2: Results and Overview, Part V

- In our dataset, the vocabulary size was 5,410
- Parts of its contents and stem occurrences are:
  - “Catch” 926 times
  - “Full” 2,000 times
  - “Gopdeb” 2,104 times
  - Etc.

Projects 1.1 and 1.2: Results and Overview, Part VI

- Finally, we choose a classifier adequate for this feature space, *random forest*, for which we split the dataset in training and test
- The contingency matrix for the trained classifier is:
  [[157 29]
  [30 145]]
  with an overall accuracy above 83%

Projects 1.1 and 1.2 Results and Overview

The End

Lessons Learned

Claudio Delrieux

Lessons Learned, Part I

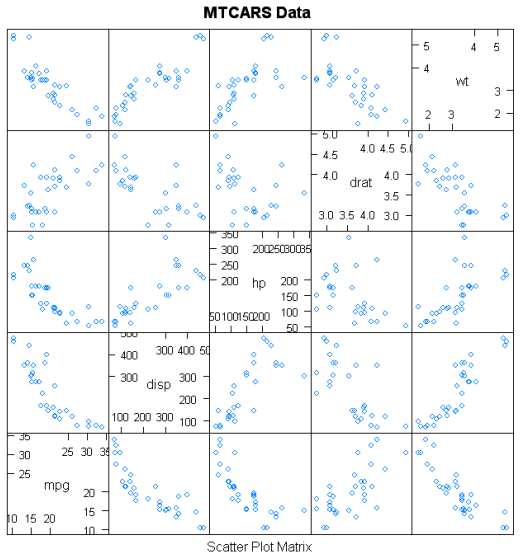
- As per project 1, we note that sometimes summary statistics and simple data models (for example, regressions) can be misleading.
- For this reason, it is advisable to invest some extra time using visual inspection of the datasets prior to the data analysis tasks.

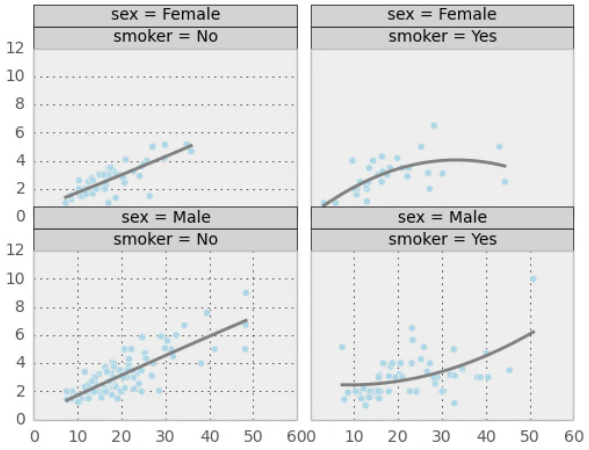
Lessons Learned, Part II

- This prior visual analysis can be performed manifold, for example, examine some of the variables (column-wise) to see if some remarkable feature arises.
- Other typical visual analysis is to represent records in a projected parameter space to see if they cluster together in some manner.

Lessons Learned, Part III

- In the lattice of scatterplots, several variables are contrasted all-against-all to see their pairwise correlation.
- In the trellis, three or more variables are represented together.





Lessons Learned, Part IV

- Regarding project 2 and unstructured data, free text for instance, this requires some intermediate wrangling steps, mostly on an *ad hoc* basis.
- For most available datasets, however, there are freely available libraries with adequate functions to cope with most curation issues.

Lessons Learned, Part V

- In next weeks, we will devote some of our time together to presenting some helpful libraries, as well as use examples in datasets frequently required for DS projects.
- Also, we will discuss the application of the most useful data analysis tools and visual representation techniques for typical DS analysis.

The End