Data Management and Data Cleaning for Scientists I

Mark van der Loo

Contents

Materials for these lectures

github.com/markvanderloo/UFPEL2019

The faces of data

What do we mean when we say ‘data’?

Data are a representation of information.

What a user of data wants

CO_2 emission (fictional)

fuel	emission
Petrol	215
– of which bio	75
Diesel	456
– of which bio	89

Example: IBGE

	A	B	C	D	E	F	G	H	I
1	**Tabela 1.2 - Número de empresas, pessoal ocupado total, pessoal ocupado assala**								
2	**segundo as seções da classificação de atividades e as faixas**								
3									
4	Seções da classificação de atividades	Faixas de pessoal ocupado assalariado	Número de empresas, por tipos de eventos demográficos						
5			Total	Entradas			Saídas	Sobreviventes	Total
6				Total	Nascimentos	Reentradas			
7	**Total**	**Total**	**4 458 678**	**676 444**	**503 212**	**173 232**	**699 376**	**3 782 234**	**38 354**
8	Total	0	2 058 400	499 557	357 848	141 709	579 351	1 558 843	2 85
9	Total	1 a 9	1 944 144	161 548	131 874	29 674	113 548	1 782 596	8 91
10	Total	10 ou mais	456 134	15 339	13 490	1 849	6 477	440 795	26 59
11	A Agricultura, pecuária, produção florestal, pesca e aquicultura	Total	33 110	5 704	4 307	1 397	5 029	27 406	48
12	A Agricultura, pecuária, produção florestal, pesca e aquicultura	0	15 040	4 138	3 019	1 119	4 201	10 902	2
13	A Agricultura, pecuária, produção florestal, pesca e aquicultura	1 a 9	13 085	1 327	1 071	256	755	11 758	6
14	A Agricultura, pecuária, produção florestal, pesca e aquicultura	10 ou mais	4 985	239	217	22	73	4 746	38
15	B Indústrias extrativas	Total	10 067	1 315	859	456	1 510	8 752	20
16	B Indústrias extrativas	0	4 266	1 056	675	381	1 298	3 210	
17	B Indústrias extrativas	1 a 9	3 862	237	165	72	188	3 625	1

Source: https://www.ibge.gov.br/en/statistics/economic/industry-and-construction/22733-demography-of-enterprises-and-statistics-of-entrepreneurship.html?=&t=resultados

What an analyst wants

fuel	emission
Petrol	215
– of which bio	75
Diesel	456
– of which bio	89

$\longrightarrow$

fuel	type	emission
Petrol	regular	140
Petrol	bio	75
Diesel	regular	367
Diesel	bio	89

What a web developer wants

```
[{"fuel":"petrol","type":"regular","emission":140},
 {"fuel":"petrol","type":"bio","emission":75},
 {"fuel":"diesel","type":"regular","emission":367},
 {"fuel":"diesel","type":"bio","emission":89}]
```

Example: IBGE

{} [{"id":"1501","nome":"Belém","nivel":{"id":"7","nome":"Região metropolitana"}},
{"id":"2301","nome":"Fortaleza","nivel":{"id":"7","nome":"Região metropolitana"}},
{"id":"2601","nome":"Recife","nivel":{"id":"7","nome":"Região metropolitana"}},
{"id":"2901","nome":"Salvador","nivel":{"id":"7","nome":"Região metropolitana"}}, {"id":"3101","nome":"Belo Horizonte","nivel":{"id":"7","nome":"Região metropolitana"}},
{"id":"3301","nome":"Rio de Janeiro","nivel":{"id":"7","nome":"Região metropolitana"}}, {"id":"3501","nome":"São Paulo","nivel":{"id":"7","nome":"Região metropolitana"}},
{"id":"4101","nome":"Curitiba","nivel":{"id":"7","nome":"Região metropolitana"}}, {"id":"4301","nome":"Porto Alegre","nivel":{"id":"7","nome":"Região metropolitana"}}]

Source: https://servicodados.ibge.gov.br/api/v3/agregados/1705/localidades/N7

```
[{"fuel":"petrol","type":"regular","emission":140},
 {"fuel":"petrol","type":"bio","emission":75},
 {"fuel":"diesel","type":"regular","emission":367},
 {"fuel":"diesel","type":"bio","emission":89}]
```

↓

fuel	type	emission
Petrol	regular	140
Petrol	bio	75
Diesel	regular	367
Diesel	bio	89

What a database designer sees

Fuel

id	name
11	petrol
12	diesel

,

Type

id	name
1	regular
2	bio

,

Emission

id	fuel	type	amount
120	11	1	140
121	11	2	75
123	12	1	367
124	12	2	89

Fuel

id	name
11	petrol
12	diesel

,

Type

id	name
1	regular
2	bio

,

Emission

id	fuel	type	amount
120	11	1	140
121	11	2	75
123	12	1	367
124	12	2	89

$\downarrow$

fuel	type	emission
Petrol	regular	140
Petrol	bio	75
Diesel	regular	367
Diesel	bio	89

Summarizing

Presentation

- Convey a (single) message
- Human-readable

Analyses

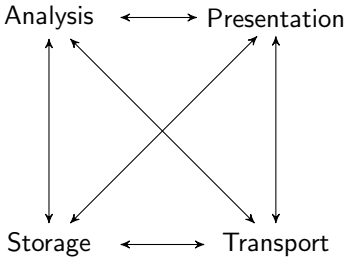
- Reusable for (interactive) analyses
- Machine-readable, easy to manipulate

Transport

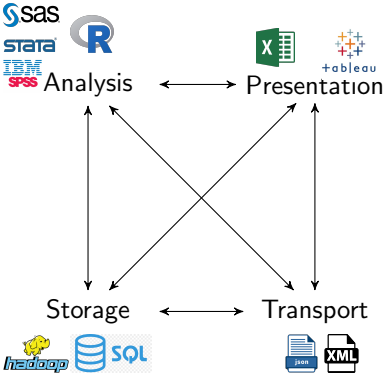
- Machine-readable
- Generic, language-independent format

Storage

- Create, Read, Update, Delete (CRUD)



Examples of tools



Why choosing the right tool is important (NYT, 2013)

The New York Times

Opinion

PAUL KRUGMAN

The Excel Depression

By Paul Krugman

April 18, 2013

Why choosing the right tool is important (Nature, 2019)

nature

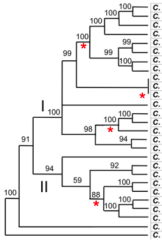
6. Protect raw data

All data are precious, but raw data are irreplaceable: the only way to recreate them is to run the experiment again. These must therefore be backed up – and kept as read-only files. Wickes once had to kill a project because she opened a crucial file in Microsoft Excel, which automatically formatted a column, changing the values and ruining the underlying data set. So, protect your raw data, says Martinez, “no matter what”.

Why choosing the right tool is important (Science, 2016)

Science Contents News Careers Journals

SHARE



	feeding (and soaking)			injections into gonad (and body cavity)		
gene	Ole-actin	Cel-unc-22	Cel-pos-1, or ama-1	Ole-actin	Cel-unc-22	ama-1
expression site	soma		germ line		soma	germ line
JU1325	-					
AF16	-		-			+++
JU727	-					
PB4641	-		-			+++
JU1333	-					
PB2801	-		-			-
JU1373	-			+++		
JU1873	+					
N2	++	+++	+++		+++	+++
CB4856		+	-			
JU1580	-	-	+++		-	
QG122	+++					
EG6142	-					
JU1825	-					
JU1857	-					
EG5716	+++					
JU1199	+++					
NIC113	+					
EG4788	+++					
JU1528	+++					
DF5070	-		-			+
DF5077	-		-	+		
RGD1	-		+++	+		++
JU1426	-					
QX1182	-					
SB355	+/-		+/-			+++
SB341			+++			+++

PLOS ONE PHYLOGENY/FLICKR (CC BY 2.0)

One in five genetics papers contains errors thanks to Microsoft Excel

By Jessica Boddy | Aug. 29, 2016 , 1:45 PM

Ziemann, Eren, El-Osta *Genome Biology* **17** 177 (2016)

Homework assignment

Google ‘excel disasters’

and spend 30 minutes reading what you find

Presentation versus analysis

Spreadsheet software is unsuited for analysis because

- Autoformatting
- Does not force consistency
- Hard to analyze: code is hidden
- Hard to test
- You see a ‘state’ not the process

A bit of terminology

Computer scientists	$\leftrightarrow$	**Statisticians**
Entity type	$\leftrightarrow$	Population
Entity	$\leftrightarrow$	Population unit
Attribute	$\leftrightarrow$	(Stochastic) variable
Value	$\leftrightarrow$	Value

How to recognize whether data is suited for analysis[1]

Boxes to tick

1. Does each row correspond to one entity?
2. Are all entities of the same type?
3. Is every entity represented only once?
4. Does every column correspond to a single property for each entity?
5. Are all elements of each column of the same and the correct type?
6. Is the data valid?

Rule of thumb

Can you make meaningful summary statistics over each column?

[1] Only for simple rectangular data sets

Quizz (1): Ready for analyes?

	Alice	Bob	Carol
Shoe size	38	43	41
Income	3300	2800	4000

Quizz (2): Ready for analyses?

	Shoe size	Income
Alice	38	3300
Bob	43	2800
Carol	41	4000

Quizz (3): Ready for analyses?

Income distribution

	€14k – €20k	€20k – €40k	€40k – €80k	€80K+
Amsterdam	20%	40%	35%	5%
Rotterdam	30%	30%	38%	2%
Den Haag	25%	35%	30%	10%

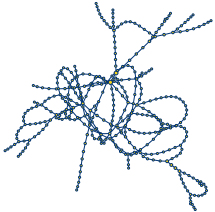
Quizz (4): Ready for analyses?

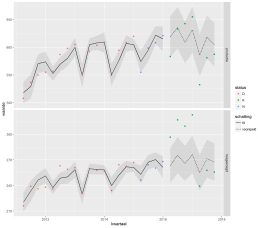
	Age	has job
Dave	36	No
Eve	5	Yes

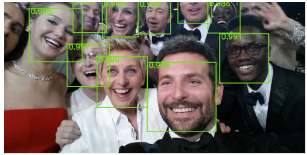
Quizz (5): Ready for analyses?

	costs	profit
Retailers	50	10
Wholesalers	20	5
Total	70	15

Not all data is ‘simple rectangular’







Each data type consists of particular basic elements and is manipulated with particular basic operations.

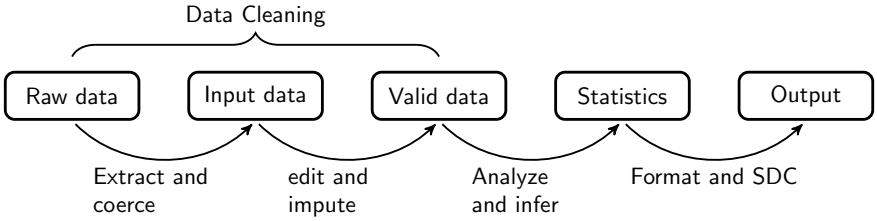
The Statistical Value Chain

Value Chains

Porter's value chain (1985)

The idea of the value chain is based on the process view of organizations, the idea of seeing a manufacturing (or service) organization as a system, made up of subsystems each with inputs, transformation processes and outputs.

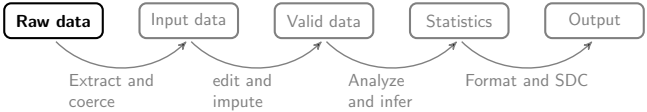
Statistical Value Chain



Notes

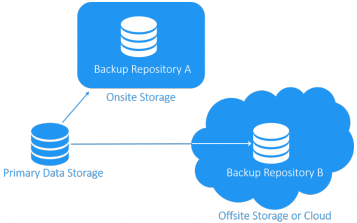
- This part only pertains to the data processing stage. Collection, design of experiments, dissemination of results, and so on are not included.
- The fixed points (half-fabricates) are well-defined statistical products.

Raw data



Your most valueable resource!

- Hard/expensive to obtain
- Keep unchanged
- Backup according to 3-2-1 principle



The importance of backups

Backblaze Lifetime Hard Drive Annualized Failure Rates

For hard drive models in service as of June 30, 2019
Reporting period April 2013 - June 2019 inclusive

MFG	Model	Drive Size	Drive Count	Average Age	Drive Days	Drive Failures	AFR*
Toshiba	MG07ACA14TA	14TB	1,220	8.85	328,960	7	0.78%
HGST	HUH721212ALE600	12TB	520	4.47	61,360	2	1.19%
HGST	HUH721212ALN604	12TB	9,609	3.52	976,794	10	0.37%
Seagate	ST12000NM0007	12TB	34,710	13.58	14,245,745	737	1.89%
Seagate	ST10000NM0086	10TB	1,200	21.29	787,144	12	0.56%
HGST	HUH728080ALE600	8TB	1,001	19.29	654,219	15	0.84%
Seagate	ST8000DM002	8TB	9,875	33.26	10,003,569	280	1.02%
Seagate	ST8000NM0055	8TB	14,380	23.90	10,532,321	336	1.16%
Seagate	ST6000DX000	6TB	886	50.85	2,739,695	79	1.05%
HGST	HMS5C4040ALE640	4TB	2,639	39.01	11,174,488	155	0.51%
HGST	HMS5C4040BLE640	4TB	12,752	32.53	17,236,735	214	0.45%
Toshiba	MD04ABA400V	4TB	99	49.23	216,631	5	0.84%
Seagate	ST4000DM000	4TB	19,570	44.41	49,043,264	3,652	2.72%
		Totals	108,461		118,000,925	5,504	1.70%

* AFR - Annualized Failure Rate

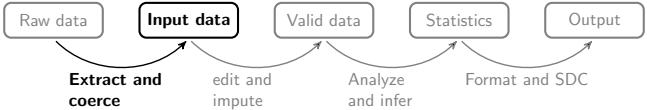
BACKBLAZE

backblaze.com/blog/backblaze-hard-drive-stats-q2-2019/

Optimistically:

- On average $P(failure) \approx 0.001/\text{year}$.
- UFPEL has $\approx 20\,000$ students.
- We expect ≈ 20 failures/year.

Input data



Technically 'clean' data

- File type is known and can be read
- Data structured for analyses
- Variables are of correct type (number/date/text/categorical...)
- Records identified with statistical objects
- Variables identified with statistical properties

Rule of thumb

You can read this data into your favorite analyses tool, without errors, with a single expression.

From raw to input, an example from the LATTES system

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

Currículo Lattes

XML ?

Dados gerais | Formação | Atuação | Projetos | Produções | Patentes e Registros | Inovação | Eventos | Orientações | Bancas | Citações | +

Tatiana Pereira Cenci

Bolsista de Produtividade em Pesquisa do CNPq - Nível 2

Endereço para acessar este CV: http://lattes.cnpq.br/6217846985830016

ID Lattes: **6217846985830016**

Última atualização do currículo em 15/10/2019

Professora Associada, Departamento de Odontologia Restauradora da Faculdade de Odontologia da Universidade Federal de Pelotas. Graduada em Odontologia pela FOB/USP (2001), Especialista em Prótese Dentária pelo HRAC/USP (2004), Mestre (2006) e Doutora (2008) em Clínica Odontológica/UNICAMP, com PDEE-CAPES na ACTA/Holanda (Nov/2006 a. Out/2007). Tem experiência nos seguintes temas: biofilme, ensaios clínicos randomizados e revisões sistemáticas. É membro do The BRIGHTER (Bias, Reporting, Implementation, Guidance, ETHics, InIEgrity and Reproducibility in Research) Meta-Research Group Initiative. **(Texto informado pelo autor)**

Identificação

Nome Tatiana Pereira Cenci

Nome em citações bibliográficas PEREIRA-CENCI,T.;PEREIRA, T.;Pereira-Cenci, Tatiana;PEREIRA, T;PEREIRACENCI, T;Pereira, Tatiana;Pereira-Cenci, T.;MULO PATIAS, MAURO ELIAS MESKO

Lattes iD http://lattes.cnpq.br/6217846985830016

Endereço

Endereço Profissional Universidade Federal de Pelotas, Faculdade de Odontologia.
Rua Gonçalves Chaves, 457
Centro
96015560 - Pelotas, RS - Brasil
Telefone: (53) 32256741
Ramal: 135

Formação acadêmica/titulação

2006 - 2008 Doutorado em Clínica Odontológica (Conceito CAPES 7).
Universidade Estadual de Campinas, UNICAMP, Brasil.
com **período sanduíche** em Academisch Centrum Tandheelkunde Amsterdam (Orientador: Jacob Martien ten Cate).
Título: Avaliação da formação de biofilme de espécies de Candida formados sobre a superfície de resinas acrílicas para base e reembasamento de próteses removíveis. Ano de obtenção: 2008.

UNIVERSIDADE FEDERAL DE PELOTAS EM BRASIL

50 ANOS 1969 2019 UFPEL

From raw to input, an example from the LATTES system

From raw to input, an example from the LATTES system


```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?><CURRICULO-VITAE SISTEMA-ORIGEM-XML="LATTES_OFFLINE" NUMERO-IDENTIF
ICADOR="6217846985830016" DATA-ATUALIZACAO="15102019" HORA-ATUALIZACAO="140645"><DADOS-GERAIS NOME-COMPLETO="Tatiana Pereira C
enci" NOME-EM-CITACOES-BIBLIOGRAFICAS="PEREIRA-CENCI,T.;PEREIRA, T.;Pereira-Cenci, Tatiana;PEREIRA, T;PEREIRACENCI, T;Pereira,
Tatiana;Pereira-Cenci, T.;MULO PATIAS, MAURO ELIAS MESKO" NACIONALIDADE="B" PAIS-DE-NASCIMENTO="Brasil" UF-NASCIMENTO="SP" CID
ADE-NASCIMENTO="Santos" PERMISSAO-DE-DIVULGACAO="NAO" DATA-FALECIMENTO="" SIGLA-PAIS-NACIONALIDADE="BRA" PAIS-DE-NACIONALIDADE
="Brasil"><RESUMO-CV TEXTO-RESUMO-CV-RH="Professora Associada, Departamento de Odontologia Restauradora da Faculdade de Odonto
logia da Universidade Federal de Pelotas. Graduada em Odontologia pela FOB/USP (2001), Especialista em Prótese Dentária pelo H
RAC/USP (2004), Mestre (2006) e Doutora (2008) em Clínica Odontológica/UNICAMP, com PDEE-CAPES na ACTA/Holanda (Nov/2006 a Out
/2007). Tem experiência nos seguintes temas: biofilme, ensaios clínicos randomizados e revisões sistemáticas. É membro do The
BRIGHTER (Bias, Reporting, Implementation, Guidance, ETHics, IntEgrity and Reproducibility in Research) Meta-Research Group In
itiative." TEXTO-RESUMO-CV-RH-EN= ....
```

LATTES XML format: nodes and attributes

```
<CURRICULO-VITAE SISTEMA-ORIGEM-XML="LATTES_OFFLINE" ...>
   |                  |                    |
  NODE NAME        ATTRIBUTE NAME       ATTRIBUTE VALUE

<DADOS-GERAIS NOME-COMPLETO="Tatiana Pereira Cenci" ...>
   |                  |                    |
  NODE NAME        ATTRIBUTE NAME       ATTRIBUTE VALUE
```

Structuring XML data using R

```
library(xml2)
xml  <- read_xml("curriculo.xml")
node <- xml_find_first(xml,"/CURRICULO-VITAE")

d <- data.frame(
    LattesId  = xml_attr(node, "NUMERO-IDENTIFICADOR")
  , Updated   = xml_attr(node, "DATA-ATUALIZACAO")
)
print(d)
```

```
##           LattesId  Updated
## 1 6217846985830016 15102019
```

Structuring XML data using R

```r
library(lubridate)

# convert from text to proper data-time format
d$Updated <- dmy(d$Updated)

print(d)
```

```
##           LattesId    Updated
## 1 6217846985830016 2019-10-15
```

```r
# export to CSV format
write.csv(d, file="CV.csv")
```

With a little more work

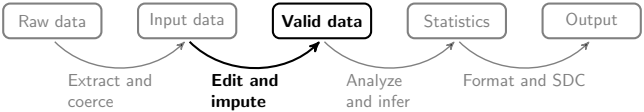
```r
node <- xml_find_first(xml, "//ENDERECO-PROFISSIONAL")
d$Institute <- xml_attr(node, "NOME-INSTITUICAO-EMPRESA")
d$Faculty   <- xml_attr(node, "NOME-ORGAO")

nodes <- xml_find_all(xml, "//ARTIGO-PUBLICADO")
d$Articles <- length(nodes)
nodes <- xml_find_all(xml, "//CAPITULO-DE-LIVRO-PUBLICADO")
d$BookChapters <- length(nodes)

print(d)
```

```
##           LattesId    Updated                        Institute
## 1 6217846985830016 2019-10-15 Universidade Federal de Pelotas
##                     Faculty Articles BookChapters
## 1 Faculdade de Odontologia      116            2
```

Valid data



Satisfy domain knowledge constraints

- The last update can not be in the future
- Full professorship under 24 is highly unlikely
- More than n papers/year is unlikely (depending on field)
- ...

Justification

Invalid data leads to invalid statistical results.

Statistics

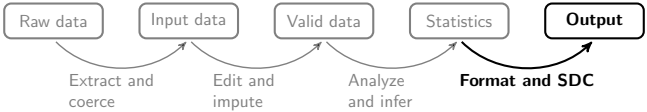
Target output values (aggregates)

- The resulting numbers for publication

Note

- These also need to satisfy domain knowledge constraints.

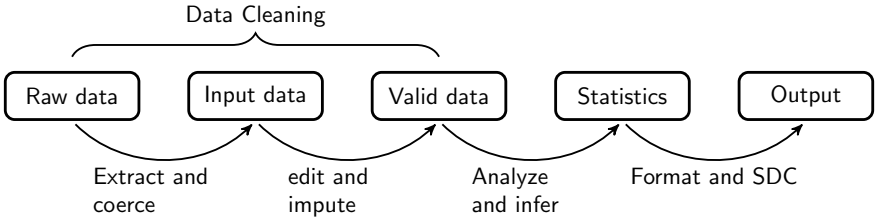
Output



Your paper!

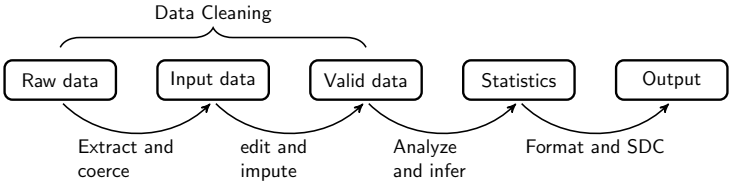
- Formatted, annotated
- Data possibly treated with anonymization techniques (SDC = statistical disclosure control)

The SVC: Remarks



- Actual data processing is not linear, you will go round a few times.
  - Build up the SVC as your research project progresses.
- Add or remove stages as needed.
- This general idea scales really well.

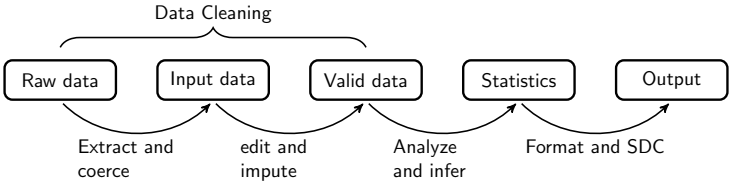
Quizz (1)



Where does the following activity take place?

Formatting date-time variable to `ISO8106` format.

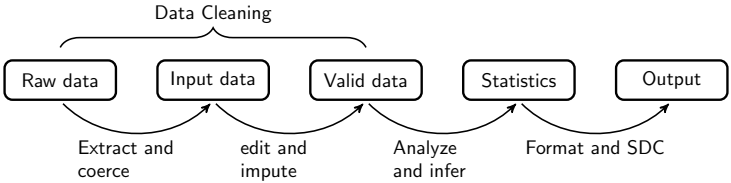
Quizz (2)



Where does the following activity take place?

Estimating effect of internationalization academic output.

Quizz (3)

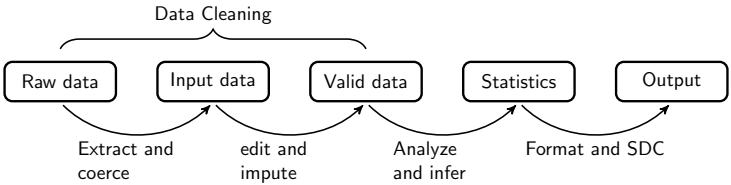


Where does the following activity take place?

Standardizing miss-spelled categories, e.g.

- `"Sim","si"` $\rightarrow$ `"sim"`
- `"NO", "Nao"` $\rightarrow$ `"não"`

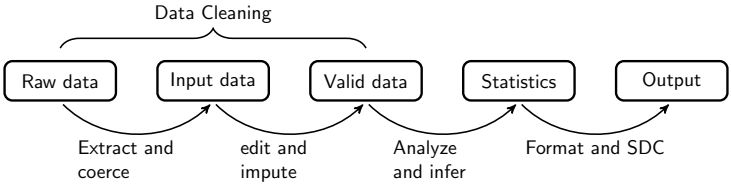
Quizz (4)



Where does the following activity take place?

Removing of fixing records where unemployed persons have a positive income from employment.

Quizz (5)



Where does the following activity take place?

Join data with a backbone using probabilistic linkage, based on approximate matches between various columns of the data and the backbone.

Implementation

Demo

Summary

1. Data represents information
2. It is important to choose a representation that suites analyses.
3. Obtaining, cleaning, analyzing data and reporting on results follow a value chain structure. It is useful to separate tasks accordingly.

There are free and open source tools supporting all necessary methods and transformations.

For the next lecture: please install R and RStudio

Instructions:

github.com/markvanderloo/UFPEL2019