

Improving Open Data Quality using Python

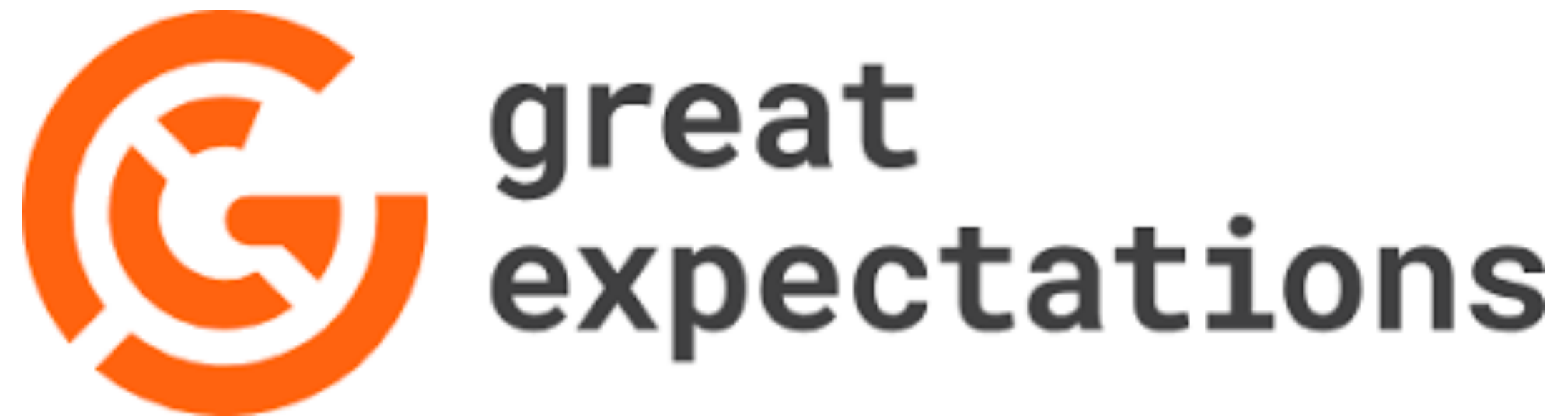
**How to measure and improve data quality
using Open Source Tools**

César García Sáez - @lahoramaker - 06/12/2023 - PyData Global 2023

Agenda

- Introduction to Open Data and Data Quality concepts (10 min)
- Validation of single datasets (40 min)
- Preparing longitudinal data (30 min)
- Q& A (10 min)

Tools we will use today: GX, pandas, sweetviz and more!



Getting your environment ready

Decide either local or cloud environment

- Repository: <https://github.com/elsatch/pydata-global-2023-Improving-Open-Data-Quality-using-Python>
- Full instructions are available in the README.md file
- TDLR; instructions:
 - Local: Clone the repo and follow the instructions to create a virtual environment and prepare it
 - Cloud: Open Google Colab, paste the repo URL to open it up in Colab
 - Note: Great Expectations is not officially supported in Windows

Intro to Data Quality concepts

Summary



`df.isnull().sum()`



**ISO
25012**

Data Cleaning using pandas

df.isnull.sum()

df.fillna(value)

df.dropna()

df.isna().any().sum()

df['col'].astype(correct_data_type)

Data Quality

International Standards, best practices and requirements

- ISO 25012
- ISO 25024
- DAMA DMBOK2
- New requirements added to the incoming EU Artificial Intelligence law
- Artificial Intelligence - Data quality for analytics and machine learning (ISO/IEC 5259-X por desarrollar)

For Spanish speakers

New norms recently published

- Norma UNE 0077:2023 - Gobierno del dato
- Norma UNE 0077:2023 - Gestión del dato
- Norma UNE 0079:2023 - Gestión de la calidad del dato
- Norma UNE 0080:2023 - Guía de evaluación del Gobierno, Gestión y Gestión de la Calidad del Dato
- Norma UNE 0081:2023 - Guía de la evaluación de la Calidad de un Conjunto de Datos

Ahora mismo: El acceso a estas Especificaciones UNE está patrocinado por la Oficina del Dato de la SEDIA, siendo su descarga gratuita.

The Quality of a Data Product may be understood as the degree to which data satisfy the requirements defined by the product-owner organization

ISO/IEC 25012

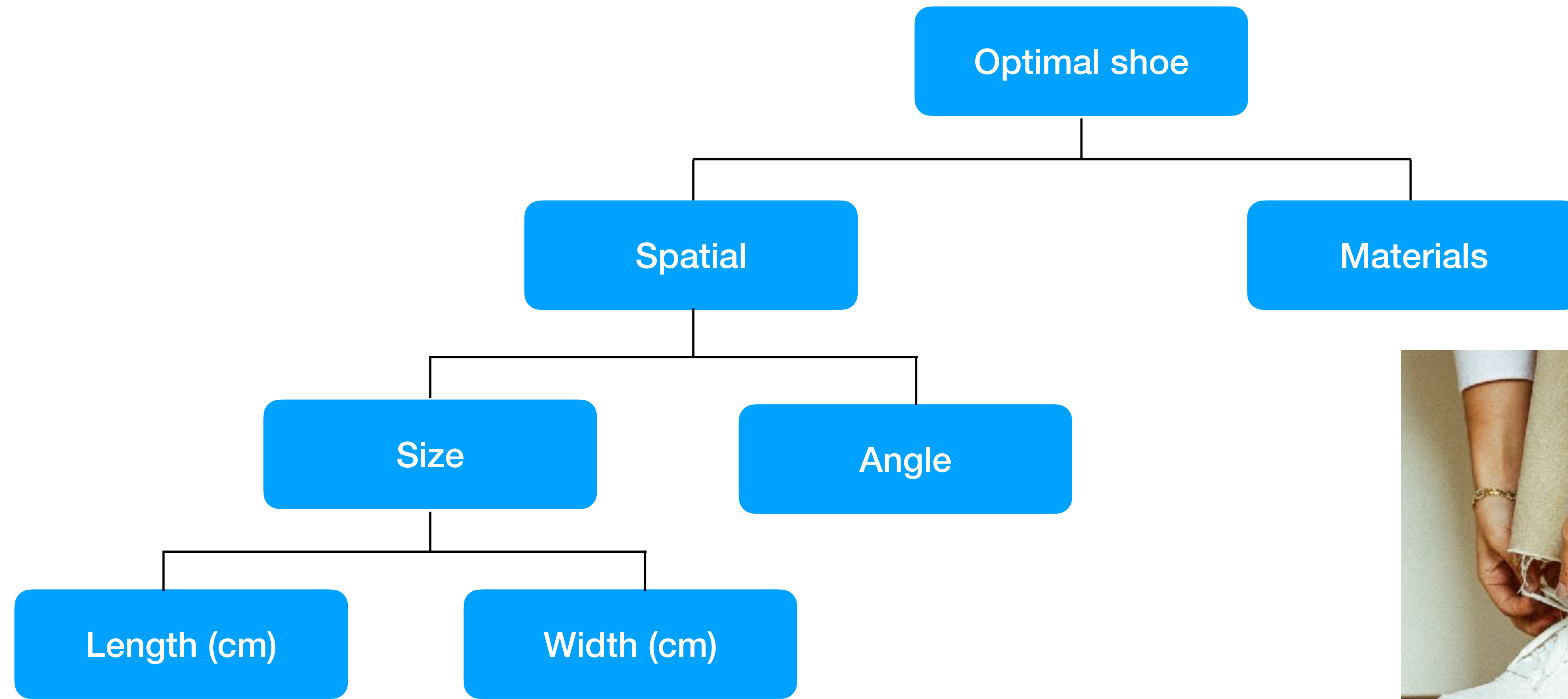
Reference: <https://iso25000.com/index.php/en/iso-25000-standards/iso-25012>

**Extent to which data characteristics are fit
for purpose.**

Personal version

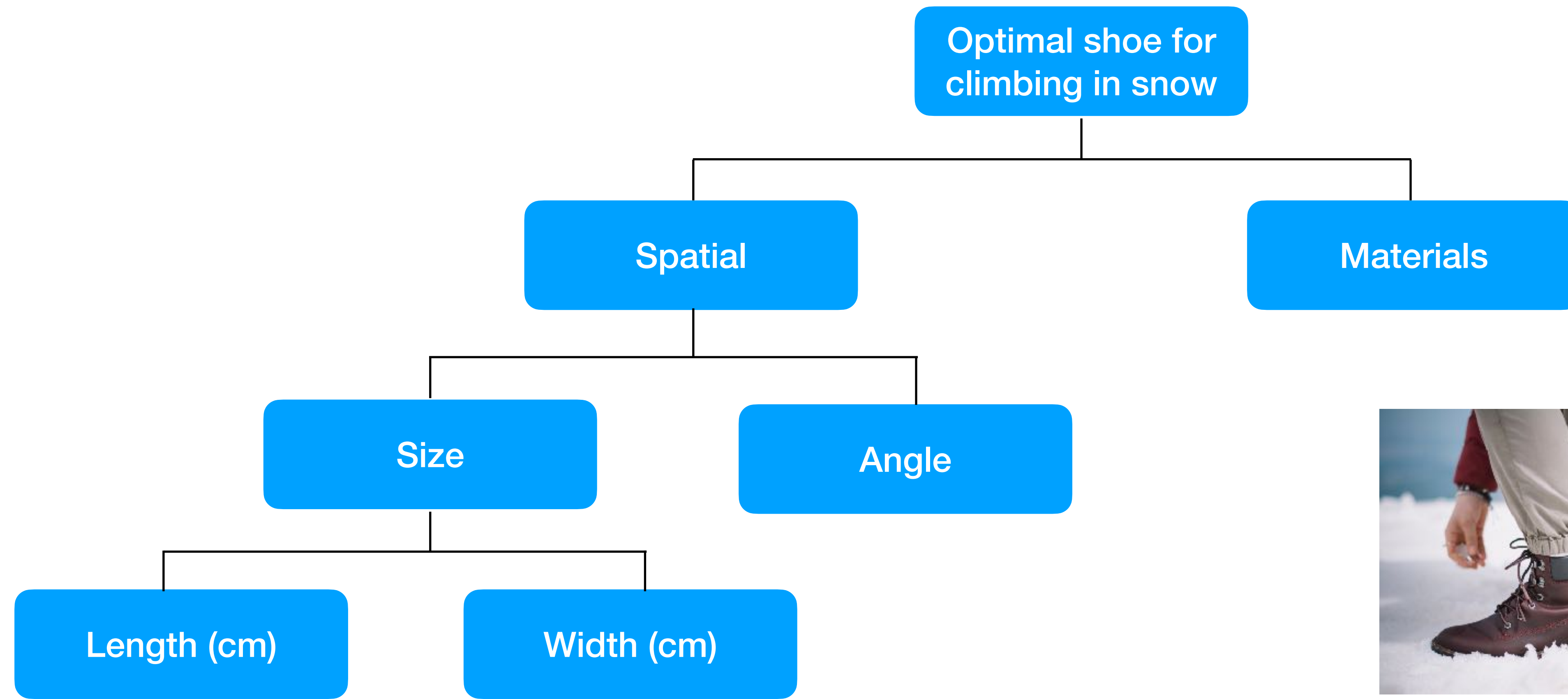
Choosing a shoe

Is it fit for purpose?



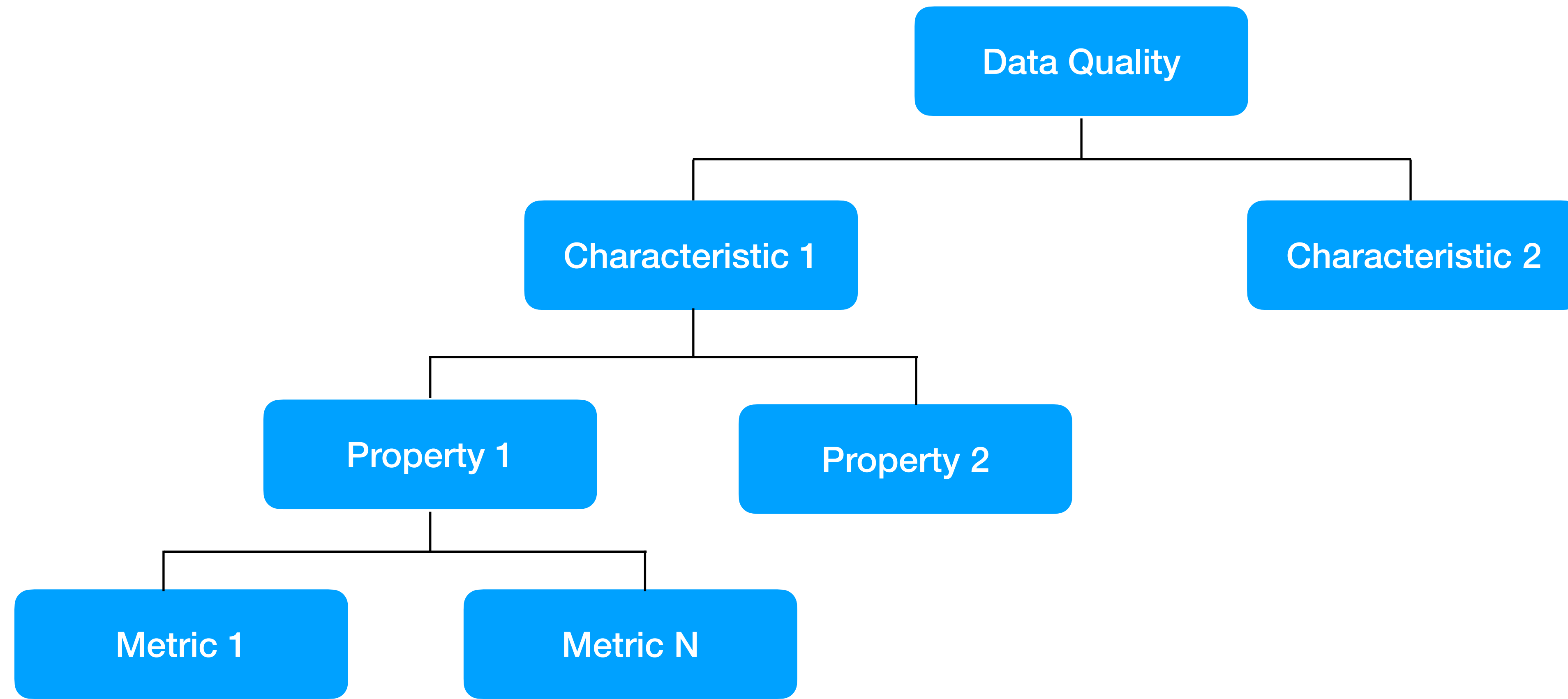
Choosing a shoe

Is it fit for purpose?



Data Quality characteristics / dimensions

Hierarchical approach

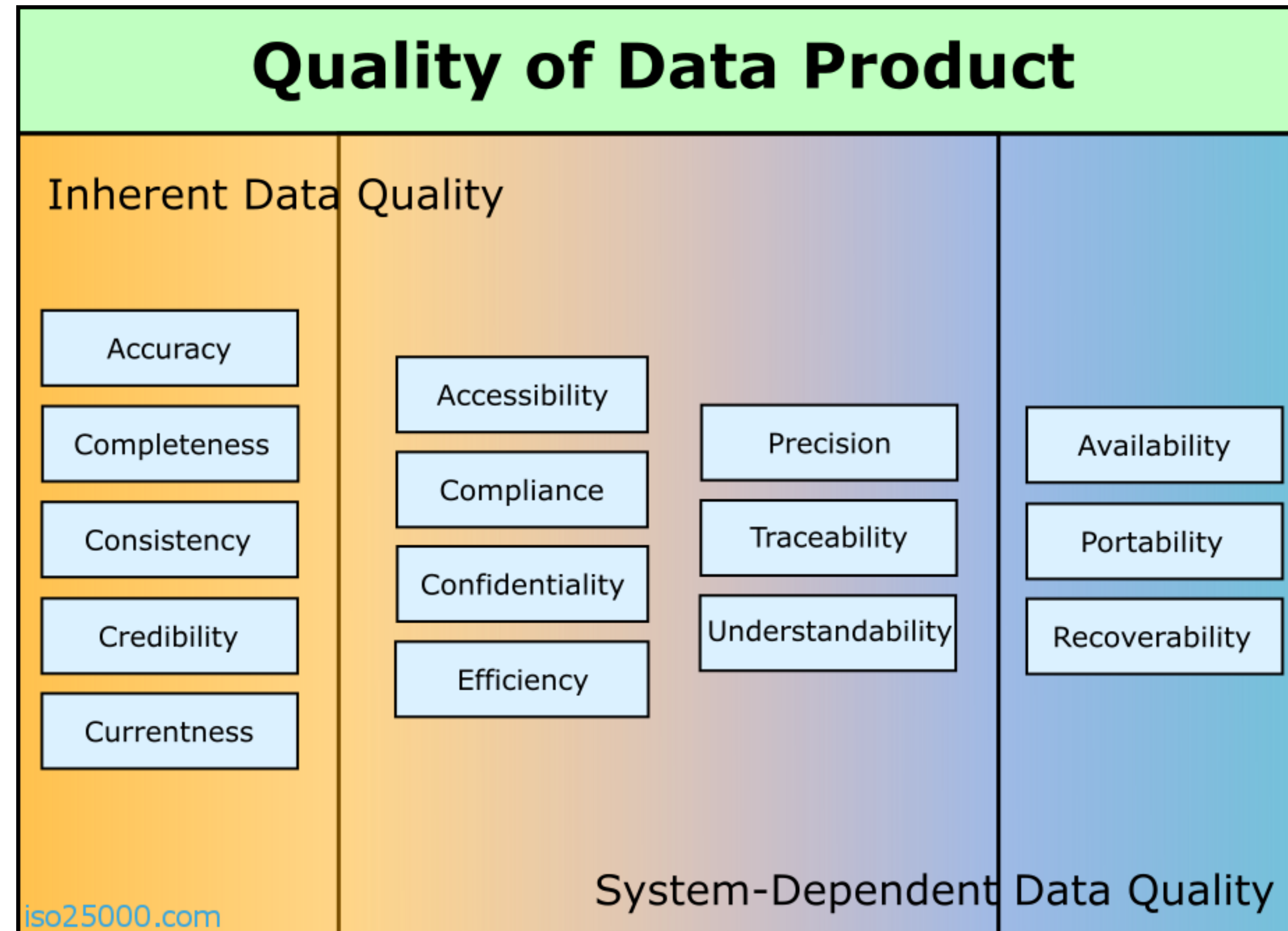


What is our purpose?

**What are data quality
dimensions?**

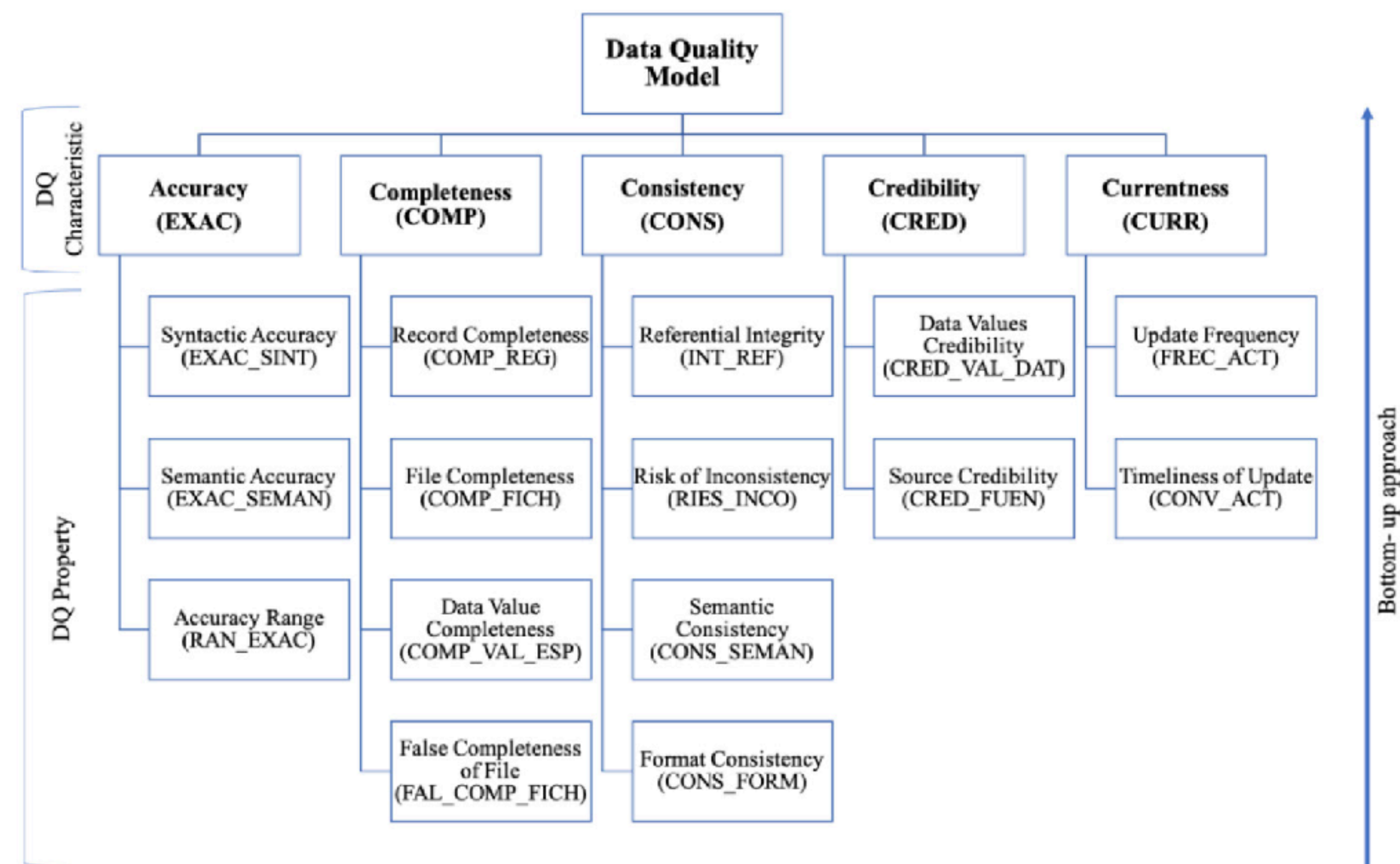
Characteristics ISO 25012

Inherent and System-Dependent



Inherent Characteristics

Can be extracted from the datasets



Simplified Data Quality definition process

From business requisites to data quality requisites

- Business **data requisite**: Each user account in the bank will have associated one or more IBAN associated
- **Data quality requisites**:
 - IBAN won't be null for any account
 - All IBAN numbers will be well formatted
 - (Or alternatively) At least 95% of the accounts would have a well formatted IBAN

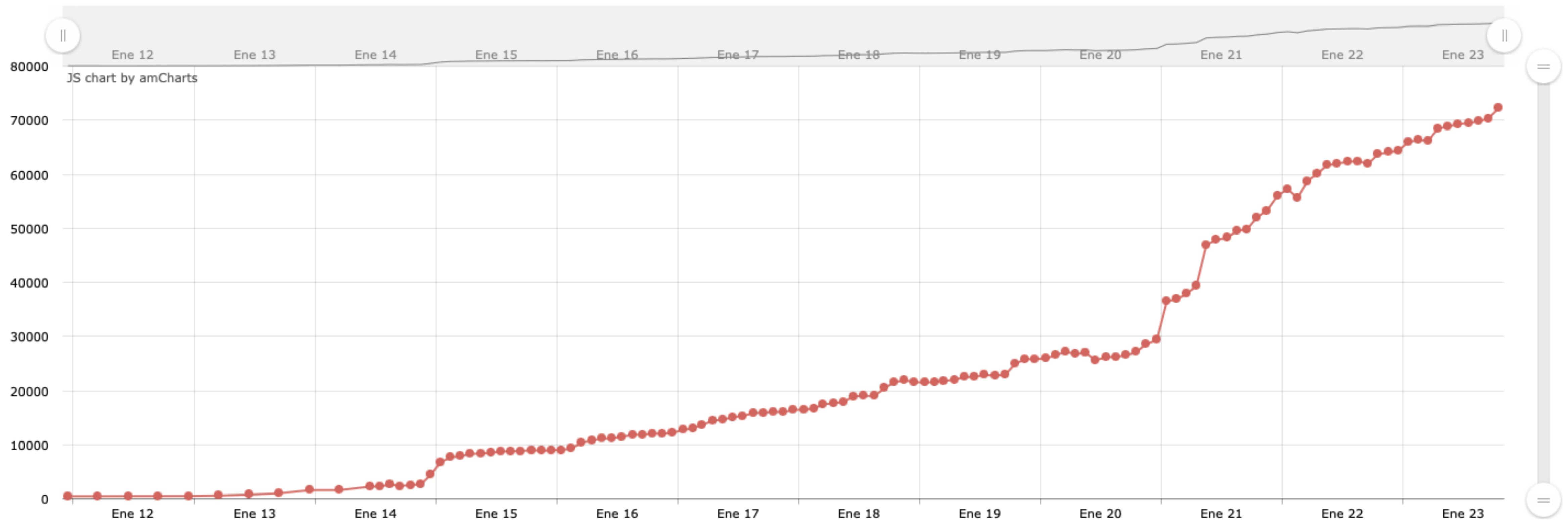
What is open data?

Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.

Open Data Handbook - Open Knowledge Foundation

Dataset growth in the last ten years

October 2023 - 72.230 datasets (Spain)



Source: <https://datos.gob.es/es/dashboard>

1.527.490

Datasets at data.europa.eu

**Quantity >
Quality**

Data quality dependent on purpose

Product owner determines data quality



Todos los registros tienen una fecha. No hay fechas en blanco.



conjunto_datos_2019.xls

2019			
ID	Fecha	Importe	Adjudicatario
1	5/01/2019		
2	4 Marzo 2019		
3	12/31/2019		
4	1.Oct.19		



Los formatos de las fechas son todos diferentes y dificulta el análisis.



Open Data Toronto - Data Quality Score (DQS)



OPEN DATA

[Data Catalogue](#) [Knowledge Centre](#) [About](#) [Gallery](#) [Contact](#)

Search

SEARCH

[OPEN DATA PORTAL HOME](#) / [OPEN DATA CATALOGUE](#) / [CATALOGUE QUALITY SCORES](#) DetailsData quality score beta

 Gold

Data last refreshed

Dec 17, 2019

Refreshed

Weekly

Data type

Table

Topics

City government

License

Open Government License -
Toronto

Publisher

Published by

Information & Technology

Contact

carlos.hernandez@toronto.ca

About Catalogue quality scores

The Data Quality Score reflects, in the form of a Gold, Silver or Bronze badge, how valuable a dataset is based on a set of characteristics that increase its potential to be used for addressing civic issues such as how usable, timely, complete, and well-described it is. High quality data enables high quality impact.

This dataset contains the current and historic Data Quality Score results for the Open Data Toronto catalogue, as well as the versions of the algorithm used for scoring.

Collection Method

Querying the Open Data Portal via the CKAN API.

Limitations

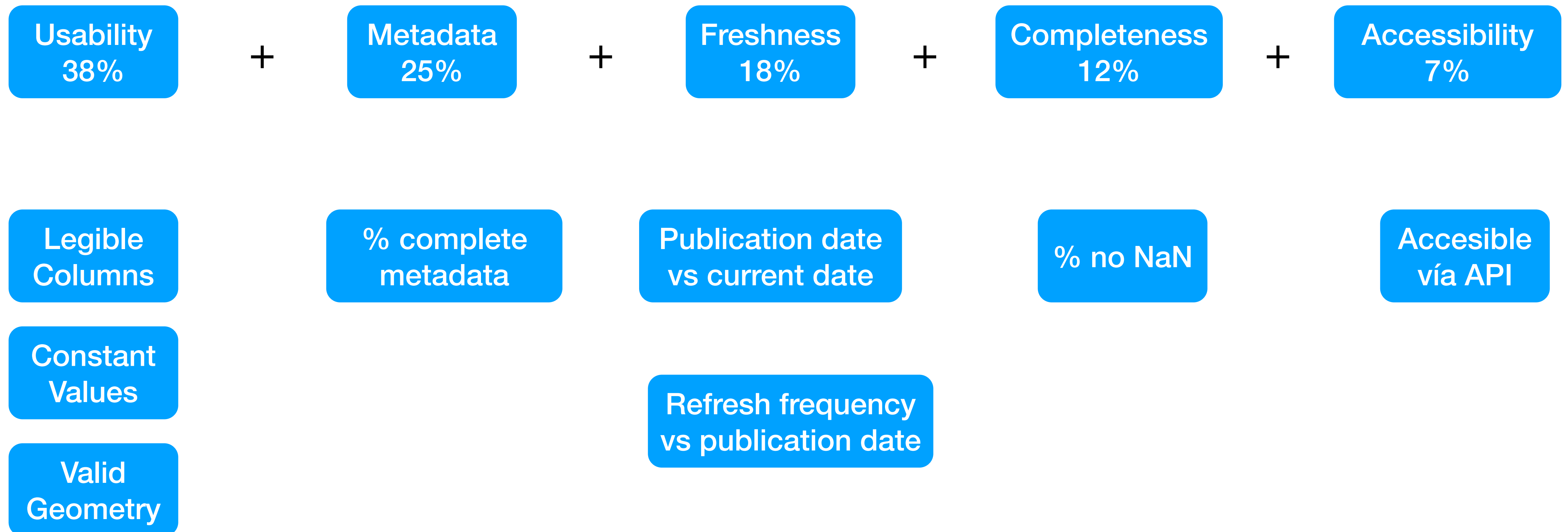
1. Data Quality Score applies only to resources in the CKAN datastore, which is a SQL database. Static files are not scored due to lack of standardization and inability to readily read the data. This means datasets containing only files, such as Excel or Zip, are not scored.
2. There is no distinction between "Read Me" and "data" resources. They are both assessed and weighted equally when calculating the final score.

DATA PREVIEW

_id	package	accessibility	completeness	freshness	metadata	usability	score
186	air-conditioned-and-cool-	1	0.69	0.5	0.84	0.86	0.78

How is a data quality score calculated?

Open Data Toronto



**pandas is part of the solution,
but not the whole picture**

Root cause remediation: Improving the quality of data goes beyond correcting errors. Problems with the quality of data should be understood and addressed at their root causes, rather than just their symptoms. Because these causes are often related to process or system design, improving data quality often requires changes to processes and the systems that support them.


DAMA DMBOK2 - Data Quality Programs principles

**We are aiming for root cause
solutions based on documented
observations**

Validation of single datasets

Choosing our dataset


Open Data Toronto - Bike Theft




OPEN DATA [Data Catalogue](#) [Knowledge Centre](#) [About](#) [Gallery](#) [Contact](#)

SEARCH

[OPEN DATA PORTAL HOME](#) / [OPEN DATA CATALOGUE](#) / [CATALOGUE QUALITY SCORES](#)

 **Details**

Data quality score beta
 Gold


Data last refreshed
Dec 17, 2019

Refreshed
Weekly

Data type
Table

Topics
[City government](#)

License
[Open Government License - Toronto](#)

 **Publisher**

Published by
Information & Technology

Contact
carlos.hernandez@toronto.ca

About Catalogue quality scores

The Data Quality Score reflects, in the form of a Gold, Silver or Bronze badge, how valuable a dataset is based on a set of characteristics that increase its potential to be used for addressing civic issues such as how usable, timely, complete, and well-described it is. High quality data enables high quality impact.

This dataset contains the current and historic Data Quality Score results for the Open Data Toronto catalogue, as well as the versions of the algorithm used for scoring.

Collection Method

Querying the Open Data Portal via the CKAN API.

Limitations

1. Data Quality Score applies only to resources in the CKAN datastore, which is a SQL database. Static files are not scored due to lack of standardization and inability to readily read the data. This means datasets containing only files, such as Excel or Zip, are not scored
2. There is no distinction between "Read Me" and "data" resources. They are both assessed and weighted equally when calculating the final score.

DATA PREVIEW

_id	package	accessibility	completeness	freshness	metadata	usability	score
186	air-conditioned-and-cool-	1	0.69	0.5	0.84	0.86	0.78

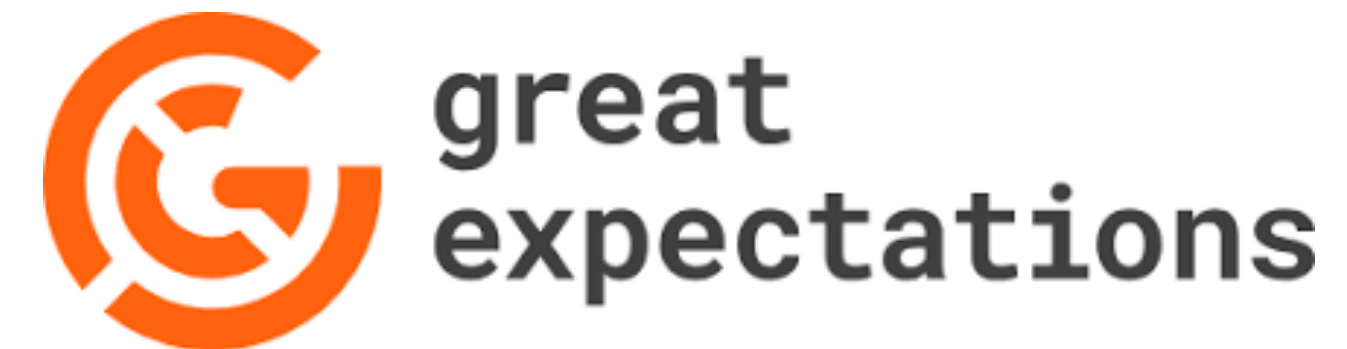
Data Quality Workflow - ISO 8000:61

Plan

Data Requirements

Data Quality Requirements

Check

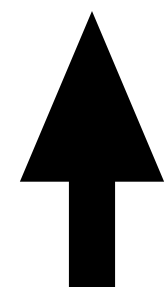
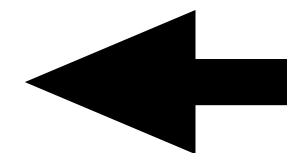
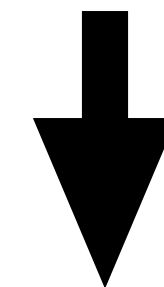
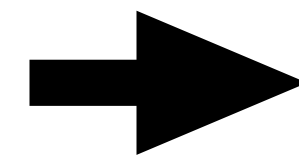


Do



Act

Improve the validations,
based on feedback



Time to open your notebooks!
single_datasets.ipynb

Detective mode

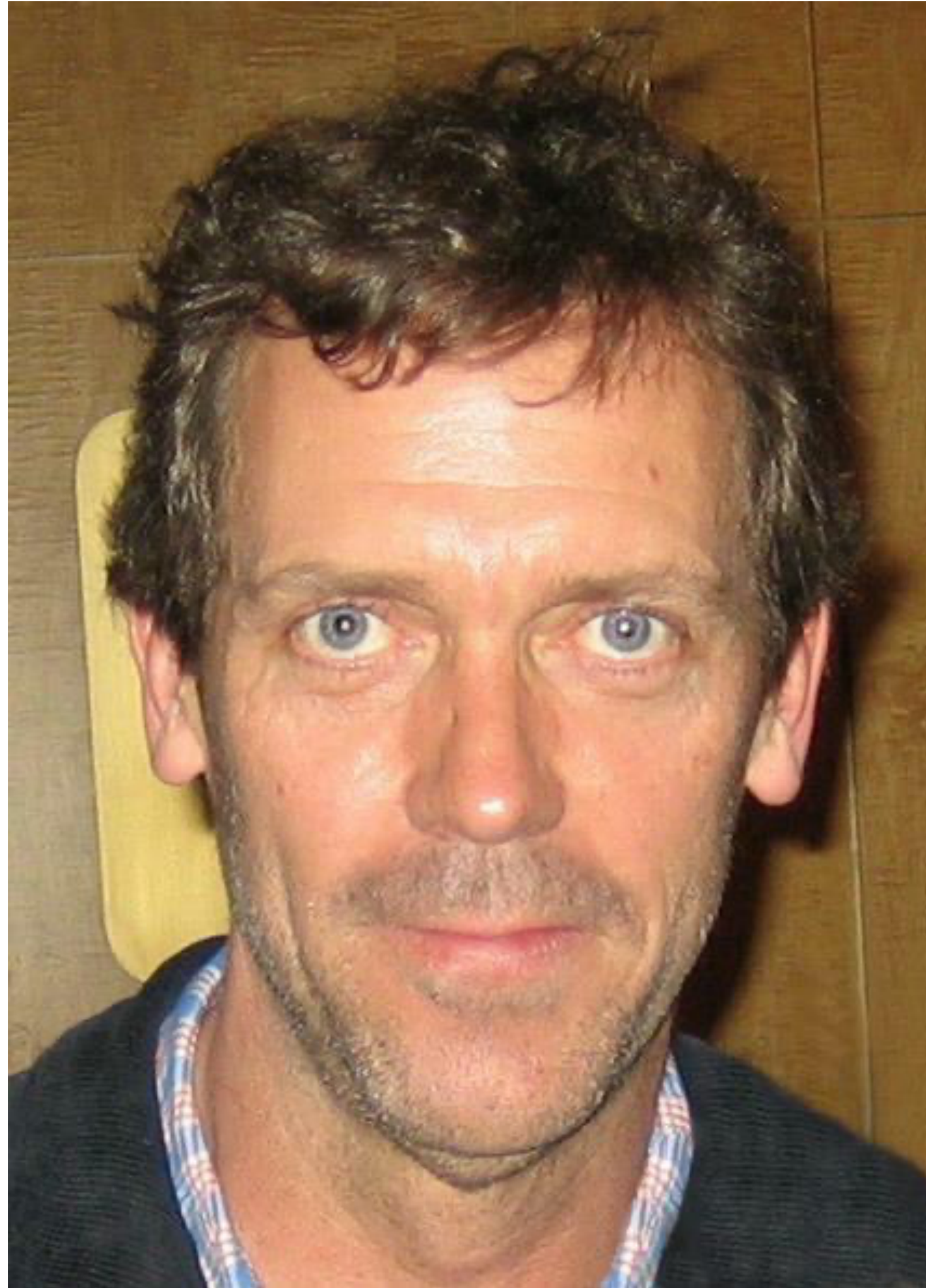


Photo credits Alan Light, CC BY 2.0, <https://commons.wikimedia.org/w/index.php?curid=1792538>

Plan

Explore Data Domain + Exploratory Data Analysis

- First step is to learn more about the domain of your dataset. Find documents, definitions, metadata, etc.
- Then perform an exploratory data analysis
- Tools we will use:
 - Web browser
 - Sweetviz (Python library for exploratory data analysis)



**Don't trust anything besides data
(Everything else is lies)**

(Well, technically not lies, but language is not very accurate most of the times, updates might take longer than expected, errors happen, etc.)

Plan

What do we want to measure?

- Data requirements
- Data quality requirements

Check

Expectations



`df.isnull().sum()`



`expect_column_values_to_not_be_null()`

Great Expectations

What does this library provide?

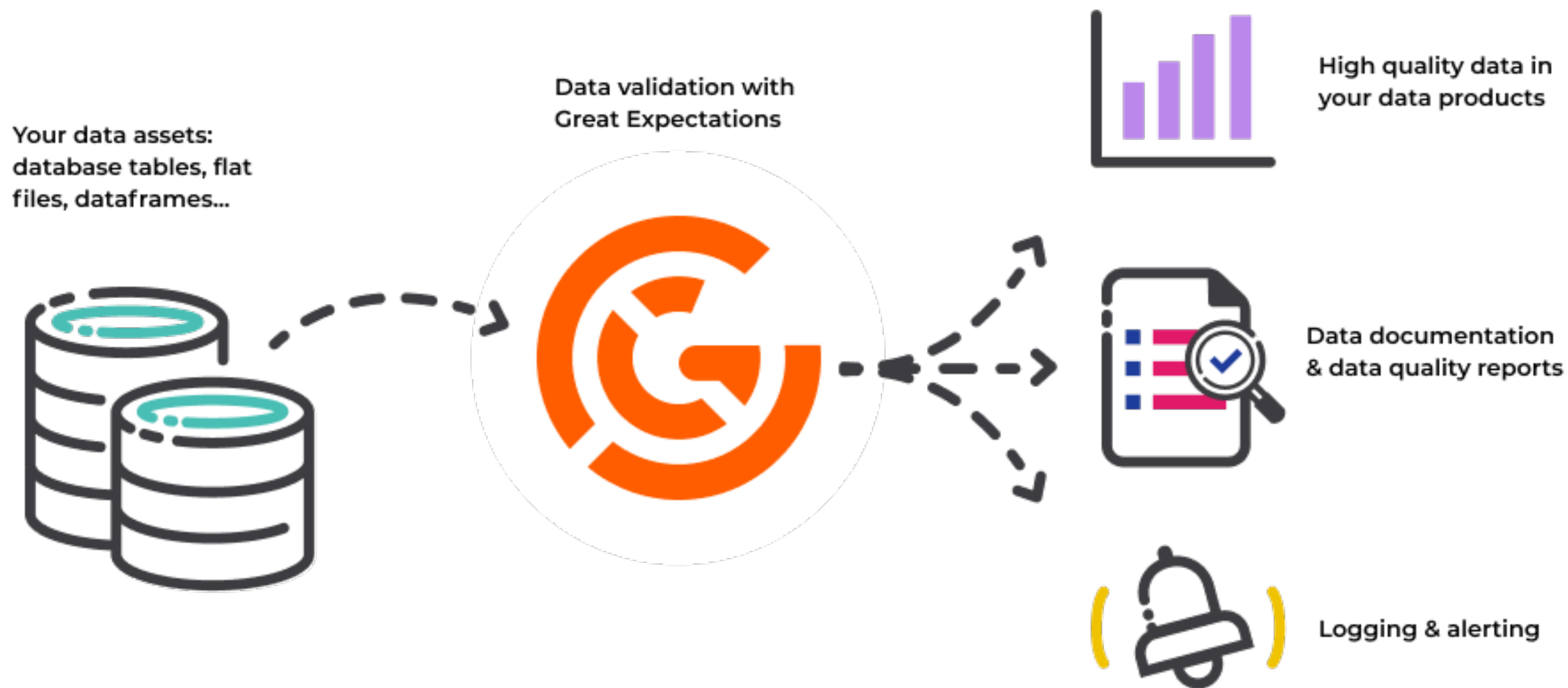
It facilitates the generation of data quality rules, called expectations, in pseudo-natural language.

ej. `expect_columns_not_to_be_null()`

It allows you to validate and document data, but the means for remediation are outside its scope.



Great Expectations




Sample expectations

Great Expectations Gallery

expect_column_values_to_not_be_null

● This expectation level is PRODUCTION

Contributors:

 @great_expectations

Tags:

core expectation

column map expectation

Metrics:

column_values.nonnull.unexpected_count

table.row_count

column_values.nonnull.unexpected_values

Backend support:

🔗 Pandas

🔗 Spark

🔗 SQLite

🔗 PostgreSQL

🔗 MySQL

🔗 MSSQL



















🔗 Trino

🔗 Redshift

🔗 BigQuery


















































































































































🔗 Snowflake

Core + Contrib - 324 in total

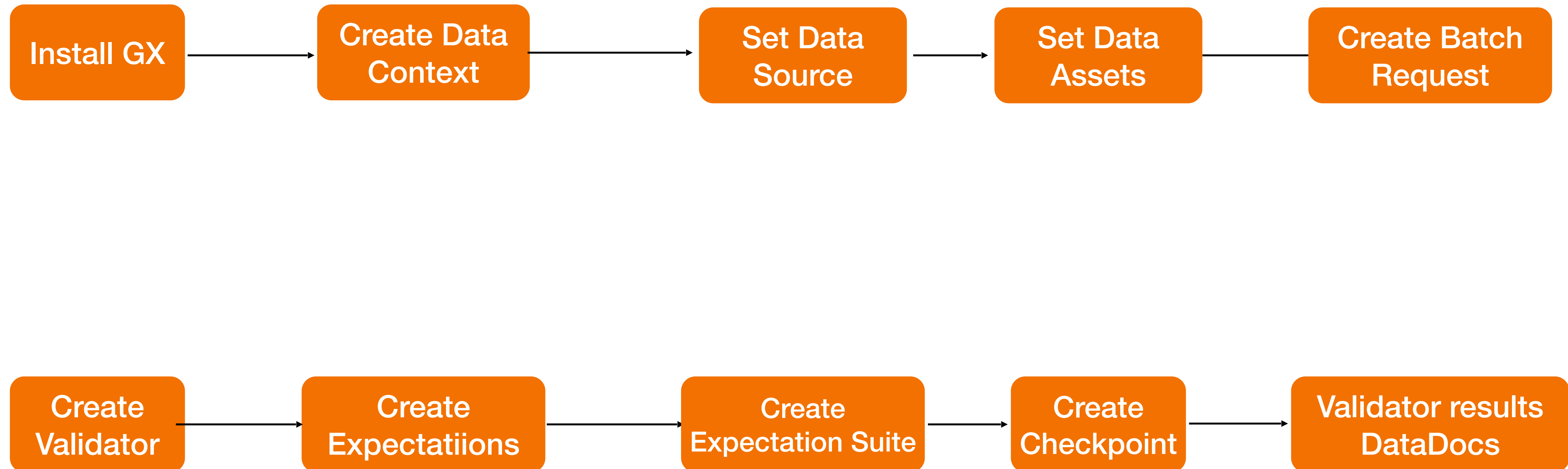
<div> <div>○</div> <div> expect_column_distinct_values_to_be_continuous (Contrib ColumnAggregateExpectation)  </div> </div> <div>See More</div> <p>Expect the set of distinct column values to be continuous.</p> <div> <div>Tags:</div> <div> <div>core expectation</div> <div>column aggregate ex...</div> </div> <div>Support:</div> <div></div> </div> <div>Contribution status:</div> <div> <div> <div>●</div> <div>Experimental</div> <div>4 / 4</div> </div> <div> <div>●</div> <div>Beta</div> <div>3 / 3</div> </div> <div> <div>○</div> <div>Production</div> <div>1 / 2</div> </div> </div>	
<div> <div>○</div> <div> expect_column_distinct_values_to_be_in_set (Core ColumnAggregateExpectation)  </div> </div> <div>See More</div> <p>Expect the set of distinct column values to be contained by a given set.</p> <div> <div>Tags:</div> <div> <div>core expectation</div> <div>column aggregate ex...</div> </div> <div>Support:</div> <div>        <div>+3</div> </div> </div> <div>Contribution status:</div> <div> <div> <div>●</div> <div>Experimental</div> <div>4 / 4</div> </div> <div> <div>●</div> <div>Beta</div> <div>3 / 3</div> </div> <div> <div>●</div> <div>Production</div> <div>2 / 2</div> </div> </div>	
<div> <div>○</div> <div> expect_column_distinct_values_to_contain_set (Core ColumnAggregateExpectation)  </div> </div> <div>See More</div> <p>Expect the set of distinct column values to contain a given set.</p> <div> <div>Tags:</div> <div> <div>core expectation</div> <div>column aggregate ex...</div> </div> <div>Support:</div> <div>        <div>+3</div> </div> </div> <div>Contribution status:</div> <div> <div> <div>●</div> <div>Experimental</div> <div>4 / 4</div> </div> <div> <div>●</div> <div>Beta</div> <div>3 / 3</div> </div> <div> <div>●</div> <div>Production</div> <div>2 / 2</div> </div> </div>	

Massive support for pandas

But also for other popular backends

Filter by:		Backend support										
		Select Items										
												
	expect_batch_row_count_to_match_prophet_date_model (Contrib BatchExpectation)											
	expect_column_average_lat_lon_pairwise_distance_to_be_less_than (Contrib ColumnAggregateExpectation)											
	expect_column_average_to_be_within_range_of_given_point (Contrib ColumnAggregateExpectation)											
	expect_column_chisquare_simple_test_p_value_to_be_greater_than (Contrib BatchExpectation)											
	expect_column_discrete_entropy_to_be_between (Contrib ColumnAggregateExpectation)											
	expect_column_distinct_values_to_be_continuous (Contrib ColumnAggregateExpectation)											
	expect_column_distinct_values_to_be_in_set (Core ColumnAggregateExpectation)											
	expect_column_distinct_values_to_contain_set (Core ColumnAggregateExpectation)											
	expect_column_distinct_values_to_equal_set (Core ColumnAggregateExpectation)											
	expect_column_distribution_to_match_benfords_law (Contrib ColumnAggregateExpectation)											
	expect_column_kl_divergence_to_be_less_than (Core ColumnAggregateExpectation)											
	expect_column_kurtosis_to_be_between (Contrib ColumnAggregateExpectation)											
	Experimental	4 / 4			Beta		3 / 3			Production		2 / 2

GX Interactive Data Validation Workflow



Data Context

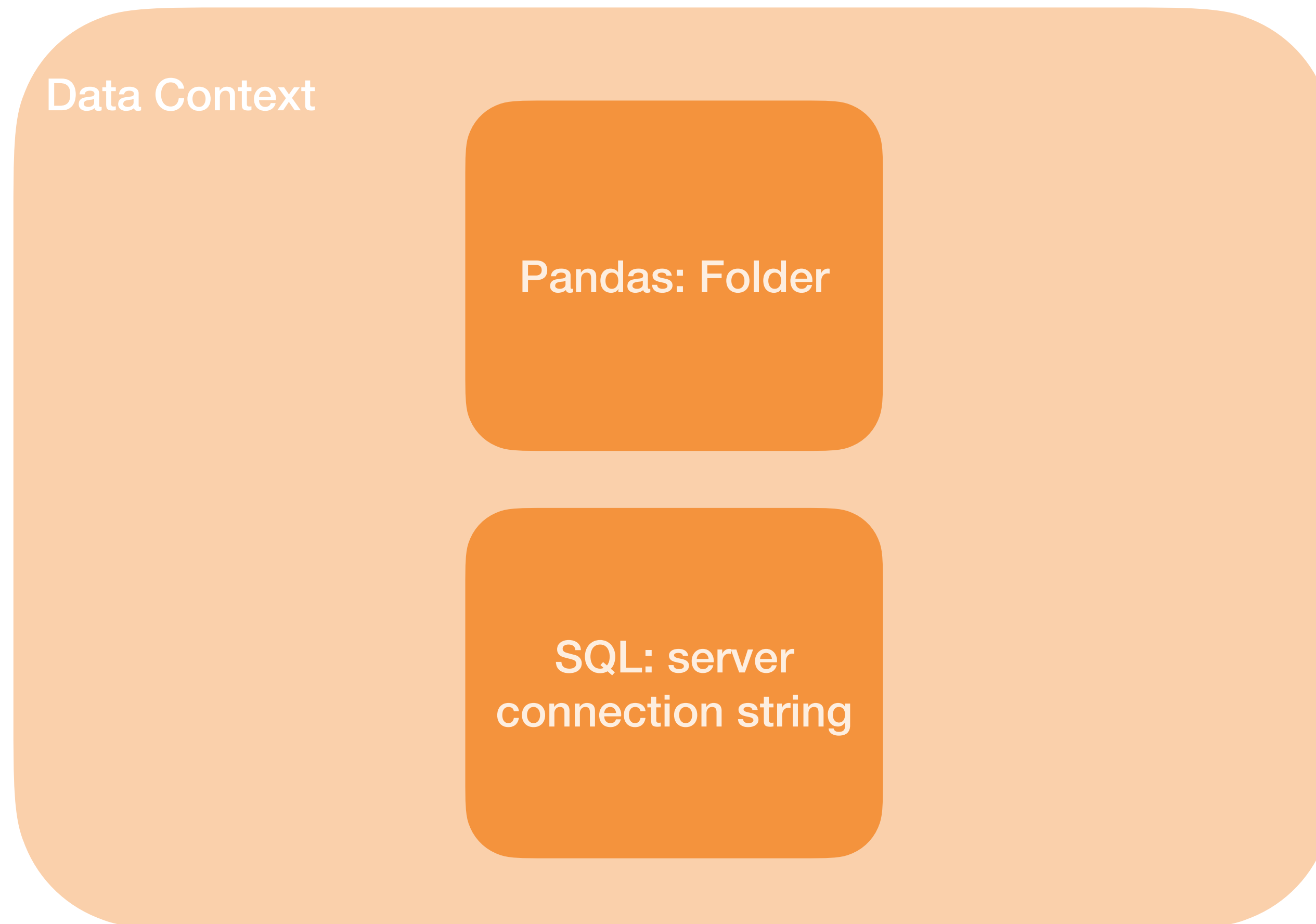
High level container for GX objects related to single topic



Data Context

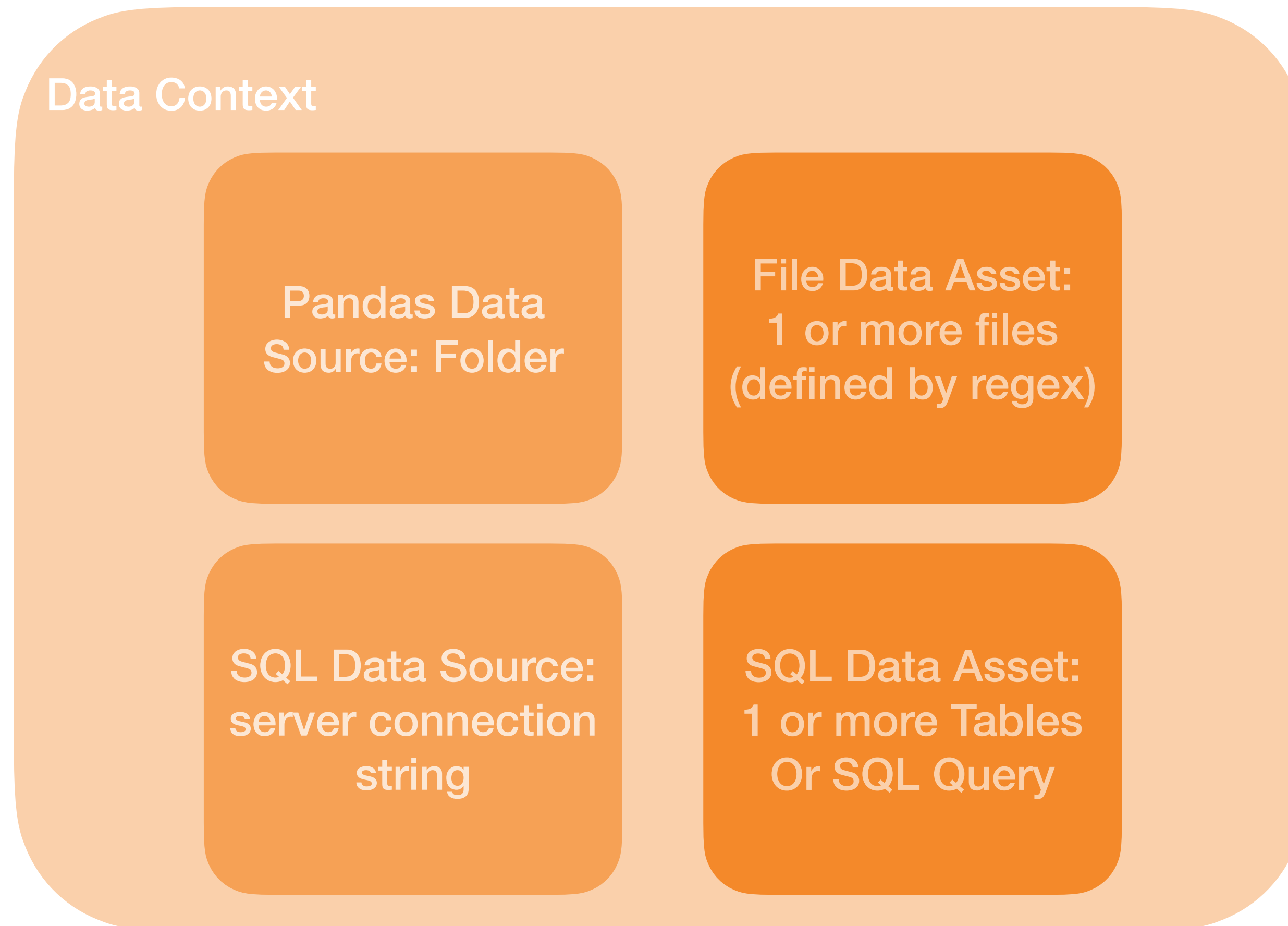
Data Source

Defines where is data located (pandas, postgresql, Snowflake, etc).



Data Asset

A collection of records within a data source



Batch Request

Creates a batch of data from a defined data asset

Data Context

Pandas Data
Source: Folder

File Data Asset:
1 or more files
(defined by regex)

Slice of records
retrieved from the
files

SQL Data Source:
server connection
string

SQL Data Asset:
1 or more Tables
Or SQL Query

Subset of the
records retrieved
from the Data
Asset

Expectation

A verifiable assertion about data

`expect_column_values_to_match_regex`

`expect_column_values_to_be_unique`

Expectation Suite

A collection of verifiable assertions about data

```
graph TD; A[Expectation Suite] --- B[expect_column_values_to_match_regex]; A --- C[expect_column_values_to_be_unique];
```

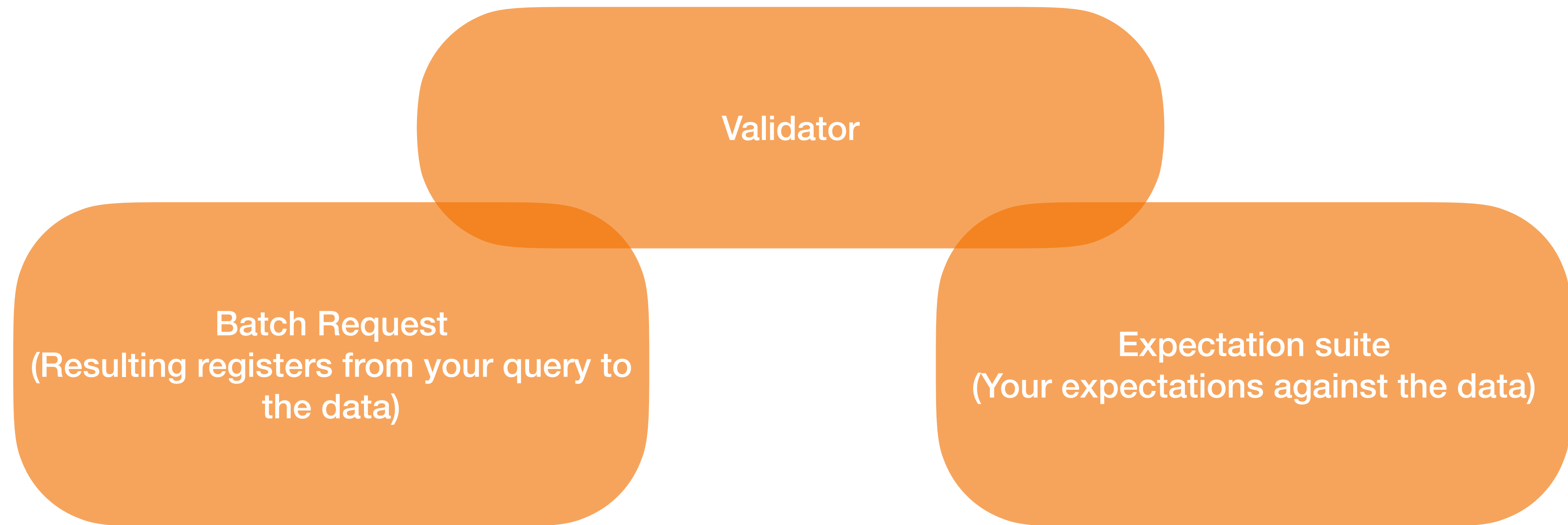
Expectation Suite

`expect_column_values_to_match_regex`

`expect_column_values_to_be_unique`

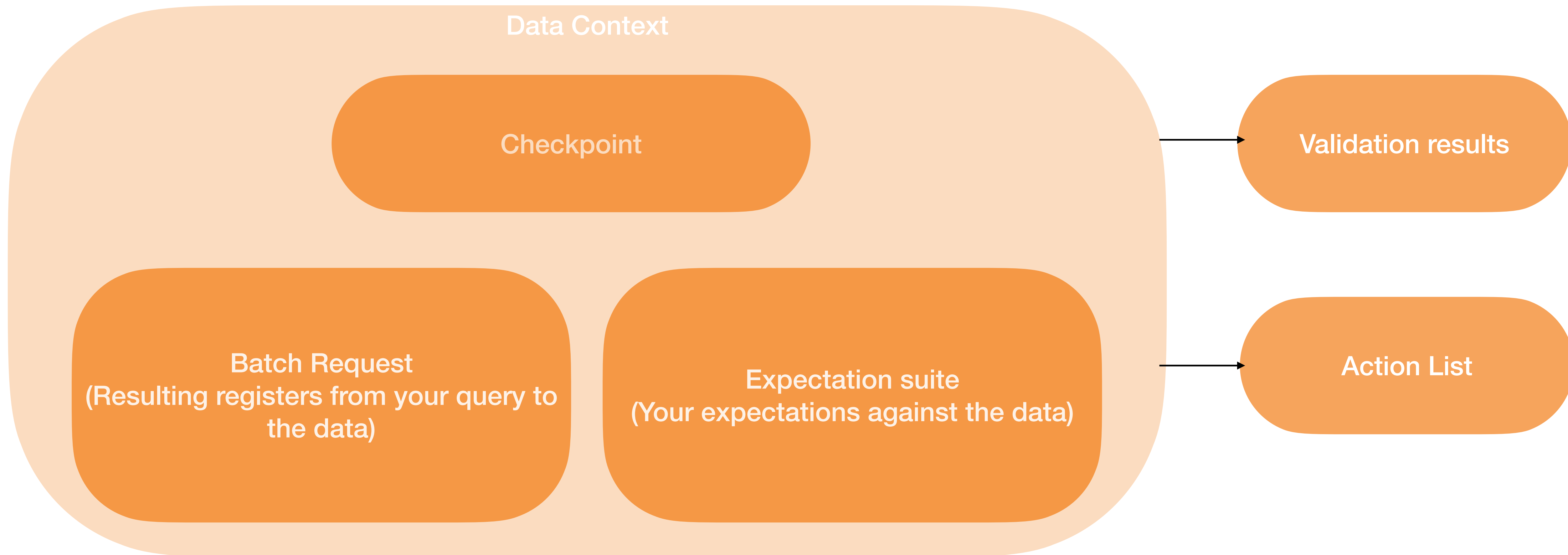
Validator

An object to facilitate interactive validation of data



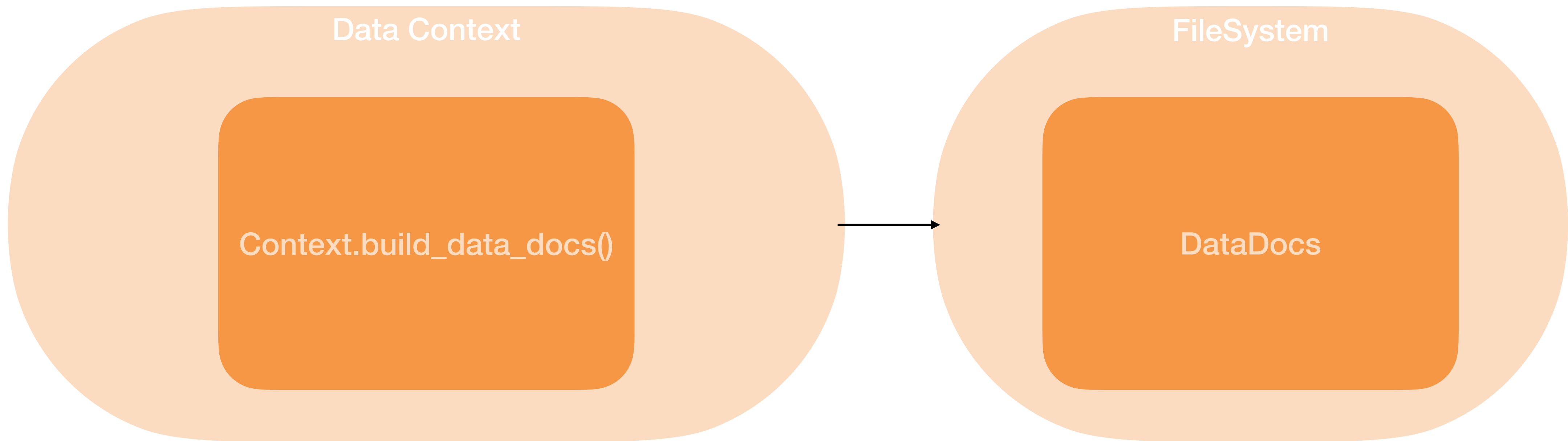
Checkpoint

Primary mean to validate data in production



Data Docs

HTML Docs that display all checkpoint results



Pros

- You can use the same function against multiple backends:: Pandas, Spark, SQLite, Postgresql, MySQL, MSSQL, Trino, Redshift, BigQuery, Snowflake
- Integrations: Airflow, Databricks, Meltano, Datahub y most of cloud services (AWS, GCP, Azure, etc.)
- The expectations are quite readable and verbose
- Interpretability
- Open Source Solution
- GX Cloud for a managed solution

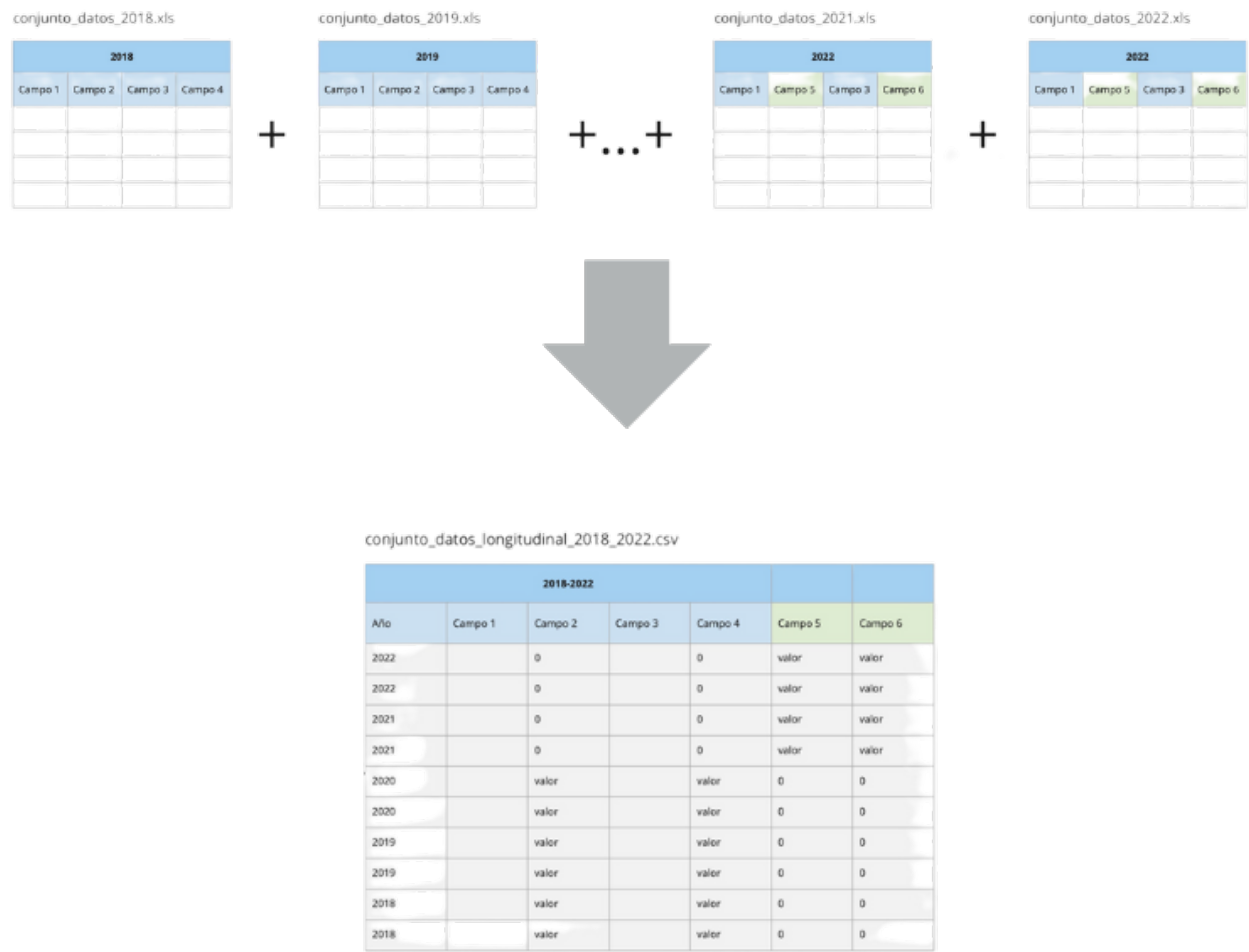
Cons

- Configuration might not be trivial
- Configuration format has evolved lots over time (now: no CLI, no blocks, no Jupyter based config)
- Requires other tools to remediate data quality issues
- It's quite hard for non technical profiles

Preparing longitudinal data

What is longitudinal data?

Multiple yearly datasets are very common in open data



Why does longitudinal data matter?

Required to compare performance over time


- Enables check and balances
 - Analyze the evolution of public services
 - Being able to pinpoint any mis-practice around public contracting
 - Track law changes to “accidental fires”
 - Allows training of machine learning predictive systems
 - Etc.

Common problematic patterns and pythonic solutions

Goal: Being able to `pd.concat()`

Inverse ladder pattern

Problem description

 Ajuntament de Barcelona												<p>Nota: Sempre cal informar les modificacions dels contractes d'acord amb els continguts a l'art. 304 LCCRP (gerència) i a l'art. 305 LCCRP (ex previsions). Respecte als contractes signats després de l'admissió, amb el requeriment dels terminis d'execució de les modificacions, en no s'ha considerat l'eliminació el preu del contracte, no es consideren modificacions en el mateix ítem a l'art. 304 i 305 de la LCCRP. Les informacions que requereix d'un contracte no són modificacions del preu del contracte.</p>									
CONTRACTES AMB MODIFICACIONS 2021 (1 de gener a 31 de desembre)																					
GERÈNCIES I DISTRICTES		Data Actualització de les dades: 14/06/2022																			
Òrgan contractant	Contracte (nòm., serv., subministraments...)	Típus de contracte (llogar, servei, subministraments...)	Objecte del contracte			NOM ARQUELCATARI (RAÓ SOCIAL)		NIF ARQUELCATARI (Presència Plànolge associat/licitat)	Data formalització contracte	Preu contracte (sense IVA)	Preu contracte (IVA inclòs)	Data formalització modificació	Import modificació (sense IVA)	Import modificació (IVA inclòs)	Típus modificació (sempre amb una "E") Presència en el plec (art. 304 LCCRP)	% variació sobre el preu del contracte (sense IVA)					
															No presència en el plec (art. 305 LCCRP)						

[illegible]



Ajuntament de Barcelona

CONTRACTES AMB MODIFICACIONS 2019 (1 gener a 31 desembre)

GERÈNCIAES I DISTRICTES DE L'AJUNTAMENT DE BARCELONA

* Iva inclòs

Òrgan contractant	Contracte sol·licitat	Tipus de contracte (Obres, serveis, subministrament materials...)	Objecte contracte	Data adjudicació / formalització	Preu* contracte	Data formalització modificació	Import * modificació	Tipus modificació (margen amb IVA)		% variació sobre el preu del contracte	
								Prescrita en el plec	No prescrita en el plec		
Gerència de Recursos	15002241-002	Privat d'Administració Pública	Cessió obra artística Bombers de Barcelona	26/10/15	40.000,00	NO CONSTA	93.000,00	SI		25,00	Ampliació de temps- Reajustament d'anualitats

[illegible][illegible]

Inverse ladder pattern

Remediation

- Do an EDA on the whole set of files
- Select one of the files as the golden standard (usually the most recent year)
- Extract the column names in a set
- Create an expectation to check if columns are in set
- Validate against the golden standard
- Validate the expectation against previous years until it breaks down massively

Inconsistent formatting

Problem description

2021 Llistat Ampliat Contractistes de Contractes Públics Adjudicats en 2021 - Ajuntament de Barcelona (Gerències i Districtes)									
Nom Adjudicatari	NIF (per Publicar: PF anonimitzat)	Tipus Contracte	Objecte		Procediment (detall per a Publicar)	Import Adjudicació Net (sense IVA)	Import Adjudicació amb IVA	Nombre de Contractes	
Saltos de linea									
2020 Llistat Ampliat Contractistes de Contractes Públics Adjudicats en 2020 - Ajuntament de Barcelona (Gerències i Districtes)									
Nom Adjudicatari	NIF (per a Publicar: PF anonimitzat)	Tipus Contracte	Objecte		Procediment (detall per a publicar)	Import Adjudicació Net (sense IVA)	Import Adjudicació amb IVA	Nombre de Contractes	
Cambios de nombre									
2019 Llistat Ampliat Proveïdors Contractes Públics Ajuntament de Barcelona (Gerències i Districtes)									
Nom Proveïdor	NIF (per Publicar: PF anonimitzat)	Tipus Contracte	Objecte		Procediment (per Publicar)	Import d'Adjudicació Net	Import d'Adjudicació IVA	Nombre de Contractes (Cte+ NIF)	
Espacios en blanco									
Nom Proveïdor	NIF Proveïdor	Tipus de Contracte	Objecte		Procediment (per Publicar)	Import Adjudicació Net	Import d'Adjudicació+IVA	Nombre de Contractes (Cte+NIF)	
Cambios de nombre									
Proveïdor	NIF Proveïdor	Tipus Contracte	Objecte		Procediment	Import Adjudicació	Nombre de contractes		

Inconsistent formatting

Problem description

Gerència Drets de Ciutadania, Participació i Transparència	17002834-001	Serveis	Gestió, impuls i dinamització Centre Recursos DH	31/01/18	249.853,60	15/03/19	0,00	SI		0,00
Gerència Drets de Ciutadania, Participació i Transparència	18002708-001	Serveis	Servei atenció telefònica, gestió i tram. 010	19/02/19	11.021.334,34	NO CONSTA	-642.911,17	SI		-5,83
Gerència Drets de Ciutadania, Participació i Transparència	19001461-001	Serveis	Serveis de traducció i correcció de textos	02/04/19	18.876,00	NO CONSTA	0,00	0		0,00
Gerència Drets de Ciutadania, Participació i Transparència	19002112-001	Subministraments	Trobada BCN Ciutat Diversa (2019-2020)	26/08/19	164.424,86	10/12/19	12.829,33	SI		7,80
Gerència de Seguretat i Prevenció	18003713	Serveis	Neteja, descontaminació equips respiracio SPEIS	31/7/19	503.536,10	17/10/19	177.706,79	justament anualitats		35%
Gerència de Seguretat i Prevenció	19002901	Serveis	Suport organització comunitats veïns i veïnes	30/10/19	332.000,00	19/11/19	6916,66	justament anualitats		2%
Gerència de Seguretat i Prevenció	19002770	Serveis	Mant.preventiu i correctiu eines hidràuliques SPEIS Lot 1	5/12/19	85.498,60	19/12/19	9406,88	justament anualitats		11%
Gerència d'Ecologia Urbana	18002285	Subministraments	Subministrament de material d'impremta per a les campanyes de comunicació de l'Àrea d'Ecologia Urbana 2018-2019	08.01.2019	120.028,50 €	27.08.2019	24.005,70	X		20,00%
Gerència d'Ecologia Urbana	16001721	Serveis	Assistència Tècnica i Control de Qualitat del Contracte de Conservació de l'Enllumenat Públic de Barcelona (2016-19)	14.12.2019	775.473,85 €	31.10.2019	21.000,00	X		2,71%
Gerència d'Ecologia Urbana	18004961	Mixte	Subministrament i serveis de col·locació i retirada de plafons electorals durant el període 2019-2021	05.04.2019	73.928,79 €	25.10.2019	20.500,00	X		27,73%
Gerència d'Ecologia Urbana	16002516	Serveis	Conservació de les instal·lacions d'enllumenat - Lot 1 (Ciutat Vella, Eixample i Sants-Montjuïc)	05.12.2016	17.374.923,70 €	09.12.2019	40.000,00	X		
Gerència d'Ecologia Urbana	16002517	Serveis	Conservació de les instal·lacions d'enllumenat - Lot 2 (Les Corts, Sarrià-Sant Gervasi, Gràcia i Horta-Guinardó)	05.12.2016	17.641.863,60 €	09.12.2019	160.000,00	X		
Gerència d'Ecologia Urbana	16002518	Serveis	Conservació de les instal·lacions d'enllumenat - Lot 3 (Nou Barris, Sant Andreu i Sant Martí)	05.12.2016	21.083.212,70 €	09.12.2019	20.000,00	X		
Gerència d'Ecologia Urbana	16005605	Serveis	Serveis de recollida i trasllat d'animals de companyia perduts o abandonats a la ciutat de Barcelona al Centre d'Acolliment d'Animals de Companyia de Barcelona (CAACB),	06.10.2017	323.000,00 €	16.12.2019	32.300,00	X		10,00%
Gerència d'Ecologia Urbana	14003093	Serveis	Manteniment del clavegueram de Barcelona (2015-2022)	16.02.2015	103.512.021,28 €	23.12.2019	54.132,10	X		0,05%
Gerència d'Ecologia Urbana	17006454	Serveis	Servei de Manteniment d'Escales Mecàniques, Ascensors Verticals i Ascensors Inclinat 2018-2020	13.07.2018	3.384.216,51 €	23.12.2019	100.000,00	X		2,95%

Inconsistent formatting

Problem description

Gerència Drets de Ciutadania, Participació i Transparència	17002834-001	Serveis	Gestió, impuls i dinamització Centre Recursos DH	31/01/18	249.853,60	15/03/19	0,00	SI		0,00
Gerència Drets de Ciutadania, Participació i Transparència	18002708-001	Serveis	Servei atenció telefònica, gestió i tram. 010	19/02/19	11.021.334,34	NO CONSTA	-642.911,17	SI		-5,83
Gerència Drets de Ciutadania, Participació i Transparència	19001461-001	Serveis	Serveis de traducció i correcció de textos	02/04/19	18.876,00	NO CONSTA	0,00	0		0,00
Gerència Drets de Ciutadania, Participació i Transparència	19002112-001	Subministraments	Trobada BCN Ciutat Diversa (2019-2020)	26/08/19	164.424,86	10/12/19	12.829,33	SI		7,80
Gerència de Seguretat i Prevenció	18003713	Serveis	Neteja, descontaminació equips respiracio SPEIS	31/7/19	503.536,10	17/10/19	177.706,79	justament anualitats		35%
Gerència de Seguretat i Prevenció	19002901	Serveis	Suport organització comunitats veïns i veïnes	30/10/19	332.000,00	19/11/19	6916,66	justament anualitats		2%
Gerència de Seguretat i Prevenció	19002770	Serveis	Mant.preventiu i correctiu eines hidràuliques SPEIS Lot 1	5/12/19	85.498,60	19/12/19	9406,88	justament anualitats		11%
Gerència d'Ecologia Urbana	18002285	Subministraments	Subministrament de material d'impremta per a les campanyes de comunicació de l'Àrea d'Ecologia Urbana 2018-2019	08.01.2019	120.028,50 €	27.08.2019	24.005,70	X		20,00%
Gerència d'Ecologia Urbana	16001721	Serveis	Assistència Tècnica i Control de Qualitat del Contracte de Conservació de l'Enllumenat Públic de Barcelona (2016-19)	14.12.2019	775.473,85 €	31.10.2019	21.000,00	X		2,71%
Gerència d'Ecologia Urbana	18004961	Mixte	Subministrament i serveis de col·locació i retirada de plafons electorals durant el període 2019-2021	05.04.2019	73.928,79 €	25.10.2019	20.500,00	X		27,73%
Gerència d'Ecologia Urbana	16002516	Serveis	Conservació de les instal·lacions d'enllumenat - Lot 1 (Ciutat Vella, Eixample i Sants-Montjuïc)	05.12.2016	17.374.923,70 €	09.12.2019	40.000,00	X		
Gerència d'Ecologia Urbana	16002517	Serveis	Conservació de les instal·lacions d'enllumenat - Lot 2 (Les Corts, Sarrià-Sant Gervasi, Gràcia i Horta-Guinardó)	05.12.2016	17.641.863,60 €	09.12.2019	160.000,00	X		
Gerència d'Ecologia Urbana	16002518	Serveis	Conservació de les instal·lacions d'enllumenat - Lot 3 (Nou Barris, Sant Andreu i Sant Martí)	05.12.2016	21.083.212,70 €	09.12.2019	20.000,00	X		
Gerència d'Ecologia Urbana	16005605	Serveis	Serveis de recollida i trasllat d'animals de companyia perduts o abandonats a la ciutat de Barcelona al Centre d'Acolliment d'Animals de Companyia de Barcelona (CAACB),	06.10.2017	323.000,00 €	16.12.2019	32.300,00	X		10,00%
Gerència d'Ecologia Urbana	14003093	Serveis	Manteniment del clavegueram de Barcelona (2015-2022)	16.02.2015	103.512.021,28 €	23.12.2019	54.132,10	X		0,05%
Gerència d'Ecologia Urbana	17006454	Serveis	Servei de Manteniment d'Escales Mecàniques, Ascensors Verticals i Ascensors Inclinat 2018-2020	13.07.2018	3.384.216,51 €	23.12.2019	100.000,00	X		2,95%

Inconsistent formatting

Problem description

31/01/18	249.853,60	15/03/19	0,00	Sl		0,00
19/02/19	11.021.334,34	NO CONSTA	-642.911,17	Sl		-5,83
02/04/19	18.876,00	NO CONSTA	0,00	0		0,00
26/08/19	164.424,86	10/12/19	12.829,33	Sl		7,80
31/7/19	503.536,10	17/10/19	177.706,79	justament anualitats		35%
30/10/19	332.000,00	19/11/19	6916,66	justament anualitats		2%
5/12/19	85.498,60	19/12/19	9406,88	justament anualitats		11%
08.01.2019	120.028,50 €	27.08.2019	24.005,70	X		20,00%
14.12.2019	775.473,85 €	31.10.2019	21.000,00	X		2,71%
05.04.2019	73.928,79 €	25.10.2019	20.500,00	X		27,73%
05.12.2016	17.374.923,70 €	09.12.2019	40.000,00	X		
05.12.2016	17.641.863,60 €	09.12.2019	160.000,00	X		
05.12.2016	21.083.212,70 €	09.12.2019	20.000,00	X		
06.10.2017	323.000,00 €	16.12.2019	32.300,00	X		10,00%
16.02.2015	103.512.021,28 €	23.12.2019	54.132,10	X		0,05%
13.07.2018	3.384.216,51 €	23.12.2019	100.000,00	X		2,95%

Inconsistent formatting

Remediation

- Do an EDA on the whole set of files
- Select one of the files as the golden standard (usually the most recent year)
- Generate an expectation suite for the golden standard
- Validate against the golden standard
- Validate the expectation suite against previous years, remediating inconsistencies when needed

Non tidy data

Problem description

022.csv

INVITADOS A PRESENTAR OFERTA

(A79054748) TECNICOS ASOCIADOS INFORMATICA, S.A. PYME: SI (B83628339) REDCOM CIBERNÉTICO, S.L.. PYME: SI (B86061249) ZENER SOFTCONSULTING S.L.. PYME: SI

(A79054748) TECNICOS ASOCIADOS INFORMATICA, S.A. PYME: SI (B86061249) ZENER SOFTCONSULTING S.L.. PYME: NO (B98198286) INFORMÁTICA FAER, S.L.. PYME: NO

(B85635910) EDICIONES EL PAIS, S.L.. PYME: NO

(A79102331) UNIDAD EDITORIAL; S.A.. PYME: NO

(B81511834) DIARIO AS; S.L.. PYME: NO

(A82031329) AUDIOVISUAL ESPAÑOLA 2000, S.A.. PYME: NO

(B82824194) DIARIO ABC S.L.. PYME: NO

(00837947B) ALBERTO VIDAL SILGADO. PYME: SI

(A28760692) VALORIZA SERVICIOS MEDIOAMBIENTALES, S.A.. PYME: NO

(B85653350) LEGAL PLANNING, SL. PYME: SI

(B84943091) FALCONERS IBERIA SL. PYME: SI

(A78510963) FORMULARIOS EUROPEOS, S.A.. PYME: SI

(B61220521) SALVETTI LLOMBART, S.L.. PYME: SI (B82095522) INFORMATION RESOURCES ESPAÑA, S.L.. PYME: SI (G08557985) ASOCIACION ESPAÑOLA DE CODIFICACIONCOME

(A28057529) FERROVIAL S.A.. PYME: NO (A28649911) INLLAMA S A. PYME: SI (B28986511) CIMA EXTINTORES. PYME: SI (B82205170) NORMA FIRE, S.L.. PYME: SI (B84194034) I

(A81286759) SIADDE SOLUCIONES S.A.. PYME: SI (B40165920) INV SISTEMAS Y SOLUCIONES DE SEGURIDAD, S. L.. PYME: SI (B63809560) WALLNER EUROPA S.L.. PYME: SI (B8

Tidy Data

Published by Hadley Wickman (2014)

- Every column is a variable
- Every row is an observation
- Every cell is a single value

Non tidy data

Remediation

- Create an expectation suite using regex to match the tidy data version
- Split multiple records per cell into individual columns
- Split multiples values per cell into individual values
- Validate the file against the expectation suite

Alternative approaches for unfixable issues in data

Divide and conquer / Reduce scope

- Required data might not be available over the years. Two ways to approach it:
 - Divide and conquer, splitting the dataset in parts and remediating them individually
 - Reduce scope, targeting only the data available over the years
- Notify data publisher about your requirements, intended usage (if possible)

Q&A time!

How to check the batch request contents

- https://docs.greatexpectations.io/docs/guides/connecting_to_your_data/fluent/data_assets/how_to_organize_batches_in_a_file_based_data_asset#use-a-batch-request-to-verify-the-data-asset-works-as-desired

How to retrieve the failed rows (up to 10k)

- https://docs.greatexpectations.io/docs/guides/expectations/advanced/identify_failed_rows_expectations/

Thanks for your attention!

César García Sáez - @lahoramaker - cesar@lahoramaker.com

This presentation is available under a CC-BY-SA 4.0 International license