# DATA ENGINEERING CHALLENGE

Gokul Kannan

Master of Science, Computer Science

# CHALLENGE 1: Skyline Operator

# Skyline Operator : Overview

- **Skyline** query returns a set of points P, such that any point pi = (xi; yi) in P is not **dominated** by any other point in the dataset.
- Different methods to identify Skyline points are
  - Block Nested Loop
  - Divide and Conquer
  - Plane-sweep

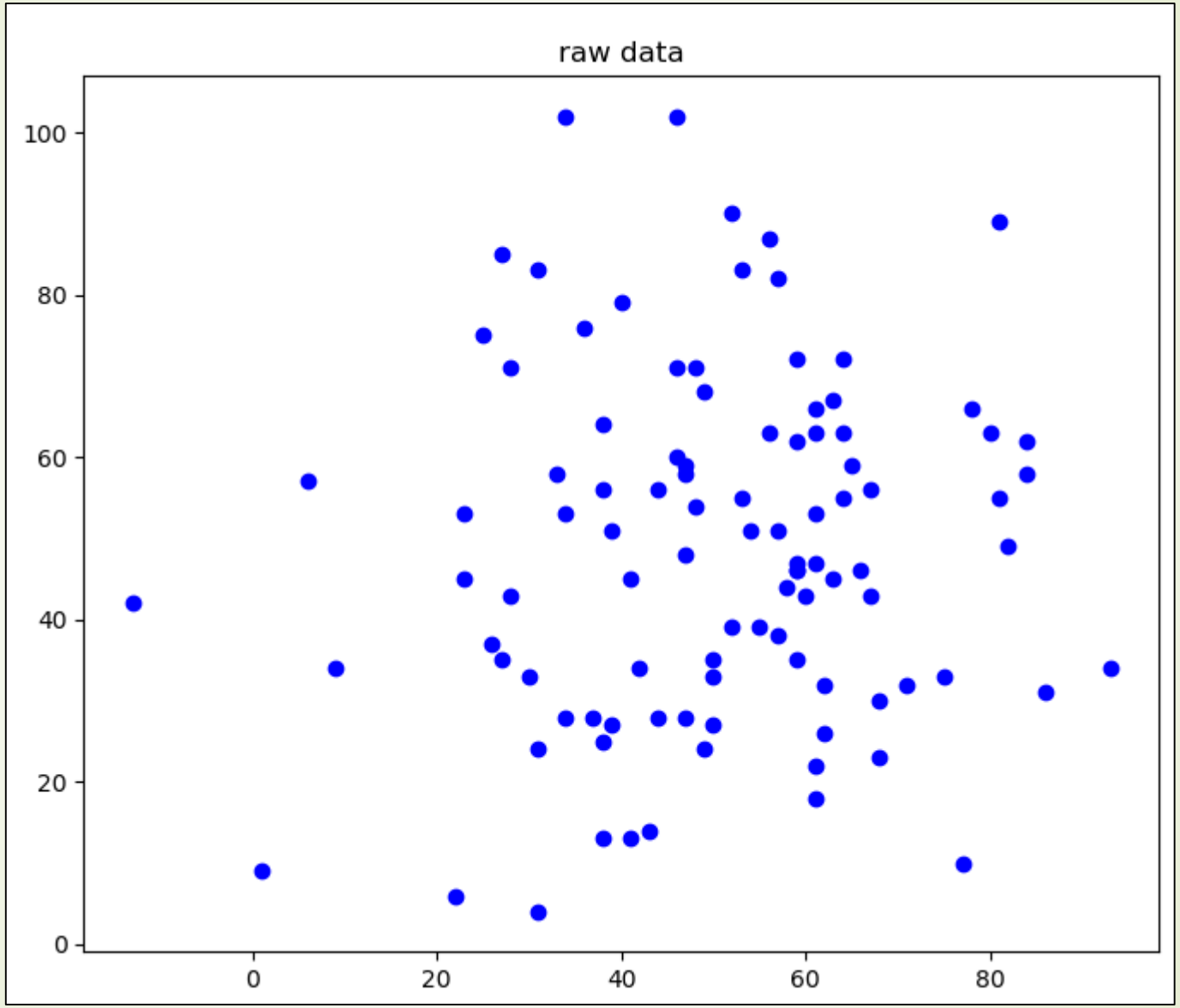
# Skyline Operator : Goal

| | |
|---|---|
| Objective | Identify all the skyline points |
| Input | skylinepoints.csv file<br>A set of two-dimensional points pi = (xi; yi) as input |
| Output | • The set of all skyline points that is not worse than any other point in data set with respect to the dimensions x and y.<br>• The program execution time must be calculated. |
| Assumptions | Higher values are preferred in both dimensions x and y |
| Constraints | 1. Efficiency : Overall program execution time should be less.<br>2. Algorithm shall not be tailored to the provided data set.<br>3. Any Algorithm shall be used to identify the skyline points |
| Programming Language | The algorithm shall be implemented in Java 8 |

# Skyline Operator : Algorithm

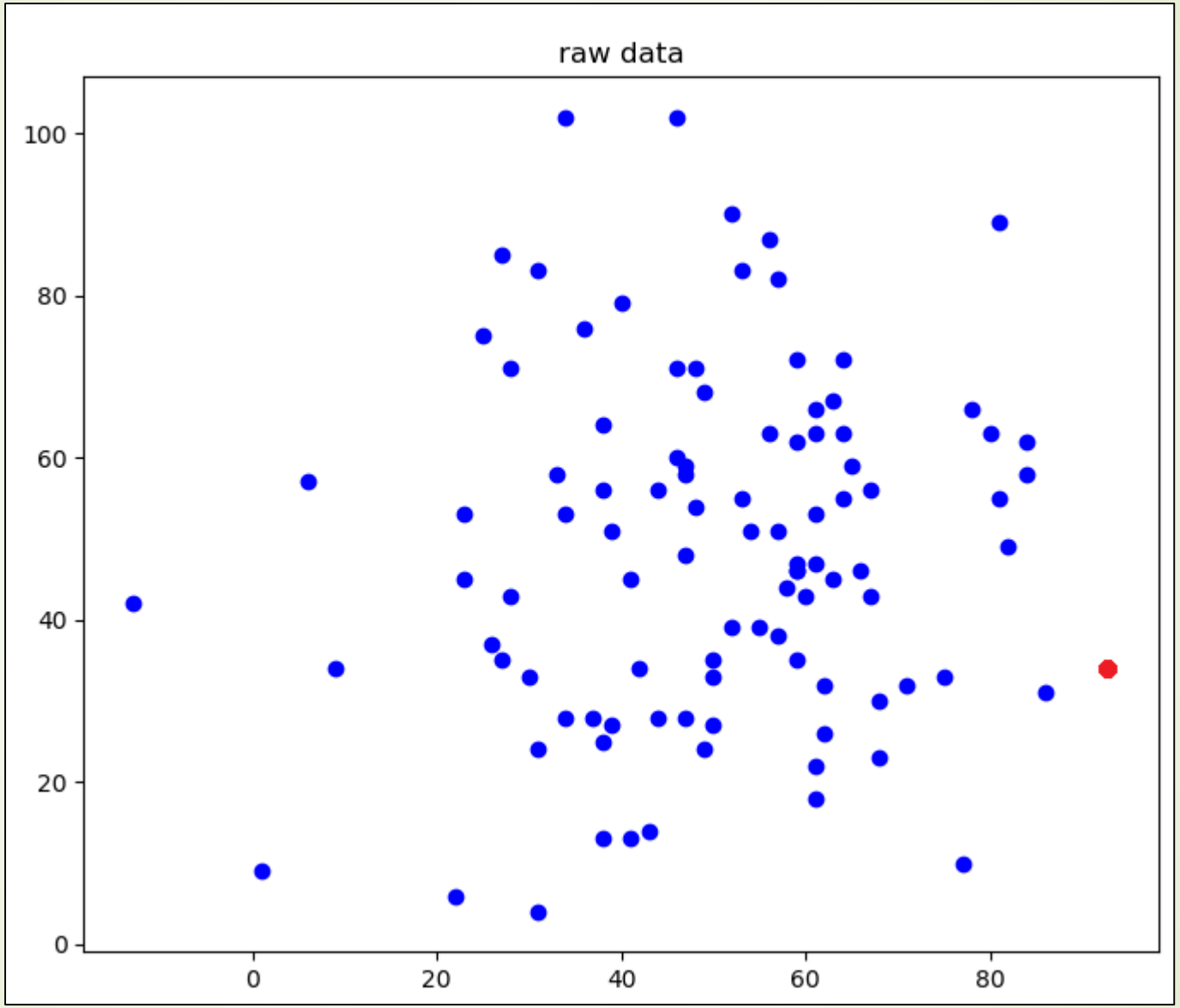
- Sort the input data pair (X, Y) with respect to data point X in **descending order**.
  - If X has same values, sort according to the values of Y.
  - Remove duplicate points
  - The first point in the sorted list is an Skyline point.
  - Store the Y value of this point (Max_Y).
- Iterate through the sorted array.
- If there is an y value in the dataset, which is greater than the stored y value (Max_Y),
  - Add this point as Skyline point.
  - Update the Max_Y value
  - Repeat this process for all the sorted points.

# Skyline Operator : Input Data Points

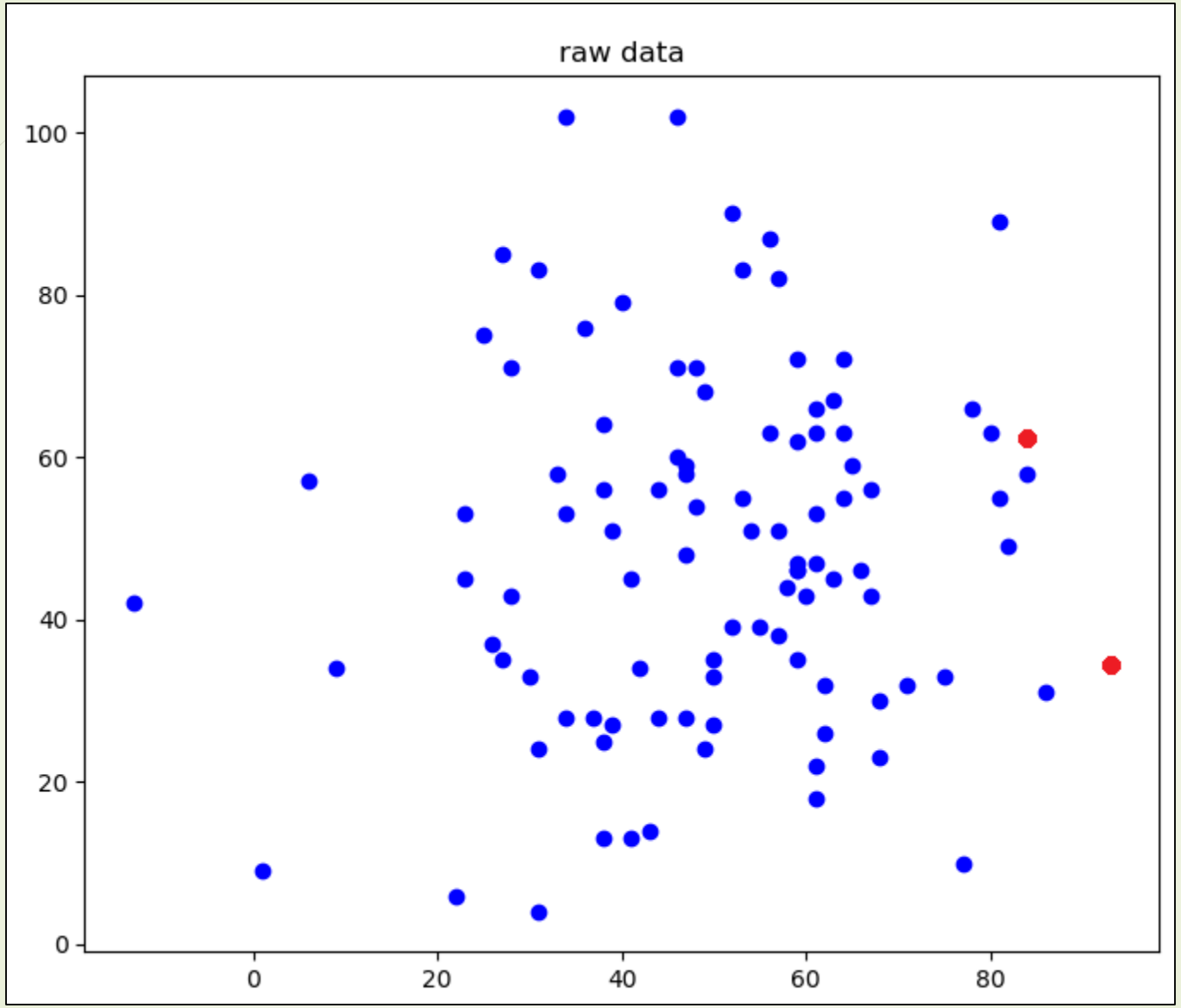


# Skyline Operator : Skyline Point 1

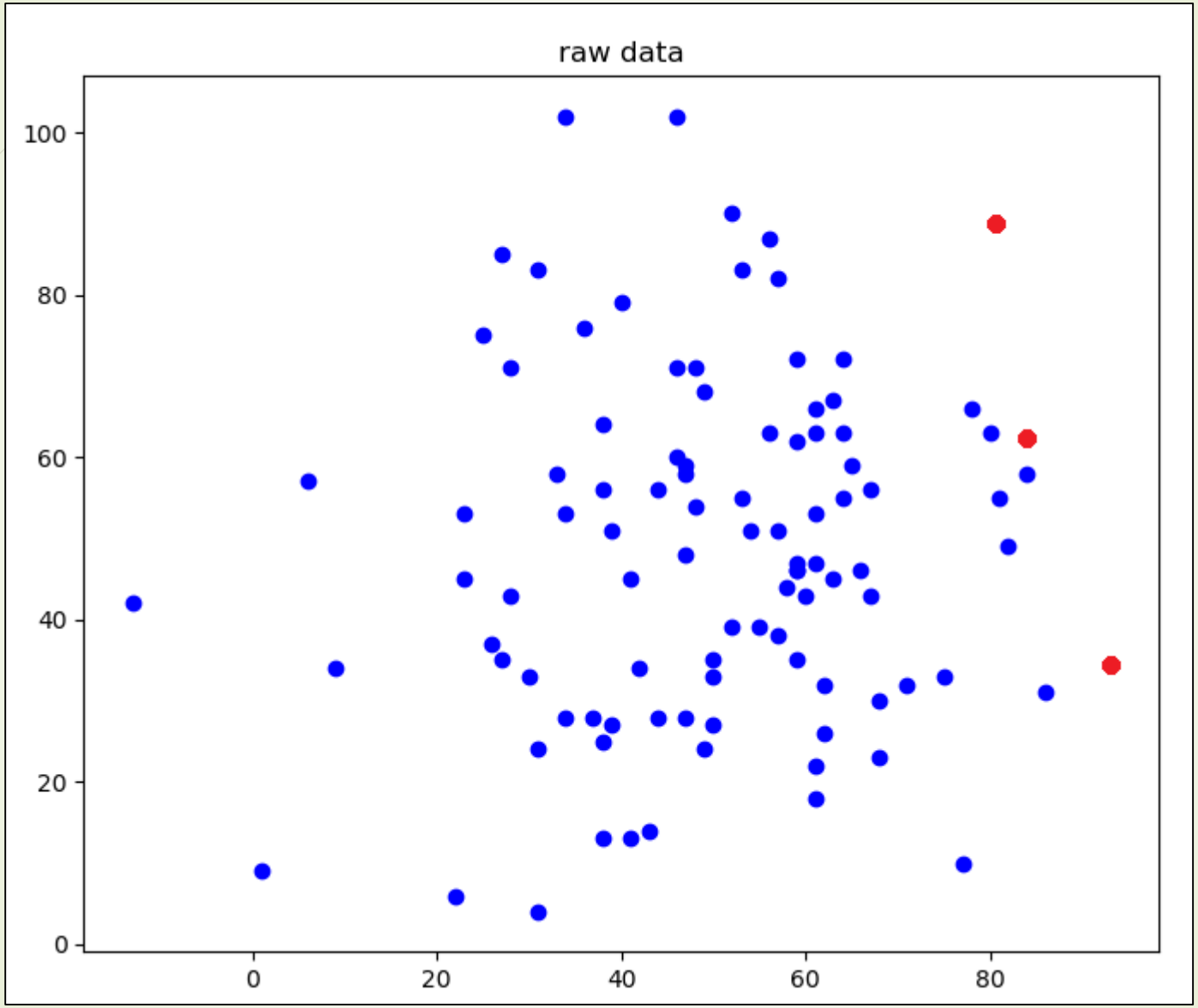




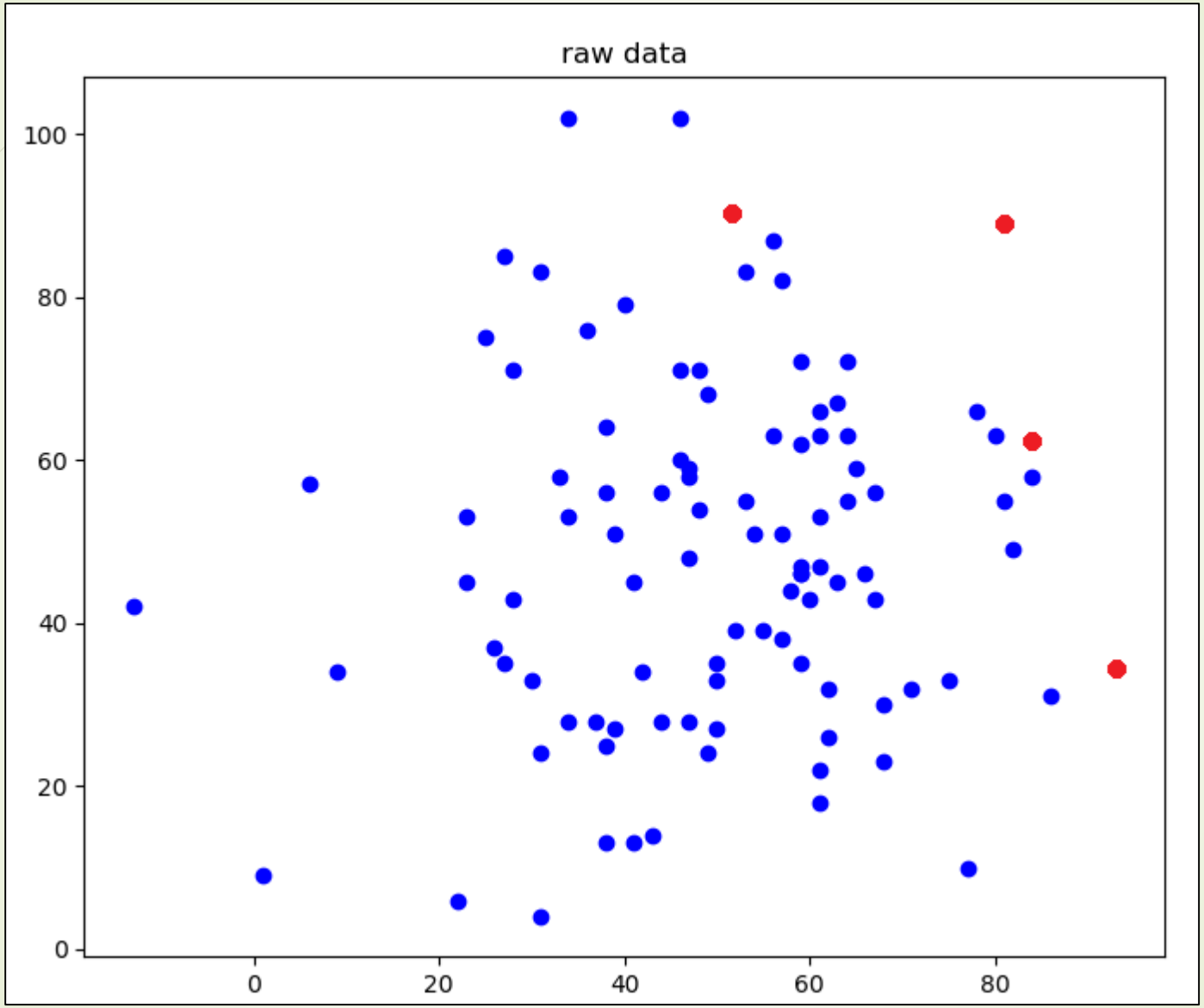
# Skyline Operator : Skyline Point 2



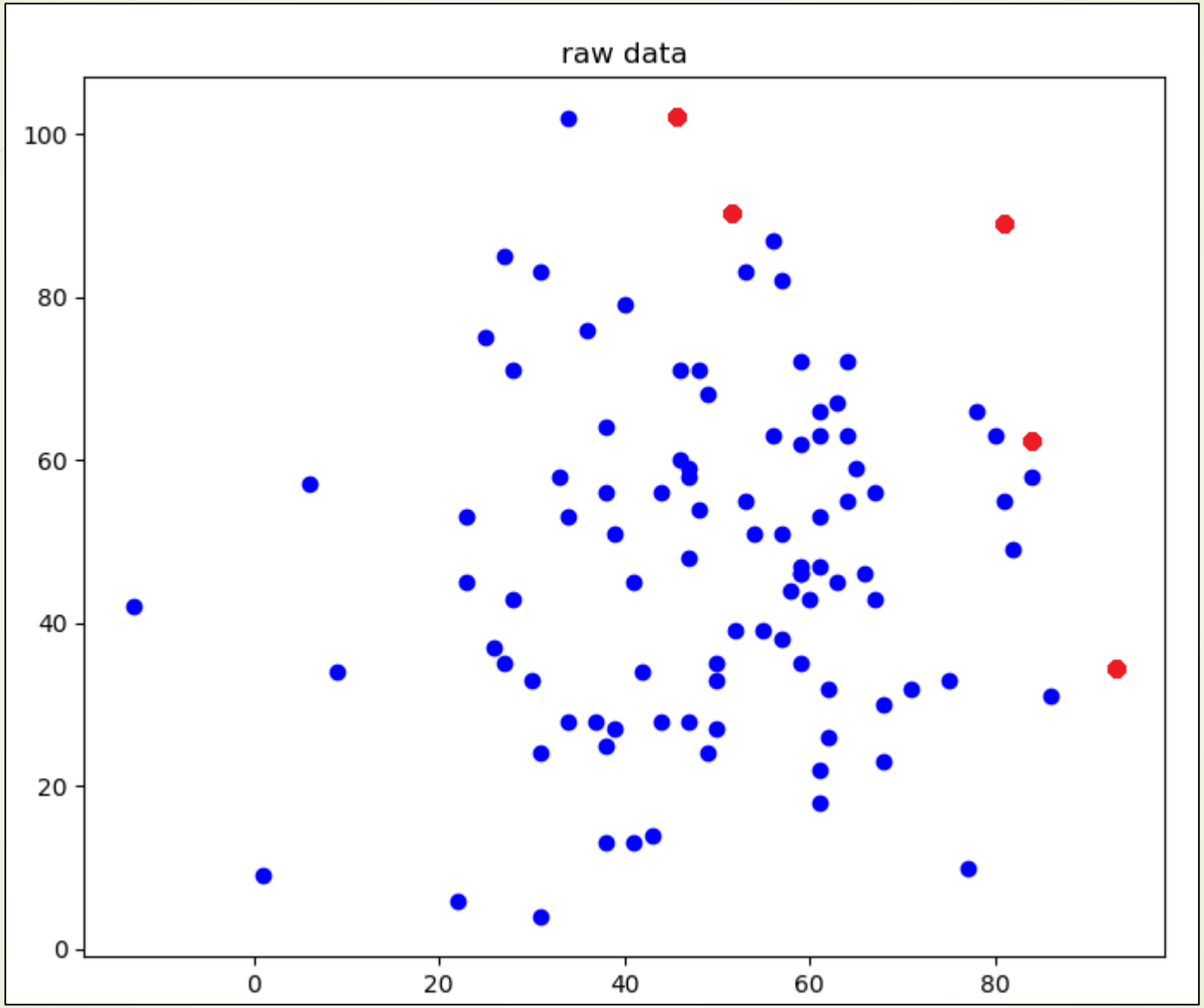
# Skyline Operator : Skyline Point 3



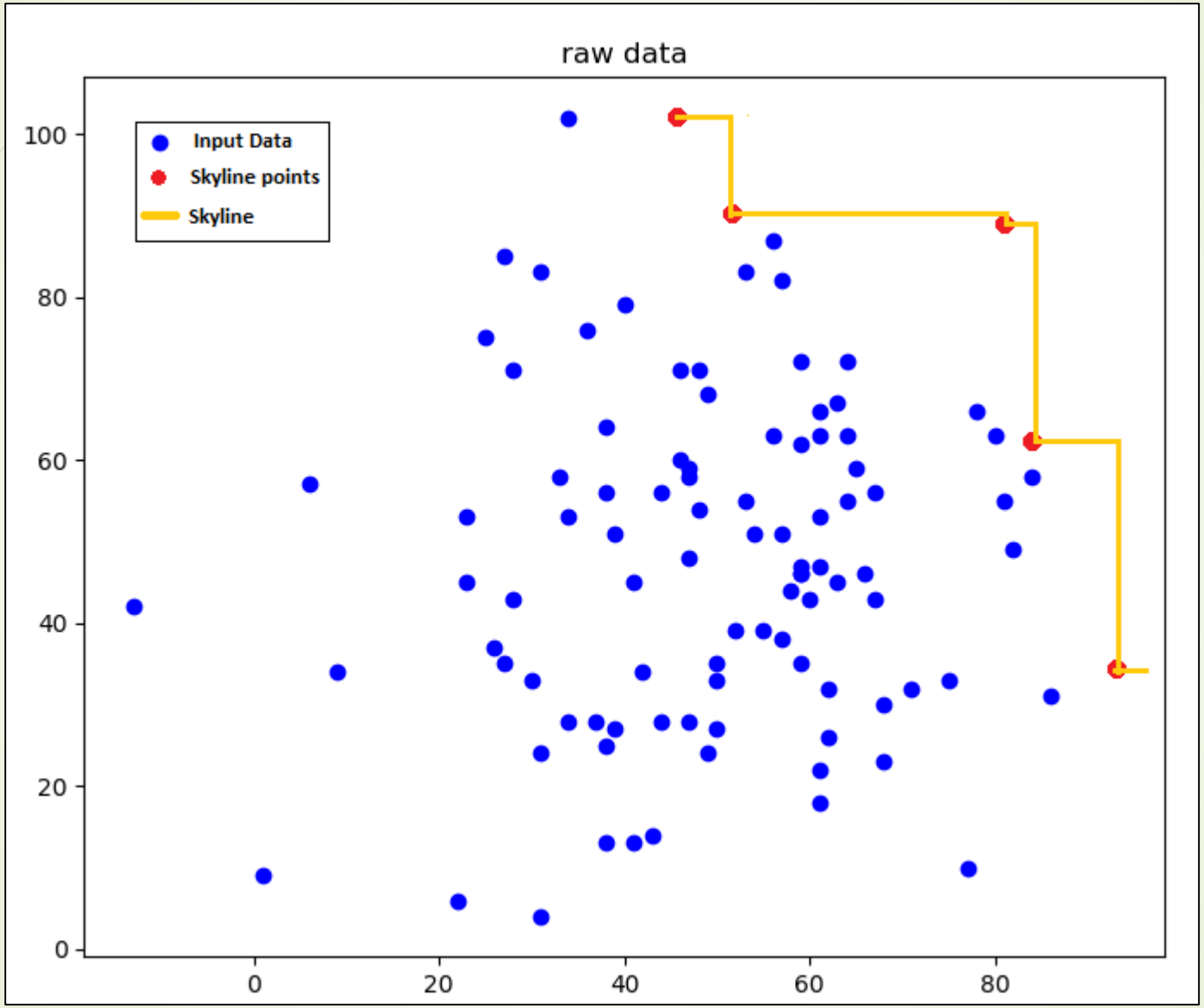
# Skyline Operator : Skyline Point 4



# Skyline Operator : Skyline Point 5



# Skyline Operator : Skyline



# Skyline Operator : Implementation (Sorting)

```
for(int i = 0; i<Data.size(); i++)
{
    for(int j=i+1;j<Data.size();j++)
    {
        if(Integer. parseInt(Data.get(i).get(0)) == Integer. parseInt(Data.get(j).get(0)))
        {
            // Remove Duplicate Points
            if(Integer. parseInt(Data.get(i).get(1)) == Integer. parseInt(Data.get(j).get(1)))
            {
                System.out.println("Removed Data Point : " + Data.get(j));
                Data.remove(j);
            }
            // If X value is same, Sort using Y value
            else if(Integer. parseInt(Data.get(i).get(1)) < Integer. parseInt(Data.get(j).get(1)))
            {
                tempdata = Data.get(i);
                Data.set(i, Data.get(j));
                Data.set(j, tempdata);
            }
        }
        else if(Integer. parseInt(Data.get(i).get(0)) < Integer. parseInt(Data.get(j).get(0)))
        {
            tempdata = Data.get(i);
            Data.set(i, Data.get(j));
            Data.set(j, tempdata);
        }
    }
}
```

# Skyline Operator : Implementation

```
Max_Y = Integer. parseInt(Data.get(0).get(1));
Output.add(Data.get(0));
for(int i = 1; i<Data.size(); i++)
{
    if(Max_Y < Integer. parseInt(Data.get(i).get(1))){
        Max_Y = Integer. parseInt(Data.get(i).get(1));
        Output.add(Data.get(i));
    }
}
System.out.println(Data.size());
return Output;
```

# Skyline Operator : Output

Problems @ Javadoc Declaration Console

```
<terminated> SkylineOperator [Java Application] C:\Program Files\Java\jre1.8.0_102\bin\javaw.exe (25-Jun-2019, 8:01:51 pm)
The Skyeline Points are :
46;102
52;90
81;89
84;62
93;34

Program Execution Time in milliseconds : 19.0 ms
```

# CHALLENGE 2: Entity Resolution

# Entity Resolution : Overview

- Entity resolution is used to identify duplicates within a single or among two datasets using similarity functions

# Entity Resolution : Goal

| | |
|---|---|
| Objective | Perform entity resolution on the CORA data set.<br>The program must identify and return duplicate pairs of publication, author and venue entities. |
| Input | CORA Dataset in XML format<br>https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/projekte/repeatability/CORA/cora-all-id.xml |
| Output | duplicates_publications.csv – List of all publications duplicates<br>duplicates_authors.csv – List of all authors duplicates<br>duplicates_venues.csv – List of all venues duplicates<br>Ex: ahlskog1994a;ahlskog1994a |
| Constraints | 1. Id fields shall not be used to solve this task<br>2. Efficiency : accuracy of dataset in terms of the F-measure. |
| Programming Language | Python 3.6 |

# Entity Resolution : Input CORA dataset


```
▼<coraRADD>
 ▼<publication id="ahlskog1994a">
    <author id="199">M. Ahlskog</author>
    <author id="74"> J. Paloheimo</author>
    <author id="64"> H. Stubb</author>
    <author id="103"> P. Dyreklev</author>
    <author id="54"> M. Fahlman</author>
    <title>Inganas</title>
    <title>and</title>
    <title>M.R.</title>
   ▼<venue>
    ▼<venue pubid="ahlskog1994a" id="1">
       <name>Andersson</name>
       <name> J Appl. Phys.</name>
       <vol>76</vol>
       <date> (1994). </date>
     </venue>
    </venue>
  </publication>
 ▼<publication id="ahlskog1994a">
    <author id="199">M. Ahlskog</author>
    <author id="74"> J. Paloheimo</author>
    <author id="64"> H. Stubb</author>
    <author id="103"> P. Dyreklev</author>
    <author id="54"> M. Fahlman</author>
    <author id="101"> O. Inganas and M.R. Andersson</author>
   ▼<venue>
    ▼<venue pubid="ahlskog1994a" id="1">
       <name>J Appl. Phys.</name>
       <vol>76</vol>
       <date> (1994). </date>
     </venue>
    </venue>
  </publication>
 ▼<publication id="ahlskog1994a">
```

# Entity Resolution : Algorithm

- Data Collection – Get the structured CORA data from the provided URL link (XML file)
- Data Extraction – Extract the XML data into list as an usable format
- Data Merging – Merge meaningful data which are seperated into multiple tags

```
<title>Provably</title>
<title>correct</title>
<title>compiler</title>
<title>development</title>
<title>and</title>
<title>implementation.</title>
```

→ "Provably correct compiler development and implementation."

```
▼<venue pubid="ahlskog1994a" id="1">
   <name>J Appl. Phys.</name>
   <vol>76</vol>
   <date> (1994). </date>
  </venue>
</venue>
```

→ "J Appl. Phys. 76 (1994)."

- Identify the similarity between the each elements.
  - Levenshtein Distance (Edit Distance)
- Add the elements which have high similarity to the corresponding CSV files

# Entity Resolution : Data Collection

```python
def GetWebData(WebURL):
    file = urllib.request.urlopen(WebURL)
    data = file.read()
    file.close()

    data = xmltodict.parse(data)
    return data
```

```
▼<coraRADD>
  ▼<publication id="ahlskog1994a">
     <author id="199">M. Ahlskog</author>
     <author id="74"> J. Paloheimo</author>
     <author id="64"> H. Stubb</author>
     <author id="103"> P. Dyreklev</author>
     <author id="54"> M. Fahlman</author>
     <title>Inganas</title>
     <title>and</title>
     <title>M.R.</title>
   ▼<venue>
     ▼<venue pubid="ahlskog1994a" id="1">
        <name>Andersson</name>
        <name> J Appl. Phys.</name>
        <vol>76</vol>
        <date> (1994). </date>
      </venue>
     </venue>
   </publication>
```

```
OrderedDict([('@id', 'ahlskog1994a'), ('author', [OrderedDict([('@id', '199'), ('#text', 'M. Ahlskog')]), OrderedDict([('@id', '74'), ('#text', 'J. Paloheimo')]), OrderedDict([('@id', '64'), ('#text', 'H. Stubb')]), OrderedDict([('@id', '103'), ('#text', 'P. Dyreklev')]), OrderedDict([('@id', '54'), ('#text', 'M. Fahlman')])]), ('title', ['Inganas', 'and', 'M.R.']), ('venue', OrderedDict([('venue', OrderedDict([('@pubid', 'ahlskog1994a'), ('@id', '1'), ('name', ['Andersson', 'J Appl. Phys.']), ('vol', '76'), ('date', '(1994).')])])])])
```

# Entity Resolution : Data Extraction and Merging

```
for Publication in InputData['coraRADD']['publication']:


    tempAut, tempID = GetInstanceAuthor(Publication,'author','#text','@id')
    PubIdList.append(Publication['@id'])
    TempTitleList.append(GetInstanceTitle(Publication,'title'))
    VenNameList.append(GetInstanceVenue(Publication,'venue','name'))
    VenIdList.append(GetInstanceVenue(Publication,'venue','@id'))
    VenDateList.append(GetInstanceVenue(Publication,'venue','date'))

    AuthorList.append(tempAut)
    AutID.append(tempID)

VenIdList,VenNameList,VenDateList = MergeVenueName(VenIdList,VenNameList,VenDateList)
```

# Entity Resolution : Title List

```python
def GetInstanceTitle(Pub, Param):
    OutputList = []
    if Param in Pub.keys():
        for Title in Pub[Param]:
            OutputList.append(Title)
    else:
        OutputList.append('NAN')

    return OutputList
```

```python
for Title in TempTitleList:
    if Title != ['NAN']:
        tempstr = str()
        for Entry in Title:
            tempstr = tempstr + Entry + " "
        TitleList.append(tempstr)
    else:
        TitleList.append('NAN')
```

```
Inganas and M.R.
NAN
NAN
NAN
NAN
NAN
NAN
NAN
Robots and Manufacturing Automation.
A spatial model of interaction in large virtual environments.
Viewpoints, Actionpoints and Spatial Frames for Collaborative User Interfaces,
User Embodiment in Collaborative Virtual Environments,
User Embodiment in Collaborative Virtual Environments.
User Embodiment in Collaborative Virtual Environments.
Networked Virtual realitty and Cooperative Work.
Actress: an action semantics directed compiler generator,
Actress: an action semantics directed compiler generator,
an action semantics directed compiler generator.
Provably correct compiler development and implementation.
Provably Correct Compiler Implementation,
```

# Entity Resolution : Author List

```
def GetInstanceAuthor(Pub, Param, Key, Idkey):
    OutputList = []
    TempList = []
    tempstr = str()
    KeyList = []
    if Param in Pub.keys():
        if(isinstance(Pub[Param],list) == False):
            OutputList = Pub[Param][Key]
            KeyList = Pub[Param][Idkey]
        else:
            for Aut in Pub[Param]:
                if(isinstance(Aut,str) == False):
                    OutputList = Aut[Key]
                    KeyList = Aut[Idkey]
    else:
        OutputList = 'NAN'
        KeyList = 'NAN'


    return OutputList,KeyList
```

```
M. Fahlman
O. Inganas and M.R. Andersson
O. Inganas and M.R. Andersson
O. Inganas and M.R. Andersson
O. Inganas and M.R. Andersson
O. Inganas and M.R. Andersson
O. Inganas and M.R. Andersson
O. Inganas and M.R. Andersson
C. Ray Asfahl.
Steve Benford and Lennart E. Fahlen.
and Fahln
and Snowdon
Snowdon
Snowdon
and Tom Rodden.
H. and Watt
H. and Watt
and D. A. Watt. Actress:
and M. Muller-Olm.
B. Buth et. al.
...
```

# Entity Resolution : Venue Instance

```
def GetInstanceVenue(Pub,Param,Key):
    OutputList = []
    if Param in Pub.keys():
        if Key in Pub[Param][Param].keys():
            OutputList.append(Pub[Param][Param][Key])
    else:
        OutputList.append('NAN')
    return OutputList
```

```
['(1994).']
['(1994).']
['(1994).']
['(1994).']
['(1994).']
['(1994).']
['(1994).']
['(1994).']
['1992.']
[['September', '1993.']]
[['(1994),', 'June 1994,']]
['May 7-11, 1995,']
['(1995).']
[]
['1995.']
['(1992b),']
['(1992b),']
['October 1992.']
['1992.']
['1992,']
```

```
[['Andersson', 'J Appl. Phys.']]
['J Appl. Phys.']
['J Appl. Phys.']
['J Appl. Phys.']
['J Appl. Phys.']
['J Appl. Phys.']
['J Appl. Phys.']
['Journal of Applied Physics']
[]
["In Proceedings of ECSCW'93"]
['6th ERCIM workshop']
['in Proc. ACM Conference on Human Factors in Computing Systems (CHI95)']
['In Proceedings of CHI95']
['In Proceedings of CHI95']
['Presence']
["`Proceedings of the International Workshop on Compiler Construction (CC-92)'"]
["`Proceedings of the International Workshop on Compiler Construction (CC-92)'"]
["Proceedings of the 4th International Conference on Compiler Construction (CC'92)"]
['Compiler Construction']
['Compiler Construction']
```

# Entity Resolution : Venue Name and Date

```python
def MergeVenueName(VenIdList,VenNameList,VenDateList):
    TempId = []
    TempName = []
    TempDate = []

    for Id in VenIdList:
        for Entry in Id:
            TempId.append(int(Entry))

    VenNameList = ProcessNameDate(VenNameList)
    VenDateList = ProcessNameDate(VenDateList)

    return TempId,VenNameList,VenDateList
```

```python
def ProcessNameDate(VenNameList):
    TempName = []
    TempDate = []
    TempStr = 'NAN'
    TempStrDate = 'NAN'
    for Name in VenNameList:
        if(len(Name) == 0):
            TempName.append('NAN')
        else:
            for NameEntry in Name:
                if isinstance(NameEntry,str) == True:
                    TempName.append(NameEntry)
                else:
                    TempStr = str()
                    for Index in NameEntry:
                        TempStr = TempStr + Index + " "
                    TempName.append(TempStr)
    return TempName
```

```
Andersson J Appl. Phys.
J Appl. Phys.
J Appl. Phys.
J Appl. Phys.
J Appl. Phys.
J Appl. Phys.
J Appl. Phys.
Journal of Applied Physics
NAN
In Proceedings of ECSCW'93
6th ERCIM workshop
in Proc. ACM Conference on Human Factors in Computing Systems (CHI95)
In Proceedings of CHI95
In Proceedings of CHI95
Presence
`Proceedings of the International Workshop on Compiler Construction (CC-92)'
`Proceedings of the International Workshop on Compiler Construction (CC-92)'
Proceedings of the 4th International Conference on Compiler Construction (CC'92)
Compiler Construction
Compiler Construction
```

```
(1994).
(1994).
(1994).
(1994).
(1994).
(1994).
(1994).
(1994).
1992.
September 1993.
(1994), June 1994,
May 7-11, 1995,
(1995).
NAN
1995.
(1992b),
(1992b),
October 1992.
1992.
1992,
```

# Entity Resolution : Venue Name and Date

```
def MergeVenueData(VenNameList,VenDateList):
    for i in range(0,len(VenNameList)):
        if VenNameList[i] == 'NAN':
            VenNameList[i] = VenDateList[i]
    return VenNameList
```

```
Andersson J Appl. Phys.
J Appl. Phys.
J Appl. Phys.
J Appl. Phys.
J Appl. Phys.
J Appl. Phys.
J Appl. Phys.
Journal of Applied Physics
1992.
In Proceedings of ECSCW'93
6th ERCIM workshop
in Proc. ACM Conference on Human Factors in Computing Systems (CHI95)
In Proceedings of CHI95
In Proceedings of CHI95
Presence
`Proceedings of the International Workshop on Compiler Construction (CC-92)'
`Proceedings of the International Workshop on Compiler Construction (CC-92)'
Proceedings of the 4th International Conference on Compiler Construction (CC'92)
Compiler Construction
Compiler Construction
```

# Entity Resolution : Similarity measure

```
def SimilarityMeasure(InputStr,ID,AvgLen):
    MatchList = []
    for i in range(0,len(InputStr)):
        for j in range(i+1,len(InputStr)):
            if InputStr[i] != 'NAN' and InputStr[j] != 'NAN':
                Sim = SequenceMatcher(None, InputStr[i], InputStr[j]).ratio()
                AverageLen = (len(InputStr[i]) + len(InputStr[j])) / 2
                if AverageLen < AvgLen:
                    if Sim > 0.9: # Similarity threshold is set as 0.9 for smaller strings
                        MatchList.append([ID[i],ID[j],InputStr[i],InputStr[j],Sim])
                else:
                    if Sim > 0.7: # Similarity threshold is set as 0.7 for Larger strings
                        MatchList.append([ID[i],ID[j],InputStr[i],InputStr[j],Sim])
            elif InputStr[i] == 'NAN' and InputStr[j] == 'NAN':
                MatchList.append([ID[i],ID[j],InputStr[i],InputStr[j],Sim])
            else:
                # Do nothing
                None
    return MatchList
```

```
def LD(s, t):
    if s == "":
        return len(t)
    if t == "":
        return len(s)
    if s[-1] == t[-1]:
        cost = 0
    else:
        cost = 1

    res = min([LD(s[:-1], t)+1,
               LD(s, t[:-1])+1,
               LD(s[:-1], t[:-1]) + cost])
    return res
```

# Entity Resolution : Duplicate Publications

| | | | | |
|---|---|---|---|---|
| fahlman1988b | fahlman1988b | An empirical study of learning speed in backpropagation networks. | An empirical study of learning speed in back-propagation networks. | 0.992481 |

| | | | | |
|---|---|---|---|---|
| brodley1992 | brodley1992 | Multivariate Versus Univariate Decision Trees. | Multivariate Versus Univariate Decision Trees. | 1 |
| brodley1992 | brodley1992b | Multivariate Versus Univariate Decision Trees. | Multivariate decision trees. | 0.710526 |
| brodley1992 | brodley1992b | Multivariate Versus Univariate Decision Trees. | Multivariate decision trees. | 0.710526 |
| brodley1992 | brodley1992b | Multivariate Versus Univariate Decision Trees. | Multivariate decision trees. | 0.710526 |
| brodley1992 | brodley1992b | Multivariate Versus Univariate Decision Trees. | Multivariate decision trees. | 0.710526 |

| | | | | |
|---|---|---|---|---|
| utgoff1982aaai | utgoff1982aaai | Acquisition of appropriate bias for inductive concept learning. | Acquisition of appropriate bias for inductive concept learning. | 1 |
| utgoff1982aaai | utgoff1984phd | Acquisition of appropriate bias for inductive concept learning. | Shift of bias for inductive concept learning. | 0.781818 |
| utgoff1982aaai | utgoff1984phd | Acquisition of appropriate bias for inductive concept learning. | Shift of bias for inductive concept learning. | 0.781818 |

# Entity Resolution : Duplicate Authors

| | | | | |
|---|---|---|---|---|
| 54 | 54 | M. Fahlman | S. Fahlman | 0.9 |
| 54 | 54 | M. Fahlman | S. Fahlman | 0.9 |
| 54 | 54 | M. Fahlman | Fahlman | 0.823529 |
| 54 | 54 | M. Fahlman | Fahlman | 0.823529 |
| 54 | 54 | M. Fahlman | S. E. Fahlman. | 0.75 |
| 54 | 54 | M. Fahlman | S. E. Fahlman. | 0.75 |
| 54 | 54 | M. Fahlman | S.E. Fahlman. | 0.782609 |

| | | | | |
|---|---|---|---|---|
| 117 | 21 | S. and Lebiere | C. Lebiere: | 0.72 |
| 117 | 117 | S. and Lebiere | S. Fahlman and C. Lebiere | 0.717949 |
| 117 | 117 | S. and Lebiere | S. Fahlman and C. Lebiere | 0.717949 |
| 117 | 117 | S. and Lebiere | S. Fahlman and C. Lebiere | 0.717949 |

| | | | | |
|---|---|---|---|---|
| 173 | 173 | D. Aha and D. Kibler. | David W. Aha and Dennis Kibler. | 0.769231 |
| 173 | 173 | D. Aha and D. Kibler. | D.W. Aha and D. Kibler. | 0.954545 |
| 173 | 28 | D. Aha and D. Kibler. | D. and Kibler | 0.764706 |
| 173 | 173 | D. Aha and D. Kibler. | D. Aha and D. Kibler. | 1 |
| 173 | 173 | D. Aha and D. Kibler. | D. Aha and D. Kibler. | 1 |

# Entity Resolution : Duplicate Venues

| | | | | | |
|---|---|---|---|---|---|
| 1 | 1 | Andersson J Appl. Phys. | J Appl. Phys. | 0.702703 | |
| 1 | 1 | Andersson J Appl. Phys. | J Appl. Phys. | 0.702703 | |
| 1 | 1 | Andersson J Appl. Phys. | J Appl. Phys. | 0.702703 | |
| 1 | 1 | Andersson J Appl. Phys. | J Appl. Phys. | 0.702703 | |

| | | | | |
|---|---|---|---|---|
| 3 | 4 | In Proceedings of ECSCW'93 | In Proceedings of CHI95 | 0.816327 |
| 3 | 4 | In Proceedings of ECSCW'93 | In Proceedings of CHI95 | 0.816327 |
| 3 | 20 | In Proceedings of ECSCW'93 | In Proceedings of Interchi '93 | 0.75 |
| 3 | 20 | In Proceedings of ECSCW'93 | In Proceedings of Interchi '93 | 0.75 |
| 3 | 20 | In Proceedings of ECSCW'93 | In Proceedings of Interchi '93 | 0.75 |
| 3 | 20 | In Proceedings of ECSCW'93 | In Proceedings of Interchi '93 | 0.75 |

| | | | | |
|---|---|---|---|---|
| 6 | 6 | `Proceedings of the International Workshop on Compiler Construction (CC-92)' | in Proceedings of the Fourth International Conference on Compiler Construction | 0.779221 |
| 6 | 6 | `Proceedings of the International Workshop on Compiler Construction (CC-92)' | in Proceedings of the Fourth International Conference on Compiler Construction | 0.779221 |
| 6 | 6 | `Proceedings of the International Workshop on Compiler Construction (CC-92)' | in Proceedings of the Fourth International Conference on Compiler Construction | 0.779221 |

| | | | | |
|---|---|---|---|---|
| 13 | 14 | Proceedings of the Royal Society London B (in press). | Proceedings of the Royal Society of London B | 0.857143 |
| 13 | 14 | Proceedings of the Royal Society London B (in press). | Proceedings of the Royal Society of London B | 0.857143 |

| | | | | |
|---|---|---|---|---|
| 22 | 22 | 1979 | 1979 | 1 |
| 22 | 22 | 1979 | 1979 | 1 |

# Entity Resolution : Current and Future Adaptations

- Similarity Threshold was modified to based on the average length of the comparing string to solve below mentioned issue.
  - 1992 and 1994 - 75% similarity
  - Hence the similarity threshold was modified to 90% for strings of length less than 5.
- The following characters and words shall be removed before comparing the strings to get a better comparison measure.
  - The special characters - { “ , ‘ ( ) } { [ ] .
  - Words like The, of , and , is
  - Initials, Salutation, Degrees
- The characters in uppercase shall be converted to lowercase
- Some Special processing is needed depending data.
  - Proceedings of the **Eighth** International Workshop on Machine Learning
  - Proceedings of the **Ninth** International Workshop on Machine Learning
  - In Proceedings of the **4 th** International Workshop on Machine Learning

# CHALLENGE 3: Data Cleaning

# Data Cleaning : Overview

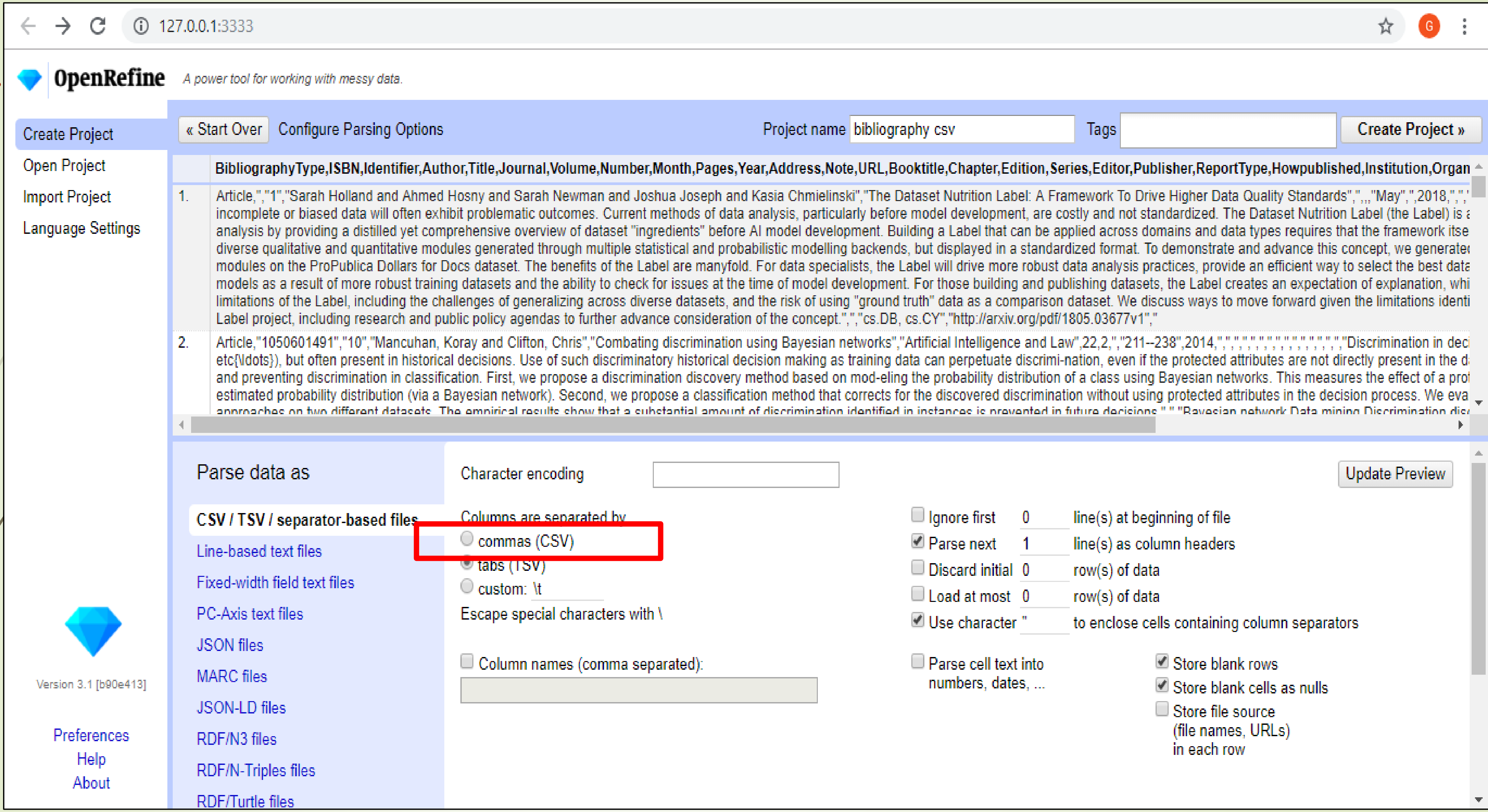
- During data cleaning, errors in data such as illegal or wrong values are resolved to improve the quality of the given data.
- **OpenRefine** is a power tool for working with messy data. Use it to improve data consistency, link it to data registries like Wikidata, augment it with data from other sources, transform it into different formats for other tools to consume, and contribute it to back to the original sources.
- OpenRefine is not a web service but a desktop app that runs on your own computer, so you can process sensitive data with privacy.
- OpenRefine was originally developed as "Freebase Gridworks" by Metaweb Technologies, Inc.. Metaweb was acquired by Google in July 2010 and they renamed the product Google Refine.
- In October, 2012, the product was renamed OpenRefine as it transitioned to a community supported project.

# Data Cleaning : Goal



| | |
|---|---|
| Objective | Clean the data in the file.<br>During the collection of the data, wrong, incomplete, and duplicate data have been introduced which should be resolved by according to the given cleaning requirements. |
| Input | bibliography.csv |
| Output | Cleaned OpenRefine Project Zip folder. |
| Constraints | 1. The cleaned data set should contain only correct publication entries.<br>2. all false, nonsense or redundant information should be removed from the data set.<br><br>**Article**: A peer-reviewed article from a journal or magazine<br>• Author, title, journal, year<br><br>**Book:** A printed and bound book with explicit publisher<br>• Title, publisher, year, author, editor<br><br>**Incollection** : Published as part of a printed and bound book with own title<br>• Author, title, booktitle, year<br><br>**Inproceedings** : An article published in a conference proceedings<br>• Author, title, booktitle, year<br><br>**Misc**: Entry type that can be used if no other entry type fits. |
| Tool | OpenRefine (http://openrefine.org) |


127.0.0.1:3333
OpenRefine
A power tool for working with messy data.
Create Project
Open Project
Import Project
Language Settings
Create a project by importing data. What kinds of data files can I import?
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all suppo
Get data from
This Computer
Web Addresses (URLs)
Clipboard
Data Package (JSON URL)
Database
Google Data
Locate one or more files on your computer to upload:
Choose Files
No file chosen
Next »
Version 3.1 [b90e413]
Preferences
Help
About



Project name: bibliography csv

Tags:

**BibliographyType,ISBN,Identifier,Author,Title,Journal,Volume,Number,Month,Pages,Year,Address,Note,URL,Booktitle,Chapter,Edition,Series,Editor,Publisher,ReportType,Howpublished,Institution,Organ**

1. Article,","1","Sarah Holland and Ahmed Hosny and Sarah Newman and Joshua Joseph and Kasia Chmielinski","The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards",",,,"May",",2018,",",'
incomplete or biased data will often exhibit problematic outcomes. Current methods of data analysis, particularly before model development, are costly and not standardized. The Dataset Nutrition Label (the Label) is a
analysis by providing a distilled yet comprehensive overview of dataset "ingredients" before AI model development. Building a Label that can be applied across domains and data types requires that the framework itse
diverse qualitative and quantitative modules generated through multiple statistical and probabilistic modelling backends, but displayed in a standardized format. To demonstrate and advance this concept, we generate
modules on the ProPublica Dollars for Docs dataset. The benefits of the Label are manyfold. For data specialists, the Label will drive more robust data analysis practices, provide an efficient way to select the best data
models as a result of more robust training datasets and the ability to check for issues at the time of model development. For those building and publishing datasets, the Label creates an expectation of explanation, whi
limitations of the Label, including the challenges of generalizing across diverse datasets, and the risk of using "ground truth" data as a comparison dataset. We discuss ways to move forward given the limitations identi
Label project, including research and public policy agendas to further advance consideration of the concept.",","cs.DB, cs.CY","http://arxiv.org/pdf/1805.03677v1","
2. Article,"1050601491","10","Mancuhan, Koray and Clifton, Chris","Combating discrimination using Bayesian networks","Artificial Intelligence and Law",22,2,","211--238",2014,",",",",",",",",",",",",","Discrimination in deci
etc{\ldots}), but often present in historical decisions. Use of such discriminatory historical decision making as training data can perpetuate discrimi-nation, even if the protected attributes are not directly present in the d
and preventing discrimination in classification. First, we propose a discrimination discovery method based on mod-eling the probability distribution of a class using Bayesian networks. This measures the effect of a prot
estimated probability distribution (via a Bayesian network). Second, we propose a classification method that corrects for the discovered discrimination without using protected attributes in the decision process. We eva
approaches on two different datasets. The empirical results show that a substantial amount of discrimination identified in instances is prevented in future decisions.","","Bayesian network Data mining Discrimination disc

Character encoding:

Columns are separated by
○ commas (CSV)
● tabs (TSV)
○ custom: \t

Escape special characters with \

☐ Column names (comma separated):

☐ Ignore first 0 line(s) at beginning of file
☑ Parse next 1 line(s) as column headers
☐ Discard initial 0 row(s) of data
☐ Load at most 0 row(s) of data
☑ Use character " to enclose cells containing column separators

☐ Parse cell text into numbers, dates, ...

☑ Store blank rows
☑ Store blank cells as nulls
☐ Store file source (file names, URLs) in each row

OpenRefine *A power tool for working with messy data.*

Create Project
Open Project
Import Project
Language Settings

« Start Over | Configure Parsing Options | Project name bibliography.csv | Tags | Create Project »

| | BibliographyType | ISBN | Identifier | Author | Title | Journal | Volume | Number | Month | Pages | Year | Address | Note | URL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | Article | | 1 | Sarah Holland and Ahmed Hosny and Sarah Newman and Joshua Joseph and Kasia Chmielinski | The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards | | | | May | | 2018 | | | |

Parse data as

- **CSV / TSV / separator-based files**
- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- JSON files
- MARC files
- JSON-LD files
- RDF/N3 files
- RDF/N-Triples files
- RDF/Turtle files

Character encoding

Update Preview

Columns are separated by
- (•) commas (CSV)
- ( ) tabs (TSV)
- ( ) custom: \t

Escape special characters with \

[ ] Column names (comma separated):

[ ] Ignore first 0 line(s) at beginning of file
[x] Parse next 1 line(s) as column headers
[ ] Discard initial 0 row(s) of data
[ ] Load at most 0 row(s) of data
[x] Use character " to enclose cells containing column separators

[ ] Parse cell text into numbers, dates, ...

[x] Store blank rows
[x] Store blank cells as nulls
[ ] Store file source (file names, URLs) in each row

Version 3.1 [b90e413]

Preferences
Help
About

OpenRefine bibliographycsv Permalink

Open... Export Help

Facet / Filter Undo / Redo 0 / 0

112 rows

Extensions: Wikidata

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 10 next › last »

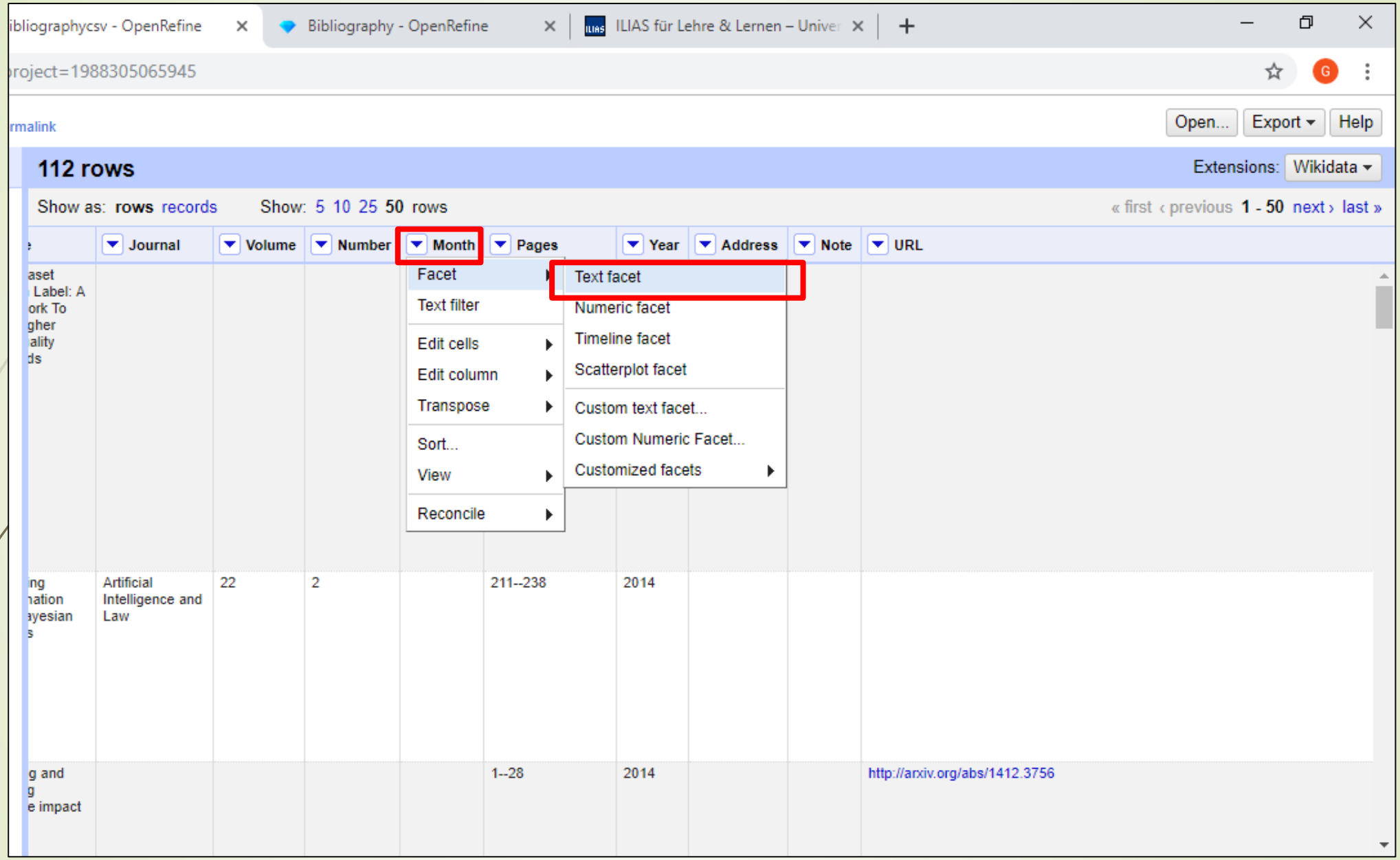
**Using facets and filters**

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

| All | BibliographyTyp | ISBN | Identifier | Author | Title | Journal | Volume | Number | Month | Pages | Year | Address | No |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | Article | | 1 | Sarah Holland and Ahmed Hosny and Sarah Newman and Joshua Joseph and Kasia Chmielinski | The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards | | | | May | | 2018 | | |

127.0.0.1:3333


ibliographycsv - OpenRefine
Bibliography - OpenRefine
ILIAS für Lehre & Lernen – Unive
project=1988305065945
rmalink
Open...
Export
Help
112 rows
Extensions: Wikidata
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 - 50 next › last »
Journal
Volume
Number
Month
Pages
Year
Address
Note
URL
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Text facet
Numeric facet
Timeline facet
Scatterplot facet
Custom text facet...
Custom Numeric Facet...
Customized facets
Artificial Intelligence and Law
22
2
211--238
2014
1--28
2014
http://arxiv.org/abs/1412.3756

Facet / Filter | Undo / Redo 0 / 0

Refresh | Reset All | Remove All

**Month** change

17 choices Sort by: **name** count | Cluster

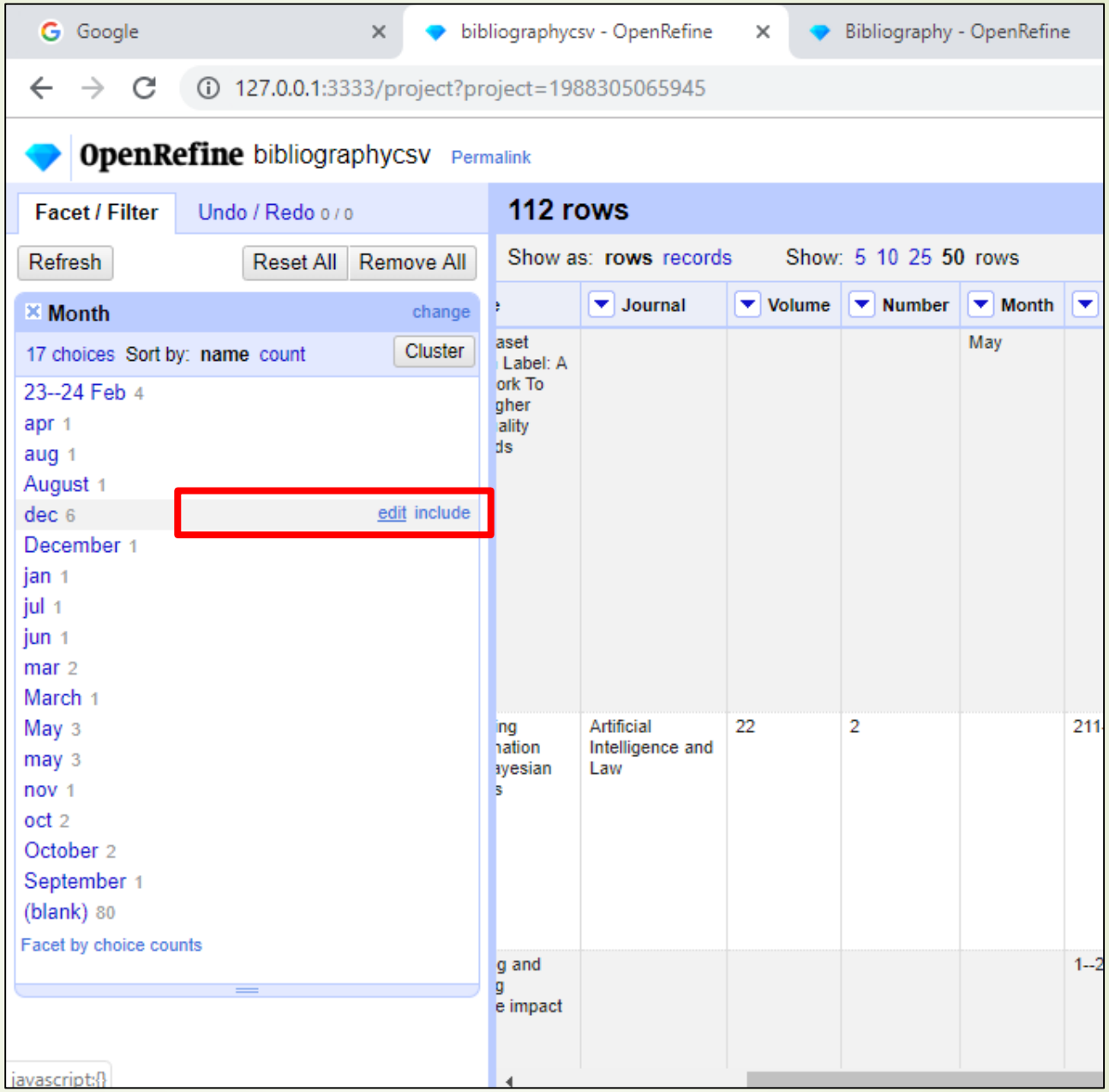
23--24 Feb 4
apr 1
aug 1
August 1
dec 6
December 1
jan 1
jul 1
jun 1
mar 2
March 1
May 3
may 3
nov 1
oct 2
October 2
September 1
(blank) 80

Facet by choice counts

**112 rows**

Show as: **rows** records  Show: 5 10 25 **50** rows

| | Journal | Volume | Number | Month | Page |
|---|---|---|---|---|---|
| aset Label: A ork To gher ality ds | | | | May | |
| ng nation ayesian s | Artificial Intelligence and Law | 22 | 2 | | 211--238 |
| g and g e impact | | | | | 1--28 |


Google
bibliographycsv - OpenRefine
Bibliography - OpenRefine
127.0.0.1:3333/project?project=1988305065945
OpenRefine bibliographycsv
Permalink
Facet / Filter
Undo / Redo 0 / 0
112 rows
Refresh
Reset All
Remove All
Show as: rows records
Show: 5 10 25 50 rows
Month
change
17 choices Sort by: name count
Cluster
23--24 Feb 4
apr 1
aug 1
August 1
dec 6
edit include
December 1
jan 1
jul 1
jun 1
mar 2
March 1
May 3
may 3
nov 1
oct 2
October 2
September 1
(blank) 80
Facet by choice counts
Journal
Volume
Number
Month
May
Artificial Intelligence and Law
22
2
javascript:{}

127.0.0.1:3333/project?project=1988305065945

**OpenRefine** bibliographycsv Permalink

Facet / Filter | Undo / Redo 0 / 0

Refresh | Reset All | Remove All

**Month** change

17 choices Sort by: name count | Cluster

- 23--24 Feb 4
- apr 1
- aug 1
- August 1
- dec 6
- December 1
- jan 1
- jul 1
- jun 1
- mar 2
- March 1
- May 3
- may 3
- nov 1
- oct 2
- October 2
- September 1
- (blank) 80

Facet by choice counts

**112 rows**

Show as: rows records  Show: 5 10 25 50 rows

| | Journal | Volume | Number | Month | Pages |
|---|---|---|---|---|---|
| aset Label: A ork To gher ality ds | | | | May | |
| dec | | | | | |
| ng nation ayesian s | Artificial Intelligence and Law | 22 | 2 | | 211--238 |
| g and g | | | | | 1--28 |

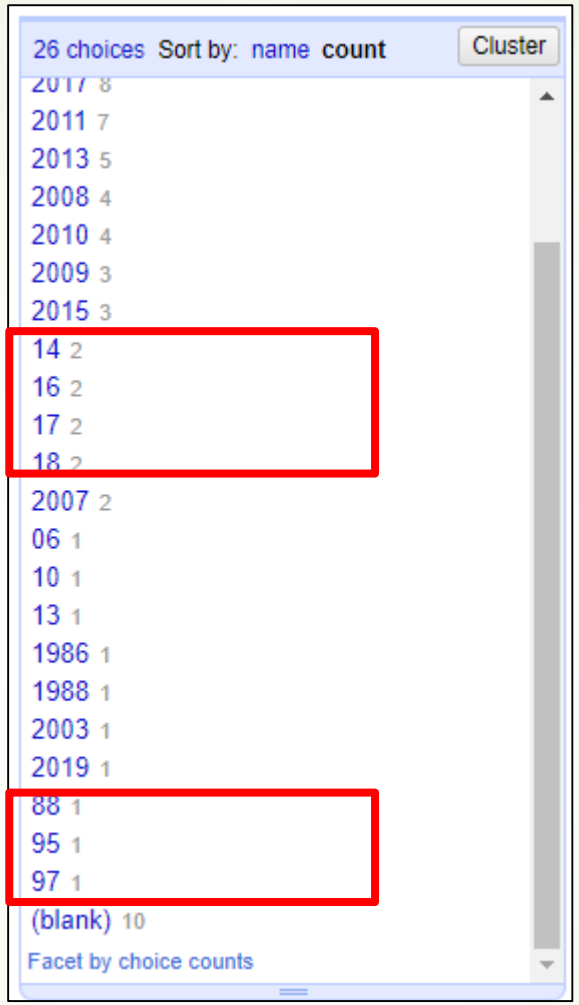
Apply (Enter) | Cancel (Esc)

**Month** change

11 choices Sort by: name count | Cluster

- April 1
- August 2
- December 7
- February 3
- January 1
- July 1
- March 3
- May 6
- November 1
- October 4
- September 1
- (blank) 75

Facet by choice counts


26 choices Sort by: name count
Cluster
2017 8
2011 7
2013 5
2008 4
2010 4
2009 3
2015 3
14 2
16 2
17 2
18 2
2007 2
06 1
10 1
13 1
1986 1
1988 1
2003 1
2019 1
88 1
95 1
97 1
(blank) 10
Facet by choice counts

Google | bibliographycsv - OpenRefine | Bibliography - OpenRefine | ILIAS für Lehre & Lernen – Unive

127.0.0.1:3333/project?project=1988305065945

**OpenRefine** bibliographycsv Permalink

Facet / Filter | Undo / Redo 0 / 0

Refresh | Reset All | Remove All

26 choices Sort by: name count | Cluster

2017 8
2011 7
2013 5
2008 4
2010 4
2009 3
2015 3
14 2
16 2
17 2
18 2
2007 2
06 1
10 1
13 1
1986 1
1988 1
2003 1
2019 1
88 1
95 1
97 1
(blank) 10

Facet by choice counts

javascript:{}

**112 rows**

Show as: rows records  Show: 5 10 25 50 rows

| | Journal | Volume | Number | Month | Pages | Year | Address | Note | URL |
|---|---|---|---|---|---|---|---|---|---|
| aset Label: A ork To gher ality ds | | | | May | | | | | |
| ng nation ayesian s | Artificial Intelligence and Law | 22 | 2 | | 211--238 | 2014 | | | |
| g and g e impact | | | | | 1--28 | 2014 | | | http://arxiv.org/abs/1412 |

Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile

Transform...
Common transforms
Fill down
Blank down
Split multi-valued cells...
Join multi-valued cells...
Cluster and edit...
Replace

```
if value == None:
  return value
elif int(value) >=0 and int(value) <20:
  value = str(int(value)+2000)
  return value
elif int(value) >19 and int(value) <99:
  value = str(int(value)+1900)
  return value
else:
  return value
```

## Custom text transform on column Year

Expression

Language: Python / Jython

```
if value == None:
  return value
elif int(value) >=0 and int(value) <20:
  value = str(int(value)+2000)
  return value
```

No syntax error.

Preview | History | Starred | Help

| | | |
|---|---|---|
| 37. | 2013 | 2013 |
| 38. | 2003 | 2003 |
| 39. | 97 | 1997 |
| 40. | 2016 | 2016 |
| 41. | 1988 | 1988 |
| 42. | 88 | 1988 |
| 43. | 10 | 2010 |
| 44. | null | null |

On error
- keep original
- set to blank
- store error

Re-transform up to 10 times until no change

OK Cancel

26 choices Sort by: name count Cluster

2017 8
2011 7
2013 5
2008 4
2010 4
2009 3
2015 3
14 2
16 2
17 2
18 2
2007 2
06 1
10 1
13 1
1986 1
1988 1
2003 1
2019 1
88 1
95 1
97 1
(blank) 10

Facet by choice counts

19 choices Sort by: name count Cluster

2018 18
2012 13
2016 12
2014 11
2017 10
2011 7
2013 6
2010 5
2008 4
2009 3
2015 3
1988 2
2007 2
1986 1
1995 1
1997 1
2003 1
2006 1
2019 1
(blank) 10

Facet by choice counts

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

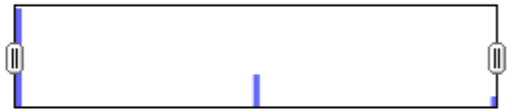
Method key collision ▾ Keying Function metaphone3 ▾ 13 clusters found

| Cluster Size | Row Count | Values in Cluster | Merge? | New Cell Value |
|---|---|---|---|---|
| 3 | 4 | • Kamiran, Faisal and Calders, Toon (2 rows)<br>• Kamiran, Faisal and Karim, Asim and Zhang, Xiangliang (1 rows)<br>• Kamiran, Faisal and {\v{Z}}liobaite, Indre and Calders, Toon (1 rows) | ☑ | Kamiran, Faisal and Calders, Toon |
| 2 | 3 | • Sara Hajian and Josep Domingo-Ferrer (2 rows)<br>• Sara Hajian and Josep Domingo-Ferrer and Antoni Mart{\'{\i}}nez-Ballest{\'{e}} (1 rows) | ☑ | Sara Hajian and Josep Domingo-Fe |
| 2 | 2 | • Ruggieri, Salvatore (1 rows)<br>• Ruggieri, Salvatore and Pedreschi, Dino and Turini, Franco (1 rows) | ☐ | Ruggieri, Salvatore |
| 2 | 2 | • Kamishima, Toshihiro and Akaho, Shotaro (1 rows)<br>• Kamishima, Toshihiro and Akaho, Shotaro and Asoh, Hideki and Sakuma, Jun (1 rows) | ☐ | Kamishima, Toshihiro and Akaho, S |
| 2 | 2 | • Hajian, Sara and Domingo-Ferrer, Josep and Martinez-Balleste, Antoni (1 rows)<br>• Hajian, Sara and Domingo-ferrer, Josep (1 rows) | ☐ | Hajian, Sara and Domingo-Ferrer, . |

**# Choices in Cluster**

2 — 3

**# Rows in Cluster**

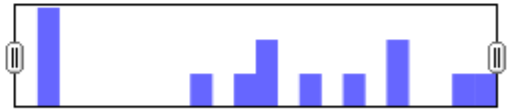


2 — 4

**Average Length of Choices**

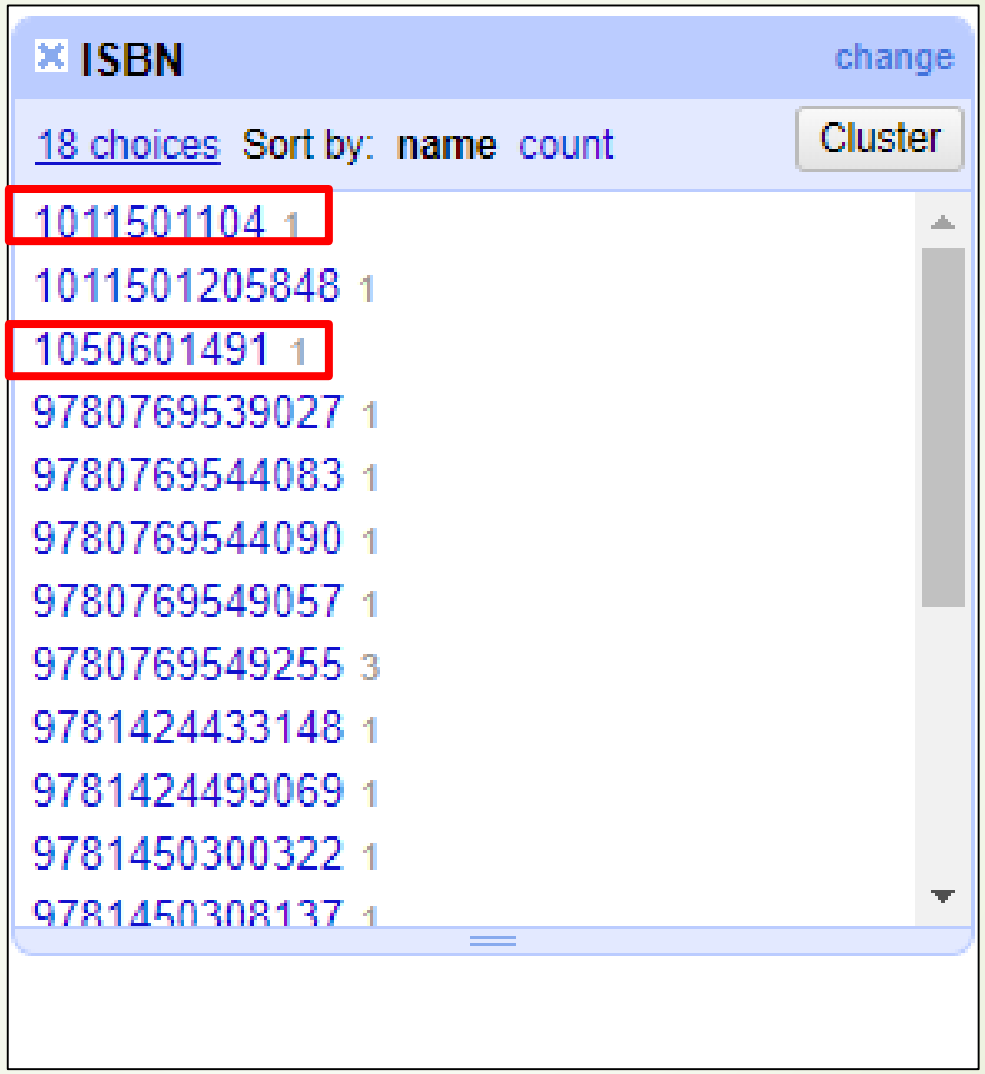
38 — 75

**Length Variance of Choices**



0 — 21

Select All | Unselect All | Export Clusters | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close


ISBN
change
18 choices
Sort by: name count
Cluster
1011501104 1
1011501205848 1
1050601491 1
9780769539027 1
9780769544083 1
9780769544090 1
9780769549057 1
9780769549255 3
9781424433148 1
9781424499069 1
9781450300322 1

| | | | | | | Booktitle | Chapter | Edition | Series | Editor | Publisher |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 87. | Book | 1011501104 | 6 | Kamiran, Faisal and Calders, Toon | Data preprocessing techniques for classification without discrimination | Knowledge and Information Systems | | | | | |

**1 matching rows** (105 total) Extensions: Wikidat

Show as: **rows** records Show: 5 10 25 **50** rows « first ‹ previous **1 - 1** next ›

| All | | BibliographyTyp | ISBN | Identifier | Author | Title | Journal | Volume | Number | Month | Pages | Year | Address | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 52. | InCollection | 0001011501104 | | 6 | Kamiran, Faisal and Calders, Toon | Data preprocessing techniques for classification without discrimination | | 33 | 1 | | 1--33 | 2012 | | |

## Custom text transform on column Bibliography Type

Expression

Language: General Refine Expression Language (GREL)

```
if(cells["Year"].value == null,"Miscc",value)
```

No syntax error.

Preview | History | Starred | Help

| row | value | if(cells["Year"].value == null ... |
|---|---|---|
| 1. | Article | Article |
| 2. | Article | Article |
| 3. | Article | Article |
| 4. | Article | Article |
| 5. | Article | Article |
| 6. | Article | Article |

On error
- keep original
- set to blank
- store error

Re-transform up to 10 times until no change

OK | Cancel

Google | bibliographycsv - OpenRefine | Bibliography - OpenRefine | ILIAS für Lehre & Lernen – U

127.0.0.1:3333/project?project=1988305065945

**OpenRefine** bibliographycsv Permalink

Facet / Filter | Undo / Redo 3 / 4

Extract... Apply...

Filter:

0. Create project
1. Text transform on 14 cells in column Year: jython:if value == None: return value elif int(value) >=0 and int(value) <20: value = str(int(value)+2000) return value elif int(value) >19 and int(value) <99: value = str(int(value)+1900) return value else: return value
2. Mass edit 14 cells in column Author
3. Star row 87
4. Star 111 rows

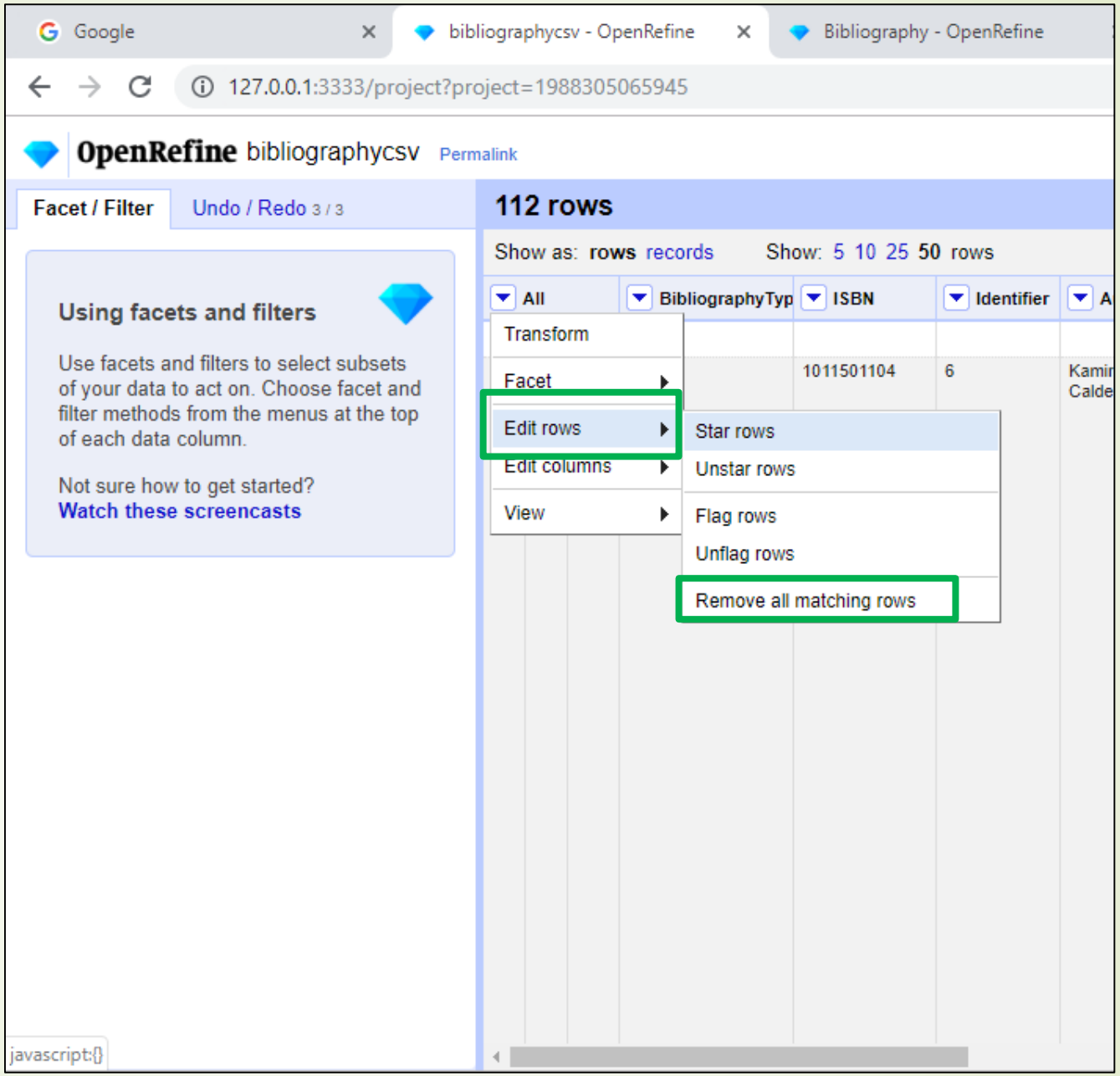
**112 rows**

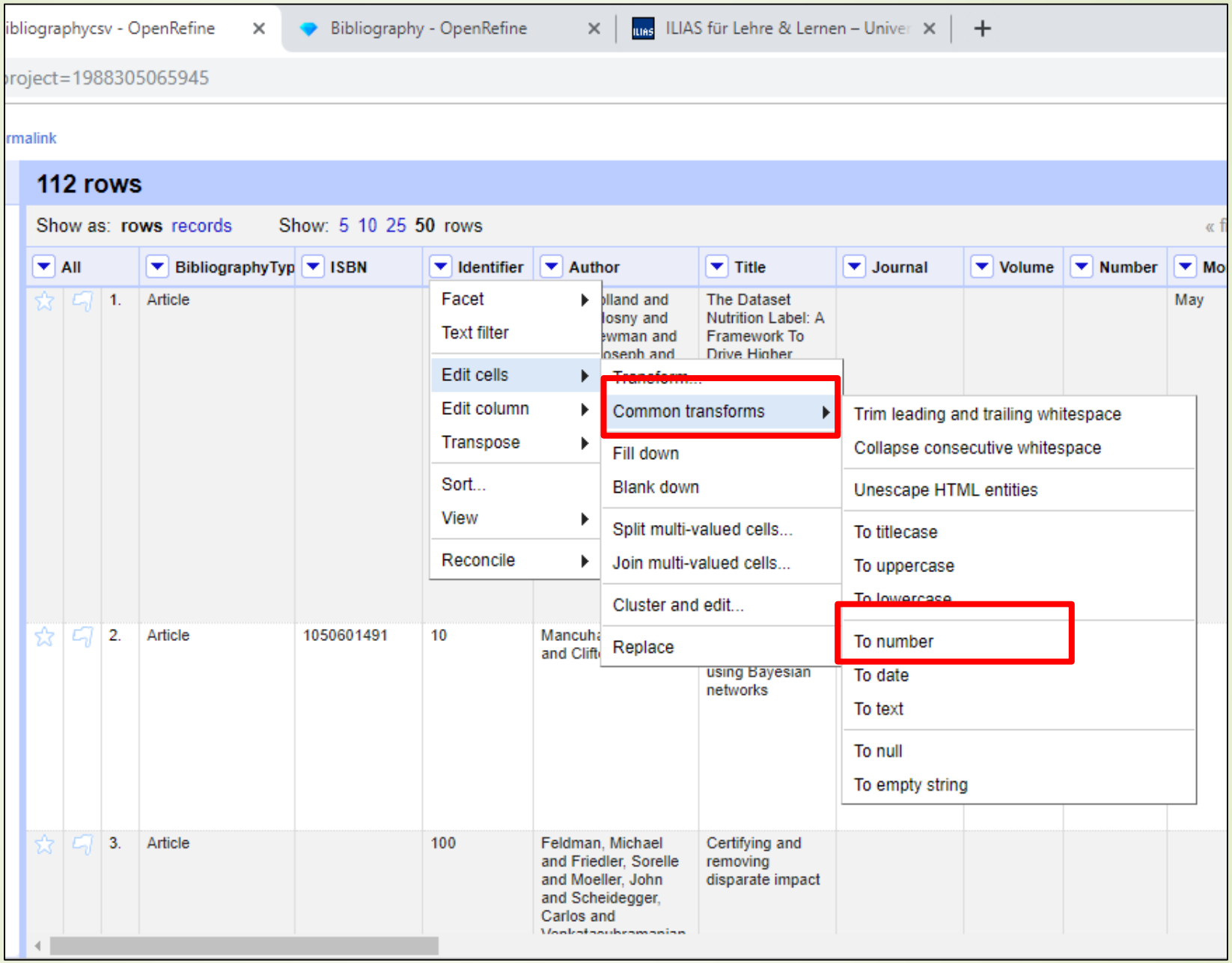
Show as: **rows** records Show: 5 10 25 **50** rows

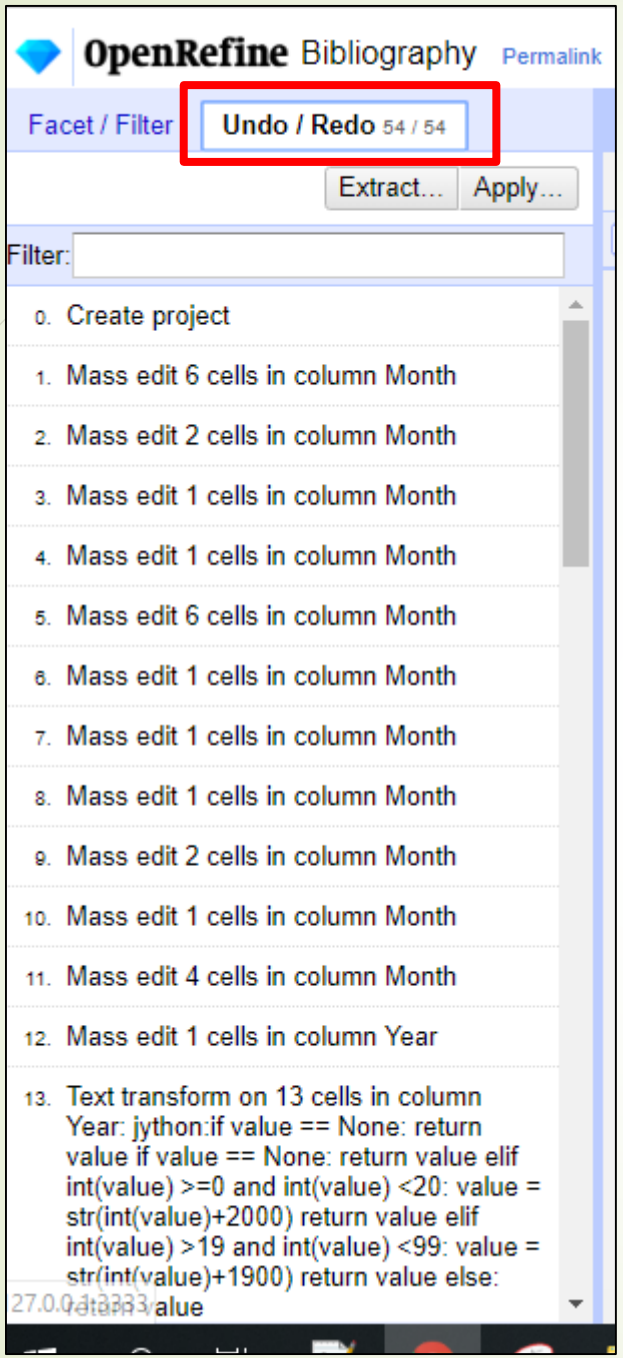
Transform
Facet ▸ Facet by star / Facet by flag
Edit rows ▸
Edit columns ▸
View ▸

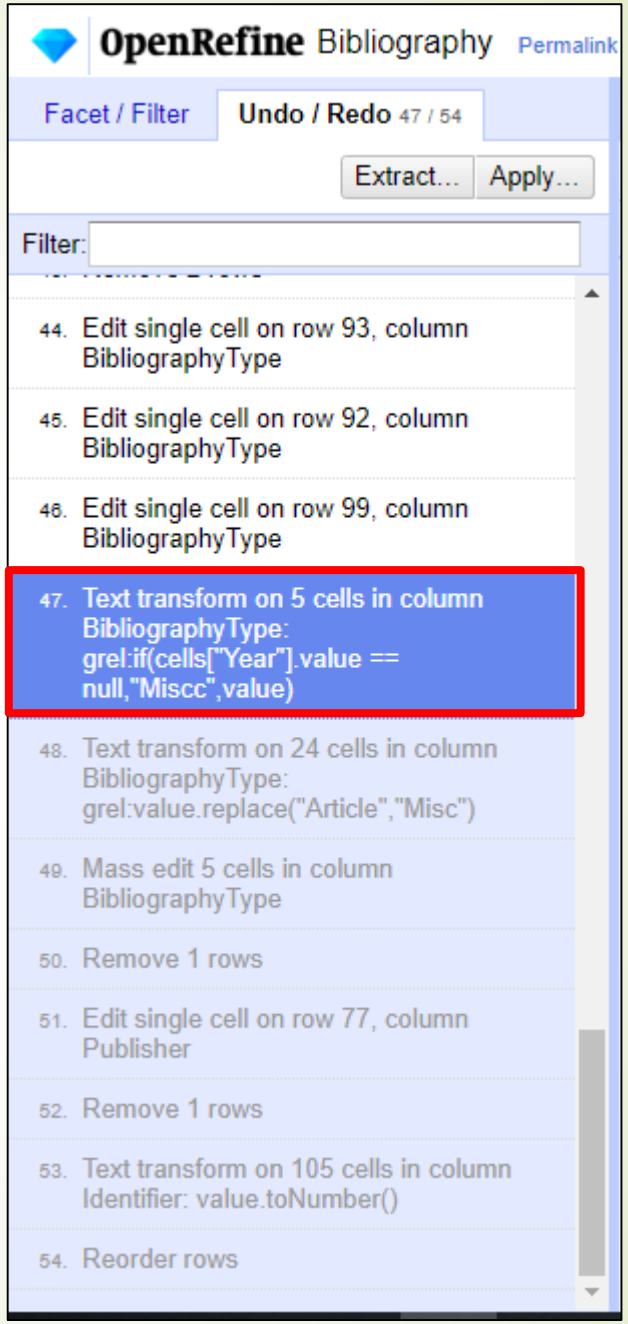
| All | BibliographyTyp | ISBN | Identifier | Author | Title |
|---|---|---|---|---|---|
| | | | 1 | Sarah Holland and Ahmed Hosny and Sarah Newman and Joshua Joseph and Kasia Chmielinski | The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards |
| 2. | Article | 1050601491 | 10 | Mancuhan, Koray and Clifton, Chris | Combating discrimination using Bayesian networks |
| 3. | Article | | 100 | Feldman, Michael and Friedler, Sorelle and Moeller, John and Scheidegger, Carlos and | Certifying and removing disparate impact |

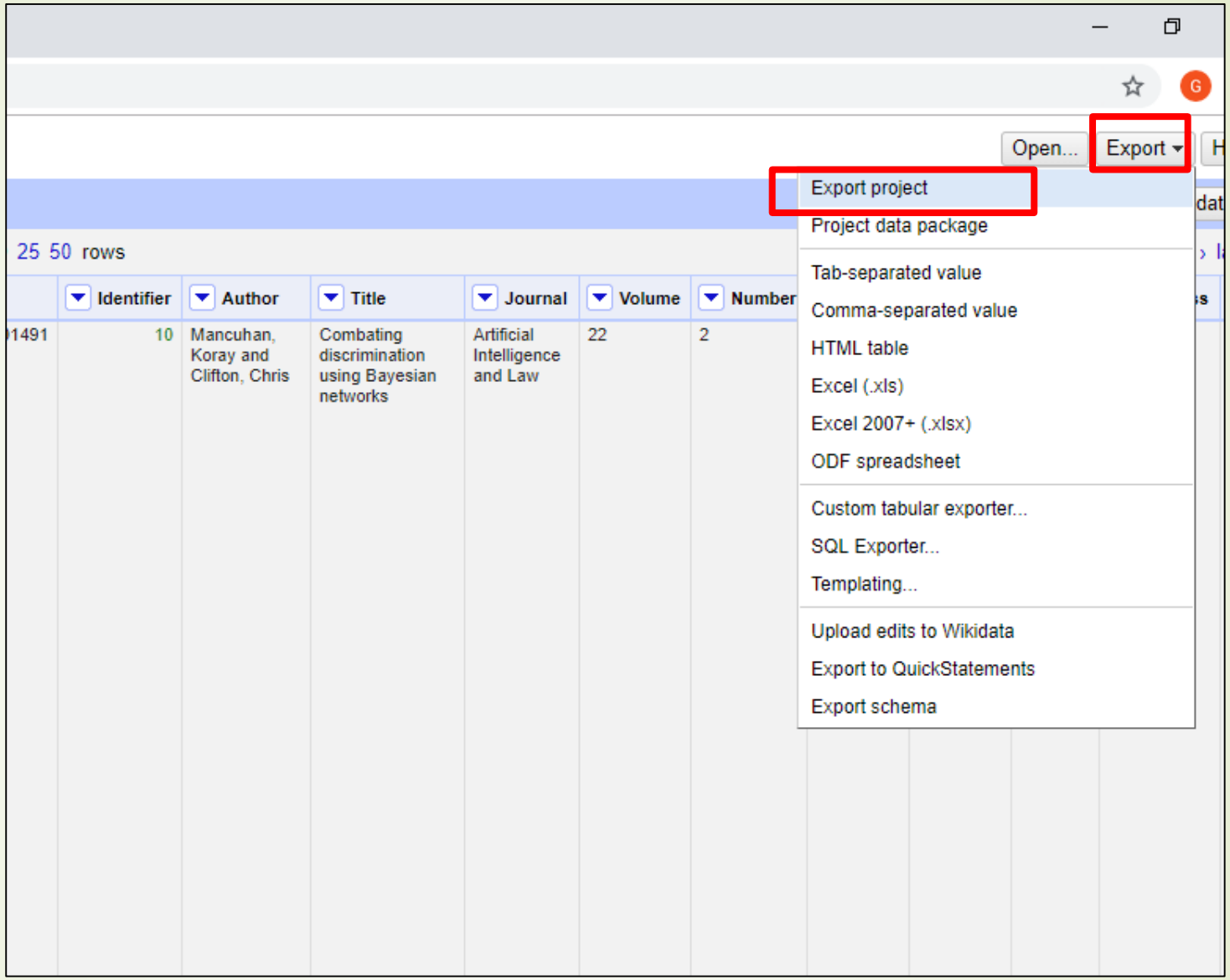
javascript:{}

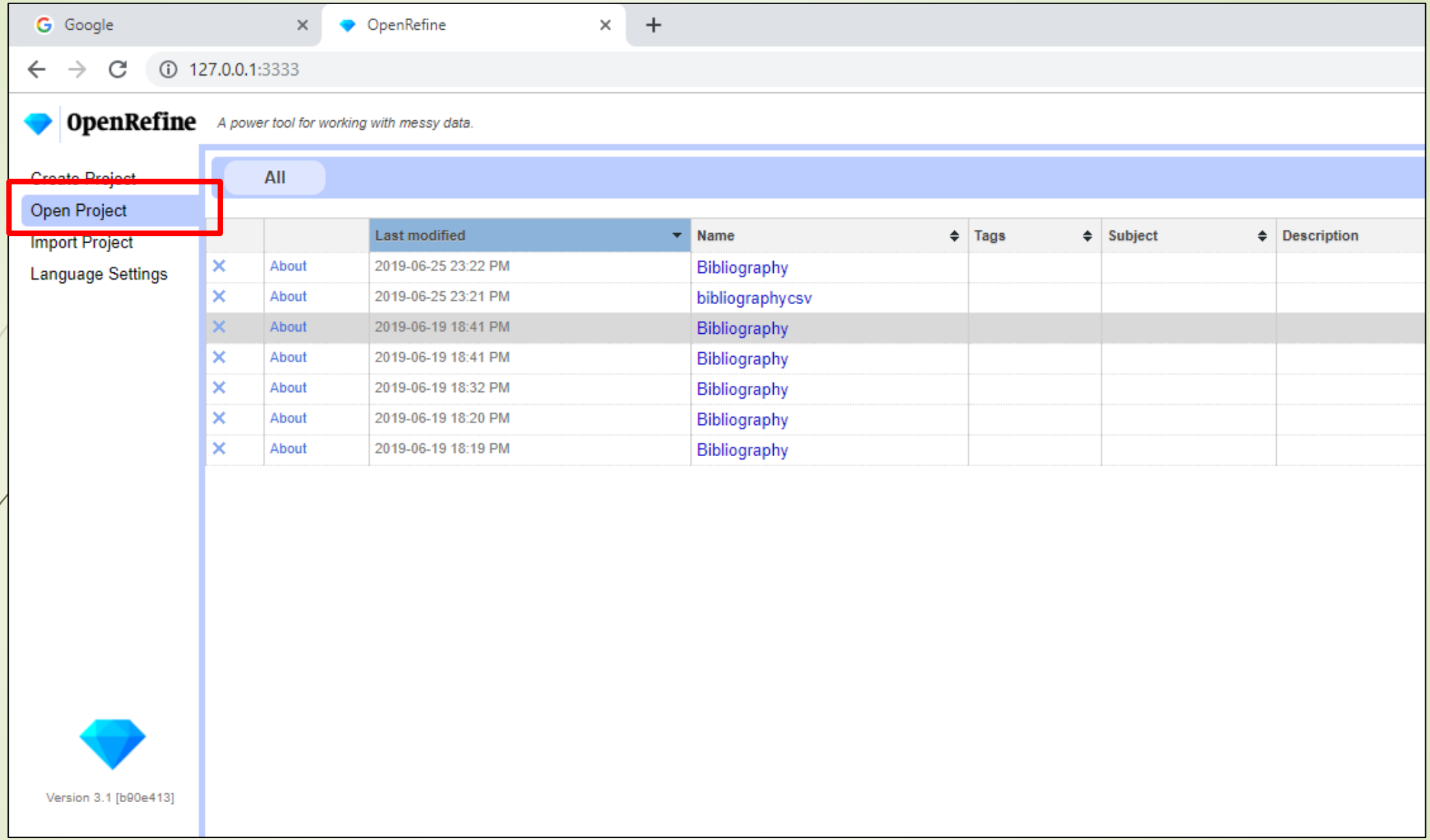

Google
bibliographycsv - OpenRefine
Bibliography - OpenRefine
127.0.0.1:3333/project?project=1988305065945
OpenRefine bibliographycsv
Permalink
Facet / Filter
Undo / Redo 3 / 3
112 rows
Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
Watch these screencasts
Show as: rows records
Show: 5 10 25 50 rows
All
BibliographyTyp
ISBN
Identifier
Transform
Facet
Edit rows
Edit columns
View
1011501104
6
Star rows
Unstar rows
Flag rows
Unflag rows
Remove all matching rows
javascript:{}

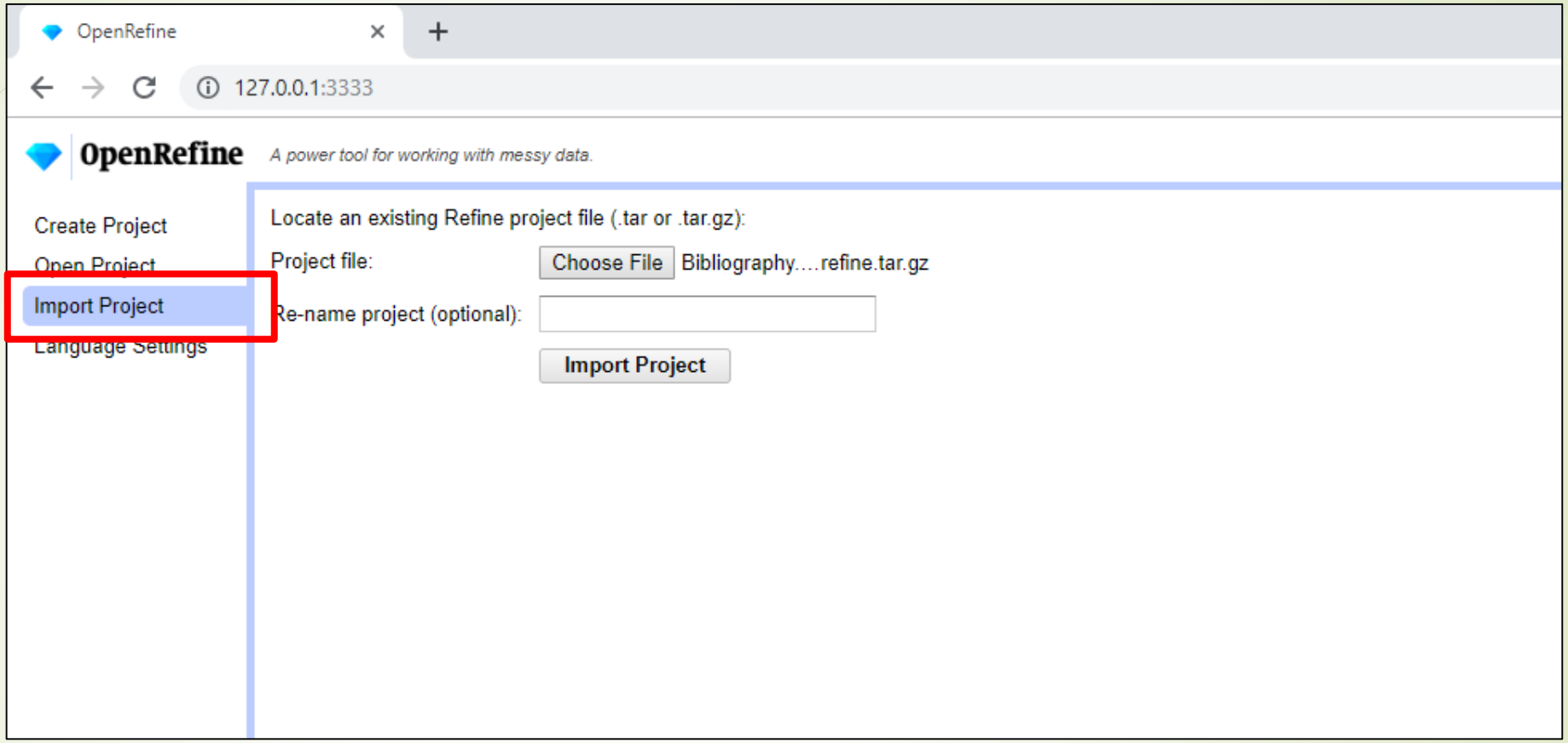

ibliographycsv - OpenRefine
Bibliography - OpenRefine
ILIAS für Lehre & Lernen – Univer
project=1988305065945
rmalink
112 rows
Show as: rows records
Show: 5 10 25 50 rows
All
BibliographyTyp
ISBN
Identifier
Author
Title
Journal
Volume
Number
Mo
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Transform...
Common transforms
Fill down
Blank down
Split multi-valued cells...
Join multi-valued cells...
Cluster and edit...
Replace
Trim leading and trailing whitespace
Collapse consecutive whitespace
Unescape HTML entities
To titlecase
To uppercase
To lowercase
To number
To date
To text
To null
To empty string
1. Article
May
2. Article 1050601491 10
using Bayesian networks
3. Article 100 Feldman, Michael and Friedler, Sorelle and Moeller, John and Scheidegger, Carlos and
Certifying and removing disparate impact


OpenRefine Bibliography
Permalink
Facet / Filter
Undo / Redo 54 / 54
Extract...
Apply...
Filter:
0. Create project
1. Mass edit 6 cells in column Month
2. Mass edit 2 cells in column Month
3. Mass edit 1 cells in column Month
4. Mass edit 1 cells in column Month
5. Mass edit 6 cells in column Month
6. Mass edit 1 cells in column Month
7. Mass edit 1 cells in column Month
8. Mass edit 1 cells in column Month
9. Mass edit 2 cells in column Month
10. Mass edit 1 cells in column Month
11. Mass edit 4 cells in column Month
12. Mass edit 1 cells in column Year
13. Text transform on 13 cells in column Year: jython:if value == None: return value if value == None: return value elif int(value) >=0 and int(value) <20: value = str(int(value)+2000) return value elif int(value) >19 and int(value) <99: value = str(int(value)+1900) return value else: return value


OpenRefine Bibliography
Permalink
Facet / Filter
Undo / Redo 47 / 54
Extract...
Apply...
Filter:
44. Edit single cell on row 93, column BibliographyType
45. Edit single cell on row 92, column BibliographyType
46. Edit single cell on row 99, column BibliographyType
47. Text transform on 5 cells in column BibliographyType: grel:if(cells["Year"].value == null,"Miscc",value)
48. Text transform on 24 cells in column BibliographyType: grel:value.replace("Article","Misc")
49. Mass edit 5 cells in column BibliographyType
50. Remove 1 rows
51. Edit single cell on row 77, column Publisher
52. Remove 1 rows
53. Text transform on 105 cells in column Identifier: value.toNumber()
54. Reorder rows


Open...
Export
Export project
Project data package
Tab-separated value
Comma-separated value
HTML table
Excel (.xls)
Excel 2007+ (.xlsx)
ODF spreadsheet
Custom tabular exporter...
SQL Exporter...
Templating...
Upload edits to Wikidata
Export to QuickStatements
Export schema
25 50 rows
Identifier
Author
Title
Journal
Volume
Number
1491
10
Mancuhan, Koray and Clifton, Chris
Combating discrimination using Bayesian networks
Artificial Intelligence and Law
22
2


Google
OpenRefine
127.0.0.1:3333
OpenRefine
A power tool for working with messy data.
Create Project
Open Project
Import Project
Language Settings
All
Last modified
Name
Tags
Subject
Description
About 2019-06-25 23:22 PM Bibliography
About 2019-06-25 23:21 PM bibliographycsv
About 2019-06-19 18:41 PM Bibliography
About 2019-06-19 18:41 PM Bibliography
About 2019-06-19 18:32 PM Bibliography
About 2019-06-19 18:20 PM Bibliography
About 2019-06-19 18:19 PM Bibliography
Version 3.1 [b90e413]


OpenRefine
127.0.0.1:3333
OpenRefine
A power tool for working with messy data.
Create Project
Open Project
Import Project
Language Settings
Locate an existing Refine project file (.tar or .tar.gz):
Project file:
Choose File
Bibliography....refine.tar.gz
Re-name project (optional):
Import Project

# Thank You