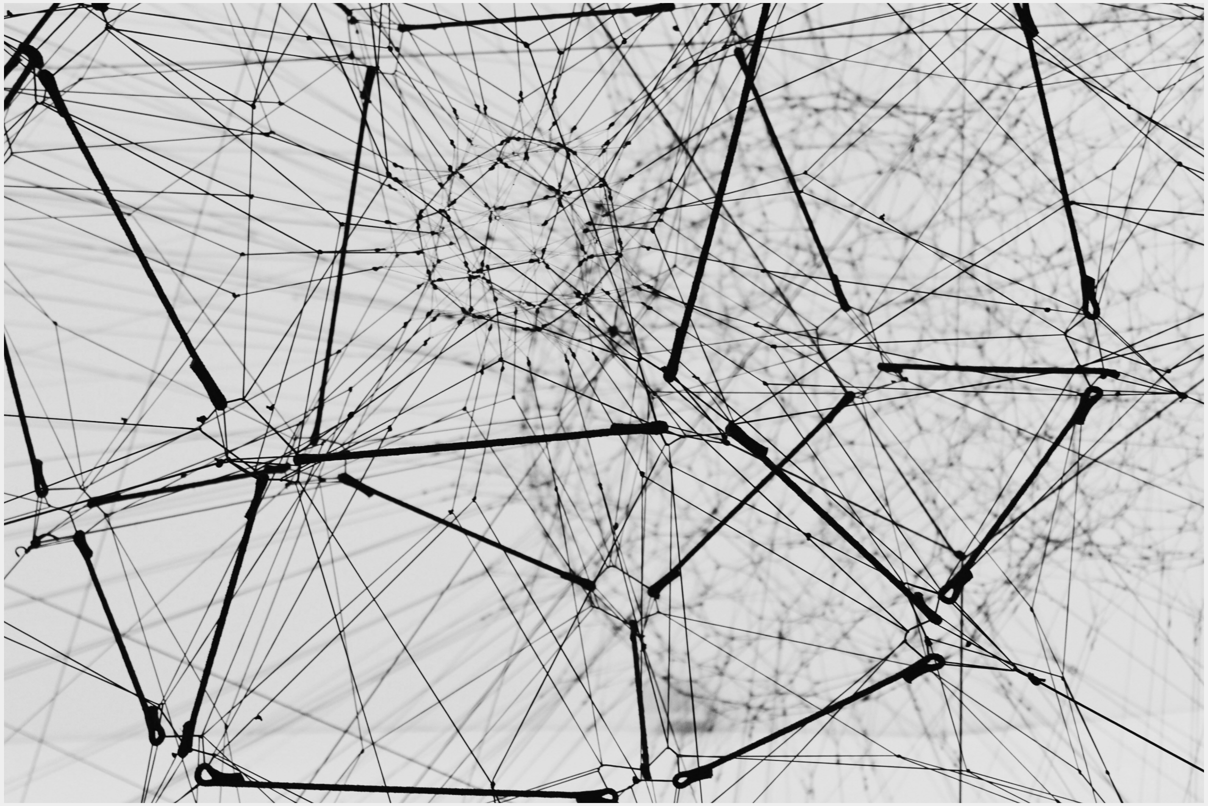
National College of Ireland
Cloud Competency Centre

Based on: González–Cebrián, A., Bradford, M., Chis, A.E., González–Vélez, H. Standardised Versioning of Datasets: a FAIR–compliant Proposal. **Scientific Data** 11, 358 (2024). https://doi.org/10.1038/s41597-024-03153-y

NCI Research Day, 18 June 2024

FAIR Versioning of Datasets



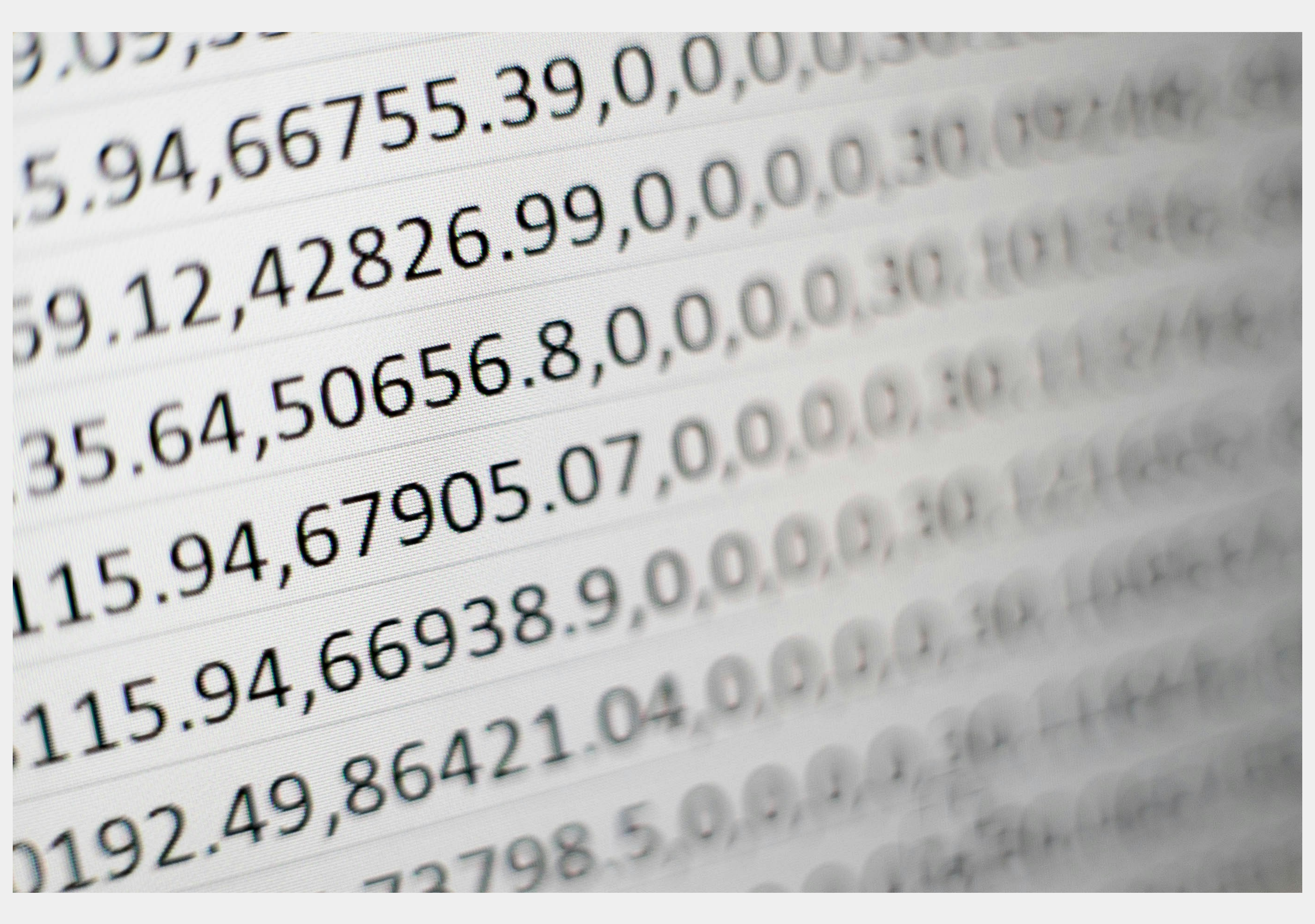
1. Background

National College of Ireland
Cloud Competency Centre

“

When is a dataset NOT the same?

Background


5.94,66755.39,0,0,0,0
59.12,42826.99,0,0,0,0,0,30
35.64,50656.8,0,0,0,0,0
115.94,67905.07,0,0,0,0
115.94,66938.9,0,0,0,0
0192.49,86421.04,0
73798.5,0

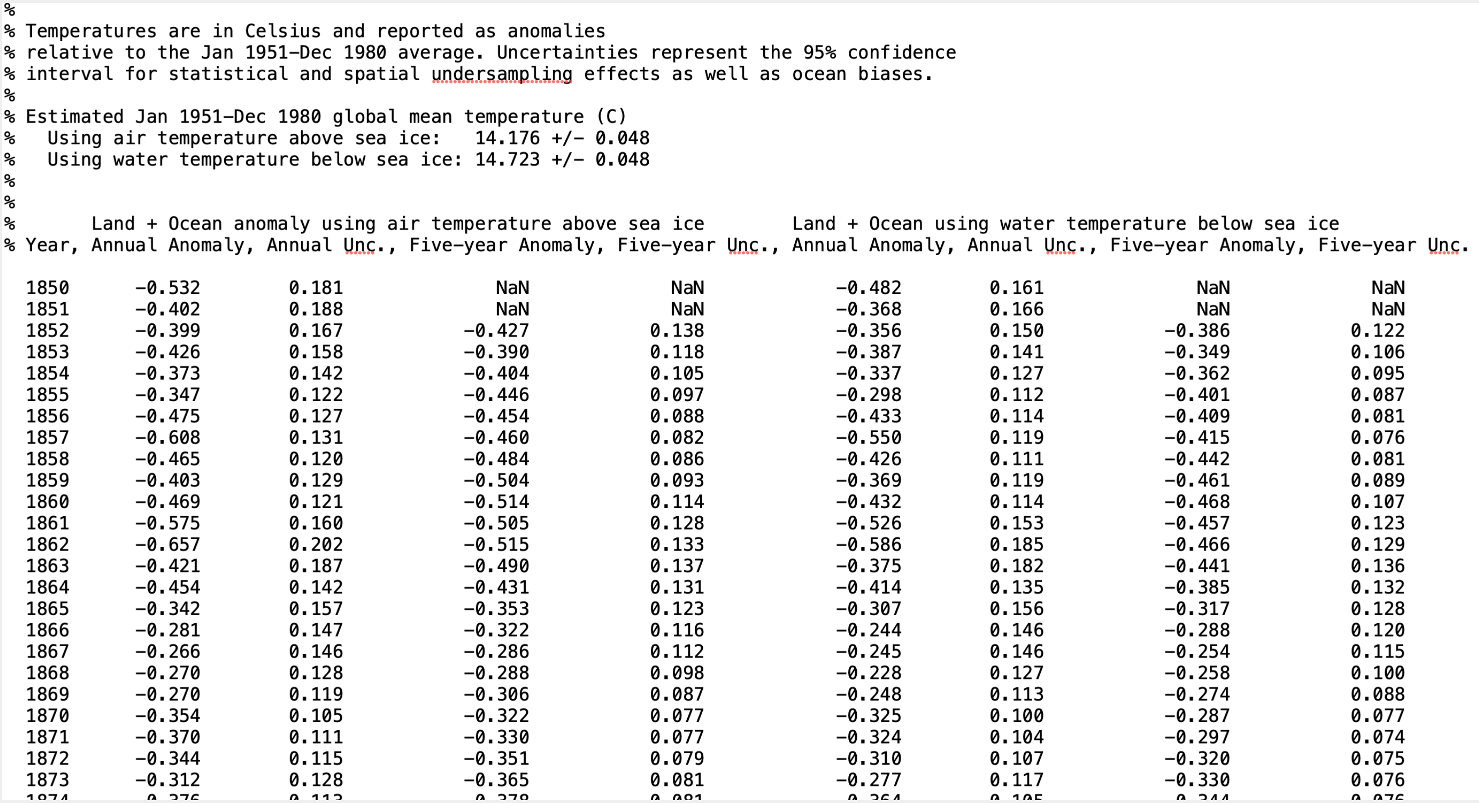
Background

a change of:

a record,
an entry,
a value, and/or
a column?

Example

Rhode R & Hausfather Z. **Berkeley Earth** Combined Land and Ocean Temperature Field, **Jan 1850-Nov 2019** https://zenodo.org/records/3634713 **Changes in Earth's global average surface temperature** estimated by combining the Berkeley Earth land-surface temperature field with a reinterpolated version of the HadSST ocean temperature field.



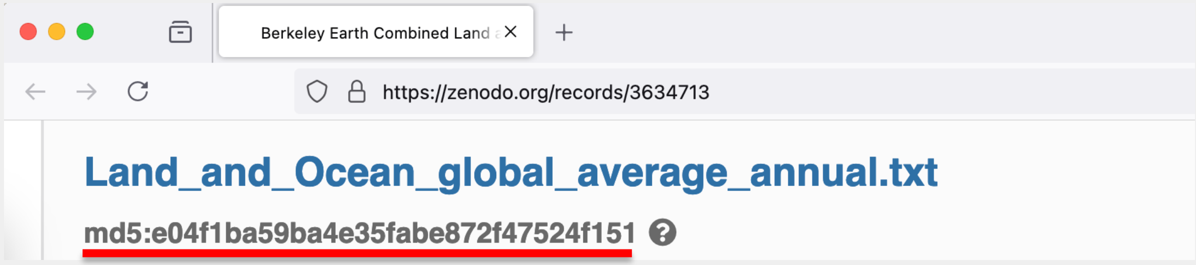
```
%
% Temperatures are in Celsius and reported as anomalies
% relative to the Jan 1951-Dec 1980 average. Uncertainties represent the 95% confidence
% interval for statistical and spatial undersampling effects as well as ocean biases.
%
% Estimated Jan 1951-Dec 1980 global mean temperature (C)
%   Using air temperature above sea ice:   14.176 +/- 0.048
%   Using water temperature below sea ice: 14.723 +/- 0.048
%
%
%       Land + Ocean anomaly using air temperature above sea ice        Land + Ocean using water temperature below sea ice
% Year, Annual Anomaly, Annual Unc., Five-year Anomaly, Five-year Unc., Annual Anomaly, Annual Unc., Five-year Anomaly, Five-year Unc.

  1850      -0.532        0.181              NaN             NaN            -0.482        0.161              NaN             NaN
  1851      -0.402        0.188              NaN             NaN            -0.368        0.166              NaN             NaN
  1852      -0.399        0.167           -0.427           0.138            -0.356        0.150           -0.386           0.122
  1853      -0.426        0.158           -0.390           0.118            -0.387        0.141           -0.349           0.106
  1854      -0.373        0.142           -0.404           0.105            -0.337        0.127           -0.362           0.095
  1855      -0.347        0.122           -0.446           0.097            -0.298        0.112           -0.401           0.087
  1856      -0.475        0.127           -0.454           0.088            -0.433        0.114           -0.409           0.081
  1857      -0.608        0.131           -0.460           0.082            -0.550        0.119           -0.415           0.076
  1858      -0.465        0.120           -0.484           0.086            -0.426        0.111           -0.442           0.081
  1859      -0.403        0.129           -0.504           0.093            -0.369        0.119           -0.461           0.089
  1860      -0.469        0.121           -0.514           0.114            -0.432        0.114           -0.468           0.107
  1861      -0.575        0.160           -0.505           0.128            -0.526        0.153           -0.457           0.123
  1862      -0.657        0.202           -0.515           0.133            -0.586        0.185           -0.466           0.129
  1863      -0.421        0.187           -0.490           0.137            -0.375        0.182           -0.441           0.136
  1864      -0.454        0.142           -0.431           0.131            -0.414        0.135           -0.385           0.132
  1865      -0.342        0.157           -0.353           0.123            -0.307        0.156           -0.317           0.128
  1866      -0.281        0.147           -0.322           0.116            -0.244        0.146           -0.288           0.120
  1867      -0.266        0.146           -0.286           0.112            -0.245        0.146           -0.254           0.115
  1868      -0.270        0.128           -0.288           0.098            -0.228        0.127           -0.258           0.100
  1869      -0.270        0.119           -0.306           0.087            -0.248        0.113           -0.274           0.088
  1870      -0.354        0.105           -0.322           0.077            -0.325        0.100           -0.287           0.077
  1871      -0.370        0.111           -0.330           0.077            -0.324        0.104           -0.297           0.074
  1872      -0.344        0.115           -0.351           0.079            -0.310        0.107           -0.320           0.075
  1873      -0.312        0.128           -0.365           0.081            -0.277        0.117           -0.330           0.076
  1874      -0.376        0.113           -0.378           0.081            -0.364        0.105           -0.344           0.076
```

Background

Example

Rhode R & Hausfather Z. **Berkeley Earth** Combined Land and Ocean Temperature Field, **Jan 1850-Nov 2019** https://zenodo.org/records/3634713 **Changes in Earth's global average surface temperature** estimated by combining the Berkeley Earth land-surface temperature field with a reinterpolated version of the HadSST ocean temperature field.

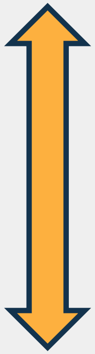


Berkeley Earth Combined Land

https://zenodo.org/records/3634713

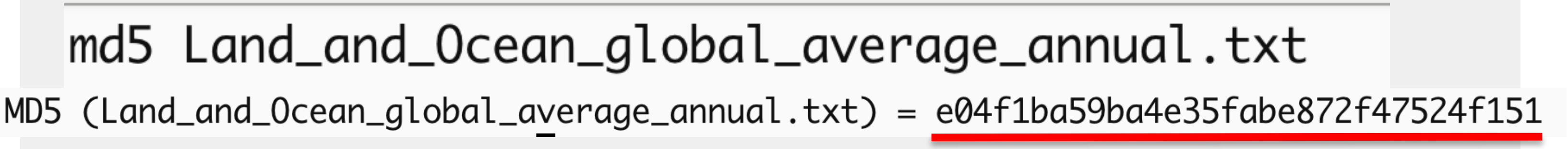
Land_and_Ocean_global_average_annual.txt

md5:e04f1ba59ba4e35fabe872f47524f151



Original dataset at Zenodo

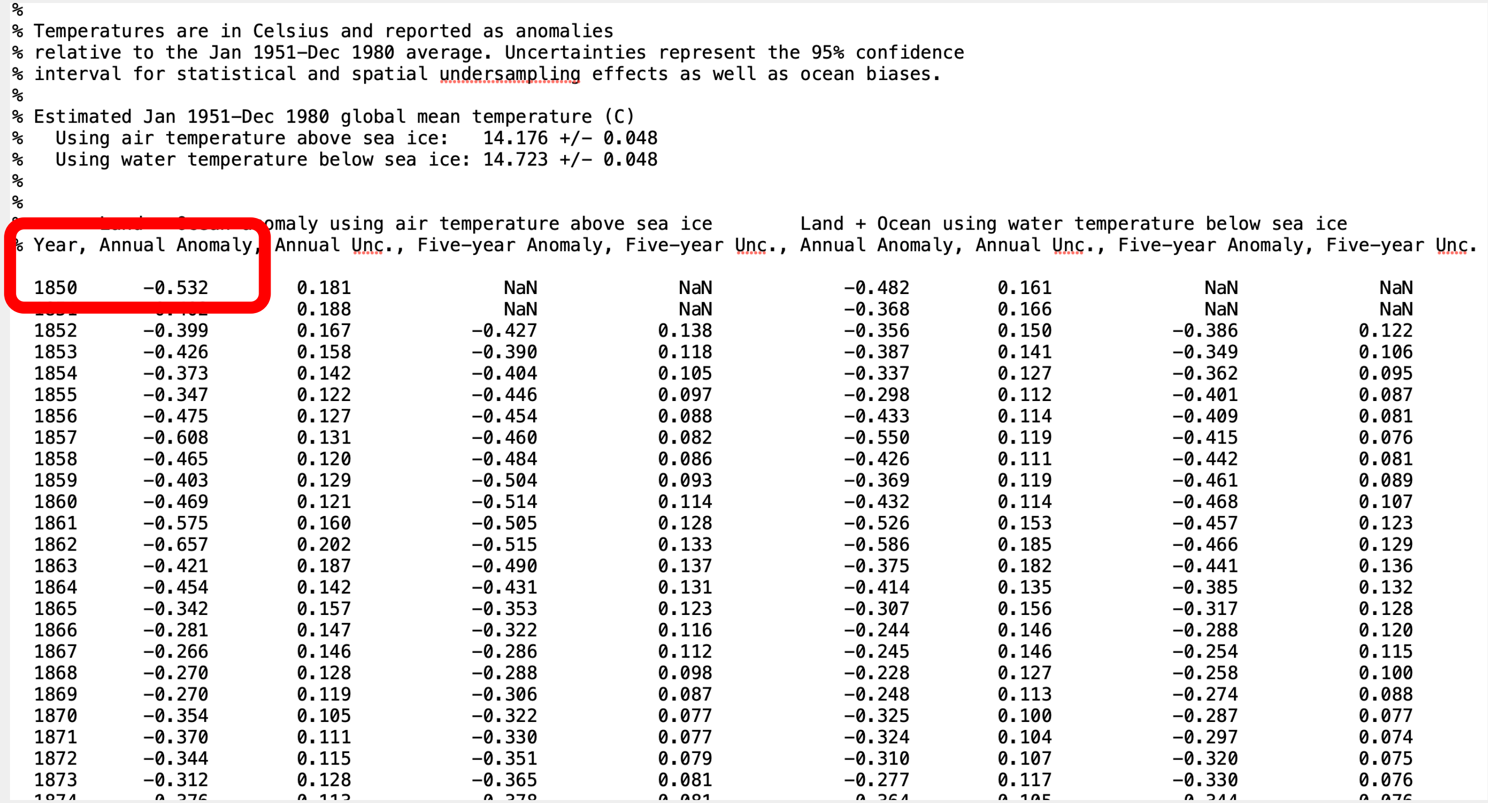
Local copy



```
md5 Land_and_Ocean_global_average_annual.txt
MD5 (Land_and_Ocean_global_average_annual.txt) = e04f1ba59ba4e35fabe872f47524f151
```

Example

Rhode R & Hausfather Z. **Berkeley Earth** Combined Land and Ocean Temperature Field, **Jan 1850-Nov 2019**
https://zenodo.org/records/3634713 **Changes in Earth's global average surface temperature** estimated by combining the Berkeley Earth land-surface temperature field with a reinterpolated version of the HadSST ocean temperature field.



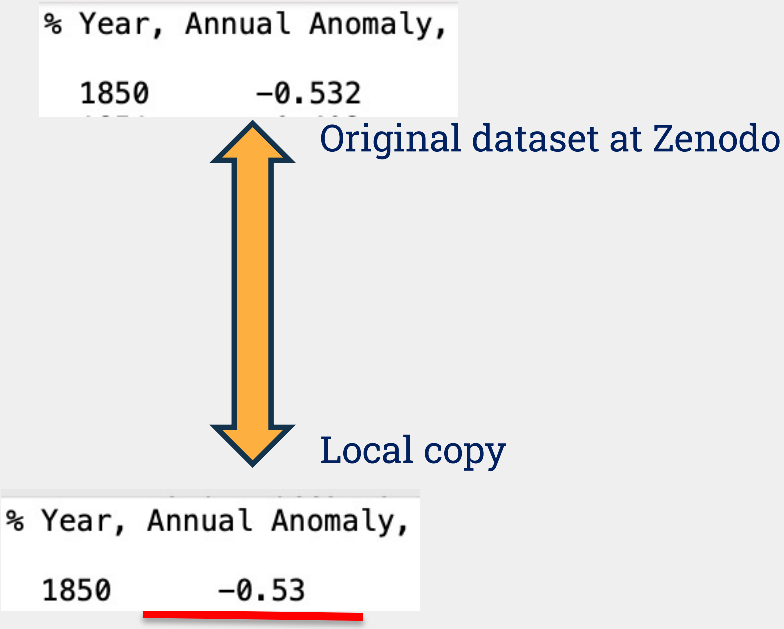
```
%
% Temperatures are in Celsius and reported as anomalies
% relative to the Jan 1951-Dec 1980 average. Uncertainties represent the 95% confidence
% interval for statistical and spatial undersampling effects as well as ocean biases.
%
% Estimated Jan 1951-Dec 1980 global mean temperature (C)
%   Using air temperature above sea ice:   14.176 +/- 0.048
%   Using water temperature below sea ice: 14.723 +/- 0.048
%
%
%      Land + Ocean anomaly using air temperature above sea ice        Land + Ocean using water temperature below sea ice
% Year, Annual Anomaly, Annual Unc., Five-year Anomaly, Five-year Unc., Annual Anomaly, Annual Unc., Five-year Anomaly, Five-year Unc.

  1850       -0.532        0.181              NaN             NaN            -0.482        0.161              NaN             NaN
  1851       -0.402        0.188              NaN             NaN            -0.368        0.166              NaN             NaN
  1852       -0.399        0.167           -0.427           0.138            -0.356        0.150           -0.386           0.122
  1853       -0.426        0.158           -0.390           0.118            -0.387        0.141           -0.349           0.106
  1854       -0.373        0.142           -0.404           0.105            -0.337        0.127           -0.362           0.095
  1855       -0.347        0.122           -0.446           0.097            -0.298        0.112           -0.401           0.087
  1856       -0.475        0.127           -0.454           0.088            -0.433        0.114           -0.409           0.081
  1857       -0.608        0.131           -0.460           0.082            -0.550        0.119           -0.415           0.076
  1858       -0.465        0.120           -0.484           0.086            -0.426        0.111           -0.442           0.081
  1859       -0.403        0.129           -0.504           0.093            -0.369        0.119           -0.461           0.089
  1860       -0.469        0.121           -0.514           0.114            -0.432        0.114           -0.468           0.107
  1861       -0.575        0.160           -0.505           0.128            -0.526        0.153           -0.457           0.123
  1862       -0.657        0.202           -0.515           0.133            -0.586        0.185           -0.466           0.129
  1863       -0.421        0.187           -0.490           0.137            -0.375        0.182           -0.441           0.136
  1864       -0.454        0.142           -0.431           0.131            -0.414        0.135           -0.385           0.132
  1865       -0.342        0.157           -0.353           0.123            -0.307        0.156           -0.317           0.128
  1866       -0.281        0.147           -0.322           0.116            -0.244        0.146           -0.288           0.120
  1867       -0.266        0.146           -0.286           0.112            -0.245        0.146           -0.254           0.115
  1868       -0.270        0.128           -0.288           0.098            -0.228        0.127           -0.258           0.100
  1869       -0.270        0.119           -0.306           0.087            -0.248        0.113           -0.274           0.088
  1870       -0.354        0.105           -0.322           0.077            -0.325        0.100           -0.287           0.077
  1871       -0.370        0.111           -0.330           0.077            -0.324        0.104           -0.297           0.074
  1872       -0.344        0.115           -0.351           0.079            -0.310        0.107           -0.320           0.075
  1873       -0.312        0.128           -0.365           0.081            -0.277        0.117           -0.330           0.076
  1874       -0.376        0.113           -0.378           0.081            -0.364        0.105           -0.344           0.076
```

Background

Example

Rhode R & Hausfather Z. **Berkeley Earth** Combined Land and Ocean Temperature Field, **Jan 1850-Nov 2019** https://zenodo.org/records/3634713 **Changes in Earth's global average surface temperature** estimated by combining the Berkeley Earth land-surface temperature field with a reinterpolated version of the HadSST ocean temperature field.



```
% Year, Annual Anomaly,

  1850        -0.532
```

Original dataset at Zenodo

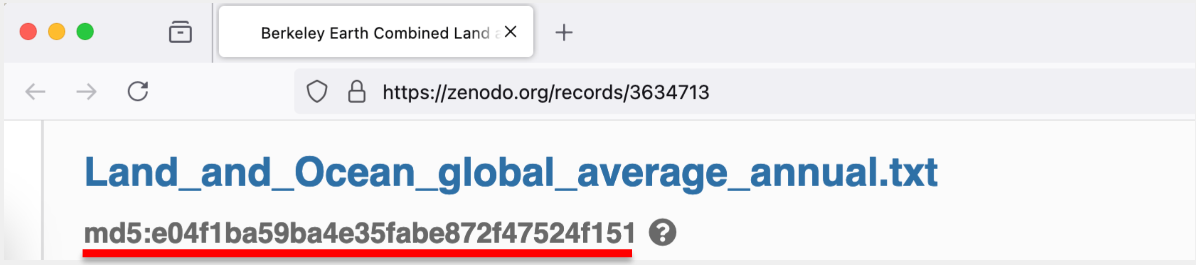
Local copy

```
% Year, Annual Anomaly,

  1850        -0.53
```

Example

Rhode R & Hausfather Z. **Berkeley Earth** Combined Land and Ocean Temperature Field, **Jan 1850-Nov 2019** https://zenodo.org/records/3634713 **Changes in Earth's global average surface temperature** estimated by combining the Berkeley Earth land-surface temperature field with a reinterpolated version of the HadSST ocean temperature field.



https://zenodo.org/records/3634713

Land_and_Ocean_global_average_annual.txt

md5:e04f1ba59ba4e35fabe872f47524f151

Original dataset at Zenodo

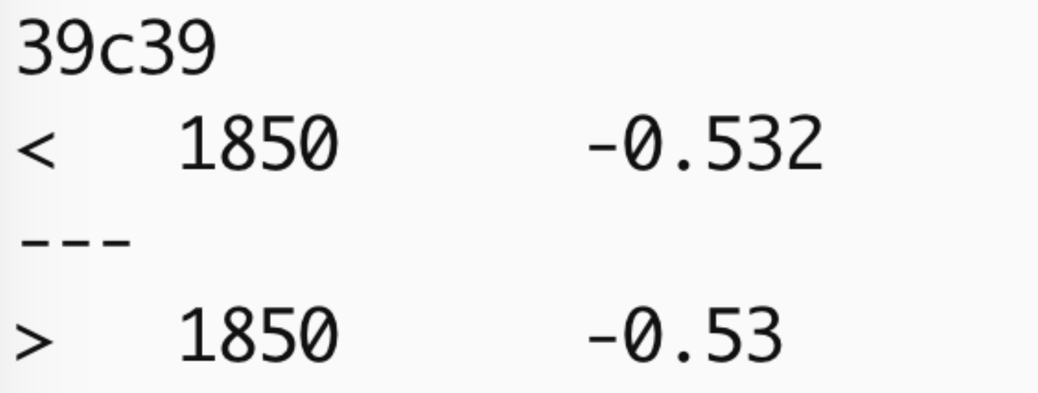
Local copy

```
md5 Land_and_Ocean_global_average_annual_R.txt
MD5 (Land_and_Ocean_global_average_annual_R.txt) = 6badea6b0a85d1fcf54c55010f09320a
```

Example

Rhode R & Hausfather Z. **Berkeley Earth** Combined Land and Ocean Temperature Field, **Jan 1850-Nov 2019** https://zenodo.org/records/3634713 **Changes in Earth's global average surface temperature** estimated by combining the Berkeley Earth land-surface temperature field with a reinterpolated version of the HadSST ocean temperature field.

```
diff Land_and_Ocean_global_average_annual.txt Land_and_Ocean_global_average_annual_R.txt
```



```
39c39
<    1850          -0.532
---
>    1850          -0.53
```

what if:

multiple records, entries, values, and /or columns CHANGE.

*Is it the **SAME** DATASET?*

2. Method

Versioning

Software-style

major.minor.patch

FAIR

Findability

Accessibility

Interoperability

Reusability

Wilkinson M *et al.* The FAIR Guiding Principles for scientific data management and stewardship. ***Scientific Data* 3** (2016).

Method

proposal

Software-style Versioning

major.minor.patch

Version	Meaning	Dataset property – FAIR
Major	significant and substantial changes involving **modifications to the data structure**, schema, or underlying data model	data **interoperability** and **reusability**
Minor	enhancements, additions, or updates that **quantify change the numerical information** but do not significantly disrupt the existing data structure	**data drift** (related to **interoperability** and **reusability**)
Patch	small, specific fixes or corrections applied to the dataset addressing inconsistencies, errors, or bugs **without significantly changing** the data structure or functionality	**timestamping** could improve **findability** according to periods of interest

Method

Version	Meaning	Dataset property – FAIR
Major	significant and substantial changes involving **modifications to the data structure**, schema, or underlying data model	data **interoperability** and **reusability**
Minor	enhancements, additions, or updates that **quantify change the numerical information** but do not significantly disrupt the existing data structure	**data drift** (related to **interoperability** and **reusability**)
Patch	small, specific fixes or corrections applied to the dataset addressing inconsistencies, errors, or bugs **without significantly changing** the data structure or functionality	**timestamping** could improve **findability** according to periods of interest

Method

National College of Ireland
Cloud Competency Centre

Data Drift

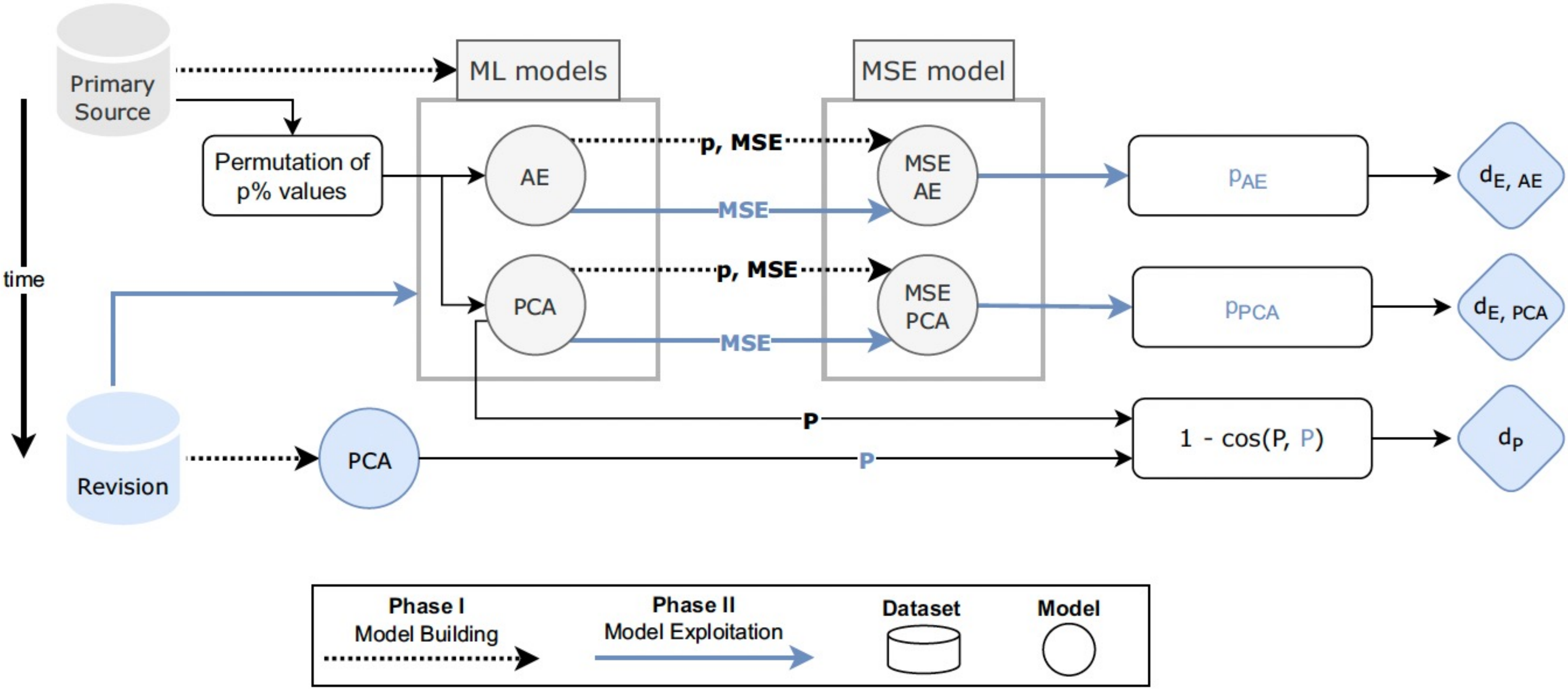
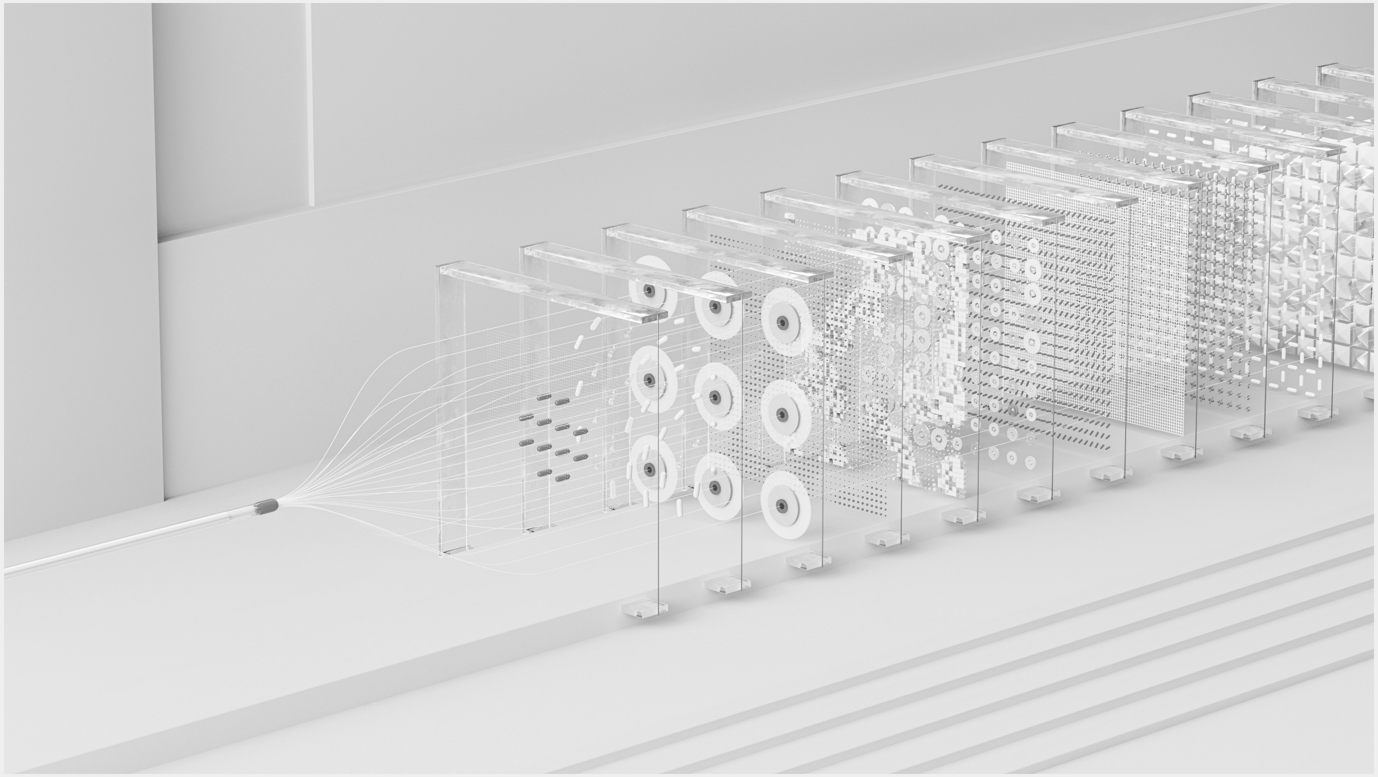


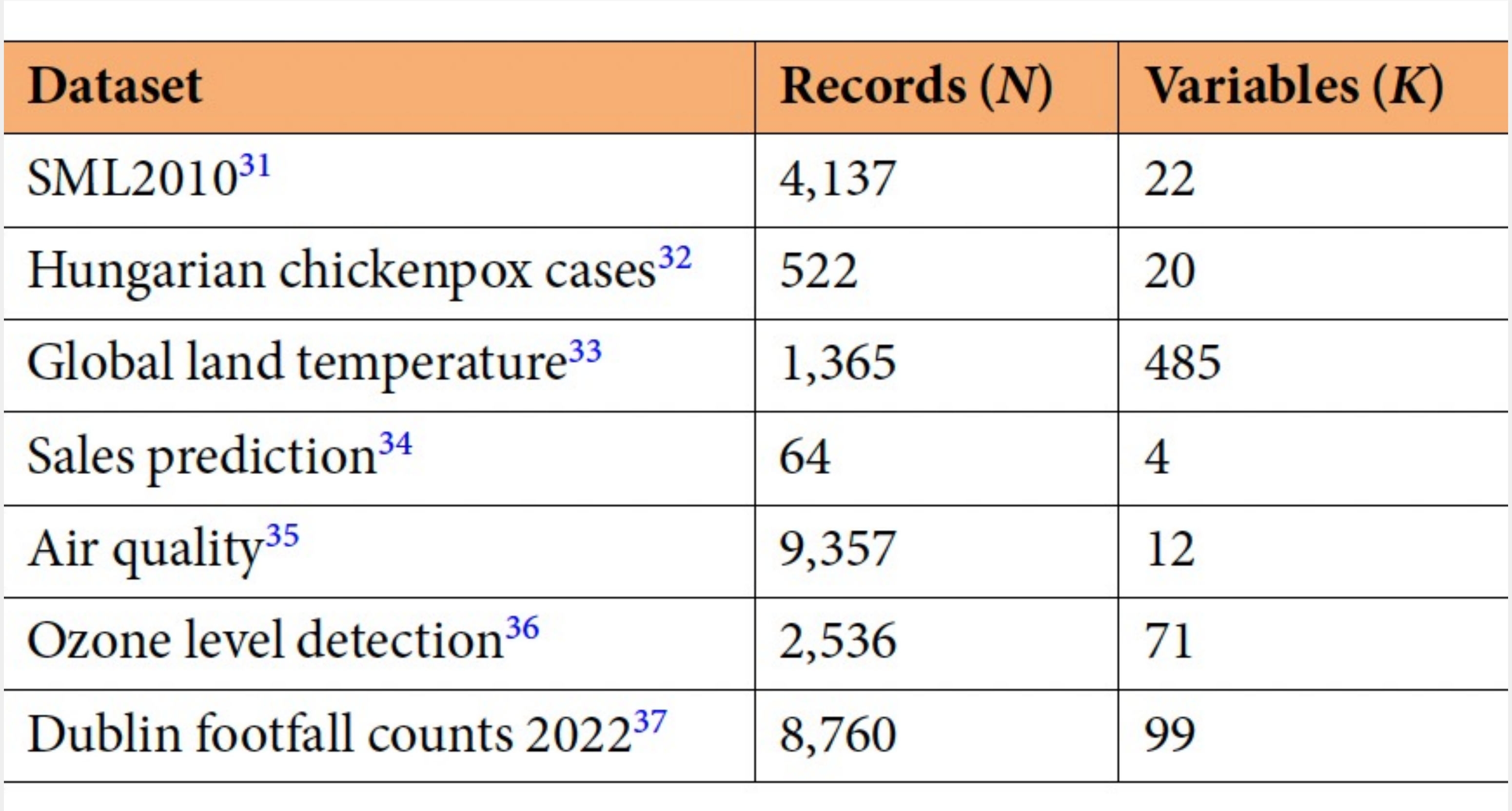
Fig. 1 Flowchart summarising our approach to quantify data drift using several strategies. Each strategy employs an ML model and an associated data drift metric. First, for a Primary Source dataset, we build ML models and predictive models based on the Mean Squared Error (MSE models), conforming the Model Building Phase. Similarly, for a Revision dataset, i.e., a new dataset version, corresponding ML models are built. These models are used during the Model Exploitation Phase to compute the associated data drift metric (e.g., $d_{E,AE}$, $d_{E,PCA}$, and d_P).

Method



3. Results

Datasets

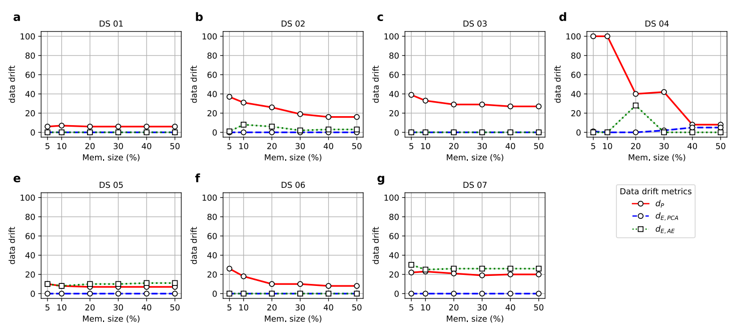


Dataset	Records (*N*)	Variables (*K*)
SML2010[31]	4,137	22
Hungarian chickenpox cases[32]	522	20
Global land temperature[33]	1,365	485
Sales prediction[34]	64	4
Air quality[35]	9,357	12
Ozone level detection[36]	2,536	71
Dublin footfall counts 2022[37]	8,760	99

N.B. Timeseries

Best Metric

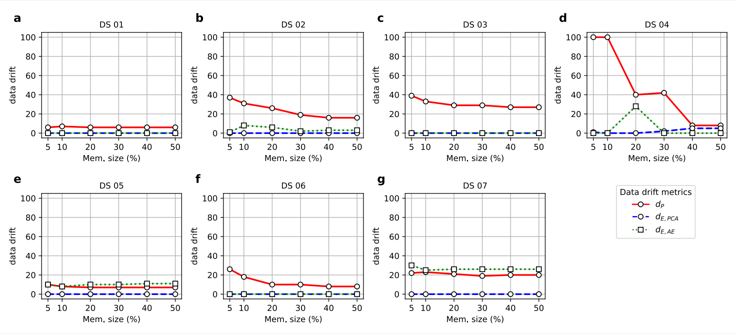
The $d_{E,PCA}$ metric was the best candidate:
Bounded (i.e., $d_{E,PCA} < 50$) and interpretable values in **creation** experiments:
Sensitive to data model changes (i.e., $d_{E,PCA}$ = 100) in **update** experiments
The most robust against the information loss (i.e., $d_{E,PCA} \approx 0$), for **deletion** experiments



Full explanation in the article

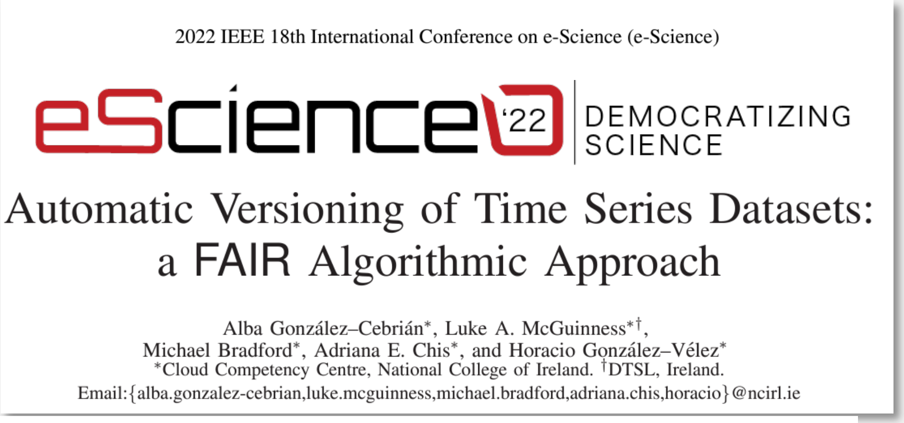
Interpretation

By calculating the data drift, the ML techniques could detect automatically a VERSION of a given datasets with variations of up to **50%** in the contents.



A TurnItIn™ for Datasets

Further Reading



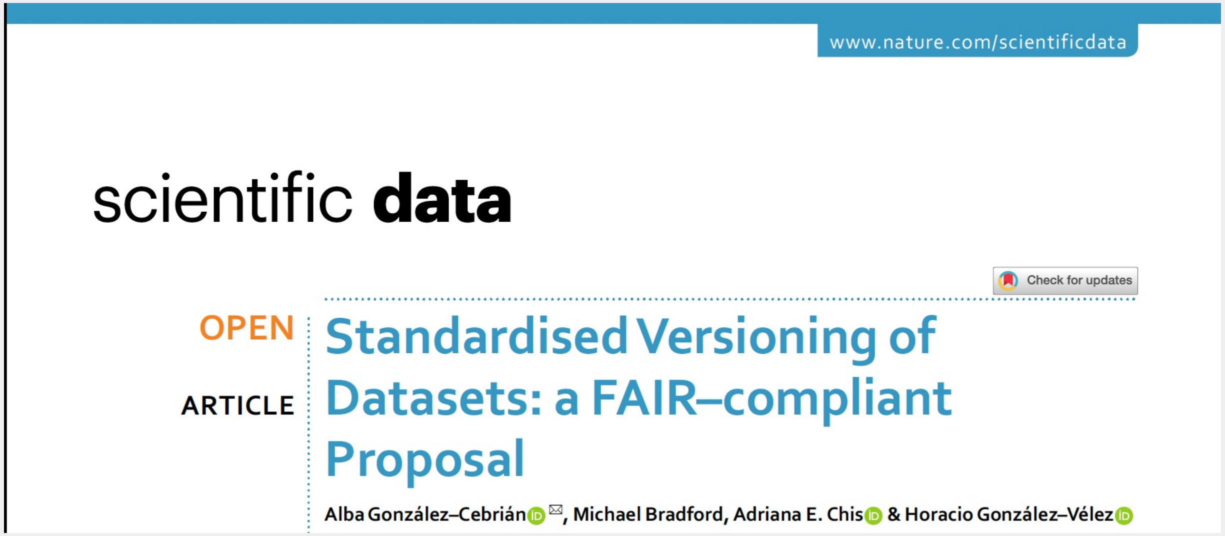
2022 IEEE 18th International Conference on e-Science (e-Science)

eScience '22 | DEMOCRATIZING SCIENCE

Automatic Versioning of Time Series Datasets: a FAIR Algorithmic Approach

Alba González–Cebrián[*], Luke A. McGuinness[*†],
Michael Bradford[*], Adriana E. Chis[*], and Horacio González–Vélez[*]
[*]Cloud Competency Centre, National College of Ireland. [†]DTSL, Ireland.
Email:{alba.gonzalez-cebrian,luke.mcguinness,michael.bradford,adriana.chis,horacio}@ncirl.ie

https://doi.org/10.1109/eScience55777.2022.00034



www.nature.com/scientificdata

scientific **data**

Check for updates

OPEN
ARTICLE

Standardised Versioning of Datasets: a FAIR–compliant Proposal

Alba González–Cebrián ✉, Michael Bradford, Adriana E. Chis & Horacio González–Vélez

https://doi.org/10.1038/s41597-024-03153-y

National College of Ireland
Cloud Competency Centre

Based on: González–Cebrián, A., Bradford, M., Chis, A.E., González–Vélez, H. Standardised Versioning of Datasets: a FAIR–compliant Proposal. **Scientific Data** 11, 358 (2024). https://doi.org/10.1038/s41597-024-03153-y

NCI Research Day, 18 June 2024

FAIR Versioning of Datasets