

Data Science 1 - SS 2020

19.05.2020

start um 13:15 Uhr

Dr. Karsten Tolle

Team Steckbriefe (Abgabe über OLAT)

Wird nicht benotet! Es geht darum zu verstehen, wer was macht, um:

- Die Teambildung zu unterstützen.
- Zu verstehen, wer Hilfe beim Definieren der Ziele und Finden von Datenquellen braucht.
- Die Mentoring-Treffen vorzubereiten.

Jeder der ein Projekt bearbeiten möchte (also CPs erhalten will), sollte eine Datei hochladen (oder erstellen).

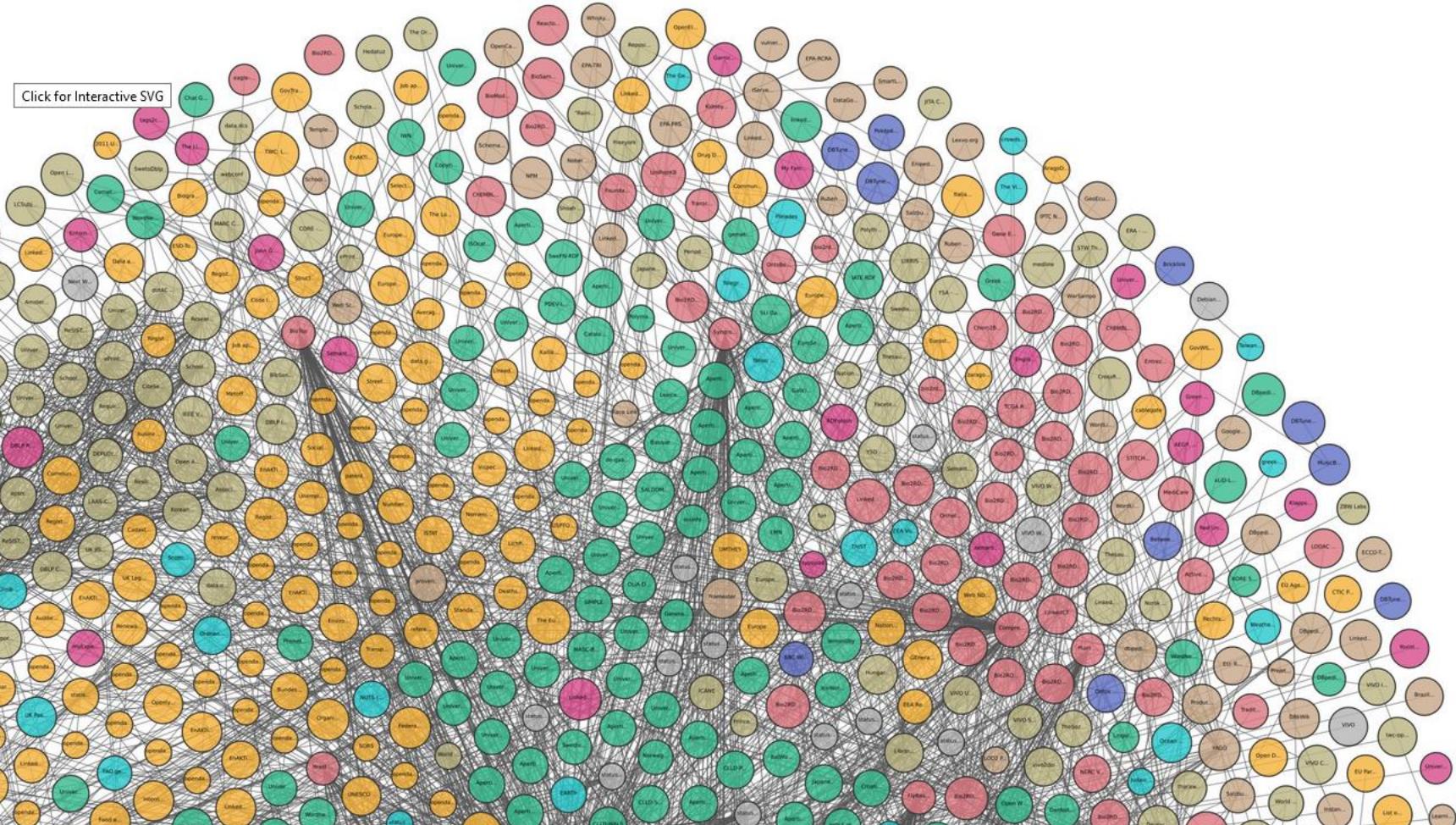
1. Falls ihr noch keinen Partner/Team habt, ladet bitte eine leere .txt-Datei mit dem Namen: ohneTeam.txt hoch.
2. Wer einen Partner/Team hat:
 - a) Wählt für Euer Team einen kurzen Namen (ein Wort)!
 - b) Einer vom Team (ihr entscheidet wer) lädt einen Steckbrief des Teams hoch (entsprechend der Vorlage unten). Die Datei soll dabei wie folgt benannt werden:
<Team-Name>_Steckbrief.txt
 - c) Alle anderen vom Team laden eine leere Datei hoch mit dem Namen: <Team-Name>.txt

**... aktuell 14 Teams (stand 18.05. 9:00 Uhr) eingetragen ...
Bitte Teams bis zum 22.05. (Freitag) eintragen!**

The Linked Open Data Cloud

Legend

- Cross Domain
- Geography
- Government
- Life Sciences
- Linguistics
- Media
- Publications
- Social Networking
- User Generated



<https://lod-cloud.net/clouds/lod-cloud.svc>

Publicly Available Datasets

- The University of California at Irvine Machine Learning Repository (UCI) hosts famous data
 - <http://archive.ics.uci.edu/ml/>
- Many governments host open data
 - <https://www.govdata.de/>
 - <http://data.gov.uk>
 - <http://data.gov>
 - <https://offenedaten.frankfurt.de>
- Kaggle Datasets
 - www.kaggle.com/
- Else
 - Linked Open Data Cloud: <https://lod-cloud.net/>
 - Open Knowledge Foundation Deutschland: <https://okfn.de/>
 - Nomisma.org: <http://nomisma.org/datasets>
 - Wikidata: <https://www.wikidata.org>
 - NASA: [http://data.nasa.gov/](http://data.nasa.gov)
 - AWS: <http://aws.amazon.com/publicdatasets/>

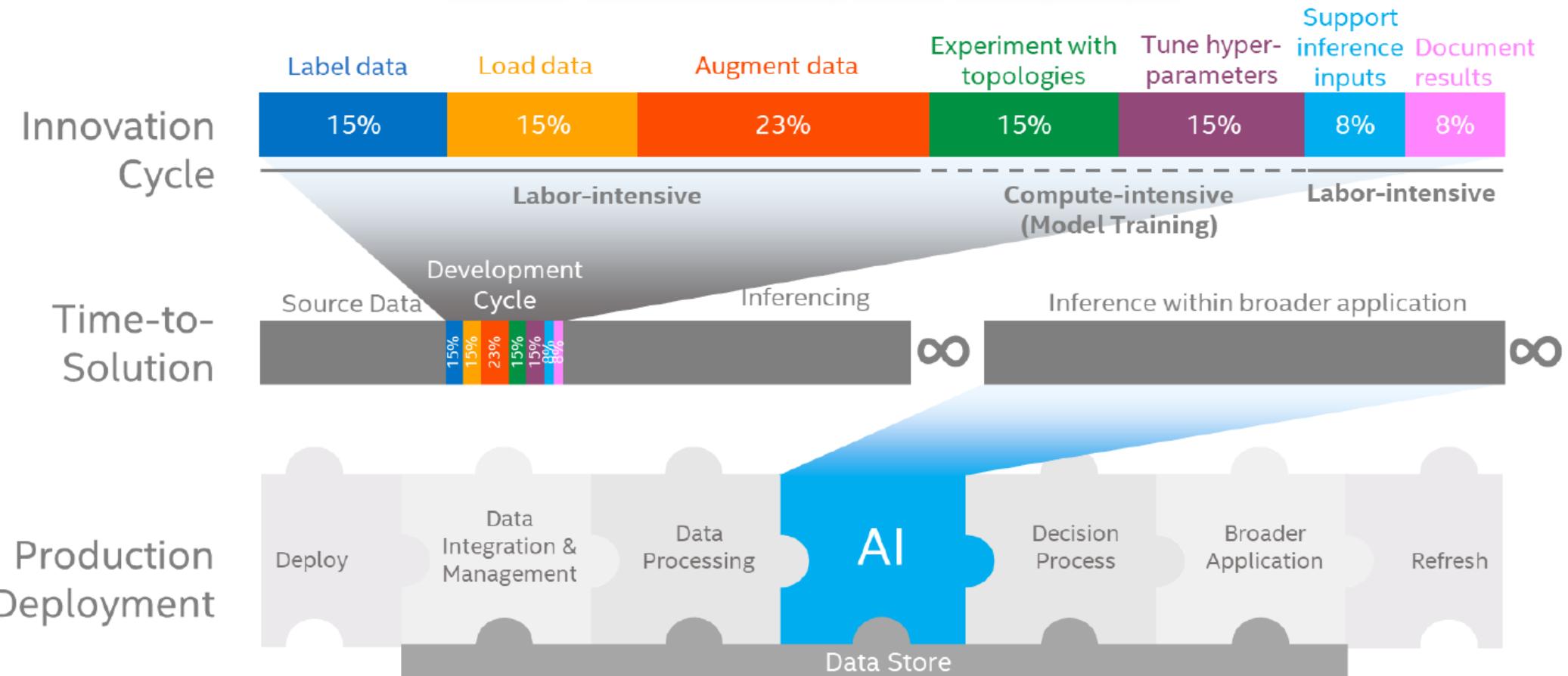
Unsupervised, Supervised, Semi-Supervised and Reinforcement Learning

... <https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>

Regression
Classification
Clustering
Decision Trees
Data Generation
Image Processing
Speech Processing
Natural Language Processing
Recommender Systems
Adversarial Networks
Reinforcement Learning

- Supervised: Input and Output (labels - aka *ground truth*)
- Unsupervised: Input is given NO output!
- Semi-Supervised: mixture of the two above
- Reinforcement Learning: no output needed, but rewards are given – typically agents taking actions (https://en.wikipedia.org/wiki/Reinforcement_learning)

DEEP LEARNING IN PRACTICE



Time-to-solution is more significant than time-to-train

Optimization Notice

Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.



Clustering

Clustering

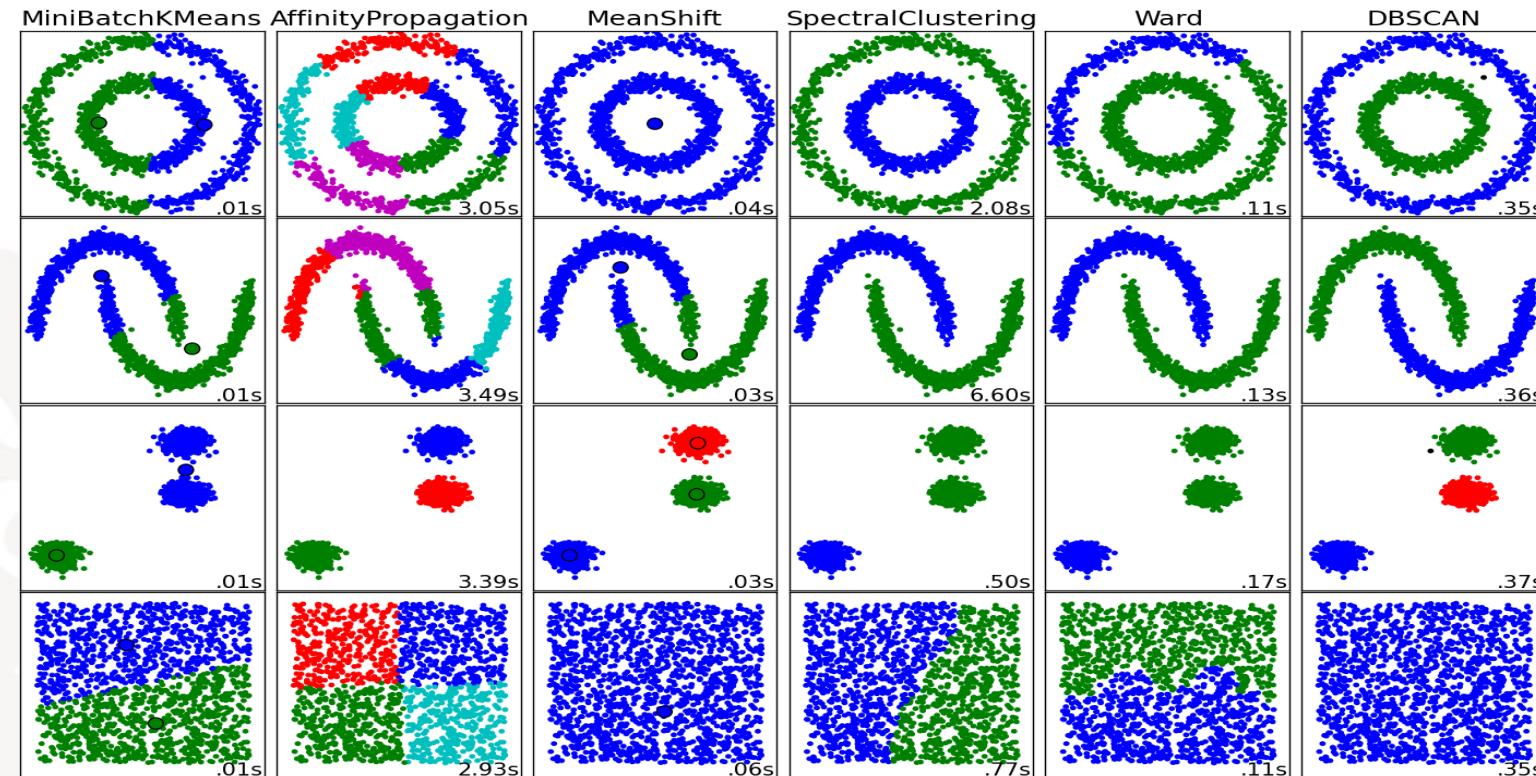
- One of Unsupervised Machine Learning Approaches
- Methods to identify possible groupings in a data-set
- Analyze the groupings automatically
- **cluster**: set of all data records that share a sub-set of the overall properties
- Infinitely many different possible clusterings (ways to cluster the data)
- Induced by the **similarity function** (some of which create the same output clusters)

Clustering Algorithms

- k-means clustering
- Hierarchical clustering
- Spectral clustering
- Density based clustering
- Graph-based clustering
- Fuzzy clustering
- Search-based clustering

Comparison of clustering algorithms

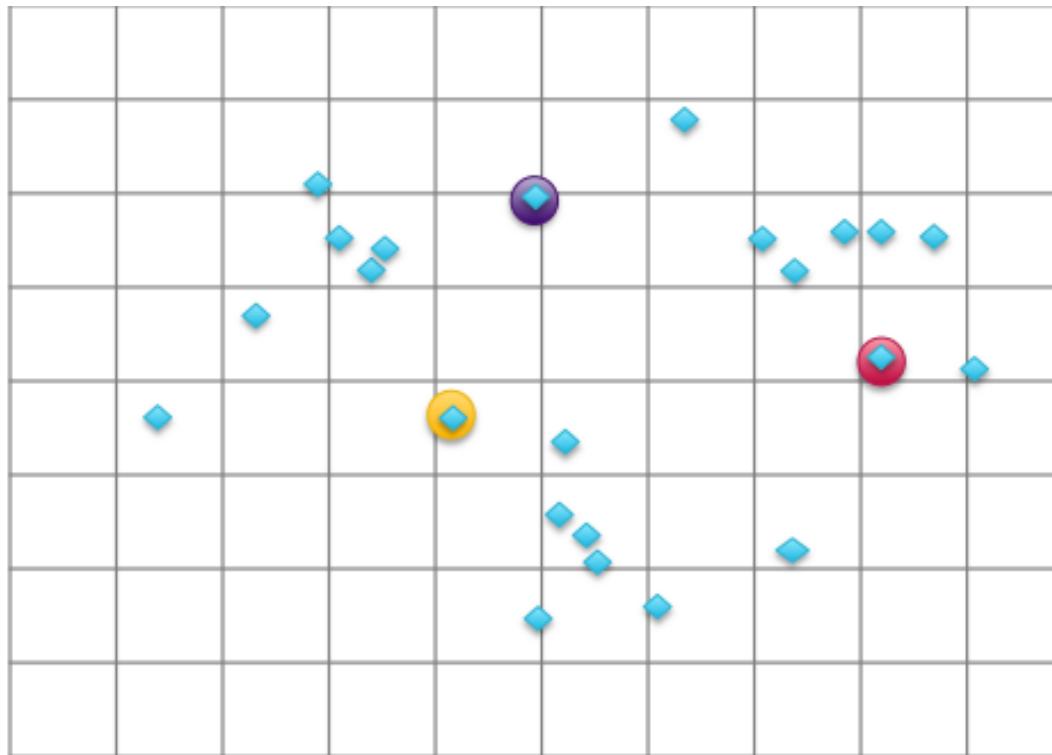
- The following image is from the Scikit-learn (<https://scikit-learn.org/stable/>) documentation and it demonstrates how the different algorithms compare to each other when executed on the same data:



k-means

- Simple and efficient; best known partitional clustering algorithm
- Sensitive to choice of seed clusters (initial centroids)
- For every point in the dataset
 - Measure distance to each centroid
 - Assign point to centroid with lowest distance
- For every cluster
 - Calculate the mean position of all points
- Repeat until convergence criteria is met
 - No points change clusters
 - Centroid changes within a short distance
 - Fixed number of iterations

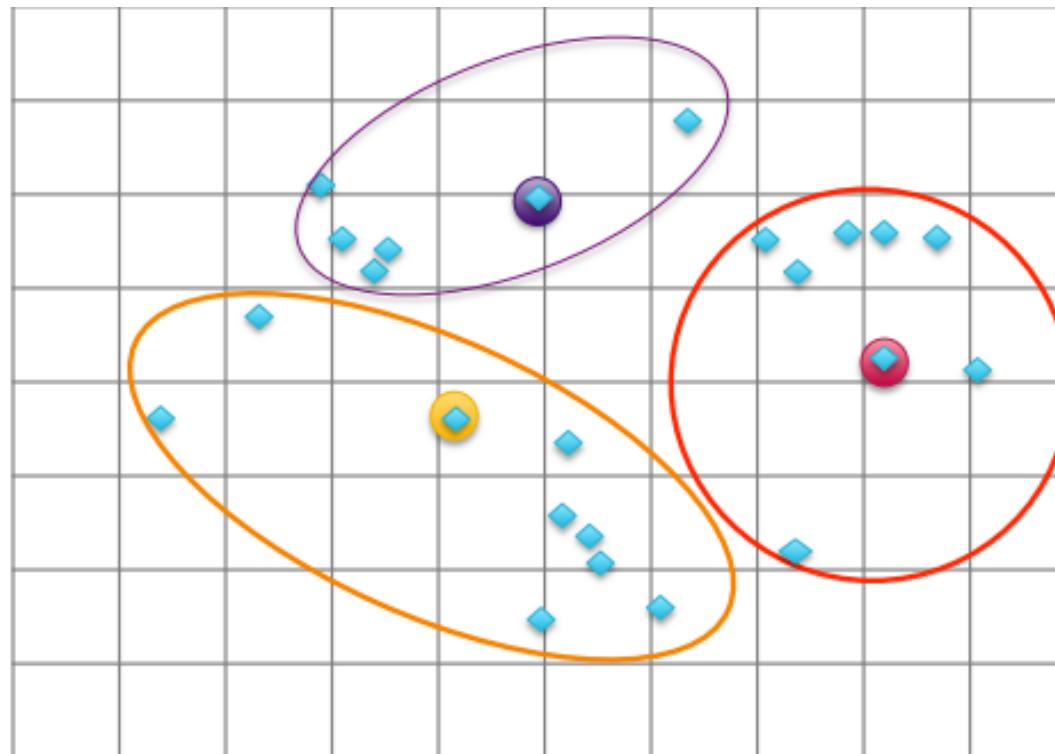
k-means demonstration



1. Choose K random points as starting centers

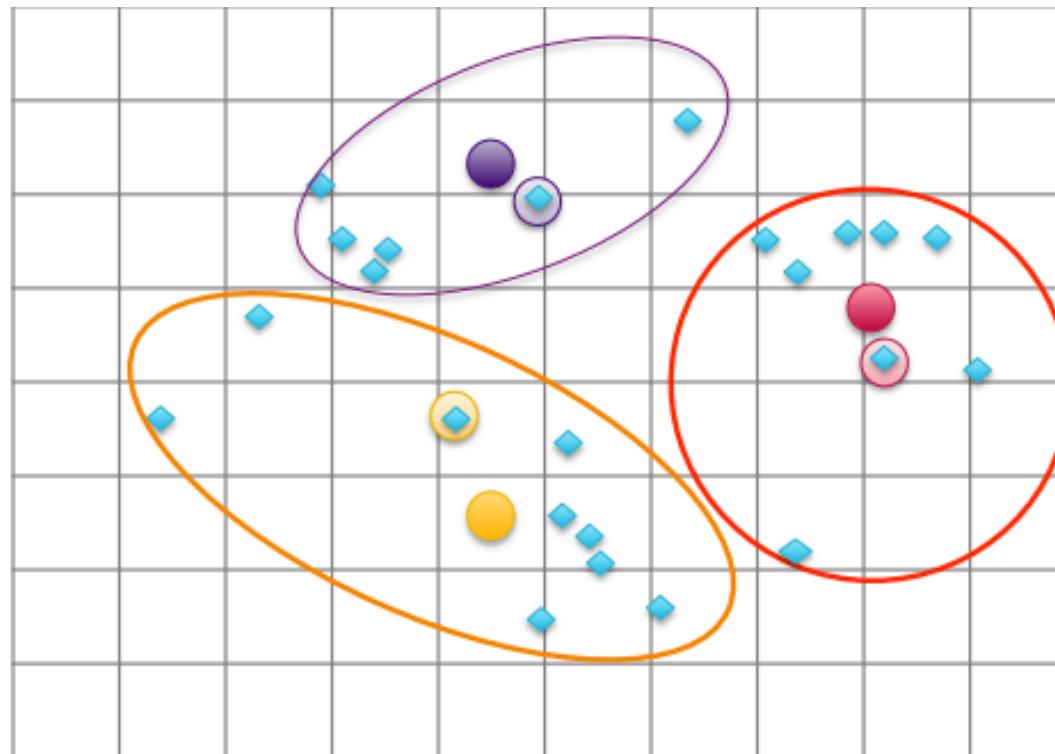
ref: Cloudera's academia teaching materials

k-means demonstration



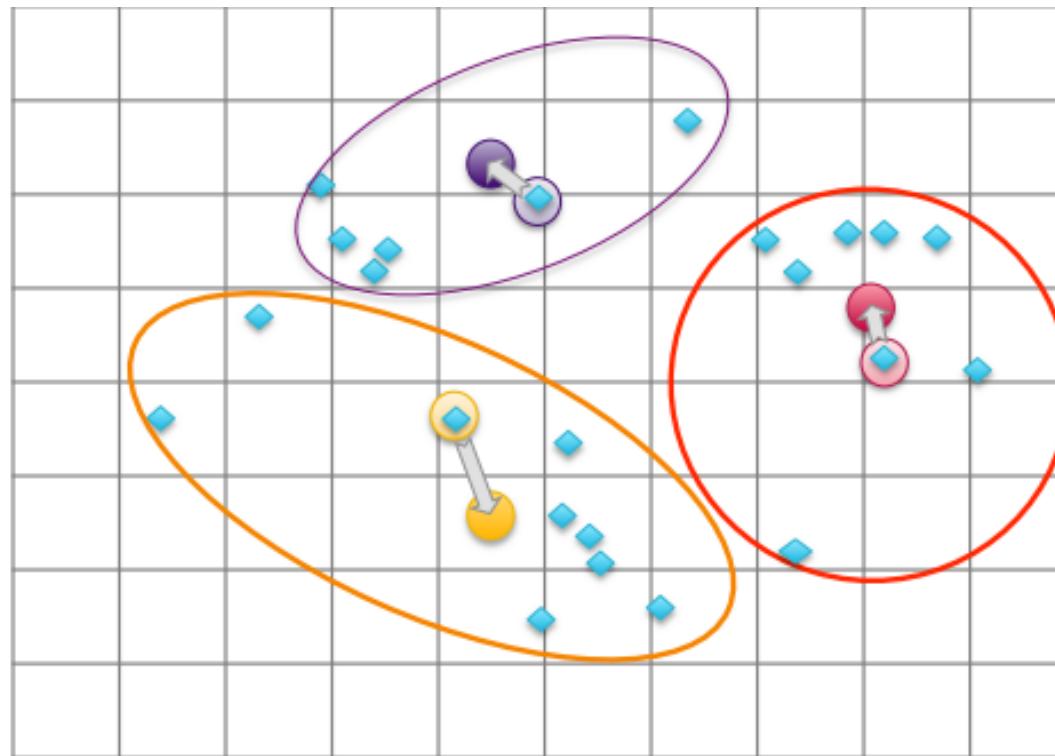
1. Choose K random points as starting centers
2. **Find all points closest to each center**

k-means demonstration



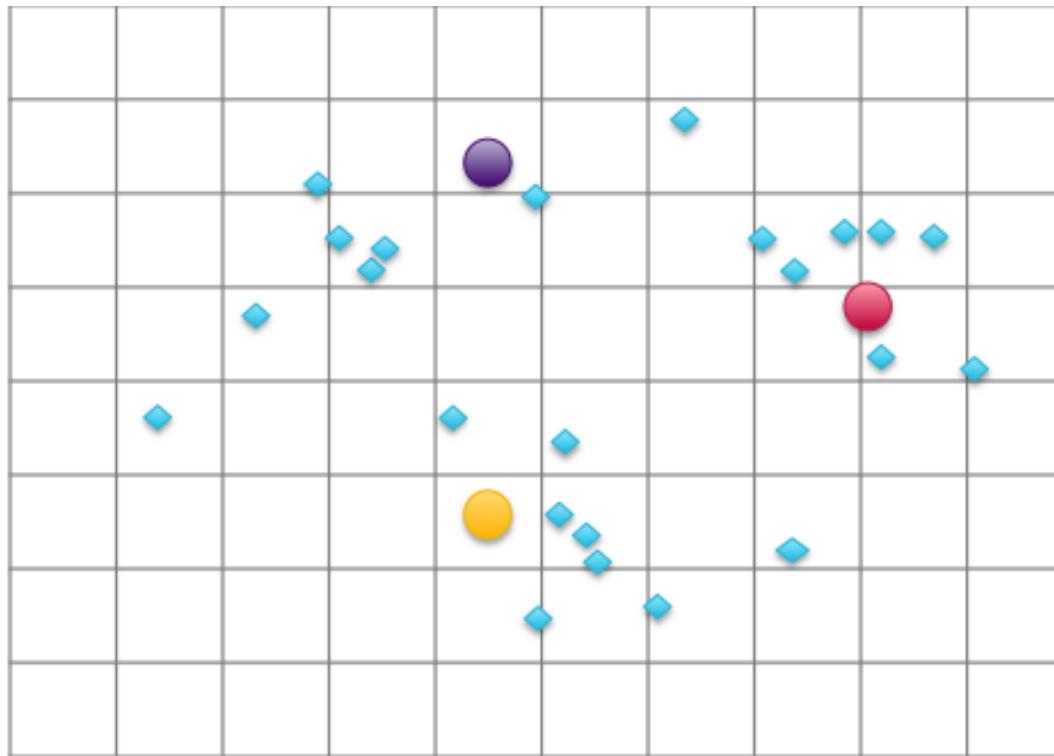
1. Choose K random points as starting centers
2. Find all points closest to each center
3. **Find the center (mean) of each cluster**

k-means demonstration



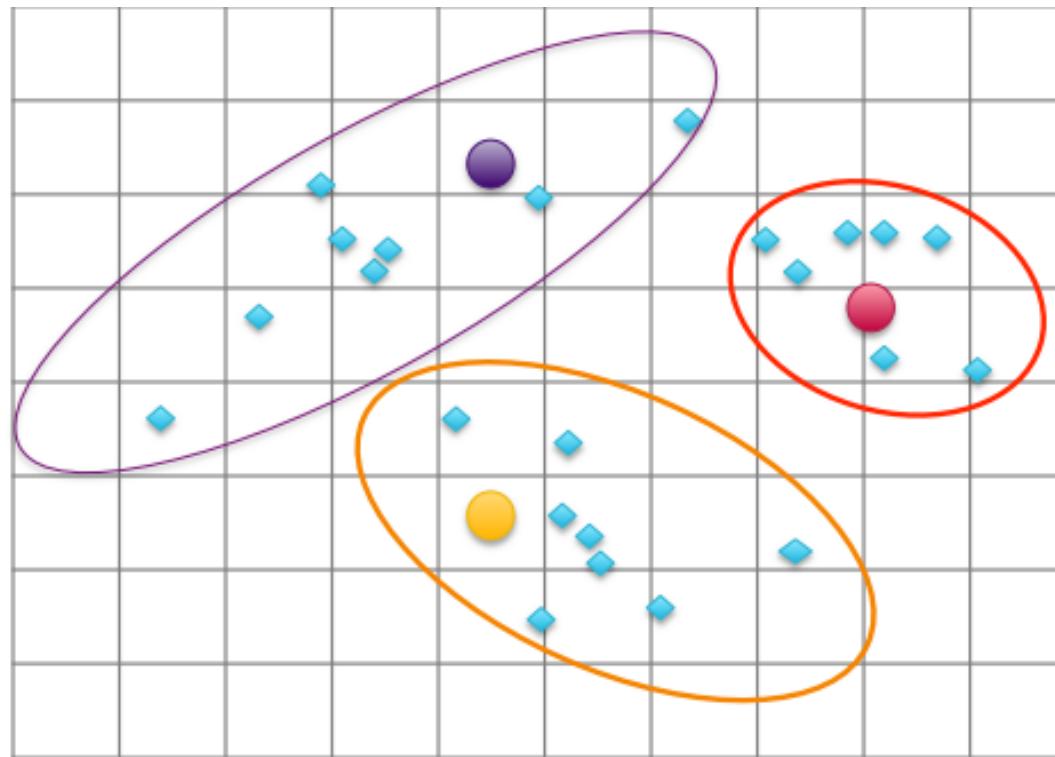
1. Choose K random points as starting centers
2. Find all points closest to each center
3. Find the center (mean) of each cluster
4. If the centers changed, iterate again

k-means demonstration



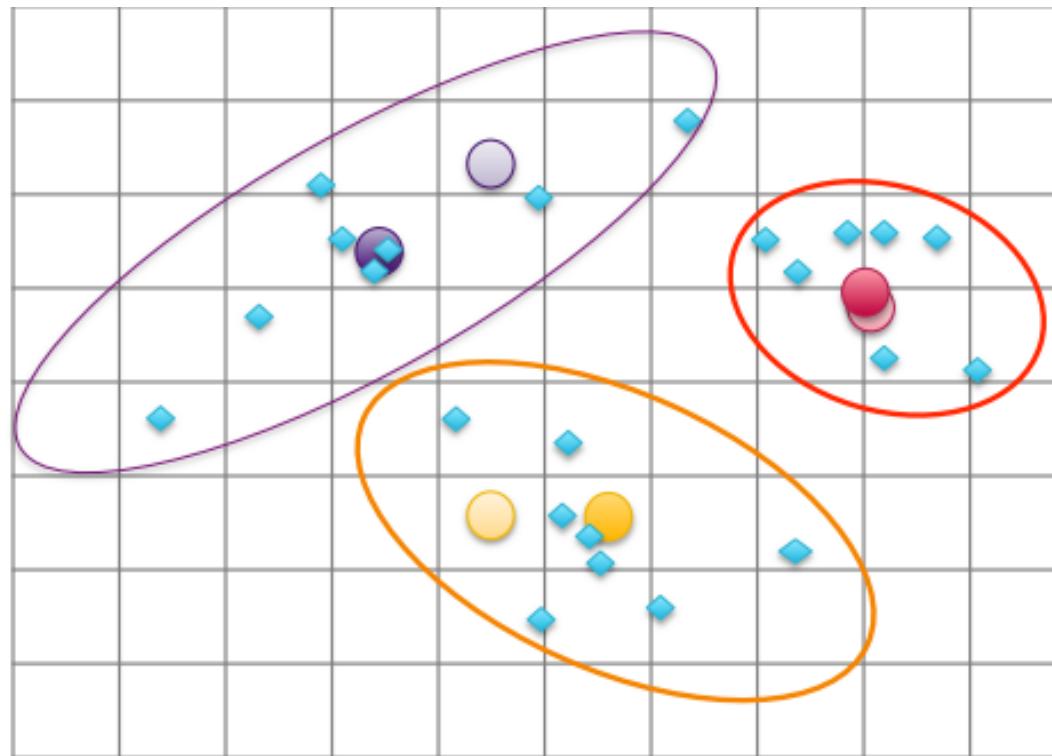
1. Choose K random points as starting centers
2. Find all points closest to each center
3. Find the center (mean) of each cluster
4. If the centers changed, iterate again

k-means demonstration



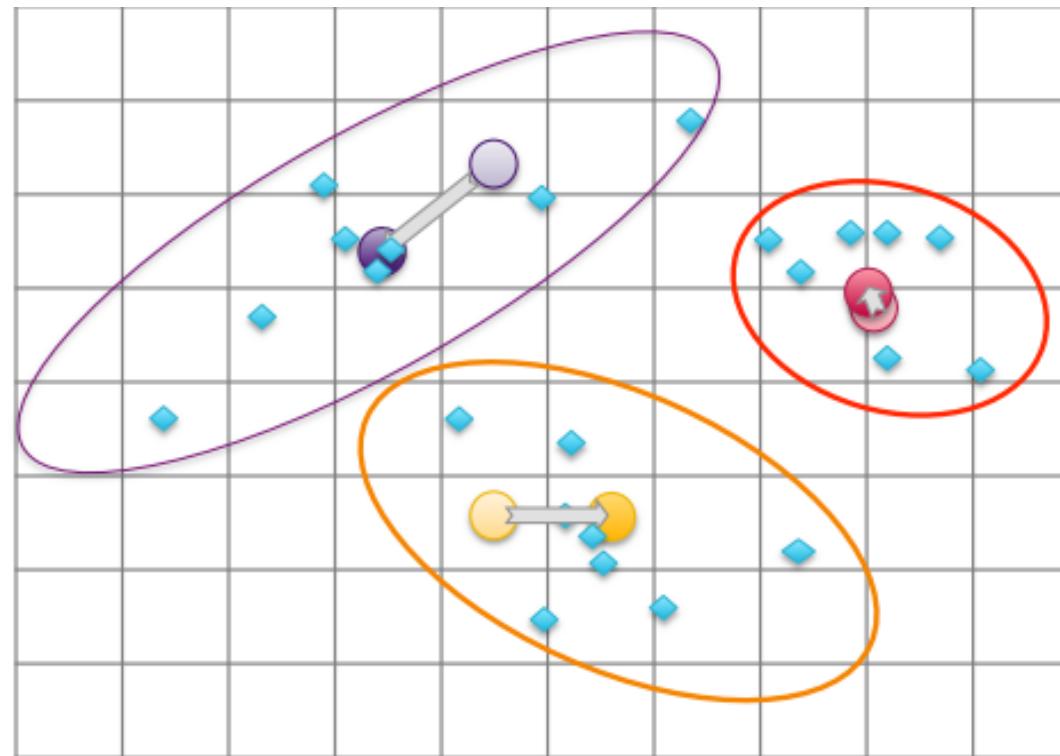
1. Choose K random points as starting centers
2. **Find all points closest to each center**
3. Find the center (mean) of each cluster
4. If the centers changed, iterate again

k-means demonstration



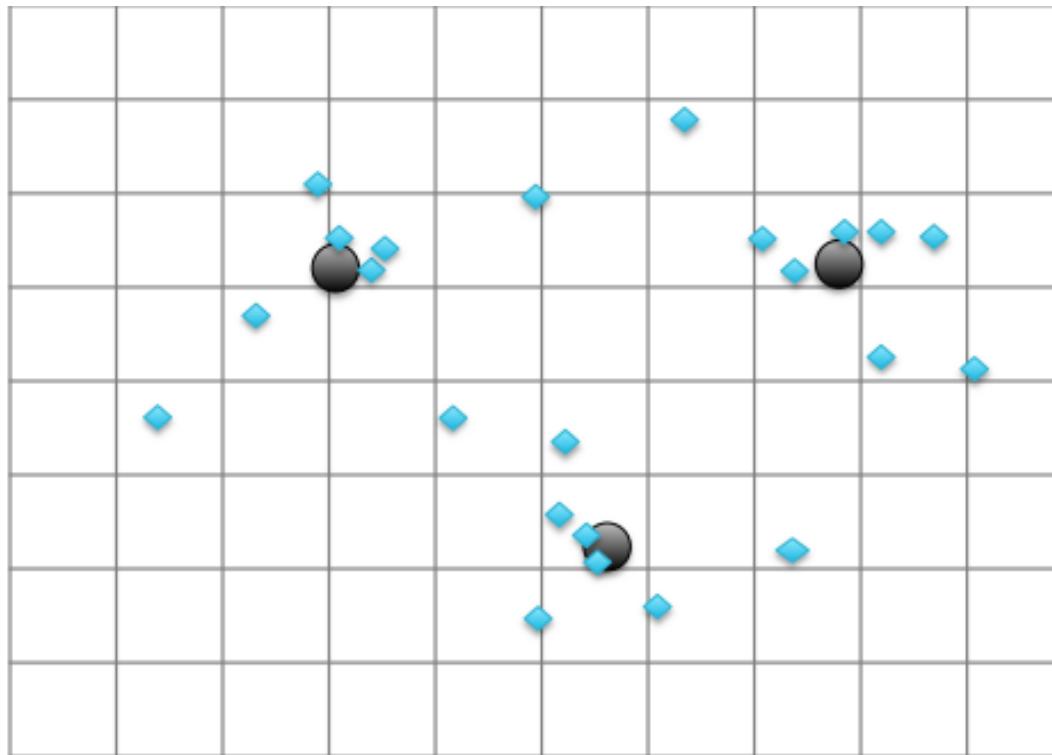
1. Choose K random points as starting centers
2. Find all points closest to each center
3. **Find the center (mean) of each cluster**
4. If the centers changed, iterate again

k-means demonstration



1. Choose K random points as starting centers
2. Find all points closest to each center
3. Find the center (mean) of each cluster
4. If the centers changed, iterate again

k-means demonstration



1. Choose K random points as starting centers
2. Find all points closest to each center
3. Find the center (mean) of each cluster
4. If the centers changed, iterate again
- ...
5. Done!

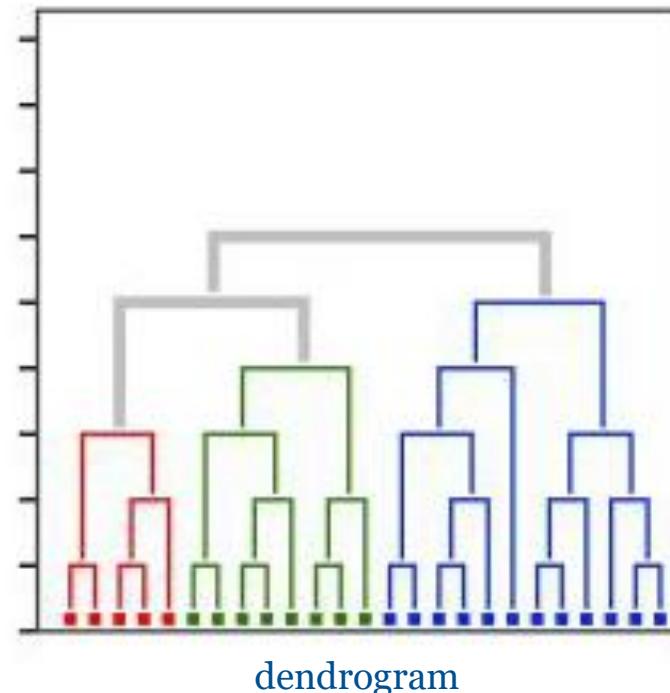
... how to find the right number of clusters?

https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

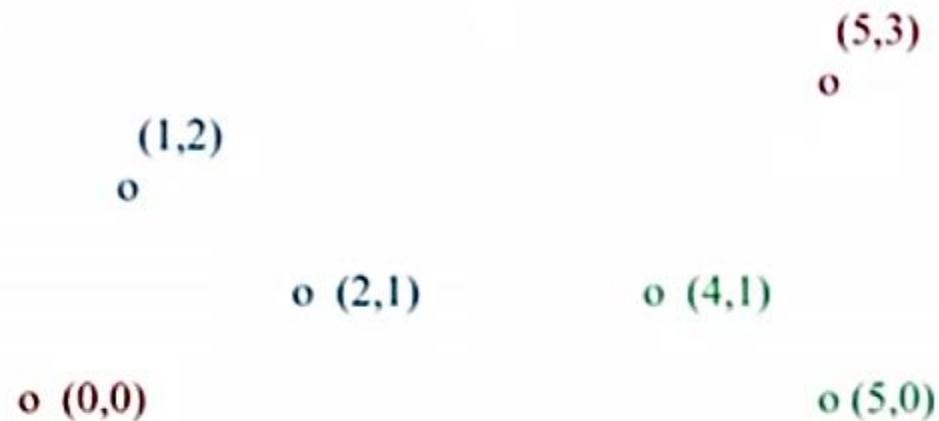


Hierarchical clustering demonstration

- Easy to understand
- A dendrogram can be used to depict this hierarchical clustering process
- Bottom-up (=agglomerative) method: iterate over each data record
- Compare two data records using a similarity function, and join most similar pair



Hierarchical clustering demonstration



Data:

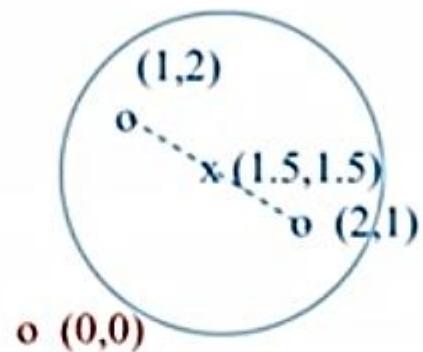
- o ... data point
- x ... centroid



Dendrogram

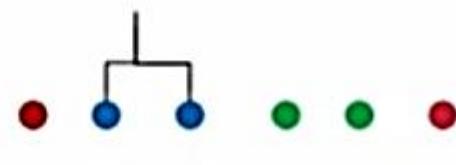
16

Hierarchical clustering demonstration



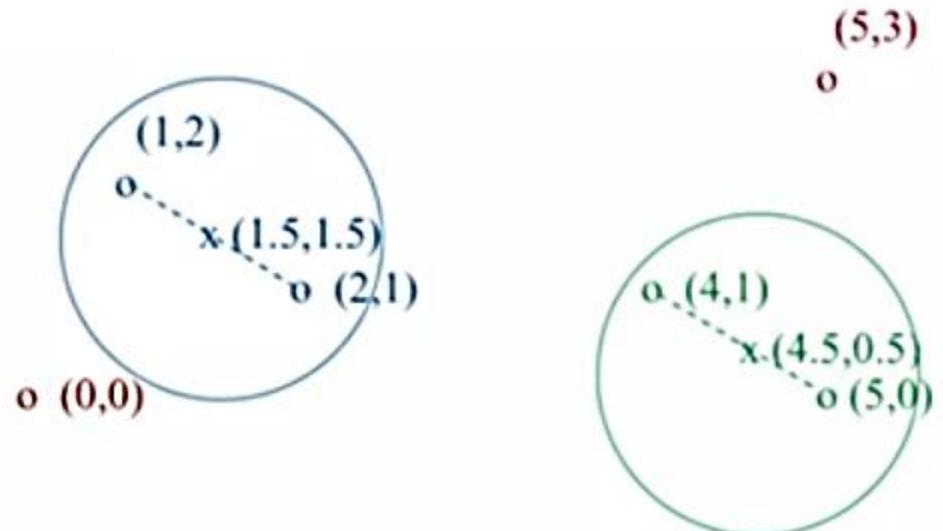
$\circ (5,3)$
 $\circ (4,1)$
 $\circ (5,0)$

Data:
 \circ ... data point
 x ... centroid

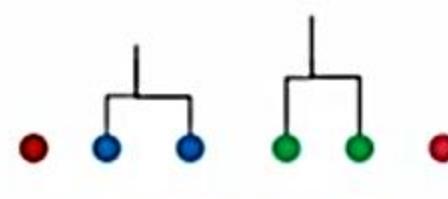


Dendrogram

Hierarchical clustering demonstration

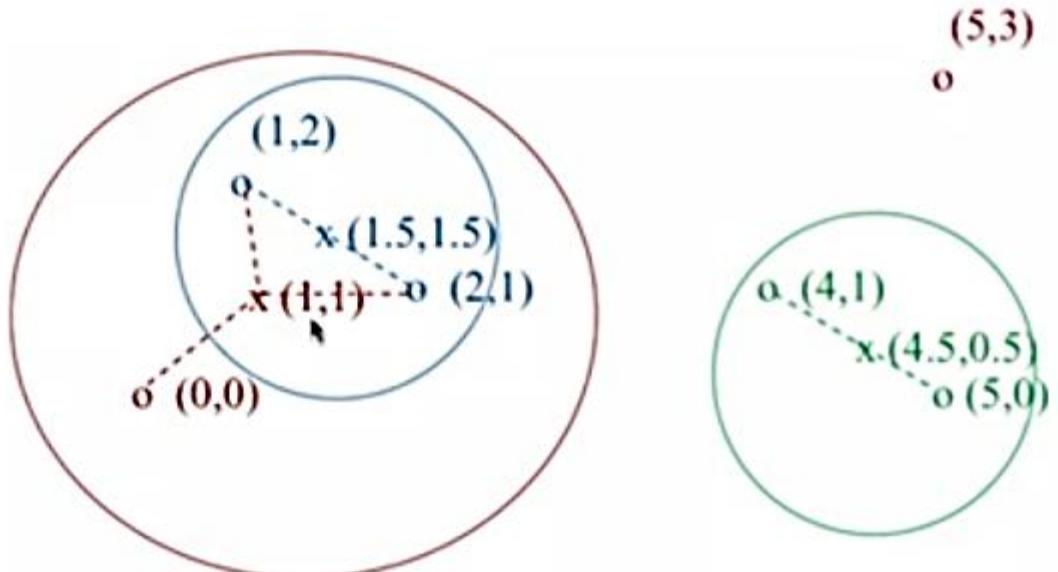


Data:
 o ... data point
 x ... centroid

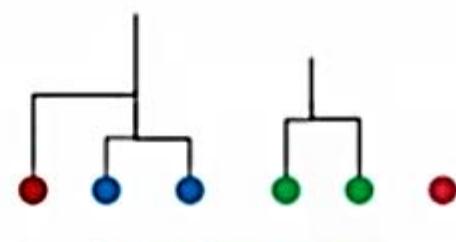


Dendrogram

Hierarchical clustering demonstration

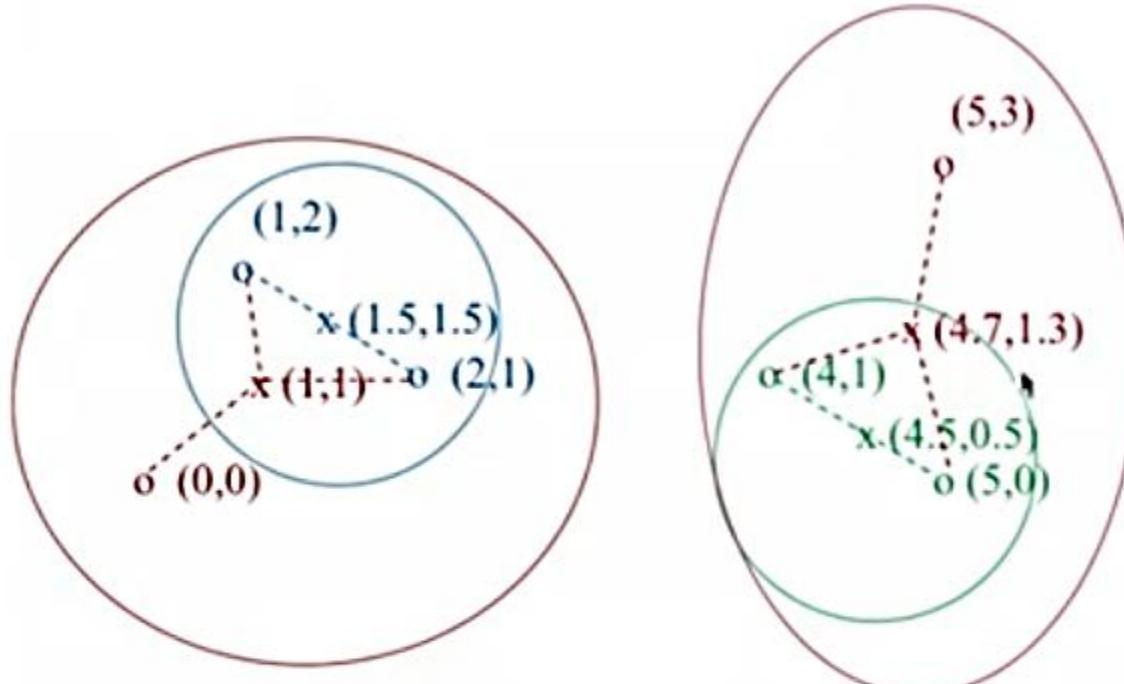


Data:
 o ... data point
 x ... centroid

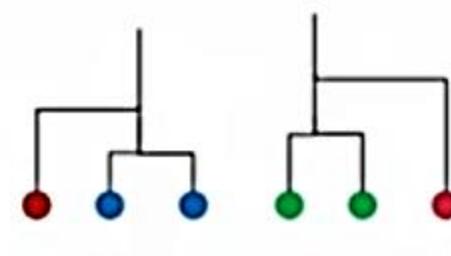


Dendrogram

Hierarchical clustering demonstration



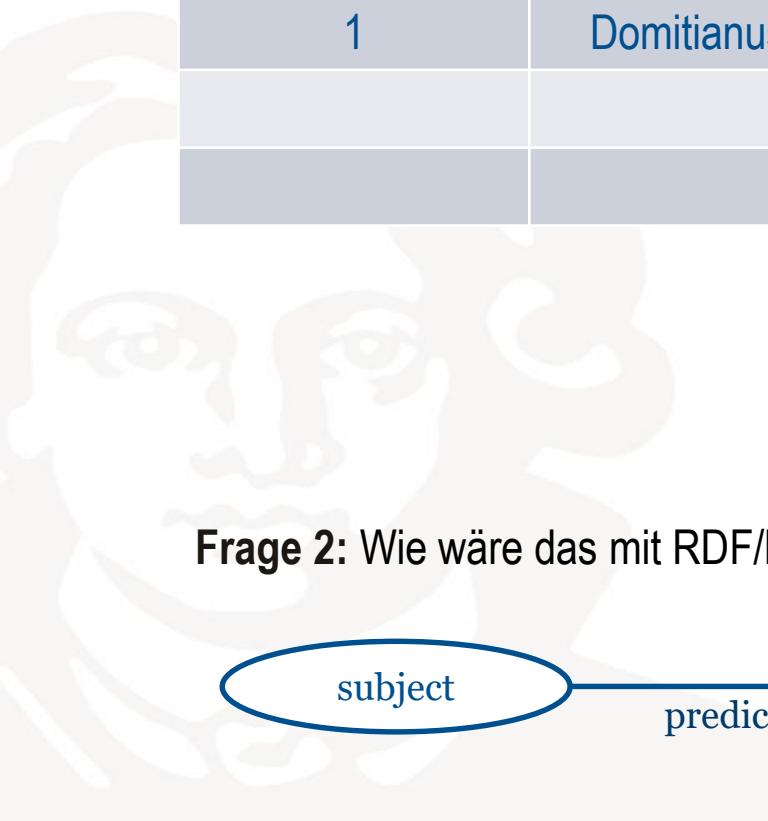
Data:
 o ... data point
 x ... centroid



Dendrogram

Unterschied zwischen: relationalen Daten und LOD

Frage 1: Wie merge ich zwei Datenbank-Tabellen aus verschiedenen Datenbanken, welche den gleichen Entitätstypen repräsentieren und die gleichen Attribute haben?
 Z.B. Prägeherr (Person)



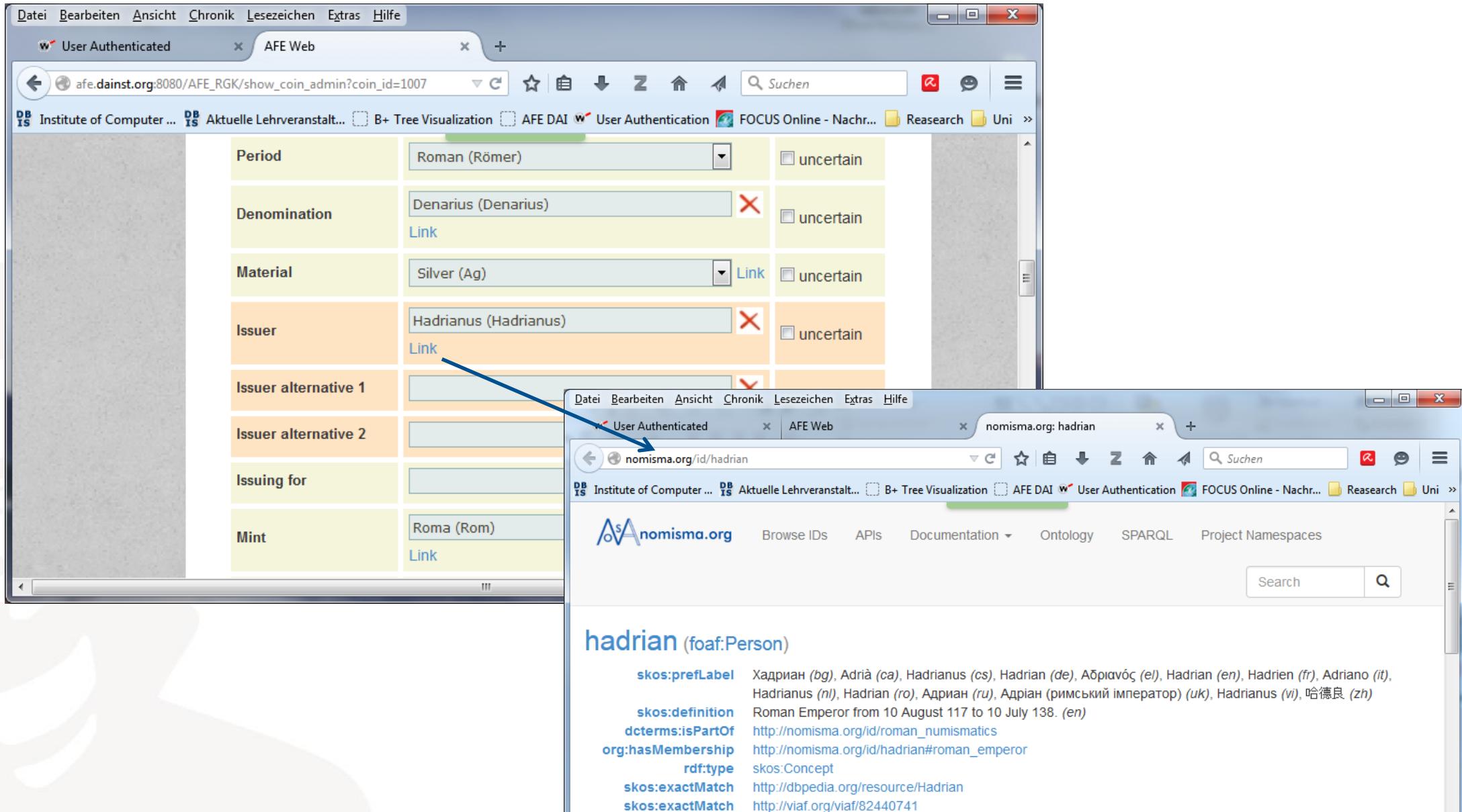
ID	Name	active_from	active_to
1	Domitianus II	271	271

ID	Name	active_from	active_to
14	Domitianus II.	270	271

Frage 2: Wie wäre das mit RDF/LOD?



Zero Star – existierende LOD nutzen



The screenshot illustrates the integration of LOD data from the nomisma.org ontology into a local application. The top window shows a form for a coin record, specifically for a Denarius of Hadrian. The bottom window shows the detailed description of the person 'hadrian' from the nomisma.org ontology.

Top Window (Local Application):

- Period:** Roman (Römer)
- Denomination:** Denarius (Denarius) [Link](#)
- Material:** Silver (Ag) [Link](#)
- Issuer:** Hadrianus (Hadrianus) [Link](#)
- Issuer alternative 1:** (empty)
- Issuer alternative 2:** (empty)
- Issuing for:** (empty)
- Mint:** Roma (Rom) [Link](#)

A blue arrow points from the 'Issuer' field in the top window to the 'nomisma.org: hadrian' entry in the bottom window, indicating a link or lookup operation.

Bottom Window (nomisma.org):

hadrian (foaf:Person)

Properties and Values:

- skos:prefLabel:** Хадриан (bg), Adrià (ca), Hadrianus (cs), Hadrian (de), Αδριανός (el), Hadrian (en), Hadrien (fr), Adriano (it), Hadrianus (nl), Hadrian (ro), Adrián (ru), Адриан (uk), Hadrianus (vi), 哈德良 (zh)
- skos:definition:** Roman Emperor from 10 August 117 to 10 July 138. (en)
- dcterms:isPartOf:** http://nomisma.org/id/roman_numismatics
- org:hasMembership:** http://nomisma.org/id/hadrian#roman_emperor
- rdf:type:** skos:Concept
- skos:exactMatch:** <http://dbpedia.org/resource/Hadrian>
- skos:exactMatch:** <http://viaf.org/viaf/82440741>

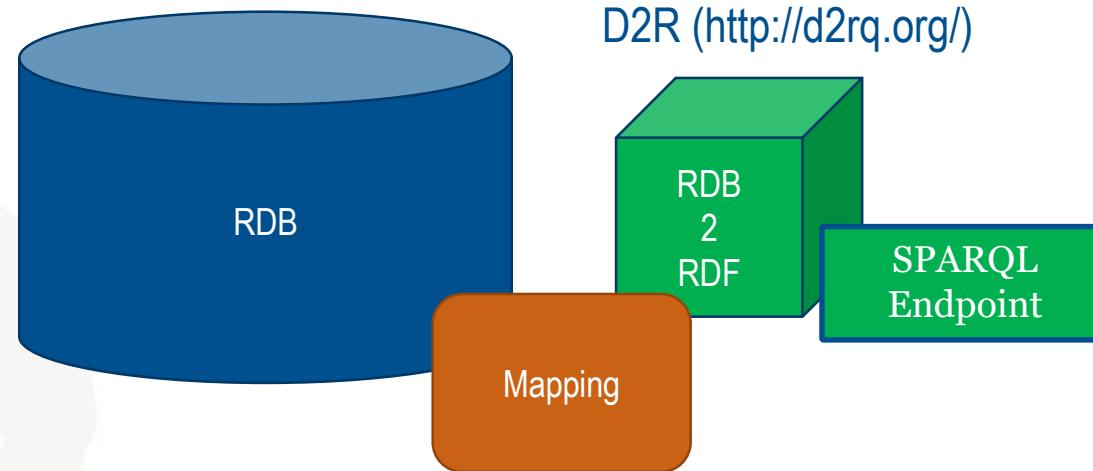
Zero Star – existierende LOD nutzen

id	name	Nomisma	active from	active to
158	Magnentius	magnentius	350	353
161	Iulianus II.	julian_the_apostate	355	363
162	Valentinianus I.	valentinian_i	364	375
163	Valens	valens		
165	Gratianus	gratian	367	383
167	Valentinianus II.	valentinian_ii	375	392
168	Valentinianische Dyna...		364	392
169	Magnus Maximus	magnus_maximus	383	388
170	Honorius	honorius	393	423
171	Theodosianische Dynastie			
173	Valentinianus III.	valentinian_iii	425	455
174	Iustinianus I.	justinian_i	527	565
182	Macrinus	macrinus	217	218
183	Flavisch		69	96
187	Maximinus II.	maximinus_daia	305	313
189	Licinius I.	licinius	308	324
194	Constantius I.	constantius_chlorus	293	306
200	Domitianus II.	domitianus	271	271
202	Aemilianus	aemilianus	253	253
203	Allectus	allectus	293	296
205	Anastasius I.	anastasius	461	518
207	Antehmius	anthemius	467	472
208	Arcadius	arcadius	383	408
209	Avitus	avitus	455	456

<http://nomisma.org/id/valens>

Existiert nicht unter Nomisma.org

Von einer SQL-Datenbank zu 4-Star LOD (oder mehr)



RDB 2 RDF: <http://www.w3.org/2001/sw/rdb2rdf/wiki/Implementations>

D2R (<http://d2rq.org/>)

Description of http://afe.dainst.org:8080/d2rq/resource/AFE_coin_1678

Resource URI: http://afe.dainst.org:8080/d2rq/resource/AFE_coin_1678

[Home](#) | [All coinfind](#)

Property	Value
skos:exactMatch	<http://afe.dainst.org:8080/AFE_RGK/show_coin?coin_id=1678>
nmo:hasDenomination	nm:denarius
nmo:hasEndDate	75 (xsd:string)
nmo:hasFindspot	db:AFE_place_131
nmo:hasIssuer	nm:vespasian
nmo:hasMaterial	nm:ar
nmo:hasMint	nm:rome
nmo:hasObjectType	nm:coin
nmo:hasStartDate	75 (xsd:string)
nmo:hasTypeSeriesItem	<http://numismatics.org/ocre/id/nc.2_1(2).ves.777>
rdf:type	nmo:NumismaticObject

The server is configured to display only a limited number of values (limit per property bridge: 100).

Metadata

<http://afe.dainst.org:8080/d2rq/data/AFE_coin_1678>
dc:date 2015-03-26T15:39:47.027Z
prv:containedBy <http://afe.dainst.org:8080/d2rq/dataset>
void:inDataset <http://afe.dainst.org:8080/d2rq/dataset>
rdf:type prv:Dataitem
rdf:type foaf:Document

Applications Item 195.37.32.42/coin?coin_id=1678 80% Suchen

DEUTSCHE
ARCHÄOLOGISCHE
INSTITUTION
RÖMISCHE-KOMMISSION

Antike Fundmünzen in Europa

Home Suche Über uns

ID: 1678

Picture AFE_22_001678.jpg
Originator: Marianne Romisch, Goethe Universität
License:

1678 Alle hervorheben Groß-/Kleinschreibung Akzente Ganze Wörter 1 von 2 Übereinstimmungen Das Seitenende wurde erreicht, Suche vom Seitenanfang fortgesetzt

Maschinen verstehbare Schnittstelle

Probleme ...

Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe

X 18013086 +

vif.org/viaf/18013086/#Augustus_Römisches_Reich,_Kaiser_v63-14

Suchen

Gaius Octavius, imperatore romano, 63 a.C.-14 d.C.

Augustus, Gaius Iulius Caesar Octavianus, imperatore romano, 63 a.C.-14 d.C.

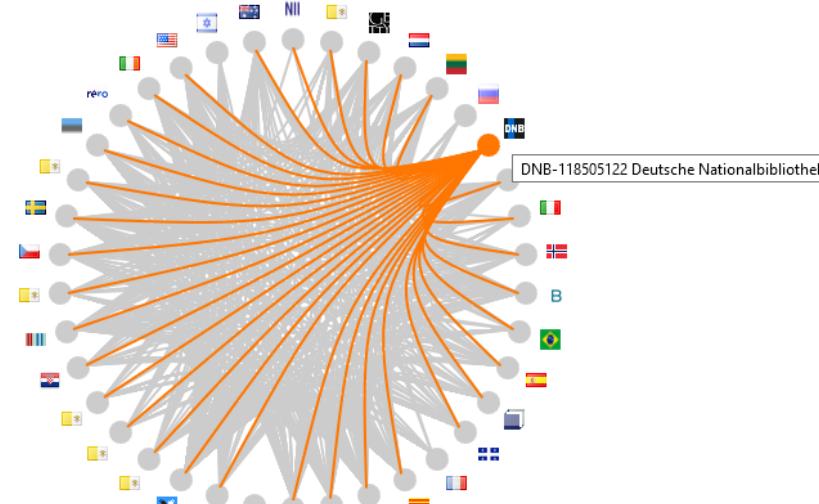
Octavius Caesar, imperatore romano, 63 a.C.-14 d.C.

VIAF ID: 18013086 (Person)

Permalink: http://vif.org/viaf/18013086

Vorzugsbezeichnungen

- 100 0 _ta Augusto_(cesar rzymski ; td 63 p.n.e.-14)
- 100 0 _ta Augusto_(cesar rzymski ; td 63 p.n.e.-14),
- 100 0 _ta Augusto_tc emperador de Roma ; td 63 aC-14 dC
- 200 _ | ta Auguste_tc empereur romain tf 0063 av. J.-C.-0014
- 100 0 _ta Auguste_tc empereur romain, td 63 av. J.-C.-14 apr. J.-C.
- 100 0 _ta Auguste_tc empereur romain
- 100 0 0 ta Augusto_tc Emperador de Roma
- 100 1 _ta Augusto_tc Imperador de Roma, td 63 A.C.-14 D.C.
- B 100 0 _ta Augustus Caesar_tc romersk keiser td 63 f.Kr.-14 e.Kr.
- H 100 0 _ta Augustus Caesar_tc romersk keiser td 63 f.Kr.-14 e.Kr.
- 200 _ 1 ta Augustus_tb ,Gaius Iulius Caesar Octavianus
- 100 0 _ta Augustus_tc Römisches Reich, Kaiser td v63-14
- DNB 100 0 _ta Augustus_tc Römisches Reich, Kaiser td v63-14
- 200 _ 0 ta Augustus_tc император римский tf 63 до н.э.- 14 н.э.
- 200 _ 0 ta Augustus_tf 63 pr.Kr.-14 po Kr._tc Romos imperatorius



Deutsche Nationalbibliothek

Die OCLC-Webseiten speichern Cookies auf Ihrem Gerät, um die Benutzerfreundlichkeit zu verbessern. Lesen Sie unsere Cookie-Richtlinie, um mehr zu erfahren.[Cookie-Richtlinie](#)

Akzeptieren

1678 Alle hervorheben Groß-/Kleinschreibung Akzente Ganze Wörter 1 von 2 Übereinstimmungen Das Seitenende wurde erreicht, Suche vom Seitenanfang fortgesetzt

... nichts ist sicherer als Veränderung!

Amtlicher Gemeindeschlüssel – dient zum Identifizieren von Bundesländern, Kreisen, Gemeinden
(https://de.wikipedia.org/wiki/Amtlicher_Gemeindeschlüssel)

Gebietsreform: z.B. https://de.wikipedia.org/wiki/Kreisgebietsreform_Mecklenburg-Vorpommern_2011

Auswirkungen und Umgang?

Wikidata Anfrage SPARQL

The screenshot shows the Wikidata Query Service interface at <https://query.wikidata.org>. The query window displays the following SPARQL code:

```
1 SELECT * WHERE { SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". } }
2 LIMIT 100
```

The interface includes a sidebar with various icons for editing and visualizing data, and a bottom bar with options like "Code", "Herunterladen", and "Link".



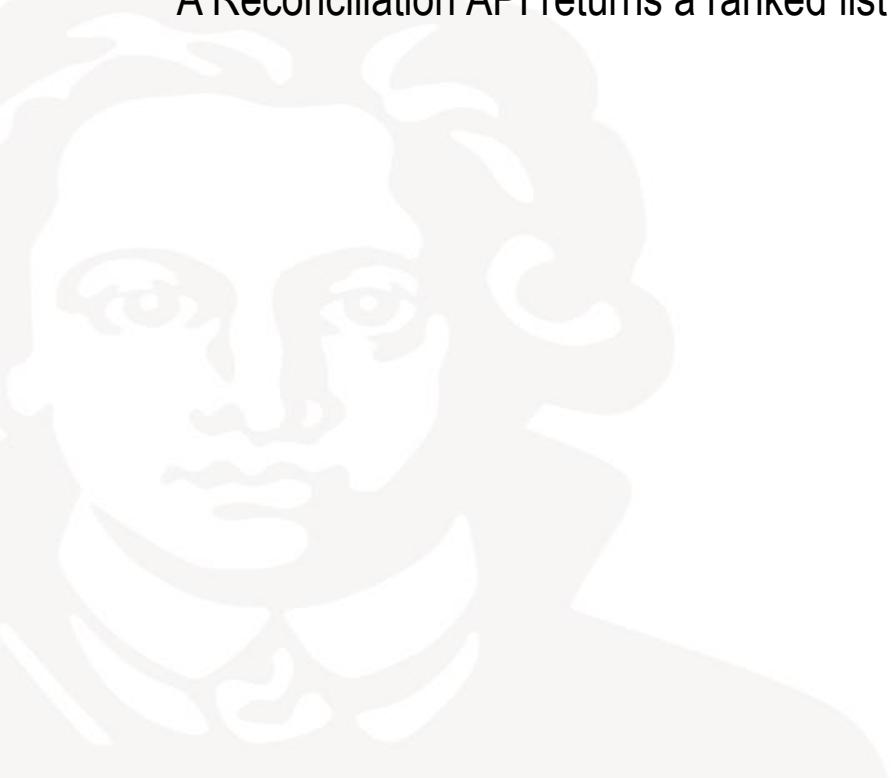
Openrefine ...

1. Basics

Home, Download, Documentation and Training: <http://openrefine.org/>

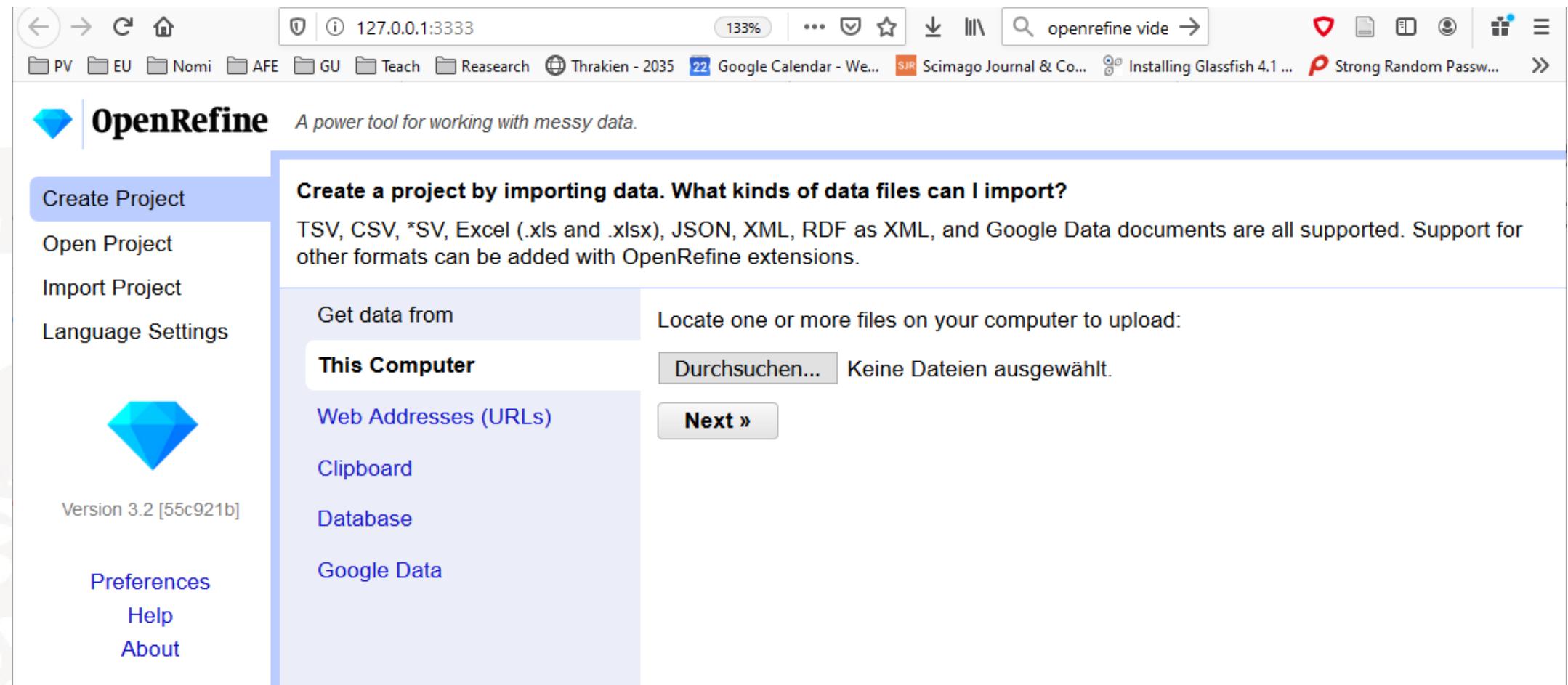
2. Reconciliation

A Reconciliation API returns a ranked list of potential entities matching input labels



<http://127.0.0.1:3333/>

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.



The screenshot shows a web browser window with the URL 127.0.0.1:3333 in the address bar. The page is titled "OpenRefine" with the subtitle "A power tool for working with messy data." A sidebar on the left contains links for "Create Project", "Open Project", "Import Project", "Language Settings", and icons for "Version 3.2 [55c921b]", "Preferences", "Help", and "About". The main content area is titled "Create a project by importing data. What kinds of data files can I import?". It explains that TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are supported, with a note that other formats can be added with OpenRefine extensions. Below this, there is a section titled "Get data from" with options: "This Computer" (selected), "Web Addresses (URLs)", "Clipboard", "Database", and "Google Data". A button labeled "Next »" is visible. A message "Locate one or more files on your computer to upload:" is displayed above a file selection input field which shows "Keine Dateien ausgewählt." and a "Durchsuchen..." button.

Reconciliation API

Reconciliation API helps to find matches between your data and existing data

Once the matches are found, you can enrich your data with extra information

Existing systems supporting the Reconciliation API:

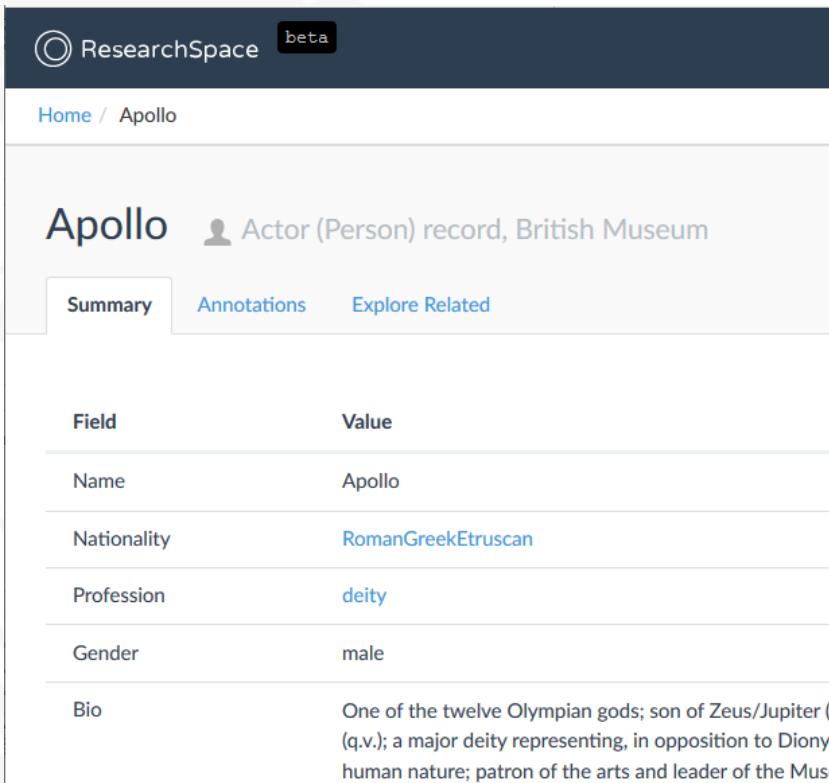
- Wikidata (www.wikidata.org)
- Getty (<https://www.getty.edu/research/tools/vocabularies/obtain/openrefine.html>)
https://www.getty.edu/research/tools/vocabularies/obtain/getty_vocabularies_openrefine_tutorial.pdf

The J. Paul Getty Trust is a cultural and philanthropic institution dedicated to the presentation, conservation, and interpretation of the world's artistic legacy.

Example: Deities from British Museum

Apollo(BM) – ???(Wikidata)

<https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56988>



The screenshot shows a ResearchSpace interface for an actor record. The title is "Apollo Actor (Person) record, British Museum". Below the title are three tabs: "Summary" (selected), "Annotations", and "Explore Related". A table displays the following fields and values:

Field	Value
Name	Apollo
Nationality	RomanGreekEtruscan
Profession	deity
Gender	male
Bio	One of the twelve Olympian gods; son of Zeus/Jupiter (q.v.); a major deity representing, in opposition to Dionysos, human nature; patron of the arts and leader of the Muse;

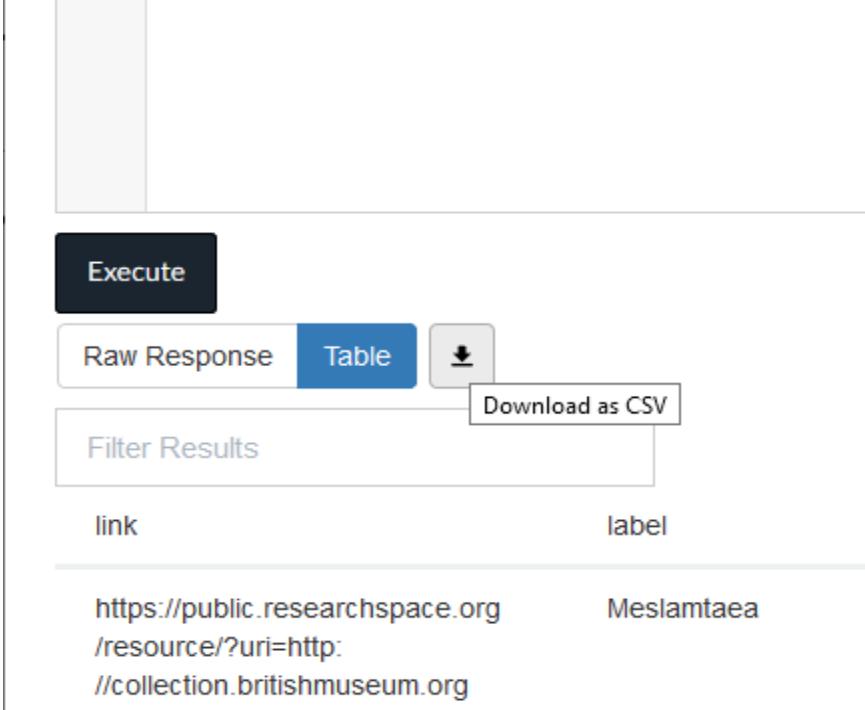
Example: Get all Deities from British Museum

SPARQL endpoint: <https://public.researchspace.org/sparql>

```
SELECT ?link ?label ?gender ?bio ?nation
WHERE {
  ?l <http://www.cidoc-crm.org/cidoc-crm/P3_has_note> ?bio.
  ?l <http://www.researchspace.org/ontology/displayLabel> ?label .
  ?l <http://www.researchspace.org/ontology/PX_profession>
<http://collection.britishmuseum.org/id/thesauri/profession/deity>.
  ?l <http://www.researchspace.org/ontology/PX_nationality> ?n.
  ?l <http://www.researchspace.org/ontology/PX_gender> ?g.
  BIND(CONCAT("https://public.researchspace.org/resource/?uri=", STR(?l)) AS ?link).
  BIND(STRAFTER(STR(?g), "gender/") AS ?gender) .
  FILTER((LANG(?label)) = "en") .
  BIND(STRAFTER(STR(?n), "nationality/") AS ?nation) .
}
```

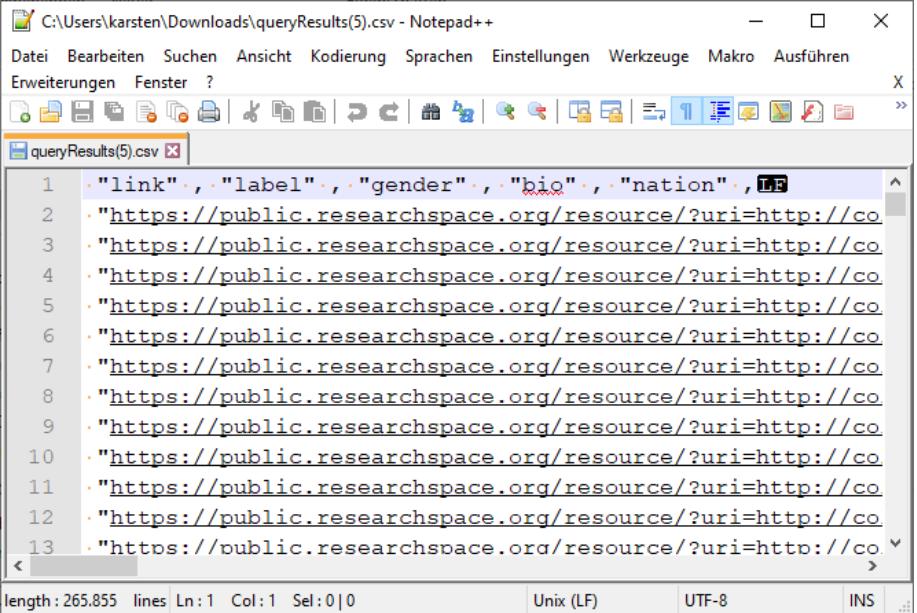
Hands on ...

1. Execute the statement ...
2. You might need to click on „Table“ in order to download the result as CSV.
3. Save CSV on your disc:



The screenshot shows a user interface for executing SPARQL queries. At the top right, there is a large "Execute" button. Below it, a row of buttons includes "Raw Response" (disabled), "Table" (selected and highlighted in blue), and "Download as CSV" with a download icon. A "Filter Results" input field is also present. The main area displays a table with two columns: "link" and "label". One visible row shows a link to a British Museum resource and its label "Meslamtaea".

link	label
https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org	Meslamtaea



The screenshot shows a Notepad++ window displaying a CSV file named "queryResults(5).csv". The file contains 13 rows of data, each consisting of three fields: "link", "label", and "bio". The "label" column contains the name "Meslamtaea". The "bio" column contains several URIs starting with "https://public.researchspace.org/resource/?uri=". The file is encoded in UTF-8.

link	label	bio
https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org	Meslamtaea	

Problems ... use Notepad++

(<https://notepad-plus-plus.org/downloads/>)

1. Since each cell is enclosed by "
"link", "label", "gender" ,...
2. Make sure the encoding fits!

C:\Users\karsten\Downloads\queryResults(5).csv - Notepad++

Datei Bearbeiten Suchen Ansicht Kodierung Sprachen Einstellungen Werkzeuge Makro Ausführen Erweiterungen Fenster ?

queryResults(5).csv

```
1 . "link" , "label" , "gender" , "bio" , "nation" , LF
2 . "https://public.researchspace.org/resource/?uri=http://co
3 . "https://public.researchspace.org/resource/?uri=http://co
4 . "https://public.researchspace.org/resource/?uri=http://co
5 . "https://public.researchspace.org/resource/?uri=http://co
6 . "https://public.researchspace.org/resource/?uri=http://co
7 . "https://public.researchspace.org/resource/?uri=http://co
8 . "https://public.researchspace.org/resource/?uri=http://co
9 . "https://public.researchspace.org/resource/?uri=http://co
10 . "https://public.researchspace.org/resource/?uri=http://co
11 . "https://public.researchspace.org/resource/?uri=http://co
12 . "https://public.researchspace.org/resource/?uri=http://co
13 . "https://public.researchspace.org/resource/?uri=http://co
```

length: 265.855 lines: Ln:1 Col:1 Sel:0|0 Unix (LF) UTF-8 INS

OpenRefine A power tool for working with messy data.

Create Project « Start Over Configure Parsing Options Project name queryResults 1 csv Tags Create Project »

link", "label", "gender", "bio", "nation", "https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/58608", "Hecate", "female", "A c

Parse data as Character encoding **UTF-8** Update Preview

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

Columns are separated by commas (CSV) tabs (TSV) custom: , 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

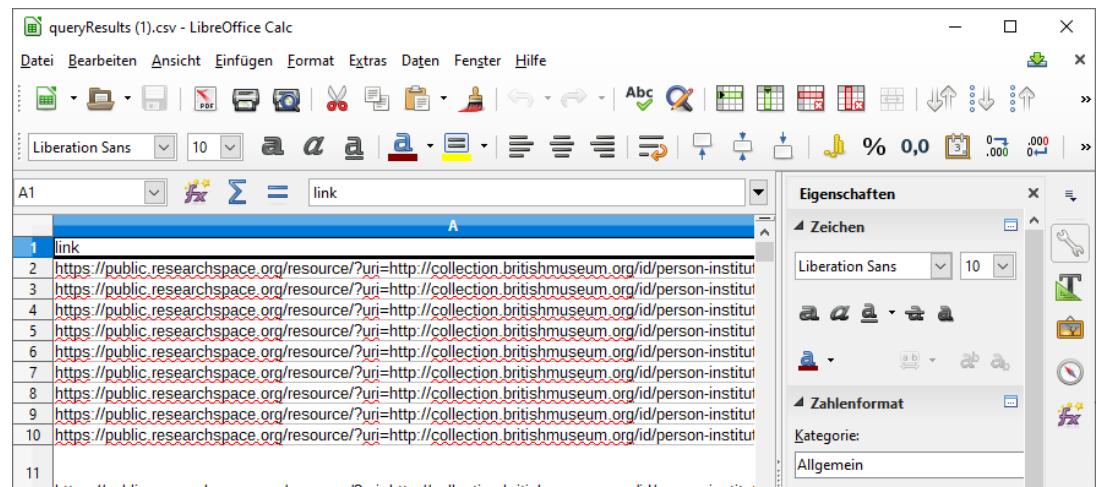
Version 3.2 [55c921b]

Preferences Help About

Solution ...

Use LibreOffice Calc to open CSV file
[\(https://www.libreoffice.org/download/download/\)](https://www.libreoffice.org/download/download/)

Save it as Text CSV

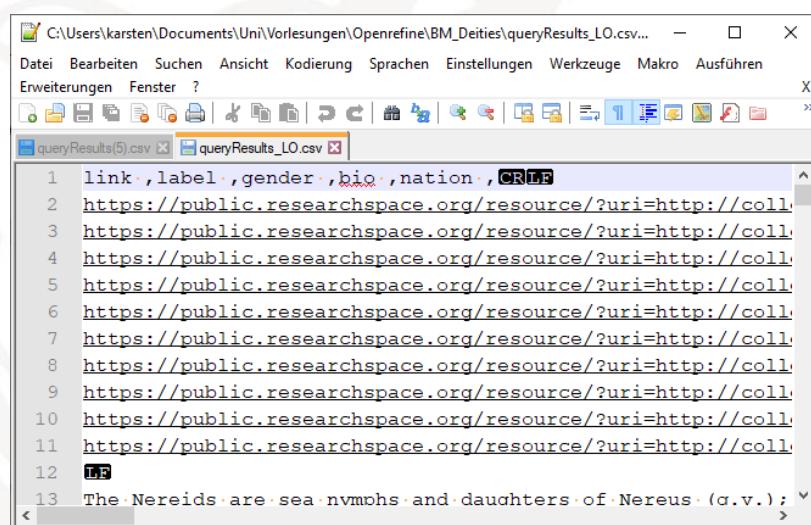


The screenshot shows a LibreOffice Calc spreadsheet titled "queryResults (1).csv". The first few rows contain URLs starting with "https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institut". The "Eigenschaften" (Properties) panel is open on the right, showing settings for the selected cell A1.

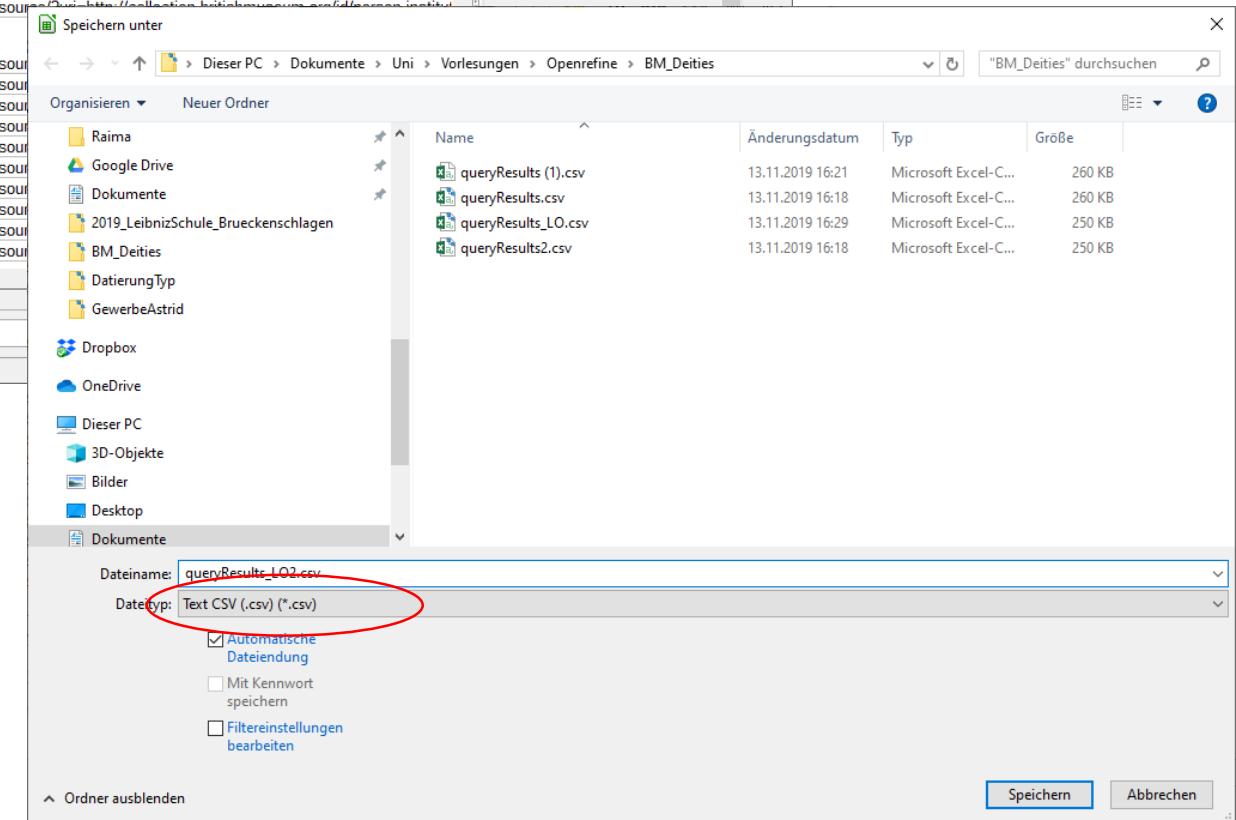
Speichern unter dialog box:

- Path: Dieser PC > Dokumente > Uni > Vorlesungen > Openrefine > BM_Deities
- Organizer: Neuer Ordner
- File list:

Name	Änderungsdatum	Typ	Größe
queryResults (1).csv	13.11.2019 16:21	Microsoft Excel-C...	260 KB
queryResults.csv	13.11.2019 16:18	Microsoft Excel-C...	260 KB
queryResults_LO.csv	13.11.2019 16:29	Microsoft Excel-C...	250 KB
queryResults2.csv	13.11.2019 16:18	Microsoft Excel-C...	250 KB



The screenshot shows the OpenRefine interface with two tabs: "queryResults(5).csv" and "queryResults_LO.csv". The "queryResults_LO.csv" tab is active, displaying a list of URLs. A large blue arrow points from this window towards the "Speichern unter" dialog box.



The screenshot shows the "Speichern unter" dialog box in OpenRefine. The "Dateiname" field is set to "queryResults_LO2.csv". The "Dateityp" field is highlighted with a red oval and contains "Text CSV (*.csv)". Other options include "Automatische Dateiendung" (checked), "Mit Kennwort speichern" (unchecked), and "Filtereinstellungen bearbeiten" (unchecked).

Create project
Select some values
via text facet ...

OpenRefine queryResults_LO csv [Permalink](#)

Facet / Filter Undo / Redo 0 / 0

Refresh Reset All Remove All

label change invert reset

621 choices Sort by: name count Cluster

label	count
Anuavis or Rosejaw	1
Anuqet	1
Aparajita Sitatapatra	2
Aphrodite-Hathor	2
Aphrodite-Isis	2
Aphrodite/Venus	3
Apis	1
Apollo	3
Apollo Helios	2
Apophis	1
Ardhanarishvaramurti	1
Ardochsho	1

10 matching rows (786 total)

Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) [rows](#)

<input checked="" type="checkbox"/> All	<input checked="" type="checkbox"/> link	<input checked="" type="checkbox"/> label	<input checked="" type="checkbox"/> gender	<input checked="" type="checkbox"/> bio	
		448. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57039	Artemis/Diana	female	One of the two sons of Apollo (q.v.) helping women and is sometimes
		449. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57039	Artemis/Diana	female	One of the two sons of Apollo (q.v.) helping women and is sometimes
		458. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	Demeter/Ceres	female	One of the two daughters of Zeus; corn; her daughter
		459. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	Demeter/Ceres	female	One of the two daughters of Zeus; corn; her daughter
		460. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	Demeter/Ceres	female	One of the two daughters of Zeus; corn; her daughter
		477. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56988	Apollo	male	One of the twelve Olympians; Artemis/Diana's brother; rational and creative; later, manifested
		478. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56988	Apollo	male	One of the twelve Olympians; Artemis/Diana's brother; rational and creative; later, manifested
		479. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56988	Apollo	male	One of the twelve Olympians; Artemis/Diana's brother; rational and creative; later, manifested
		491. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56991	Apollo Helios	male	Syncretic deity
		492. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56991	Apollo Helios	male	Syncretic deity

OpenRefine queryResults_LO csv [Permalink](#)

acet / Filter Undo / Redo 0 / 0 Extensions: Wikidata ▾

fresh Reset All Remove All

label change invert reset

choices Sort by: name count Cluster

Iris or Rosejau 1

quet 1

arajita Sitatapatra 2

irodite-Hathor 2

irodite-Isis 2

irodite/Venus 3

S 1

ollo 3

ollo Helios 2

ophis 1

hanarishvaramurti 1

ochsho 1

ipt{} =

10 matching rows (786 total) Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

		link	label	gender	bio
448.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57039	Facet			One of the twelve Olympian gods; daughter of Zeus/Jupiter (q.v.) and Leto/Latona (q.v.); a virgin huntress; symbol of chastity and protector of wildlife, also helping women at childbirth; she later became identified with the moon-goddess, and is sometimes associated with the goddess of the Underworld, Hecate (q.v.).
449.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57039	Text filter			
458.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	Edit cells			One of the twelve Olympian gods; daughter of Zeus/Jupiter (q.v.) and Leto/Latona (q.v.); a virgin huntress; symbol of chastity and protector of wildlife, also helping women at childbirth; she later became identified with the moon-goddess, and is sometimes associated with the goddess of the Underworld, Hecate (q.v.).
459.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	Edit column			
460.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	Transpose			One of the twelve Olympian gods; goddess of fertility and agriculture, particularly corn; her daughter, Persephone/Proserpine (q.v.) was abducted by Hades/Pluto (q.v.).
477.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56988	Sort...			
478.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56988	View			One of the twelve Olympian gods; goddess of fertility and agriculture, particularly corn; her daughter, Persephone/Proserpine (q.v.) was abducted by Hades/Pluto (q.v.).
479.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56988	Reconcile			
		Start reconciling...			
		Demeter/Ceres	female		
		Apollo	male		
		Apollo	male		
		Apollo	male		

Facets Reconcile text in this column with items on Freebase (q.v.) was abducted by Hades/Pluto (q.v.)

Actions

Copy reconciliation data...

Use values as identifiers

Pick a Service (only Wikidata predefined)

Pick a Service or Extension on Left

(left) select best proposed fitting classes (or enter it under „Reconcile against type“)

(right) add additional columns of „your“ data that are entered into the reconciliation process

OpenRefine queryResults_LO csv Permalink

Facet / Filter **Undo / Redo 0 / 0**

Refresh **Reset All** **Remove**

label change invert n
621 choices Sort by: name count Clus
Anubis or Rosejau 1
Anuqet 1
Aparajita Sitatapatra 2
Aphrodite-Hathor 2
Aphrodite-Isis 2
Aphrodite/Venus 3
Apis 1
Apollo 3
Apollo Helios 2
Apophis 1
Ardhanarishvaramurti 1
Ardochsho 1

Reconcile column "label"

» Access Service API

Reconcile each cell to an entity of one of these types:

- asteroid Q3863
- Greek deity Q22989102
- taxon Q16521
- goddess Q205985
- water deity Q1916821
- inner planet Q3504248
- moon of Saturn Q1972
- Anemoi Q476682
- dwarf planet Q2199

Also use relevant details from other columns:

Column	Include? As Property
link	<input type="checkbox"/>
gender	<input checked="" type="checkbox"/> sex or gender
bio	<input type="checkbox"/>
nation	<input type="checkbox"/>
Column	<input type="checkbox"/>

Column Include? As Property

link
gender sex or gender
bio
nation
Column

Reconcile each cell to an entity of one of these types:

- Reconcile against type:
- Reconcile against no particular type
- Auto-match candidates with high confidence

Maximum number of candidates to return

Add Standard Service... Start Reconciling Cancel

Extensions: Wikidata

« first < previous 1 - 10 next > last »

▼ nation ▼ Column

o/Latona (q.v.); twin sister life, also associated with oddess, Selene/Luna (q.v.) te (q.v.).	Roman
o/Latona (q.v.); twin sister life, also associated with oddess, Selene/Luna (q.v.) te (q.v.).	Greek
ticularly associated with s/Pluto (q.v.).	Roman
ticularly associated with s/Pluto (q.v.).	Greek
ticularly associated with s/Pluto (q.v.).	Etruscan
ona (q.v.); brother of s/Bacchus (q.v.), the der of the Muses (q.v.); od Helios/Sol (q.v.).	Roman
ona (q.v.); brother of s/Bacchus (q.v.), the der of the Muses (q.v.); od Helios/Sol (q.v.).	Greek
ona (q.v.); brother of s/Bacchus (q.v.), the der of the Muses (q.v.); od Helios/Sol (q.v.).	Etruscan
ios/Sol (q.v.).	Roman
ios/Sol (q.v.).	Greek

Match manually or filter and match to best candidate ...
... also possible to clear reconciliation data and start from scratch

Permalink Open... Export... Help

11 matching rows (786 total) Extensions: Wikidata ▾

Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) rows « first < previous **1 - 10** next > last »

<input type="checkbox"/> All	<input type="checkbox"/> link	<input type="checkbox"/> label	<input type="checkbox"/> gender	<input type="checkbox"/> bio	<input type="checkbox"/> nation	<input type="checkbox"/> Column
		445. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56979	<input type="checkbox"/> Facet	female	One of the twelve Olympian gods; goddess of love and fertility; mother of Eros/Cupid (q.v.), Anteros (q.v.) and Hermaphroditos (q.v.). Also used for the personification of the planet Venus. Sometimes associated with the Zoroastrian deity Anahita (q.v.).	Roman
		446. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56979	<input type="checkbox"/> Text filter	female	One of the twelve Olympian gods; goddess of love and fertility; mother of Eros/Cupid (q.v.), Anteros (q.v.) and Hermaphroditos (q.v.). Also used for the personification of the planet Venus. Sometimes associated with the Zoroastrian deity Anahita (q.v.).	Greek
		447. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56979	<input type="checkbox"/> Edit cells	female	One of the twelve Olympian gods; goddess of love and fertility; mother of Eros/Cupid (q.v.), Anteros (q.v.) and Hermaphroditos (q.v.). Also used for the personification of the planet Venus. Sometimes associated with the Zoroastrian deity Anahita (q.v.).	Etruscan
		458. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	<input type="checkbox"/> Edit column	female	One of the twelve Olympian gods; goddess of love and fertility; mother of Eros/Cupid (q.v.), Anteros (q.v.) and Hermaphroditos (q.v.). Also used for the personification of the planet Venus. Sometimes associated with the Zoroastrian deity Anahita (q.v.).	Roman
		459. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	<input type="checkbox"/> Transpose	female	One of the twelve Olympian gods; goddess of love and fertility; mother of Eros/Cupid (q.v.), Anteros (q.v.) and Hermaphroditos (q.v.). Also used for the personification of the planet Venus. Sometimes associated with the Zoroastrian deity Anahita (q.v.).	Greek
		460. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	<input type="checkbox"/> Sort...	female	One of the twelve Olympian gods; goddess of love and fertility; mother of Eros/Cupid (q.v.), Anteros (q.v.) and Hermaphroditos (q.v.). Also used for the personification of the planet Venus. Sometimes associated with the Zoroastrian deity Anahita (q.v.).	Etruscan
		477. https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org	<input type="checkbox"/> View	female	One of the twelve Olympian gods; goddess of love and fertility; mother of Eros/Cupid (q.v.), Anteros (q.v.) and Hermaphroditos (q.v.). Also used for the personification of the planet Venus. Sometimes associated with the Zoroastrian deity Anahita (q.v.).	Roman
			<input type="checkbox"/> Reconcile		<input type="checkbox"/> Start reconciling...	
					<input type="checkbox"/> Facets	One of the twelve Olympian gods; goddess of fertility and agriculture, particularly associated with corn; her daughter, Persephone/Proserpine (q.v.) was abducted by Hades/Pluto (q.v.).
					<input type="checkbox"/> Actions	<input type="checkbox"/> Match each cell to its best candidate
					<input type="checkbox"/> Copy reconciliation data...	<input type="checkbox"/> Match each cell to its best candidate in this column for all current filtered rows
					<input type="checkbox"/> Use values as identifiers	<input type="checkbox"/> Create one new item for similar cells
						<input type="checkbox"/> Match all filtered cells to...
						<input type="checkbox"/> Discard reconciliation judgments
						<input type="checkbox"/> Clear reconciliation data

Enrich your data ...

Add columns from reconciled column label

Add Property

Preview

Reset

partner

Suggested Properties

- appears in the form of
 - award received
 - based on
 - contributor(s) to the creative work or subject
 - creator
 - domain of saint or deity
 - genre
 - iconographic symbol
 - image
 - inception
 - license
 - location of formation
 - official website
 - recorded at
 - residence

label	part of <u>remove</u> <u>configure</u>	partner <u>remove</u> <u>configure</u>	
Aphrodite	Twelve Olympians	10	Configure this column
Aphrodite	Twelve Olympians	10	
Aphrodite	Twelve Olympians	10	
Demeter	Twelve Olympians	2	
Demeter	Twelve Olympians	2	
Demeter	Twelve Olympians	2	
Apollo	Twelve Olympians	45	
Apollo	Twelve Olympians	45	
Apollo	Twelve Olympians	45	
Apollo	Twelve Olympians	45	

OK

[Cancel](#)

... your enriched data ...

11 matching rows (865 total)											Extensions: Wikidata		
Show as: rows records			Show: 5 10 25 50 rows			« first < previous 1 - 10 next > last »							
		<input type="checkbox"/> link		<input type="checkbox"/> label	<input type="checkbox"/> said to be the same	<input type="checkbox"/> part of	<input type="checkbox"/> partner	<input type="checkbox"/> gender	<input type="checkbox"/> bio	<input type="checkbox"/> nation	<input type="checkbox"/> code		
		445.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56979	Aphrodite Choose new match	Venus Choose new match	Twelve Olympians Choose new match	10	female	One of the twelve Olympian gods; goddess of love and fertility; mother of Eros/Cupid (q.v.), Anteros (q.v.) and Hermaphroditos (q.v.). Also used for the personification of the planet Venus. Sometimes associated with the Zoroastrian deity Anahita (q.v.).	Roman			
exclude		456.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56979	Aphrodite Choose new match	Venus Choose new match	Twelve Olympians Choose new match	10	female	One of the twelve Olympian gods; goddess of love and fertility; mother of Eros/Cupid (q.v.), Anteros (q.v.) and Hermaphroditos (q.v.). Also used for the personification of the planet Venus. Sometimes associated with the Zoroastrian deity Anahita (q.v.).	Greek			
exclude		467.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56979	Aphrodite Choose new match	Venus Choose new match	Twelve Olympians Choose new match	10	female	One of the twelve Olympian gods; goddess of love and fertility; mother of Eros/Cupid (q.v.), Anteros (q.v.) and Hermaphroditos (q.v.). Also used for the personification of the planet Venus. Sometimes associated with the Zoroastrian deity Anahita (q.v.).	Etruscan			
exclude		488.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	Demeter Choose new match	Ceres Choose new match	Twelve Olympians Choose new match	edit	2	female	One of the twelve Olympian gods; goddess of fertility and agriculture, particularly associated with corn; her daughter, Persephone/Proserpine (q.v.) was abducted by Hades/Pluto (q.v.).	Roman		
		492.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	Demeter Choose new match	Ceres Choose new match	Twelve Olympians Choose new match	2	female	One of the twelve Olympian gods; goddess of fertility and agriculture, particularly associated with corn; her daughter, Persephone/Proserpine (q.v.) was abducted by Hades/Pluto (q.v.).	Greek			
		496.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/57930	Demeter Choose new match	Ceres Choose new match	Twelve Olympians Choose new match	2	female	One of the twelve Olympian gods; goddess of fertility and agriculture, particularly associated with corn; her daughter, Persephone/Proserpine (q.v.) was abducted by Hades/Pluto (q.v.).	Etruscan			
		516.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56988	Apollo Choose new match	Apollo Choose new match	Twelve Olympians Choose new match	45	male	One of the twelve Olympian gods; son of Zeus/Jupiter (q.v.) and Leto/Latona (q.v.); brother of Artemis/Diana (q.v.); a major deity representing, in opposition to Dionysos/Bacchus (q.v.), the rational and civilized aspects of human nature; patron of the arts and leader of the Muses (q.v.); later, manifested as Phoebus (q.v.), he became identified with the sun-god Helios/Sol (q.v.).	Roman			
		525.	https://public.researchspace.org/resource/?uri=http://collection.britishmuseum.org/id/person-institution/56988	Apollo Choose new match	Apollo Choose new match	Twelve Olympians Choose new	45	male	One of the twelve Olympian gods; son of Zeus/Jupiter (q.v.) and Leto/Latona (q.v.); brother of Artemis/Diana (q.v.); a major deity representing, in opposition to Dionysos/Bacchus (q.v.),	Greek			

Talend Open Studio for Data Integration (6.1.1.20151214_1327) | HeidelbergImport (Verbindung: Lokal)

Datei(F) Bearbeiten View Fenster Hilfe

Learn Ask Upgrade Exchange

Integration

Ablage

LOCAL: HeidelbergImport

- Business Modell
- Job Designs
 - Heidelberg_additional_im...
 - Heidelberg_Import_cointfin
 - correct_ids 0.1
 - excel_import_in_AFEDB
 - First_Import_Excel_in_A...
 - MainJob 0.1
- Kontexte
- Code
- SQL Templates
- Meta-Daten
- Documentation
- Papierkorb

Job First_Import_Excel_in_AFE_and_update_ids 0.1

Job correct_ids 0.1

HD_v6

"cofind"

tMap_1

HD_v7

row4 (Main)

row5 (Lookup)

row6 (Main)

Designer Code

Job(correct_ids 0.1)

Contexts(correct_ids)

Komponente

Starte (Job correct_ids)

correct_ids 0.1

Main	Name	correct_ids
Extra	Author	test@talend.com
Stats & Logs	Version	0.1
Version	Purpose	
Description	Status	

Palette

Komponente suchen

Favorites

Recently Used

Big Data

Google Storage

Hive

Business Intelligence

Business

Cloud

Datei

Datenbanken

Datenqualität

Matching

- tAddCRCRow
- tChangeFileEncoding
- tIntervalMatch
- tReplaceList
- tSchemaCompliance...
- tStewardshipTaskDel...
- tStewardshipTaskInput
- tStewardshipTaskOut...
- tUniqRow

DotNET

ELT

Combined SQL

Map

SQLTemplate

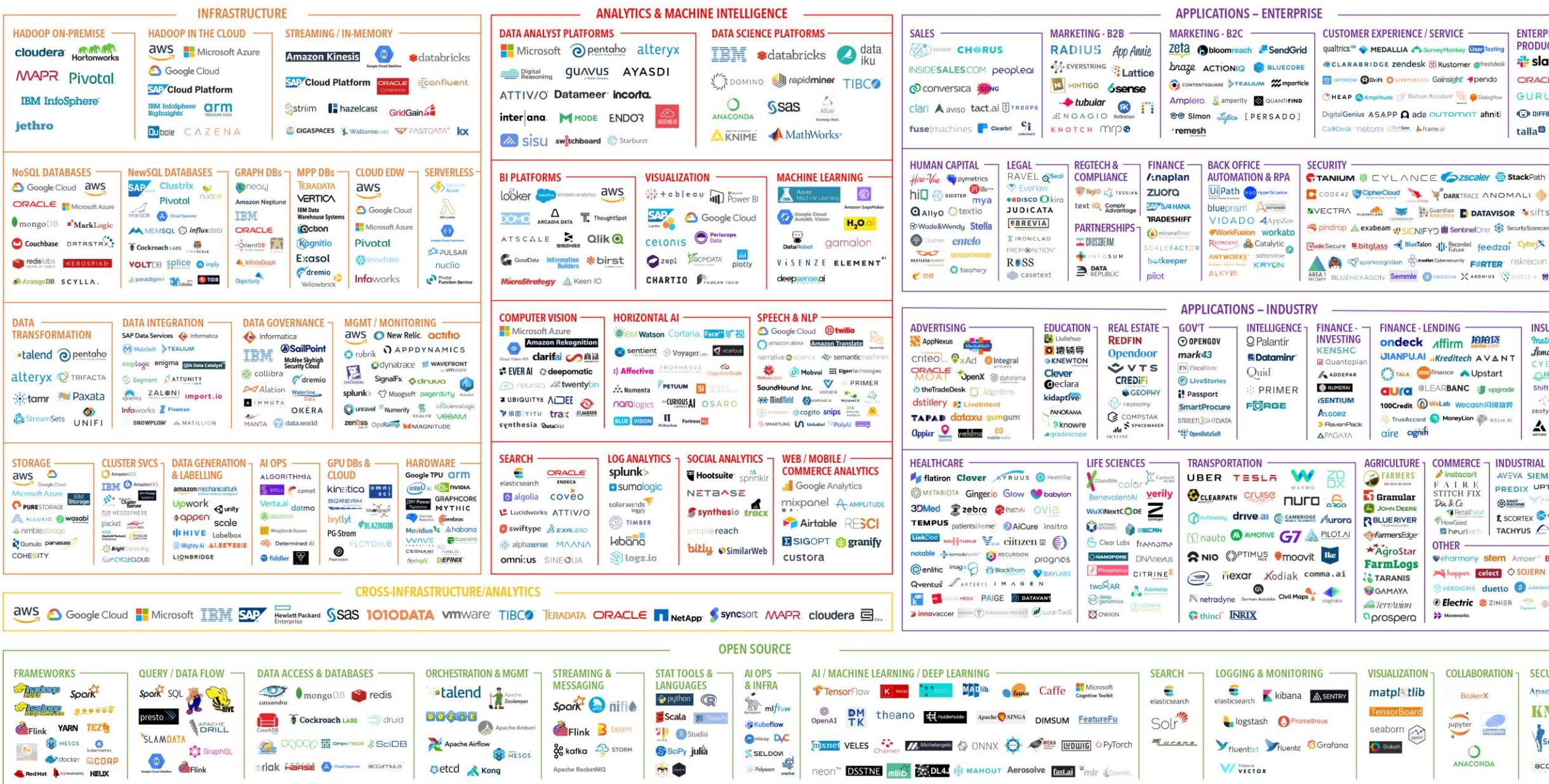
Verbindungen

EBC

Tool landscape - <https://mattturck.com/data2019/>



DATA & AI LANDSCAPE 2019



Tools

- Fuseki (RDF-Datenbank)
- D2R (Mapping zwischen relationaler DB und RDF)
- OpenRefine (Umgang mit verschiedenen Formaten, auch für große Dokumente, Reconciliation)

Weitere ...

- Notepad ++ (generischer Texteditor)
- PilotEdit Lite (wenn die Dateien zu groß werden – free bis 10GB)
- Libre Office (gut zum Einlesen von CSV-Dateien)

Hausaufgabe (bis nächste Woche):

1. Team – Steckbrief bis diesen Freitag eintragen!

2. Lesen:

https://de.wikipedia.org/wiki/Beurteilung_eines_bin%C3%A4ren_Klassifikators
bzw. (möglichst UND)

https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers

- Sie sollten die folgenden Begriffe erklären können:
 - Recall
 - Precision
 - Sensitivity
 - Specificity
 - F-Maß/F-Score

Hausaufgabe (sobald es die Zeit erlaubt)

Ethical Implications of AI

<http://www.bigdata.uni-frankfurt.de/data-challenge-ss-2020/>



Prof. Gemma Roig <http://www.cvai.cs.uni-frankfurt.de/>

Fairness, Bias and Discrimination in AI
(Prof. Gemma Roig)

[**slides** http://www.bigdata.uni-frankfurt.de/wp-content/uploads/2020/01/Fairness_Bias_Discrimination_in_AI_GemmaRoig.pdf]
[**video** <https://www.youtube.com/watch?v=LE38HWZz8NU>]