Data Management

Fabrice Rossi

CEREMADE
Université Paris Dauphine

2021

Data Management

Numerous points of view

- inclusive: Data Management comprises all disciplines related to managing data as a valuable resource (wikipedia)
- restrictive: Data Management = Database systems
- in between: Data Management = data life cycle control and data administration
- etc.

Data Management

In this course

- data science point of view
- data as a resource
- strict separation between management and analysis
- almost all the data life cycle:
  1. Data generation and acquisition
  2. Data preprocessing
  3. Data storage
  4. Data “management” and querying
- focus on transversal issues

Data Management Issues

Data as a resource

- data centered projects:
  no data $\Rightarrow$ no project!
- thus one needs data
  - security
  - quality
  - integration
  - accessibility
  - efficiency

GDPR

- compliance

Data Management Issues

Data as a resource

- data centered projects: no data $\Rightarrow$ no project!
- thus one needs data
  - security
  - quality
  - integration
  - accessibility
  - efficiency

GDPR

- compliance

Tools and methods

- standard computer engineering tools
  - validation
  - backup
  - access control
  - version control
- data management tools
  - database system
  - data warehouse
- data science based techniques
  - noise removal
  - outlier detection
  - differential privacy

Data Science Steps

Revisiting the steps

1. Data generation and acquisition
2. Data preprocessing
3. Data storage
4. Data management and querying
5. Data analysis

Data Science Steps

Revisiting the steps

1. **Data generation and acquisition**
2. Data preprocessing
3. Data storage
4. Data management and querying
5. Data analysis

Data Generation and Acquisition

Generation

- design level: what to collect?
- GDPR compliance
  - are personal data needed?
  - if they are, the system must enable
    - consent collection (potentially)
    - personal data access
    - correction and erasure
    - etc.
  - data processing activities must be recorded and documented
- more generally
  - data documentation is a key element of a data management strategy (meta data)
  - data should not be collected without a purpose

Data Generation and Acquisition

Acquisition

- measurement level: how to measure
- GDPR compliance
  - be sure to assign personal measure to a person!
  - not always obvious (see e.g. device fingerprinting)
- data quality
  - be aware of sensor limitations
  - include sensor data in the collected data
  - include context (e.g. collection data, collector's name, etc.)
- data integration
  - to enable linking between different data sources
  - to have consistent data (e.g. use UTC time!)

Manual Data Acquisition

A quite common situation

- user entered data: typical customer interaction
- data collected by human operators
  - public facing employees
  - researchers

A major source of error

- misunderstanding and typos
- inconsistent encoding (e.g. Windows-1252 rather than UTF-8)

Enforce consistency

- auto-completion
- data validation

Best practices

Specification

- what is collected and why
- thorough specification
  - entities
  - variables
    - type
    - possible values (e.g. range or modalities)
    - unit
    - encoding

Collection

- include contextual variables
  - measurement time
  - operator
  - sensor

Standards

- use standards for data values
  - UTC for time zone
  - WGS 84 for coordinates
- and for data representation
  - WGS 84 in decimal form
  - ISO 8601 for date and time
- use naming convention
  - avoid spaces in variable names
  - use only UTF-8

Example

	age	job	marital	education	default	balance	housing
1	30	unemployed	married	primary	no	1787	no
2	33	services	married	secondary	no	4789	yes
3	35	management	single	tertiary	no	1350	yes
4	30	management	married	tertiary	no	1476	yes
5	59	blue-collar	married	secondary	no	0	yes
6	35	management	single	tertiary	no	747	no
7	36	self-employed	married	tertiary	no	307	yes
8	39	technician	married	secondary	no	147	yes
9	41	entrepreneur	married	tertiary	no	221	yes
10	43	services	married	primary	no	-88	yes
11	39	services	married	secondary	no	9374	yes
12	43	admin.	married	secondary	no	264	yes
13	36	technician	married	tertiary	no	1109	no
14	20	student	single	secondary	no	502	no
15	31	blue-collar	married	secondary	no	360	yes
16	40	management	married	tertiary	no	194	no
17	56	technician	married	secondary	no	4073	no
18	37	admin.	single	tertiary	no	2317	yes
19	25	blue-collar	single	primary	no	-221	yes
20	31	services	married	secondary	no	132	no

Documentation

- sources:
  - https://archive.ics.uci.edu/ml/datasets/Bank%2BMarketing
  - http://hdl.handle.net/1822/14838
- 17 direct marketing campaigns for a Portuguese bank between May 2008 and November 2010
- data dictionary
  - age: integer
  - job: categorical with 12 possible values
  - marital: categorical with 3 values
  - balance: average yearly balance in euros
  - y: did the campaign succeeded for this client
  - etc.

Personal Data?

Are those data personal?

- no!

Personal Data?

Are those data personal?

- no!
- personal data=”any information relating to an identified or identifiable natural person”
- here we do not know:
  - the bank
  - the gender of the person
- coarse grain encoding for many attributes
- imprecise attributes
- it is unlikely that one could use external data to re-identify the persons in the data set

Problems

Missing context

- data collection time (which campaign?)
- age semantic
  - at data collection?
  - at data release?

Broken encoding

- divorced means both divorced and widowed
- admin. is abbreviated from something

Undocumented aspects

- how is the average balance computed?
- how are the job categories defined?
- what are the education levels?

Data Science Steps

Revisiting the steps

1. Data generation and acquisition
2. **Data preprocessing**
3. Data storage
4. Data management and querying
5. Data analysis

Preprocessing and Management

Preprocessing

- data transformation before storage
- data integration

Specification and collection

- specify all steps in details
- keep additional monitoring data, e.g.
  - main variable: average GPS position over one minute
  - monitoring variable: co-variance of the position over one minute
- use preprocessing to enforce standards

Data quality

Enforcing rules and constraints

- preprocessing can be used to enforce/check some rules
  - clip data from sensors to their interval
  - remove inconsistent data
- should be documented
  - both externally (in the data collection specification)
  - and internally
    - keep a raw measurement variable per variable
    - compute a cleaned variable from the raw variable
- provide a data quality report
  - missing data rate
  - rejected measure rate
  - etc.

Limitations

Preprocessing is analysis

- advanced preprocessing techniques are data science techniques
  - outlier detection
  - semantic compression
  - feature engineering
  - etc.

Consequences

- iterative process
  1. collect raw data
  2. design preprocessing techniques
  3. collect raw data and preprocess them
  4. back to 2
- documenting preprocessing is documenting data science!

Example

Bank data set

- age is not a raw data
  - the bank has generally access to a client birth date
  - turned into an age
  - undocumented
- balance is not a raw data
  - the bank has access to the whole time series of the balance
  - summarize into a single value
  - undocumented

Data science steps

Revisiting the steps

1. Data generation and acquisition
2. Data preprocessing
3. **Data storage**
4. Data management and querying
5. Data analysis

Storage Management

Reliable storage

- RAID array
- backup (at a different place)

File format

- a very complex issue
- trade off between
  - ease of use
  - efficiency (space and time)
  - longevity
  - accuracy

Storage Management

Reliable storage

- RAID array
- backup (at a different place)

File format

- a very complex issue
- trade off between
  - ease of use
  - efficiency (space and time)
  - longevity
  - accuracy

Rule of thumb

- avoid proprietary format (excel, word)
- avoid lossy compression format (jpg, mp3)
- use open formats
  - basic text in UTF-8 format!
  - OpenDocument
  - PDF/A
  - FLAC
  - WebM
- use XML or JSON for (semi)-structured data

Storing data sets

Using Comma-Separated Values

- CSV is an acceptable format for data sets with some caveats
- do not use variations
  - use a header
  - use comma for separating fields
  - enclose fields with double quote when they contain commas or double quote
  - represent a double quote by 2 double quotes
  - use a full stop (period) for the decimal separator

```
Firstname,Lastname,Nickname,Height,Weight
John,Doe,"The ""John""",1.78,72
```

- there are complex issues surrounding the textual representation of real numbers but those are taken care off in e.g. Python
- compressing CSV files is useful but must be done with a standard format such as ZIP

Storing meta data

Data Management

- documentation
- specification
- context

Data about the Data
Metadata

Core principle

- a data set without metadata is useless
- store the metadata with the data!

Storing meta data

Data Management

- documentation
- specification
- context

Data about the Data
Metadata

Core principle

- a data set without metadata is useless
- store the metadata with the data!

Structured metadata

- XML based standards
- Dublin core for high level metadata (author, title)
- specialized formats for precise specification
- e.g. the W3C model allows one to describe complex constraints on columns/variable of a CSV file
- see also DDI

Data History

Time evolving data

- data with a temporal component
- explicitly collected
  - regular snapshot
  - time-stamped events
- data collection time should be recorded for all data!

Data history

- evolution of a data set
  - bug fix
  - new observations
  - additional variables
- must be recorded but not necessarily analyzed
- basic solution: date encoded in the archive
- recommended solution
  - version control system
  - used also for software programming
  - e.g. Git

Data History

Time evolving data

- data with a temporal component
- explicitly collected
  - regular snapshot
  - time-stamped events
- data collection time should be recorded for all data!

Data collection process

- might also evolve
- should be recorded

Data history

- evolution of a data set
  - bug fix
  - new observations
  - additional variables
- must be recorded but not necessarily analyzed
- basic solution: date encoded in the archive
- recommended solution
  - version control system
  - used also for software programming
  - e.g. Git

Data science steps

Revisiting the steps

1. Data generation and acquisition
2. Data preprocessing
3. Data storage
4. **Data management and querying**
5. Data analysis

CS/CE take on data management

Implementation of DM

- systems to
  - model data
  - insure data quality
  - enforce data integrity
  - enforce data security
- essentially database management systems

Data manipulation

- filtering
- grouping
- summarizing
- merging
- reshaping
- transforming

CS/CE take on data management

Implementation of DM

- systems to
  - model data
  - insure data quality
  - enforce data integrity
  - enforce data security
- essentially database management systems

Data manipulation

- filtering
- grouping
- summarizing
- merging
- reshaping
- transforming

Manipulations must be documented

Example: Bank data set

Summarizing balance

- average: 1422.66
- median: 444
- probable outliers!
- quantiles:

0%	1%	5%	10%	25%
-3313.00	-671.40	-162.00	0.00	69.00

75%	90%	95%	99%	100%
1480.00	3913.00	6102.00	14194.60	71188.00

Example: Bank data set

Filtering the dataset

- original data: 4521 persons
- after removing the 2% extremes case: 4429 persons
- mean balance: 1243.42

Group summary

- group on the success/failure of the marketing campaign
- compute the mean value of balance per group
- with filtered data

y	mean balance
no	1210.93
yes	1492.02

Data science steps

Revisiting the steps

1. Data generation and acquisition
2. Data preprocessing
3. Data storage
4. Data management and querying
5. **Data analysis**

Data analysis

Visualization

- program generated graphics
- active documents
- portable formats
- vector graphics

Predictive models

- fully automated process
- detailed diagnostics
- save the model in a portable format
- beware of randomness!

Core principles

- documentation
- full reproducibility
- ensure a perfect match between the data set and the outputs of the analysis

Conclusion

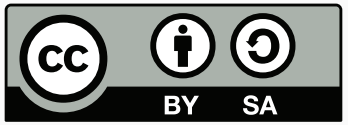
Data as a resource

- control the life cycle of the data from data oriented project design to data analysis
- documentation is the key
- together with reproducibility

Programming

- achieving reproducibility with a point and click software is very difficult
- code as documentation
- code history as project history documentation

Licence



Version

Last git commit: 2021-01-19
By: Fabrice Rossi (Fabrice.Rossi@apiacoa.org)
Git hash: a623238c82efeb5372d8b821e0e946cfd8c918cc

Changelog

- October 2019 additions:
  - details about the bank data
  - summary of DM from a CS/CE point of view
  - summary of DM for data analysis
  - conclusion
- September 2019: initial version