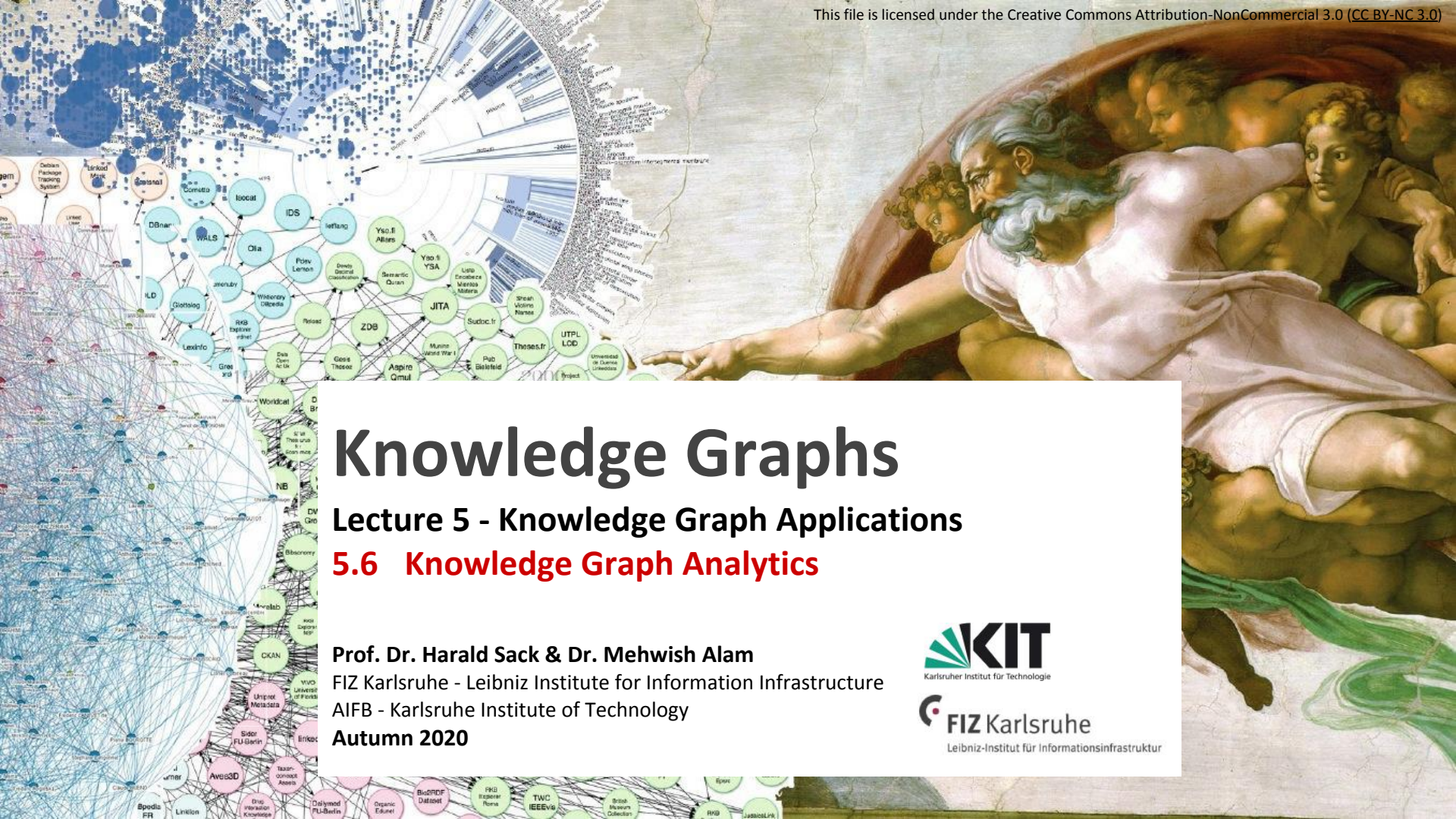


Knowledge Graphs

Lecture 5 - Knowledge Graph Applications

5.6 Knowledge Graph Analytics

Prof. Dr. Harald Sack & Dr. Mehwish Alam
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
AIFB - Karlsruhe Institute of Technology
Autumn 2020

Karlsruher Institut für Technologie

FIZ Karlsruhe
Leibniz-Institut für Informationsinfrastruktur

Knowledge Graphs

Lecture 5: Knowledge Graph Applications

Knowledge Mining and Knowledge Discovery

Knowledge Discovery [in Databases] (KDD) is the nontrivial process of identifying **valid**, **novel**, **potentially useful**, and **ultimately understandable patterns** in (massive) data sources.
(Fayyad et al, 1996)

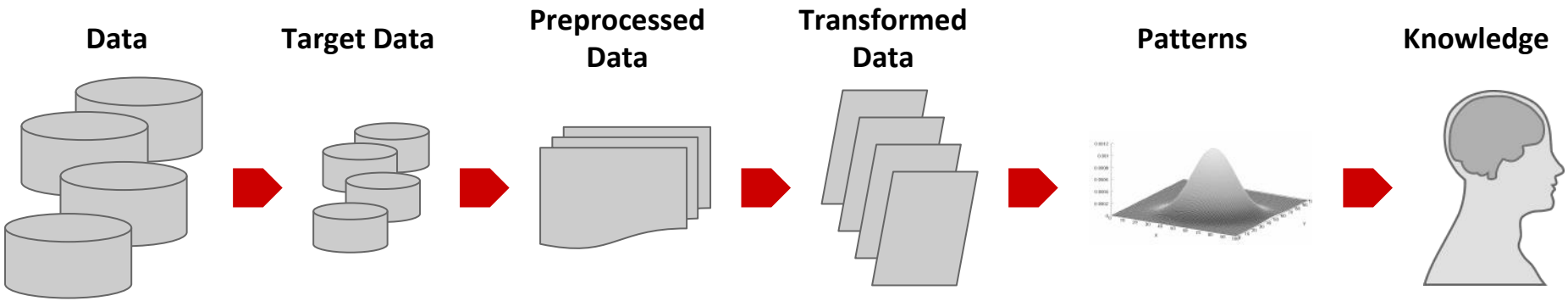
- **valid**: to a certain degree the discovered patterns should also hold for new, previously unseen problem instances.
- **novel**: at least to the system and preferable to the user.
- **potentially useful**: they should lead to some benefit to the user or task.
- **ultimately understandable**: the end user should be able to interpret the patterns either immediately or after some post-processing.

Knowledge Mining and Knowledge Discovery

Knowledge Discovery [in Databases] (KDD) is the nontrivial process of identifying **valid**, **novel**, **potentially useful**, and **ultimately understandable patterns** in (massive) data sources.
(Fayyad et al, 1996)

- **Goals:**
  - **Descriptive Modelling:** explains the characteristic and the behaviour of the observed data
  - **Predictive Modelling:** predicts the behaviour of new data based on some model
- **Important:**
  - The extracted model/pattern does not have to apply in 100% of the cases

The Knowledge Discovery Process



Selection:
Select a relevant dataset or focus on a subset of a dataset

Preprocessing/ Cleaning:
Data integration from different sources, Data Cleaning

Transformation:
Select useful features, feature transformation, dimensionality reduction

Data Mining:
Search for patterns of interest

Evaluation:
Evaluate patterns based on interestingness measures, model validation

Hands-On Example - Knowledge Graph Analytics

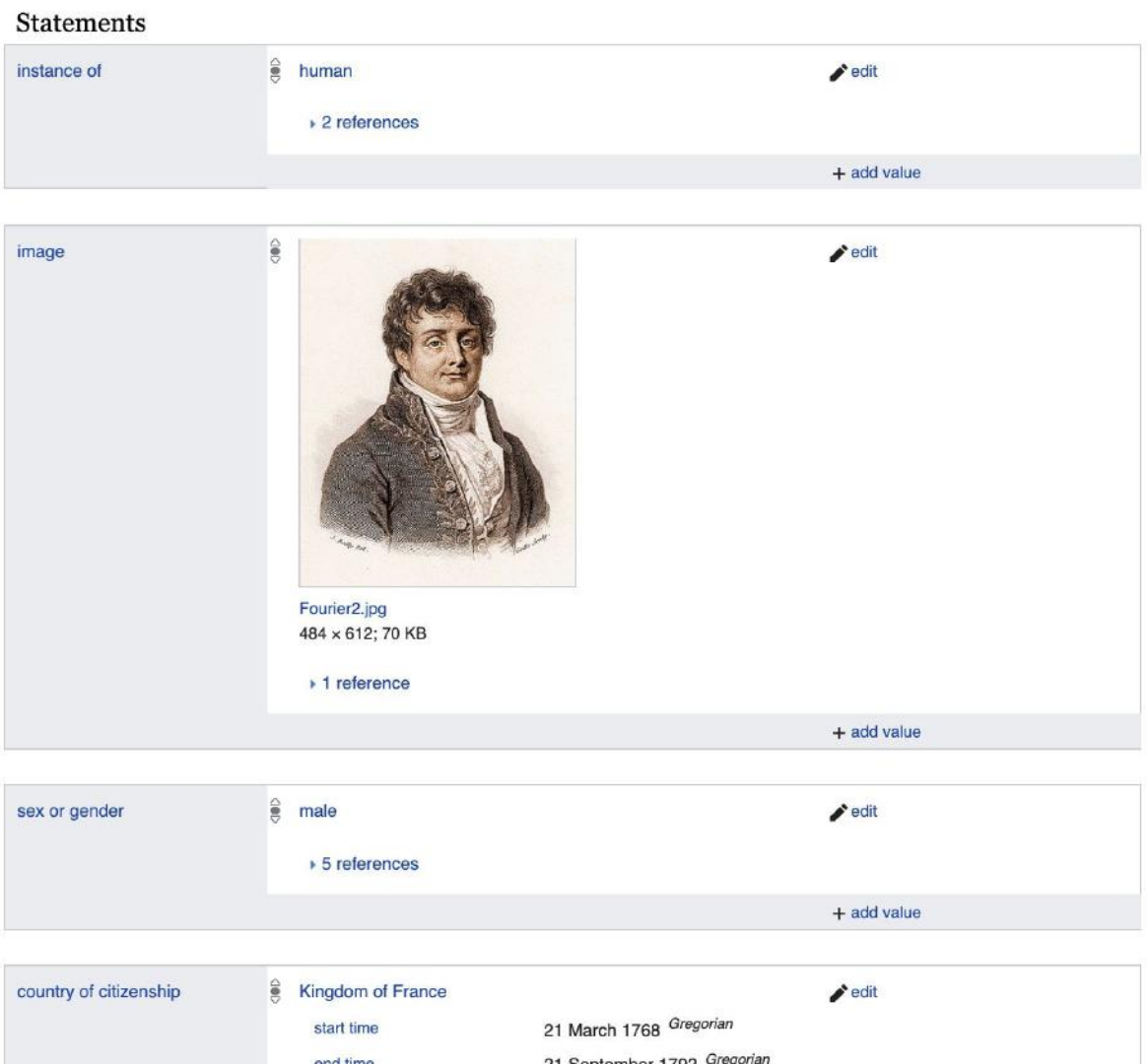
- Let's discover "interesting knowledge" about **Physicists**
- General knowledge graphs, as e.g. **DBpedia** or **Wikidata** contain data about thousands of physicists
- **Data Acquisition**
  - First, let's look at an **random example** to see what kind of data we have to expect


[1]

Joseph Fourier, physicist

1. Data Acquisition

- First, let's look at an **random example** to see what kind of data we have to expect

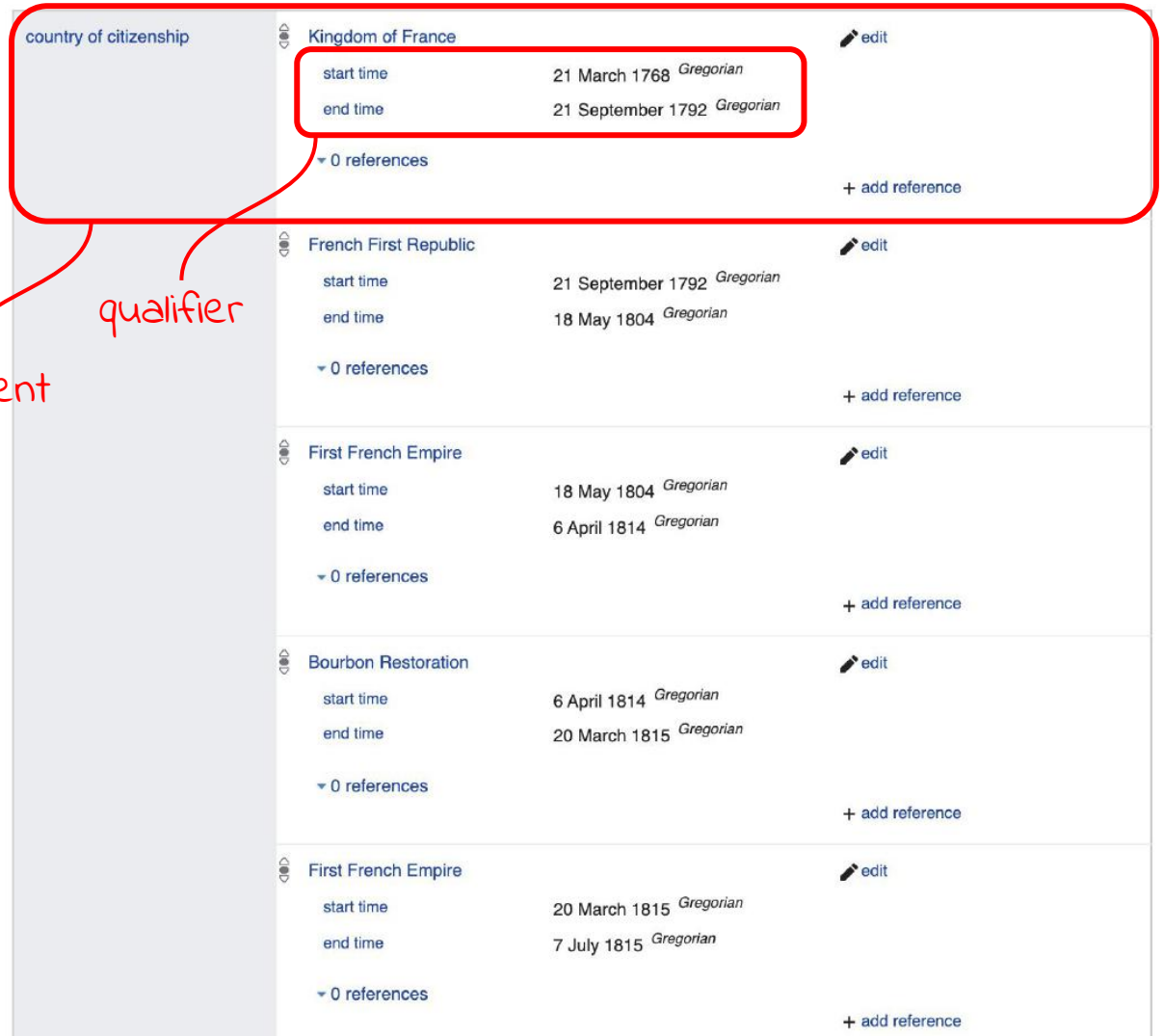


WIKIDATA

https://www.wikidata.org/wiki/Q8772

1. Data Acquisition

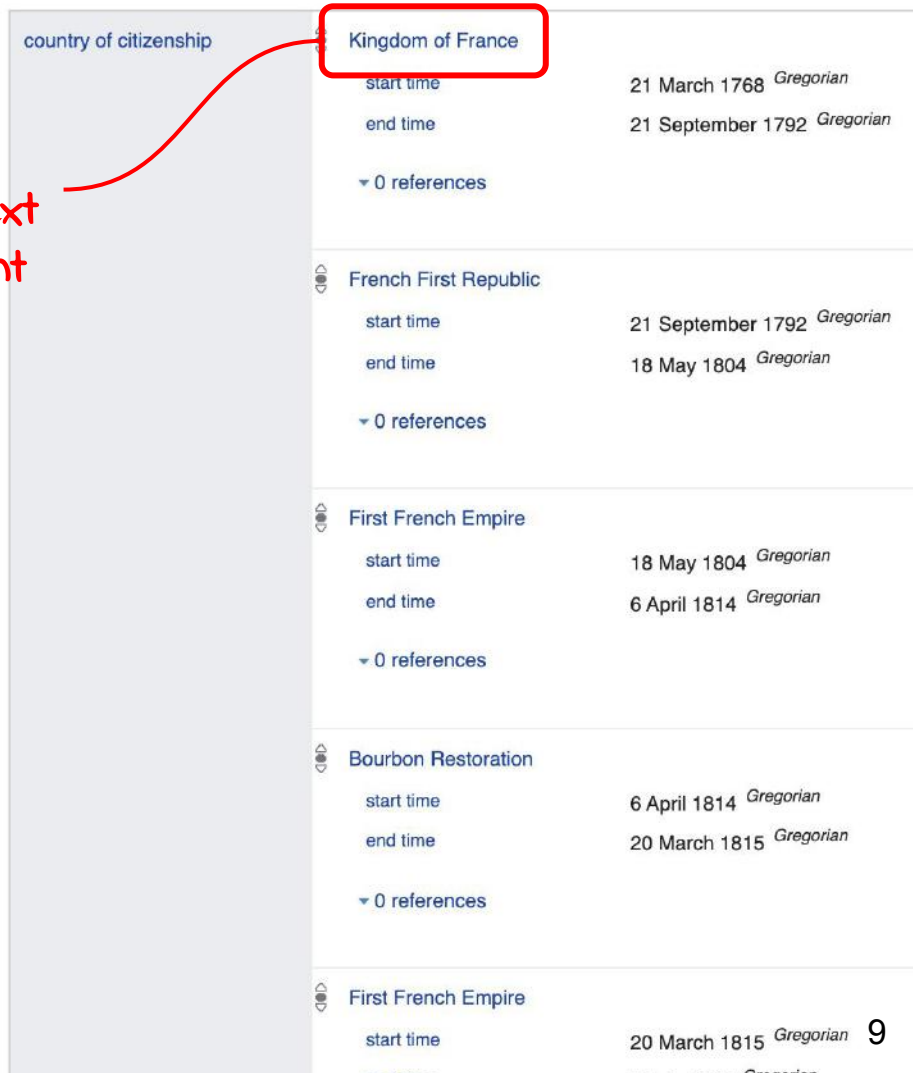
- First, let's look at an **random example** to see what kind of data we have to expect



WIKIDATA

https://www.wikidata.org/wiki/Q8772

WikiData Recap



object
=
subject/context
for statement

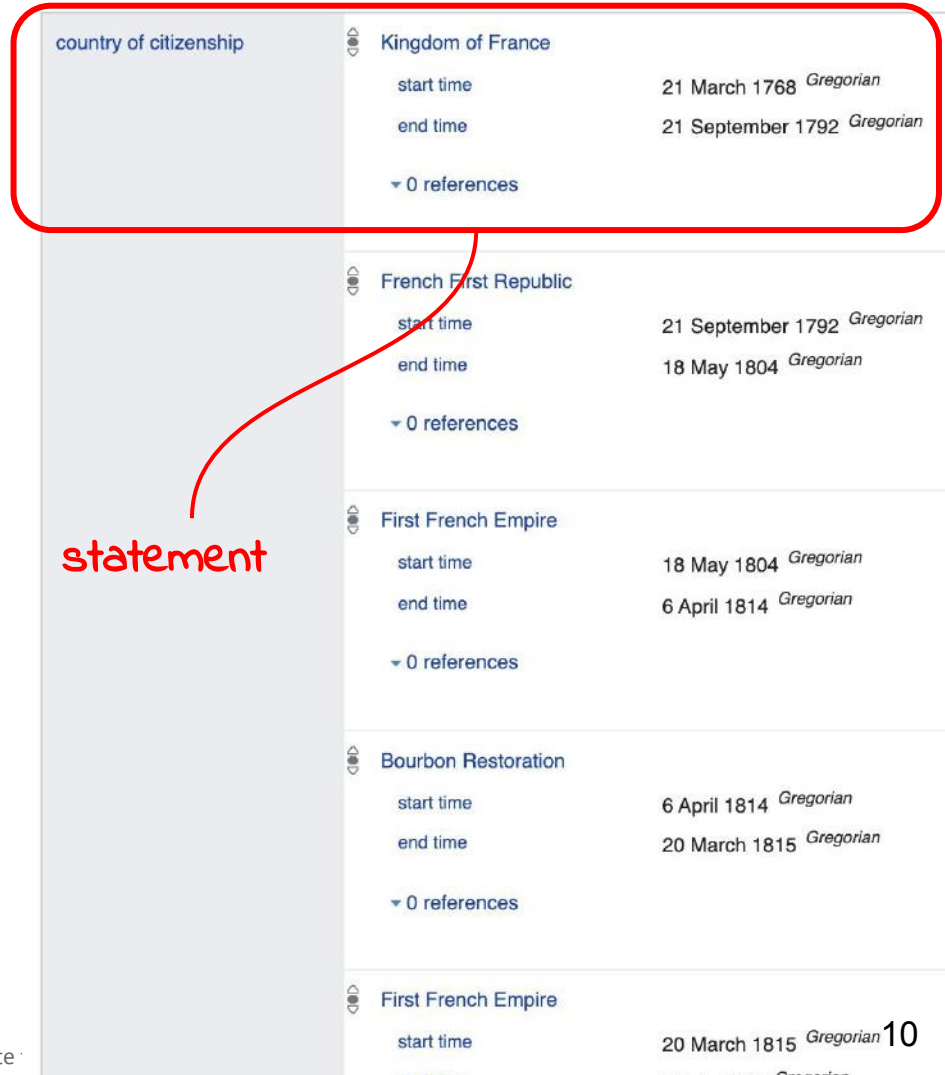
Access via **different namespaces for properties**:

- **wdt:** connects an item to a value
  wd:Q8772 wdt:P27 ?country .

WikiData Recap

Access via **different namespaces for properties**:

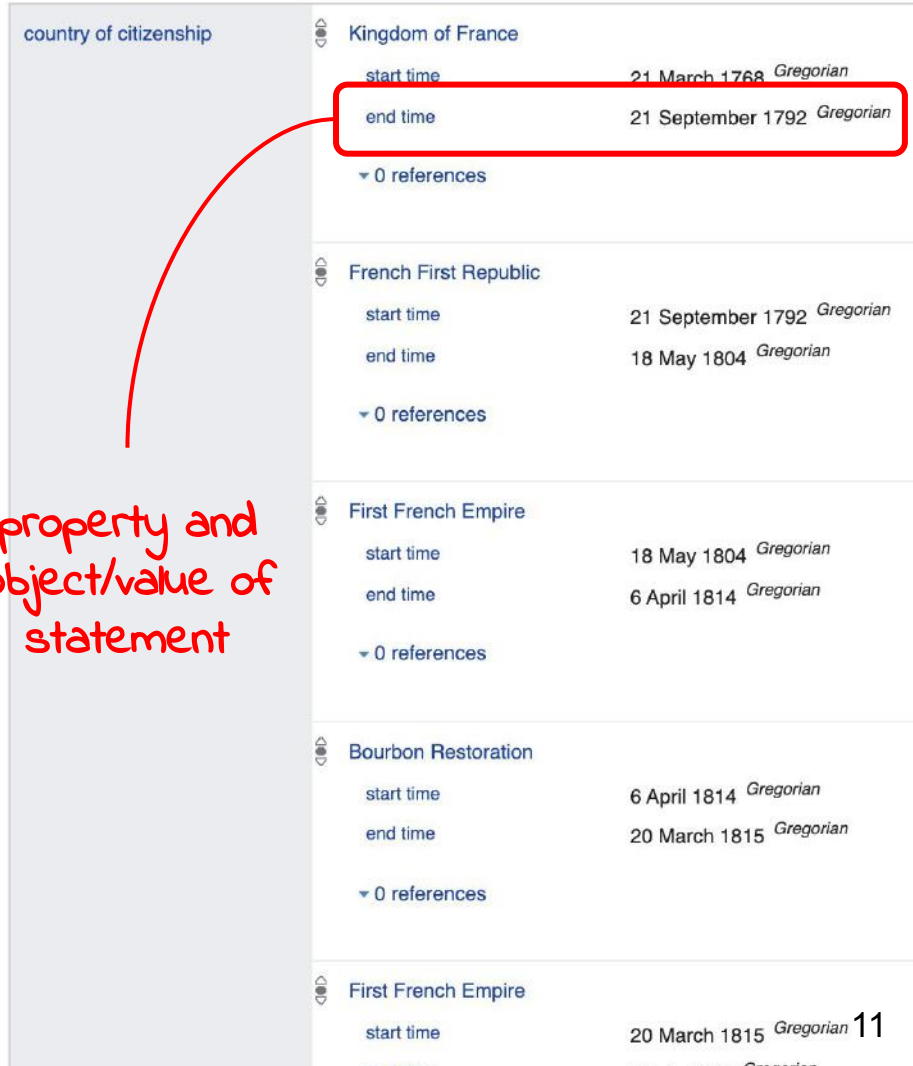
- **wdt:** connects an item to a value
  wd:Q8772 wdt:P27 ?country .
- **p:** connects a subject to a statement
  wd:Q8772 p:P27 ?country_statement .



WikiData Recap

Access via **different namespaces for properties**:

- **wdt:** connects an item to a value
  wd:Q8772 wdt:P27 ?country .
- **p:** connects a subject to a statement
  wd:Q8772 p:P27 ?country_statement .
- **pq:** connects statement to qualifier value
  ?country_statement pq:P582 ?statement_value

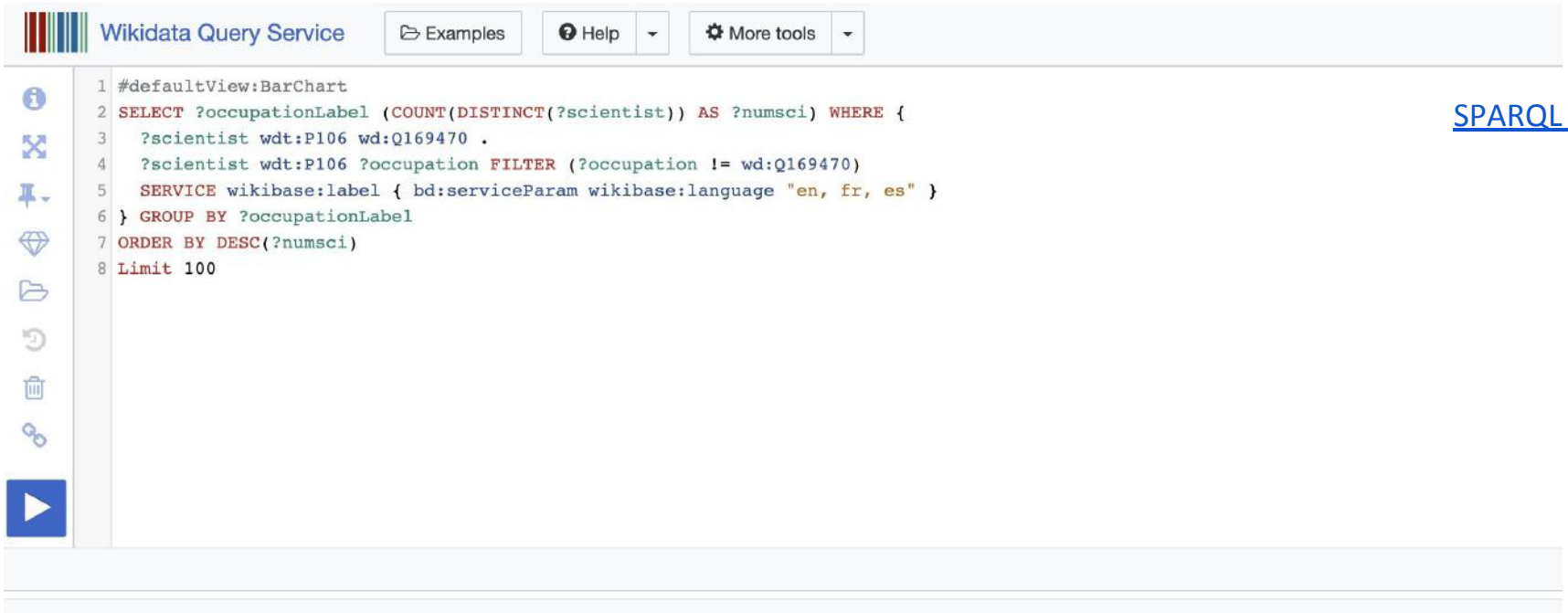


Knowledge Graph Analytics with SPARQL

- Convenient and very powerful way to analyze Knowledge Graph data
- E.g. **WIKIDATA SPARQL endpoint** provides visualization toolkit with
  - Bar plots and Histograms
  - Scatter plots
  - Timelines
  - Graph visualizations
  - etc.

Knowledge Graph Analytics with SPARQL

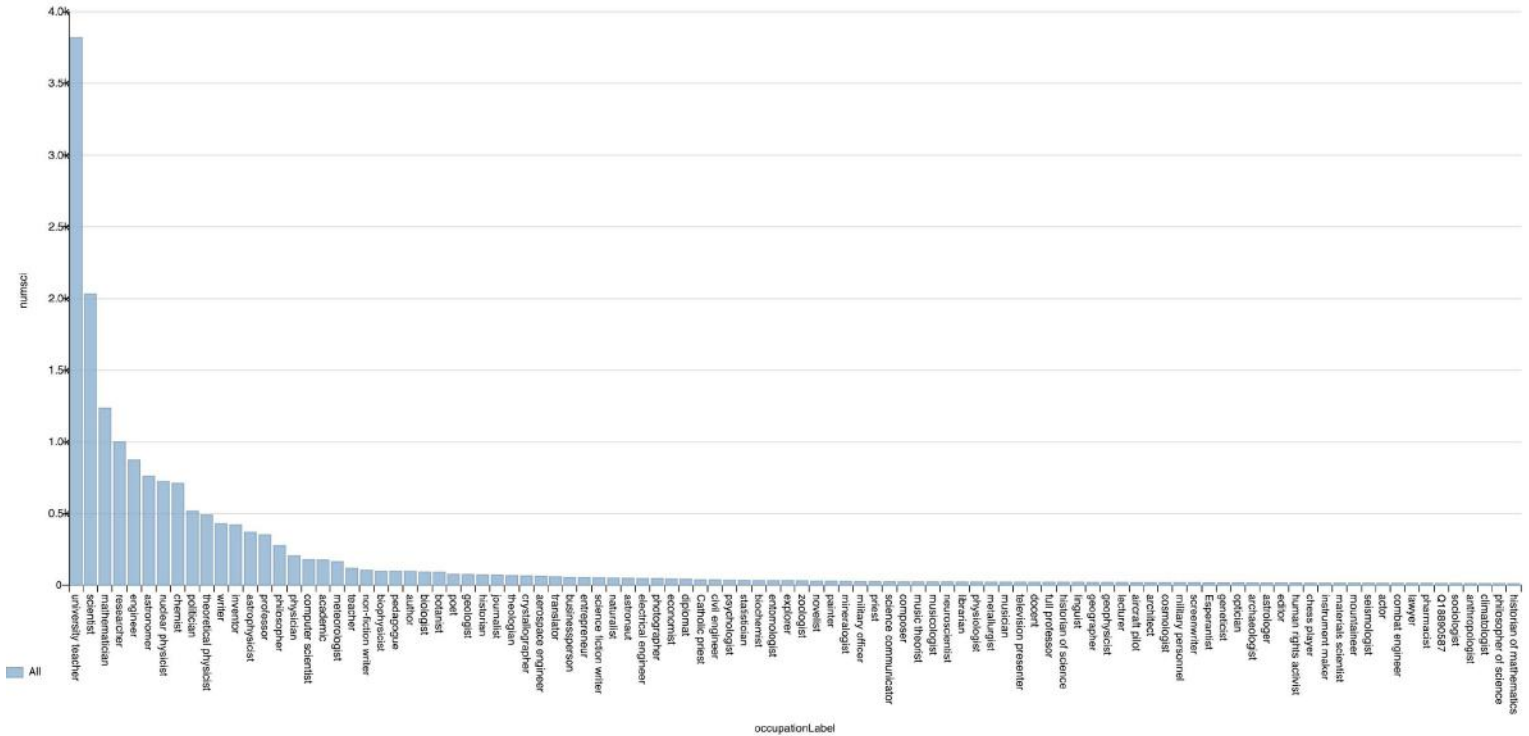
- what other occupations do Physicists have?



SPARQL query

Knowledge Graph Analytics with SPARQL

- what other occupations do Physicists have?



SPARQL query

2. Get Data

- There is so much more to know about physicists

Compose a SPARQL query to collect physicist data available in Wikidata

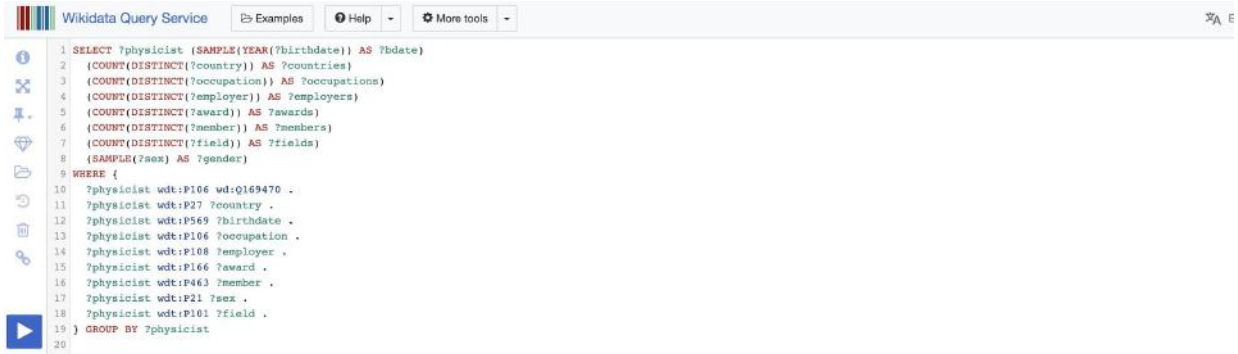
```
SELECT ?physicist (SAMPLE(YEAR(?birthdate)) AS ?bdate)
  (COUNT(DISTINCT(?country)) AS ?countries)
  (COUNT(DISTINCT(?occupation)) AS ?occupations)
  (COUNT(DISTINCT(?employer)) AS ?employers)
  (COUNT(DISTINCT(?award)) AS ?awards)
  (COUNT(DISTINCT(?member)) AS ?members)
  (COUNT(DISTINCT(?field)) AS ?fields)
  (SAMPLE(?sex) AS ?gender)
WHERE {
  ?physicist wdt:P106 wd:Q169470 .
  ?physicist wdt:P27 ?country .
  ?physicist wdt:P569 ?birthdate .
  ?physicist wdt:P106 ?occupation .
  ?physicist wdt:P108 ?employer .
  ?physicist wdt:P166 ?award .
  ?physicist wdt:P463 ?member .
  ?physicist wdt:P21 ?sex .
  ?physicist wdt:P101 ?field .
} GROUP BY ?physicist
```

Joseph Fourier, physicist

2. Get Data

- There is so much more to know about physicists

Compose a SPARQL query to collect physicist data available in Wikidata



physicist	bdate	countries	occupations	employers	awards	members	fields	gender
wd:Q937	1879	5	14	15	17	18	1	wd:Q6581097
wd:Q80	1955	1	7	6	40	5	2	wd:Q6581097
wd:Q4517	1911	3	8	1	10	3	2	wd:Q6581097
wd:Q6722	1777	3	8	1	6	12	10	wd:Q6581097
wd:Q675	1775	1	4	2	2	10	1	wd:Q6581097
wd:Q680	1745	4	4	1	3	6	1	wd:Q6581097
wd:Q464	1821	2	4	1	1	3	2	wd:Q6581097
wd:Q1585	1789	2	3	2	3	4	1	wd:Q6581097
wd:Q30693	1822	2	4	4	6	15	1	wd:Q6581097

SPARQL query

2. Get Data

[Data in Google Doc Spreadsheet]

5.6 - Physicists Data

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

	A	B	C	D	E	F	G	H
1	scientist	bdate	countries	occupations	employers	awards	members	
2	http://www.wikidata.org/entity/Q80	1955-06-08	1	7	6	40	5	
3	http://www.wikidata.org/entity/Q164384	1881-05-11	6	8	4	37	29	
4	http://www.wikidata.org/entity/Q183279	1930-03-15	2	4	2	33	12	
5	http://www.wikidata.org/entity/Q28189	1926-01-29	2	3	5	32	11	
6	http://www.wikidata.org/entity/Q71013	1941-10-30	2	3	5	31	10	
7	http://www.wikidata.org/entity/Q106624	1946-02-26	2	5	1	30	14	
8	http://www.wikidata.org/entity/Q997	1921-05-21	2	3	1	28	6	
9	http://www.wikidata.org/entity/Q83552	1916-07-11	2	3	6	28	6	
10	http://www.wikidata.org/entity/Q335213	1942-06-23	1	5	4	28	12	
11	http://www.wikidata.org/entity/Q201513	1951-08-26	1	4	1	28	8	
12	http://www.wikidata.org/entity/Q184566	1915-07-28	1	4	3	27	12	
13	http://www.wikidata.org/entity/Q487983	1903-02-13	3	3	3	26	4	
14	http://www.wikidata.org/entity/Q2587259	1946-01-23	2	1	1	26	4	
15	http://www.wikidata.org/entity/Q29573	1930-11-11	1	3	1	25	4	
16	http://www.wikidata.org/entity/Q80917	1962-09-04	1	7	3	25	5	
17	http://www.wikidata.org/entity/Q48983	1901-02-28	1	8	5	23	11	
18	http://www.wikidata.org/entity/Q91410	1962-12-23	2	4	2	23	5	
19	http://www.wikidata.org/entity/Q190697	1933-04-01	1	3	6	23	12	
20	http://www.wikidata.org/entity/Q193803	1931-08-08	1	8	7	23	7	
21	http://www.wikidata.org/entity/Q16389	1933-11-04	2	5	4	22	7	
22	http://www.wikidata.org/entity/Q1030228	1934-03-23	2	4	2	22	10	
23	http://www.wikidata.org/entity/Q148109	1910-10-19	1	5	1	22	8	
24	http://www.wikidata.org/entity/Q1975294	1921-10-15	1	6	2	22	4	
25	http://www.wikidata.org/entity/Q1248892	1934-06-30	2	2	2	22	12	
26	http://www.wikidata.org/entity/Q164396	1908-09-18	3	4	3	22	12	
27	http://www.wikidata.org/entity/Q574605	1925-06-08	2	4	1	21	7	
28	http://www.wikidata.org/entity/Q155794	1906-07-02	2	2	5	21	8	
29	http://www.wikidata.org/entity/Q172844	1894-07-08	3	3	7	21	15	
30	http://www.wikidata.org/entity/Q4106859	1935-02-02	2	4	1	21	4	
31	http://www.wikidata.org/entity/Q187224	1932-10-24	1	5	4	21	9	
32	http://www.wikidata.org/entity/Q60066	1936-10-10	1	3	4	20	12	
33	http://www.wikidata.org/entity/Q336397	1936-01-08	2	5	5	20	6	
34	http://www.wikidata.org/entity/Q237825	1929-01-23	3	3	1	20	9	
35	http://www.wikidata.org/entity/Q323320	1940-06-01	1	5	4	19	5	
36	http://www.wikidata.org/entity/Q62857	1913-06-26	1	5	4	19	5	
37	http://www.wikidata.org/entity/Q154855	1942-07-14	1	4	2	19	1	
38	http://www.wikidata.org/entity/Q172466	1948-02-28	1	3	2	19	9	

3. Clean Up Data

Can we trust this data?

[CleanedUp Data in Google Doc Spreadsheet]

5.6 - Physicists Data

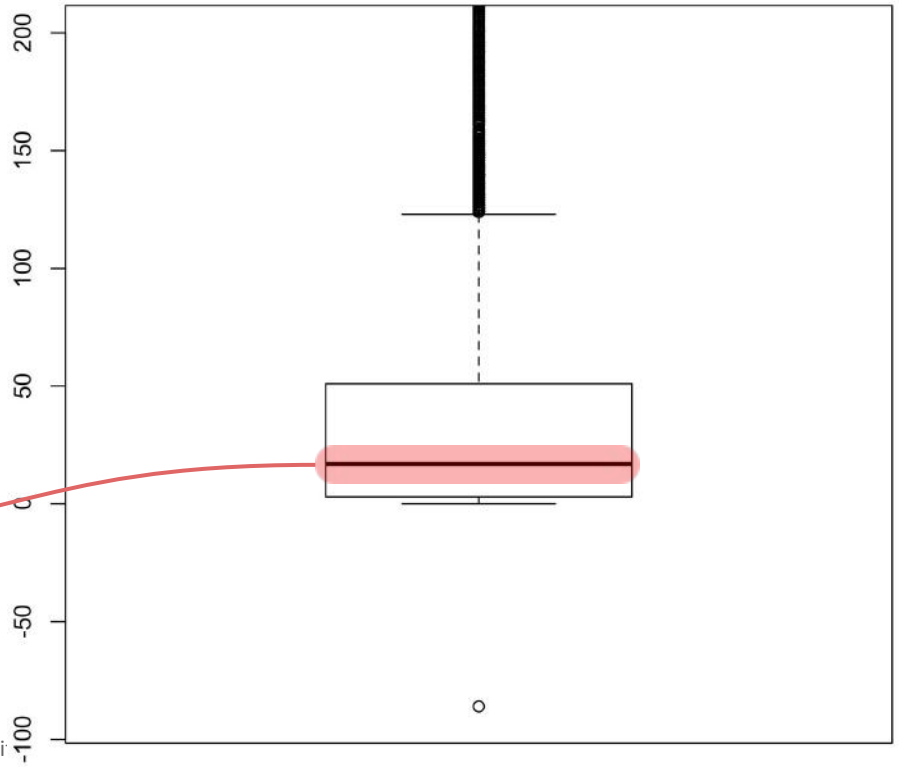
File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

	A	B	C	D	E	F	G
1	scientist	bdate	countries	occupations	employers	awards	members
2	http://www.wikidata.org/entity/Q104651	2006-08-07	1	3	1	9	3
3	http://www.wikidata.org/entity/Q5981495	2000-01-01	1	2	3	6	1
4	http://www.wikidata.org/entity/Q1479605	2000-01-01	1	2	3	5	1
5	http://www.wikidata.org/entity/Q573405	2000-01-01	1	3	4	5	3
6	http://www.wikidata.org/entity/Q656931	2000-01-01	1	2	1	4	1
7	http://www.wikidata.org/entity/Q5092271	2000-01-01	1	3	2	4	4
8	http://www.wikidata.org/entity/Q3791646	2000-01-01	1	4	1	4	2
9	http://www.wikidata.org/entity/Q1947401	2000-01-01	1	2	2	4	2
10	http://www.wikidata.org/entity/Q15432942	2000-01-01	1	4	2	4	1
11	http://www.wikidata.org/entity/Q7426990	2000-01-01	1	2	1	3	2
12	http://www.wikidata.org/entity/Q16979968	2000-01-01	1	1	1	3	1
13	http://www.wikidata.org/entity/Q2435401	2000-01-01	1	2	1	2	1
14	http://www.wikidata.org/entity/Q6776848	2000-01-01	1	2	1	2	1
15	http://www.wikidata.org/entity/Q1930849	2000-01-01	2	2	1	2	1
16	http://www.wikidata.org/entity/Q6134658	2000-01-01	1	2	1	2	1
17	http://www.wikidata.org/entity/Q15836979	2000-01-01	1	1	2	2	1
18	http://www.wikidata.org/entity/Q1516589	2000-01-01	1	2	1	1	2
19	http://www.wikidata.org/entity/Q1354655	2000-01-01	1	3	1	1	1
20	http://www.wikidata.org/entity/Q6254266	2000-01-01	1	1	1	1	1
21	http://www.wikidata.org/entity/Q26265685	2000-01-01	1	2	1	1	1
22	http://www.wikidata.org/entity/Q15990400	2000-01-01	1	4	1	1	1
23	http://www.wikidata.org/entity/Q78061707	1980-03-23	2	5	4	1	2
24	http://www.wikidata.org/entity/Q43777538	1979-01-01	1	1	1	3	2
25	http://www.wikidata.org/entity/Q21259815	1977-12-23	1	1	2	1	1
26	http://www.wikidata.org/entity/Q19502586	1977-12-18	2	2	1	2	1

Excursion: Learning How to Read Box Plots

The **median** is the value separating the higher half from the lower half of a data sample

```
       X0
Min.   : -86.00
1st Qu.:   3.00
Median :  17.00
Mean   :  38.85
3rd Qu.:  51.00
Max.   :1329.00
```

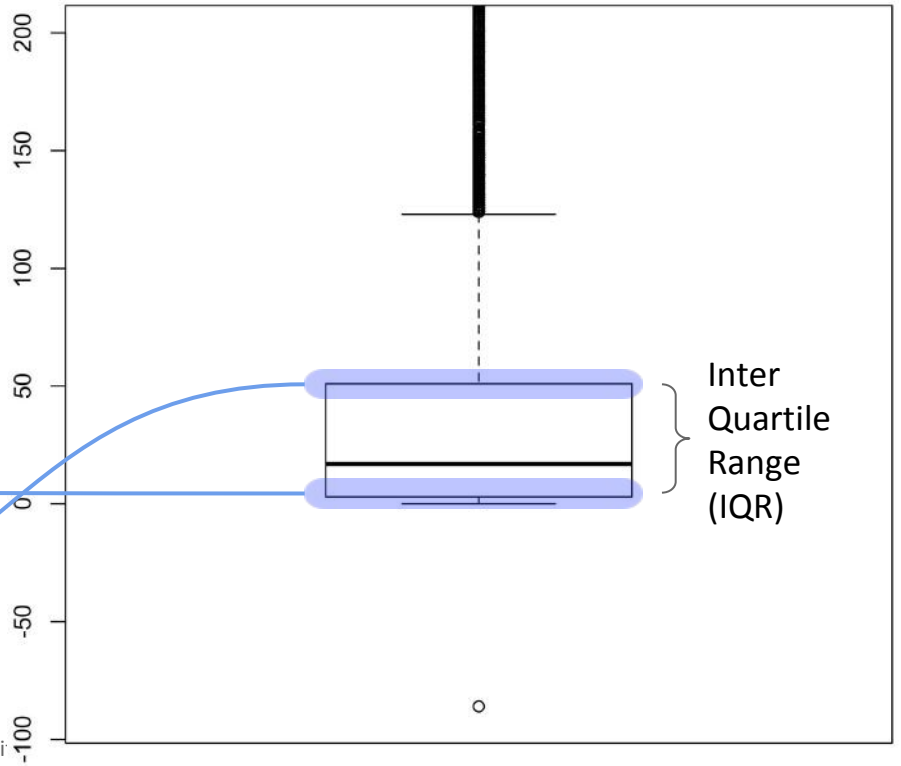


Excursion: Learning How to Read Box Plots

The **first quartile (Q_1)** is defined as the middle number between the smallest number and the median of the data set.

The **third quartile (Q_3)** is the middle value between the median and the highest value of the data set.

```
       X0
Min.   : -86.00
1st Qu.:   3.00
Median :  17.00
Mean   :  38.85
3rd Qu.:  51.00
Max.   :1329.00
```



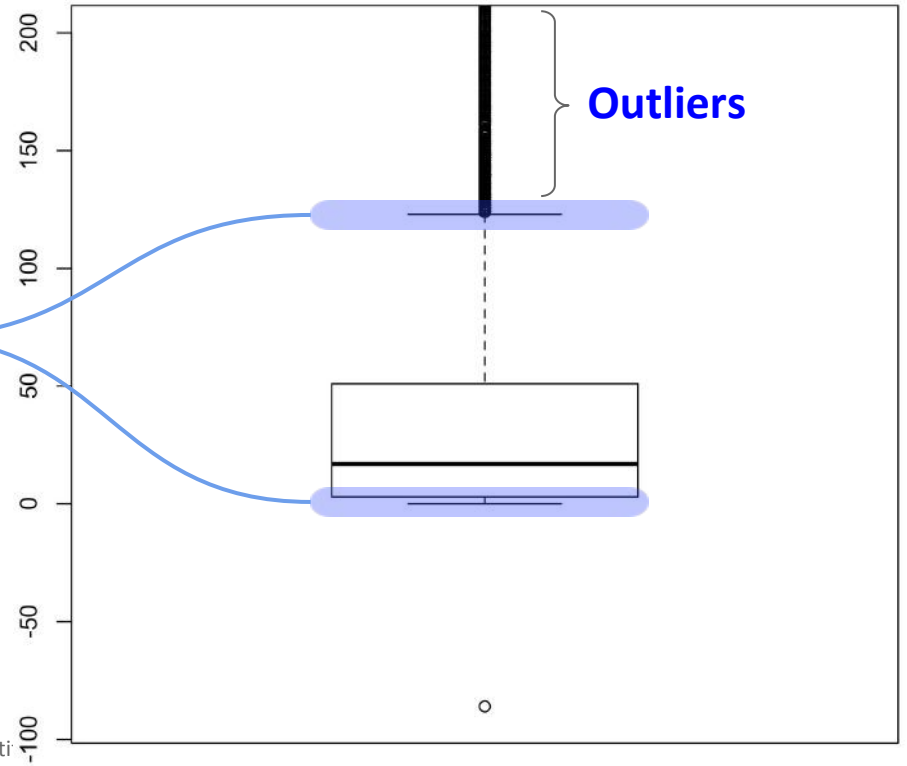
Excursion: Learning How to Read Box Plots

Whiskers are indicating variability outside the upper and lower quartiles.

Any data not included between the whiskers should be considered as an **outlier**.

Whiskers: IQR x 1.5 = (Q_3-Q_1) x 1.5

```
       X0
Min.   : -86.00
1st Qu.:   3.00
Median :  17.00
Mean   :  38.85
3rd Qu.:  51.00
Max.   :1329.00
```

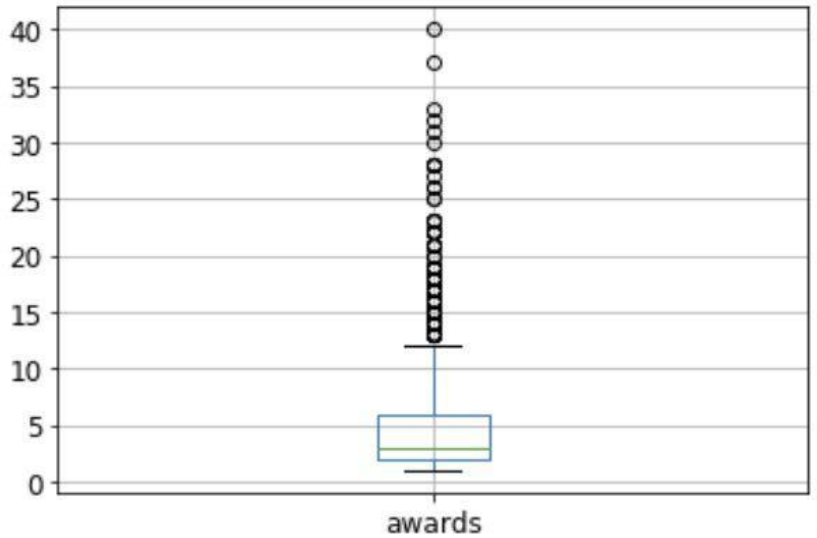


Excursion: Learning How to Read Box Plots

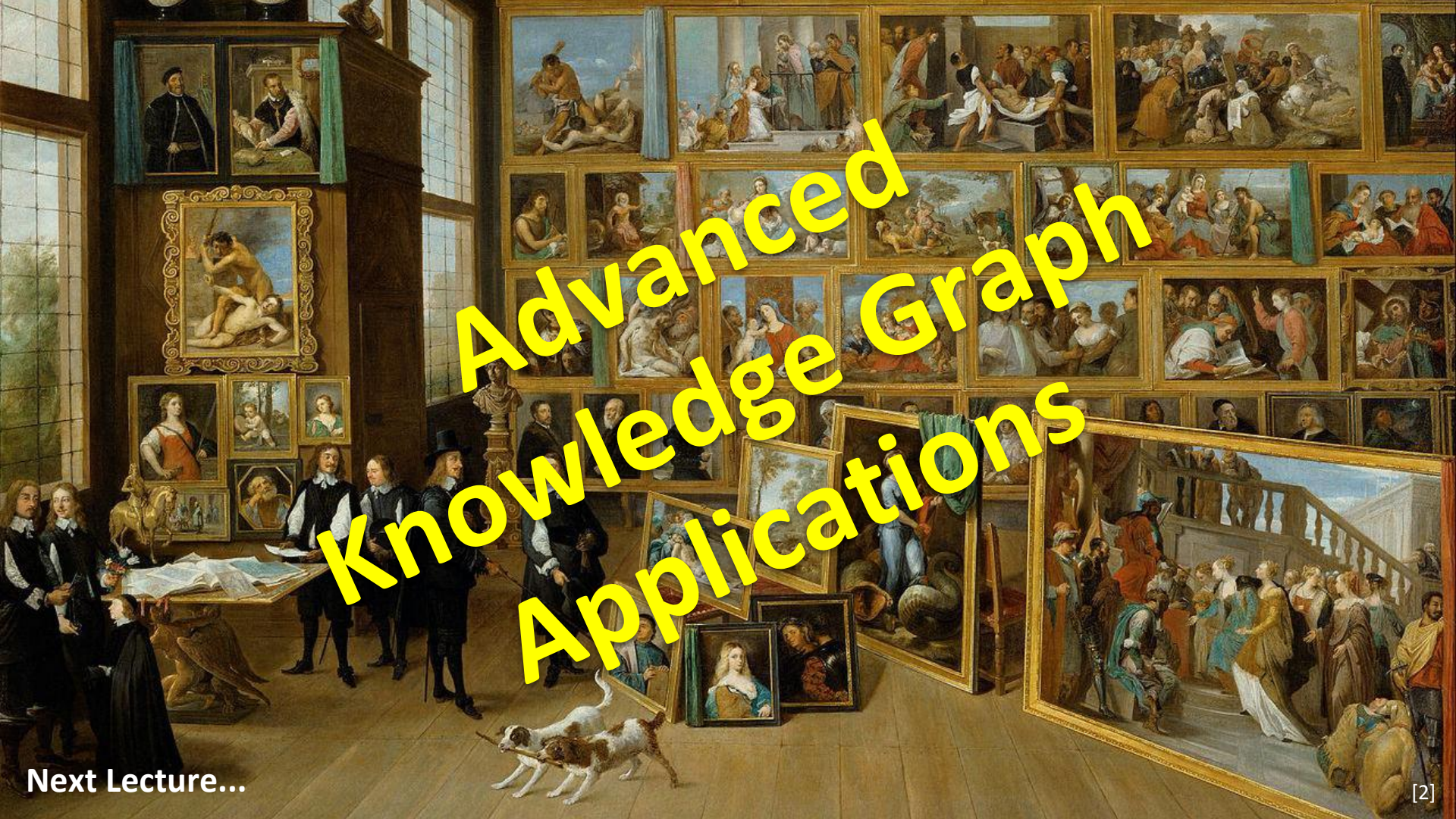
4. **Analyse the Data**
   E.g. via python

```
[29] physicists.boxplot('awards')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3c00578f28>
```



Data in Google Collab Notebook



Next Lecture...

[2]

Picture References:

- [1] Louis Leopold Bouilly, Engraved portrait of French mathematician Jean Baptiste Joseph Fourier (1768 - 1830), early 19th century. [Public Domain]
  https://commons.wikimedia.org/wiki/File:Fourier2.jpg
- [2] David Teniers the Younger, Archduke Leopold Wilhelm in his Gallery in Brussels (1651). [Public Domain]
  https://commons.wikimedia.org/wiki/File:David_Teniers_the_Younger_-_Archduke_Leopold_William_in_his_Gallery_at_Brussels_-_Google_Art_Project.jpg