Group A3D: SPARQL Queries and Analytics

Group members: Andrea Bruttomesso 2120933
Alessandro Corrò 2125034
Davide Seghetto 2122548
Andrea Stocco 2108885

Task

Provide at least 8 SPARQL queries over your RDF datasets. You may also perform advanced data analytics to uncover interesting insights from your datasets. Please submit a PDF that includes the SPARQL queries along with relevant plots or tables summarizing your analytics. For each query, provide a description that explains its purpose and overall objective.

1 Relationship between Nobel Prize winning ideas and published studies

```
PREFIX spif: <http://spinrdf.org/spif#>
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?nobelTopic ?nobel (COUNT(?paper) AS ?numPapers) WHERE {
    {
        SELECT ?paperTopic ?paper WHERE {
            ?paper :hasAbstractTopics ?topics ;
                :hasYear ?year .
            FILTER (?year = "2004"^^xsd:gYear)
            ?paperTopic spif:split(?topics ",")
        }
    }
    {
        SELECT ?nobelTopic ?nobel WHERE {
            ?nobel :hasMotivationTopics ?topics ;
                :hasYear ?year .
            FILTER (?year = "2004"^^xsd:gYear)
            ?nobelTopic spif:split(?topics ",")
        }
    }
    FILTER (?nobelTopic = ?paperTopic)
}
GROUP BY ?nobelTopic ?nobel
ORDER BY DESC (?numPapers)
```

This query shows the topics present in both Nobel Prize motivations and paper abstracts. For a given year, it returns the number of paper in which a Nobel topic appears. This query can be used to find correlations between Nobel Prize topics and research papers.
Table 1 shows the output of this query on year 2004.

Table 1: Number of papers per Nobel topic in 2004

Nobel topic	Nobel Prize	Number of papers
protein	Chemistry 2004	28
development	Peace 2004	13
flow	Literature 2004	8
interaction	Physics 2004	4
discovery	Chemistry 2004	3
discovery	Physics 2004	3
degradation	Chemistry 2004	3
asymptotic	Physics 2004	2
forces	Economics 2004	1
cycles	Economics 2004	1
olfactory	Medicine 2004	1
organization	Medicine 2004	1

Considering the limited number of papers available, the topic "protein" appeared in 28 papers. The high number of papers mentioning this Nobel topic suggests that it was widely discussed or relevant in 2004.
We cannot conclude whether the research area of molecular biology was particularly active in that year, but in Section 2 we will further investigate this.
Unfortunately, this query is not always useful. In some cases, the main topics may include words like "method" and "analysis", which are not informative enough to determine how extensively a specific topic was studied in a given year.
Due to the distribution of research papers in our dataset across different years, this query provides more meaningful results for years after 2000.

2 Most active research areas in a year

```
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT ?category (COUNT(?paper) AS ?numPapers) WHERE {
    ?paper :publishedIn ?venue ;
        :hasYear ?year .
    ?venue :hasJournalCategory ?category .
    ?category skos:broaderTransitive ?sub .
    FILTER (?year = "2004"^^xsd:gYear)
}
GROUP BY ?category
ORDER BY DESC (?numPapers)
```

This query shows the number of papers published for each journal subcategory for a given year. It can be used to identify which research areas were particularly active in that year.
Table 2 continues the analysis started in the previous section.

Table 2: Number of papers for each journal subcategory in 2004

Journal Subcategory	**Number of papers**
Biochemistry Genetics Molecular Biology	418
Social Sciences	344
Decision Sciences	125
Arts Humanities	74
Business Management Accounting	68
Physics Astronomy	65
Neuroscience	55
Health Professions	34
Psychology	27
Earth Planetary Sciences	22
Economics Econometrics Finance	14
Materials Science	12
Environmental Science	12
Agricultural Biological Sciences	10
Energy	2
Pharmacology Toxicology Pharmaceutics	2

Considering the limited number of papers in our dataset, molecular biology was the most active research area in 2004.
Building on the previous section, 28 out of 418 molecular biology papers focused on "protein".
This topic held central importance that year, which may explain why a Nobel Prize was awarded for it.

3 papersPerTopic

```
PREFIX spif: <http://spinrdf.org/spif#>
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>

SELECT ?singleTopic (COUNT(?paper) AS ?numPapers) WHERE {
    ?paper :hasAbstractTopics ?topics .
    ?singleTopic spif:split(?topics ",")
}
GROUP BY ?singleTopic
ORDER BY DESC (?numPapers)
```

4 Number of shared Nobels and number of Laureates sharing multiple Nobels

```
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>
SELECT ?share (COUNT(?nobel) AS ?nNobels) (SUM(?share) AS ?nLaureates) WHERE {
    {
        SELECT ?nobel (COUNT(?laureate) AS ?share) WHERE {
            ?laureate :hasWon ?nobel .
        }
        GROUP BY ?nobel
    }
}
GROUP BY ?share
ORDER BY ASC(?share)
```

This query shows the number of Nobel Prizes shared by multiple laureates and the number of laureates sharing Nobel Prizes.
The query provides an interesting result: 235 out of 579 Nobel Prizes (40.6%) have been shared by multiple laureates, and 567 laureates have shared different Nobel Prizes. The chart in Figure 1 summarizes the data

obtained from the query.
For each prize share on the x-axis, two columns are displayed:

- the first column represents the total number of Nobel Prizes shared among the specified number of laureates.
- the second column represents the total number of laureates who have shared these Nobel Prizes.

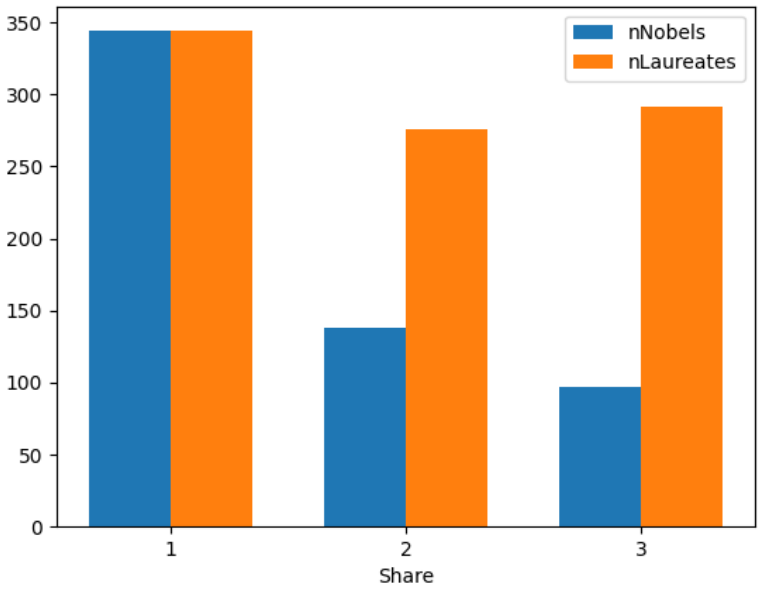


Figure 1: Graph showing the distribution of Nobel Prizes among laureates.

Another interesting fact is that the number of laureates who have won a Nobel Prize shared among three people (291) is greater than the number of laureates who have won a Nobel Prize shared between two people (276).

5 Collaborations among Nobel Laureates

```
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?title (GROUP_CONCAT(?name; SEPARATOR = ", ") AS ?laureates) WHERE {
    ?laureate rdf:type :Laureate .
    ?paper rdf:type :Paper ;
        :hasTitle ?title .
    ?laureate :hasWritten ?paper .
        foaf:name ?name .
}
GROUP BY ?title
HAVING (COUNT(DISTINCT ?laureate) > 1)
```

The goal of this query is to explore whether Nobel laureates collaborate with each other by co-authoring scientific papers. As Table 3 shows, among approximately 53,000 papers and 904 laureates, only one paper was found to have been co-authored by multiple laureates.

Table 3: Paper co-authored by multiple Nobel Laureates

Title	**Laureates**
Recursive Robust Estimation and Control Without Commitment	Lars Peter Hansen, Thomas J. Sargent

This result suggests that collaboration between Nobel laureates is extremely rare. However, it is important to note that this outcome should not be taken as definitive, as the datasets used represent only a portion of all existing papers and laureates. Nevertheless, it provides an interesting insight into the rarity of such collaborations, offering a percentage-based perspective on how seldom laureates join forces to produce scientific work.

6 How fundings in R&D affect the possibility for a country to win a Nobel?

```
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?year ?topCountry (COUNT(DISTINCT ?laureate) AS ?numLaureates) (SUM(?
    fundingAmount) AS ?totalFunding) WHERE {
    ?laureate rdf:type :Laureate ;
    :hasWon ?nobelPrize ;
        :bornIn ?city .
    ?nobelPrize :hasYear ?year .
    ?city :locatedIn ?topCountry .

    OPTIONAL {
        ?topCountry :hasFunded ?funding .
        ?funding :hasYear ?year ;
            :hasAmount ?fundingAmount .
    }
    { # Select country with most laureates
        SELECT (?country AS ?topCountry) WHERE {
        ?laureate rdf:type :Laureate ;
            :bornIn ?city .
        ?city :locatedIn ?country .
        }
        GROUP BY ?country
        ORDER BY DESC(COUNT(DISTINCT ?laureate))
        LIMIT 3
    }
}
GROUP BY ?year ?topCountry
HAVING (SUM(?fundingAmount) > 0)
ORDER BY ?year ?topCountry
```

With this query, we identified the top three countries with the highest number of Nobel laureates born there, along with the annual amount of funding allocated to research and development (R&D) by these nations. To ensure data consistency, we focused exclusively on the years from 2000 to 2016.

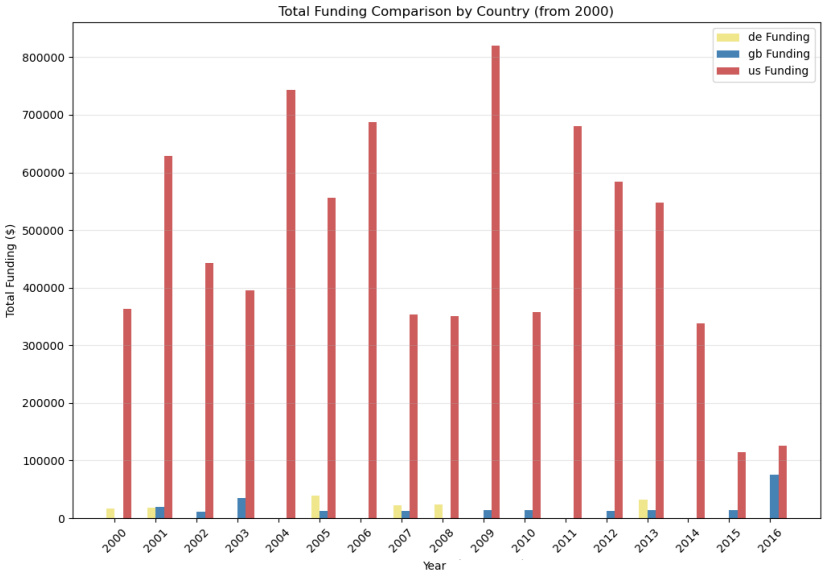


Figure 2: Funding Comparison by Country

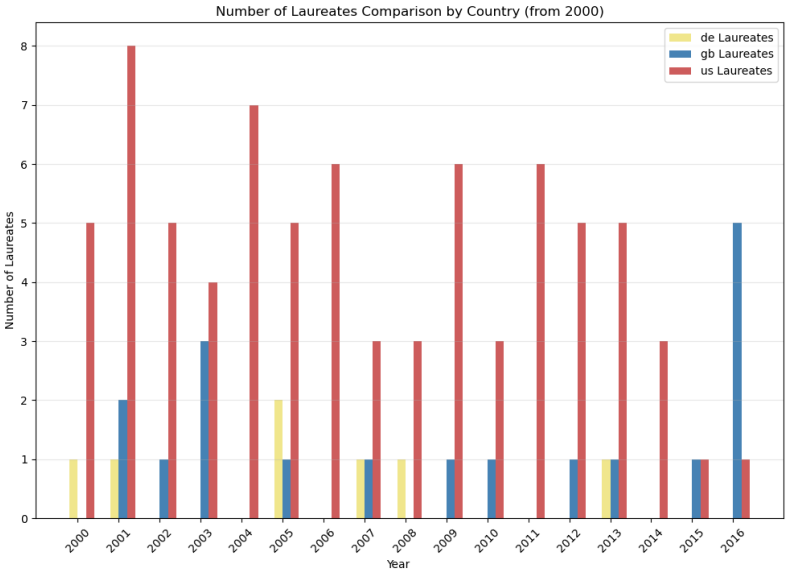


Figure 3: Laureates Comparison by Country

Charts in Figure 2 and 3 reveal a strong correlation between R&D funding and the number of Nobel laureates. In particular, the United States dominates both metrics, demonstrating how substantial investments in research directly contribute to significant achievements in this field, resulting in a higher number of laureates annually.

The situation in Great Britain, highlighted in the chart in Figure 4, is particularly curious and further supports this observation:

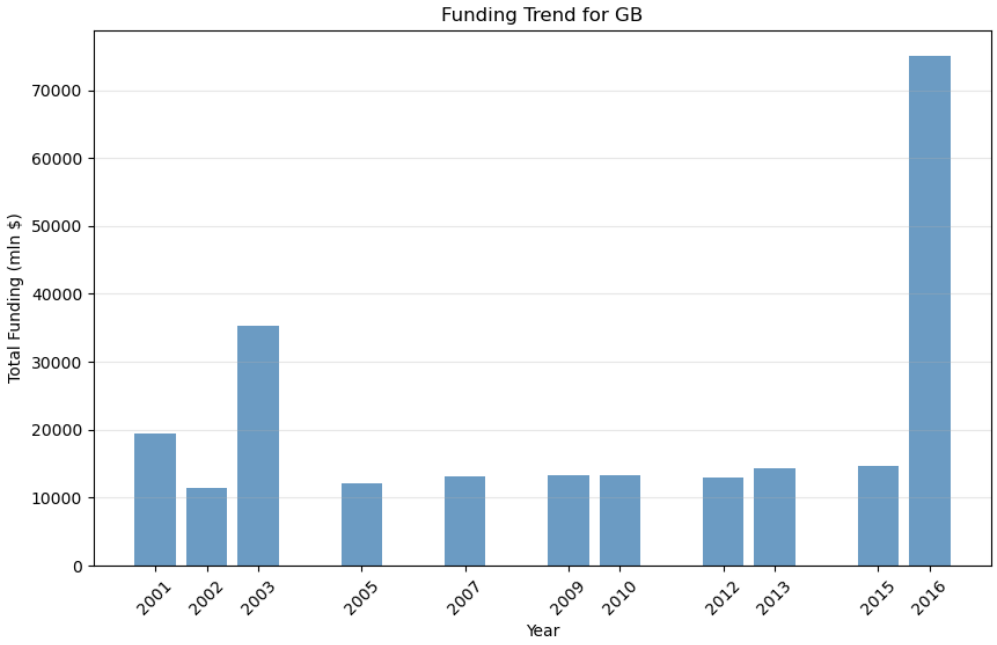


Figure 4: Great Britain R&D Funding Trend

It's clear that the trends in funding and the number of laureates mirror each other closely. From 2001 to 2003, we observe the same pattern in both metrics. Subsequently, a steady and low level of R&D funding still reflects the number of British Nobel laureates until 2016, when a sharp increase in Nobel prizes matches with a significant rise in R&D investments.

7 Laureates that won multiple Nobels

```
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?laureate (COUNT(?nobel) AS ?numNobels) (GROUP_CONCAT(DISTINCT ?category;
    SEPARATOR = ", ") AS ?categories) WHERE {
    ?laureate :hasWon ?nobel .
    ?nobel :hasNobelCategory ?category .
}
GROUP BY ?laureate
HAVING (?numNobels > 1)
ORDER BY DESC (?numNobels)
```

With this query we aim to discover Laureates that won multiple Nobel Prizes and for which categories.
Table 4 highlights that only 6 Laureates have been awarded multiple times and only 3 Nobel categories have repeated winners.

Table 4: Laureates who won more than one Nobel Prize

Laureate	**Number of Nobels won**	**Categories**
Comite International De La Croix-Rouge	3	Peace
Frederick Sanger	2	Chemistry
John Bardeen	2	Physics
Linus Carl Pauling	2	Chemistry, Peace
Marie Curie	2	Physics, Chemistry
UNHCR	2	Peace

It is interesting to note that all of these individuals or organizations won a prize in the same category, or a similar one in the case of Marie Curie, except for one. Linus Carl Pauling won both the Chemistry and Peace Nobel

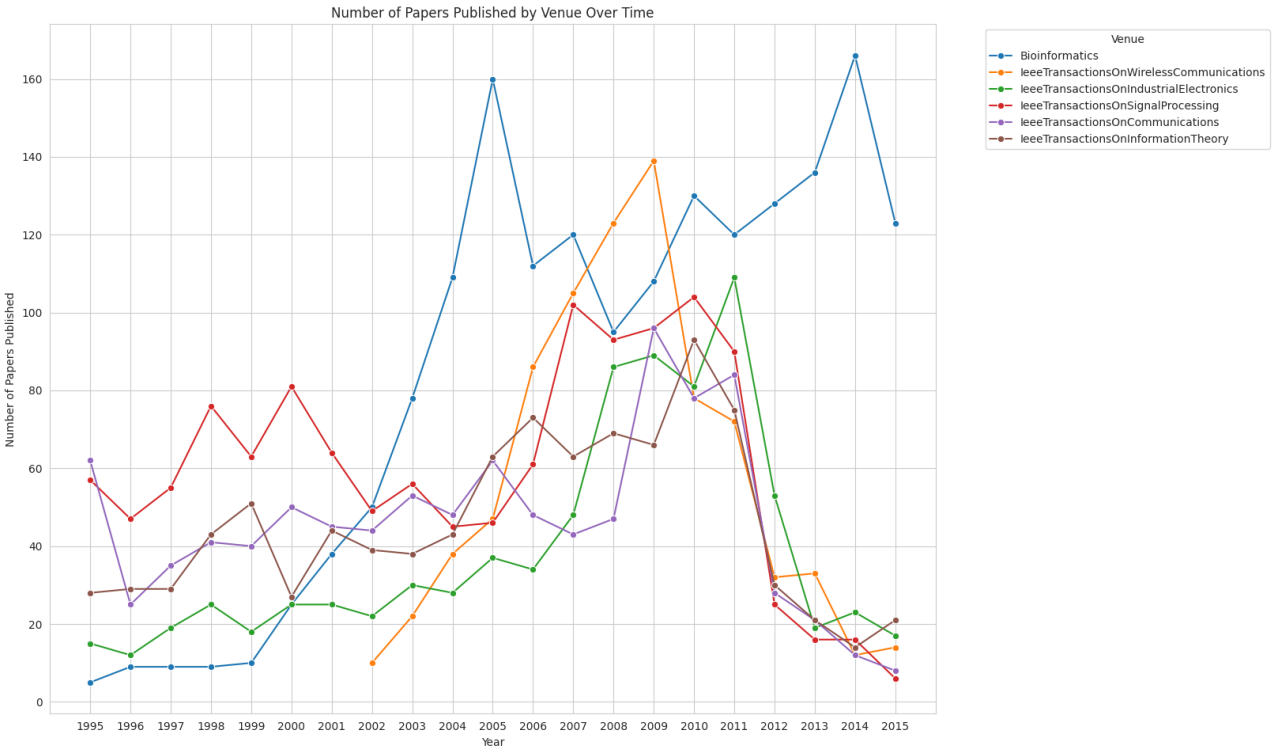
Prizes, suggesting that he was not only a scientist but also a socially engaged figure.
Additionally, no repeat winners are found in the categories of Economics, Literature and Medicine. For Economics and Literature this absence might be partially attributed to the average high age of Nobel Prize winners in these fields, as further analyzed in Section 10.

8 Number of papers published from the most important venues over the years

```
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>

SELECT ?venue ?year (COUNT(?paper) AS ?numPapers) WHERE {

    # Get the most important venues (the ones with at least 800 papers published)
    {
        SELECT ?venue (COUNT(?paper) AS ?totPapers) WHERE {
            ?paper :publishedIn ?venue .
        }
        GROUP BY ?venue
        HAVING (?totPapers > 800)
        ORDER BY DESC (?totPapers)
    }

    # get the number of paper published in the most important venues for each year
    ?paper :publishedIn ?venue ;
        :hasYear ?year .
}
GROUP BY ?venue ?year
ORDER BY ASC (?year)
```

This query returns the number of papers published over the years by major venues (those with at least 800 papers published, according to our dataset).

Figure 8 shows the trends of the six major research venues.



In recent years, Bioinformatics could be considered one of the most influential venue due to its consistently higher number of papers published compared to others. IEEE venues, are the most prominent in the fields of

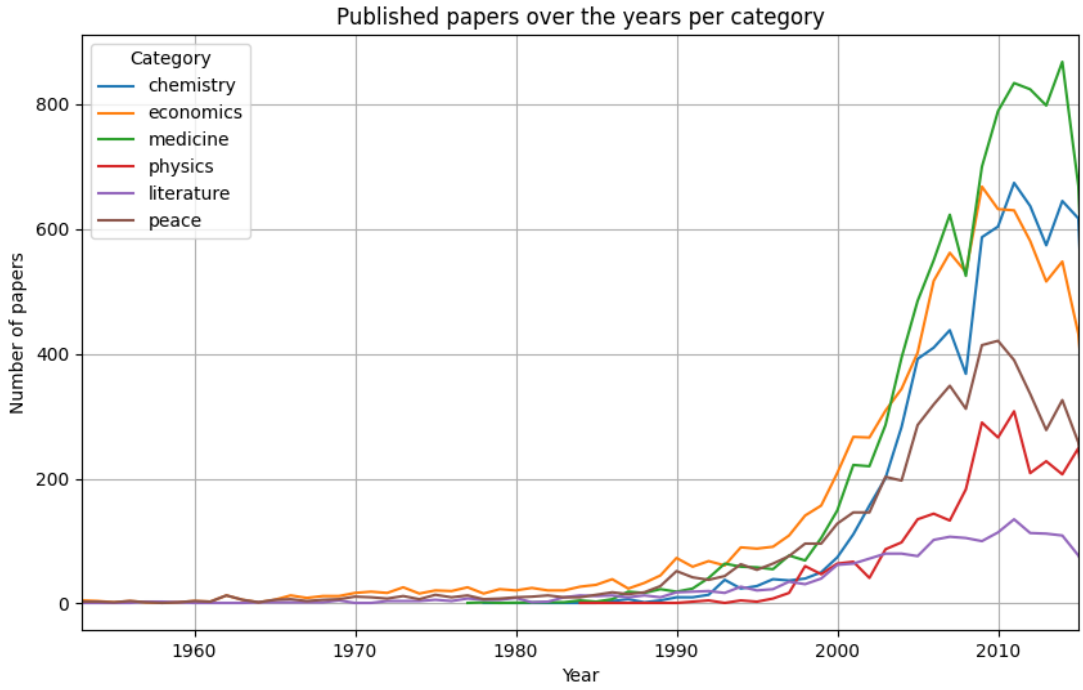
information and tecnology.
For instance, on 2009, the research community focused more on the field of communications. That same year, the Physics Nobel Prize was awarded for "groundbreaking achievements concerning the transmission of light in fibers for optical communication".

9 Number of papers published for each Nobel category

The following query allows us to extract, for each year, the number of scientific articles published in each relevant category. The categories returned as results are the TopConcepts categories of our SKOS taxonomy, and they include in the count their various subcategories. For example, in the count of papers for the medicine category, articles belonging to subcategories like neuroscience are also included.
To obtain this data, the query uses two distinct subqueries. The first subquery extracts the number of articles published for each main category (TopConcept), while the second identifies the number of articles associated with the subcategories of each main category. The sum of the results of the two subqueries, aggregated by year and category, provides the total number of articles published for each category and for each year.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>
# Extracts the number of papers we have for each category over the years --> the most studied research areas over the years
SELECT ?year ?category (SUM(?howmany) AS ?totalPapers) WHERE {
    # Inner query to extract the number of papers published in journals that have at least one category that is a top concept of our skos scheme
    {
        SELECT ?year ?category (COUNT(DISTINCT ?paper) AS ?howmany) WHERE {
            ?journal :hasJournalCategory ?category .
            :journalCategoryScheme skos:hasTopConcept ?category .
            ?paper :publishedIn ?journal ;
                :hasYear ?year .
        }
        GROUP BY ?year ?category
    }
    UNION
    # Inner query to extract the number of papers published in journals that have at least one category that is a subcategory of a top concept category
    {
        SELECT ?year ?category (COUNT(DISTINCT ?paper) AS ?howmany) WHERE {
            ?journal :hasJournalCategory ?cat .
            ?cat skos:broaderTransitive ?category .
            ?paper :publishedIn ?journal ;
                :hasYear ?year .
        }
        GROUP BY ?year ?category
    }
}
GROUP BY ?year ?category
ORDER BY DESC (?totalPapers)
```

This approach offers a comprehensive view of the distribution of published articles over time, allowing us to identify which research areas, related to Nobel categories, have attracted the most attention from scholars over the years. Below is a plot representing the trend of the number of papers published over the years, divided by their respective categories. In the recent years, the most studied field is medicine which got a big leap around 2002, while in the nineties the most studied one was economics, which reached its peak in 2008, probability due to the economic crisis of that time.



10 Age Analysis of Nobel Prize Winners

```
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?category (MIN(?age) AS ?maxAge) (ROUND(AVG(?age)) AS ?avgAge) (MAX(?age) AS ?
    minAge) WHERE {
    ?laureate a :Laureate ;
        :birthDate ?birthDate ;
        :hasWon ?prize .
    ?prize :hasYear ?prizeYear ;
        :hasNobelCategory ?category .

    BIND (YEAR(?prizeYear) - YEAR(?birthDate) AS ?age)
}
GROUP BY ?category
```

The data extracted from the query provides insights into the age at which individuals achieve great results in their fields:

Table 5: Age statistics of Nobel Prize winners by category.

Category	Min Age	Avg Age	Max Age
Chemistry	35	58	85
Economics	51	67	90
Literature	42	65	88
Medicine	32	58	87
Peace	17	61	87
Physics	25	55	88

The table 5 highlights notable trends in the ages of Nobel Prize winners:

- **Economics**: With an average age of 67, this category has the highest average age. This observation likely reflects the time required to develop groundbreaking theories and gain significant recognition in this field. Also the minimum age of 51 suggests that Nobel laureates in Economics often achieve their recognition later in life.
- **Peace**: This category includes the youngest laureate, at only 17 years old. The diversity in ages within this category (average age 61, maximum 87) reflects the variety of contributions recognized, from lifelong efforts in diplomacy to impactful single events, such as activism or humanitarian work.
- **Chemistry, Medicine, and Physics**: These science-related categories show similar average ages (55-58 years) and a range of minimum ages (from 25 to 35 years). The relatively young ages of some laureates in Physics (25) and Medicine (32) suggest that groundbreaking discoveries in these fields are sometimes made early in a researcher's career, possibly during postdoctoral experiences.
- **Literature**: With an average age of 65, second only to Economics, and a minimum age of 42, this category highlights the time often required for authors to improve and craft a masterpiece. Furthermore, the Nobel Prize in Literature is frequently awarded in recognition of an author's entire career, considering their overall contribution to the field rather than a single achievement.

By digging deeper into the age data, we can perform an additional query to identify the youngest and oldest Nobel Prize winners.

```
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?laureate ?name ?birthDate ?prize ?prizeYear ?age WHERE {
  {
    SELECT ?laureate ?name ?birthDate ?prize ?prizeYear ?age WHERE {
        ?laureate a :Laureate ;
            foaf:name ?name ;
            :birthDate ?birthDate ;
            :hasWon ?prize .
        ?prize :hasYear ?prizeYear .
        BIND (YEAR(?prizeYear) - YEAR(?birthDate) AS ?age)
    }
    ORDER BY ?age
    LIMIT 1
  }
  UNION
  {
    SELECT ?laureate ?name ?birthDate ?prize ?prizeYear ?age WHERE {
        ?laureate a :Laureate ;
            foaf:name ?name ;
            :birthDate ?birthDate ;
            :hasWon ?prize .
        ?prize :hasYear ?prizeYear .
        BIND (YEAR(?prizeYear) - YEAR(?birthDate) AS ?age)
    }
    ORDER BY DESC (?age)
    LIMIT 1
  }
}
```

The results are presented in the following table 6:

Table 6: The youngest and oldest Nobel Prize winners.

Name	Birth Date	Prize	Year	Age
Malala Yousafzai	1997-07-12	Peace 2014	2014	17
Leonid Hurwicz	1917-08-21	Economics 2007	2007	90

The youngest Nobel Prize winner is Malala Yousafzai, who received the Peace Prize in 2014 at the age of just 17. Her recognition came as a result of her activism for girls' education and human rights, which gained international attention and inspired millions worldwide.
The oldest Nobel laureate is Leonid Hurwicz, who was awarded the Nobel Prize in Economics in 2007 at the age of 90. This result demonstrates a lifetime of work in economics, highlighting how breakthroughs in theoretical fields often come from years and years of efforts and experience.

11 Nobel Laureates: Birthplace vs. Research Location by State

The following query returns, for each state, the number of laureates who were born in that state but conduct research abroad, and the number of laureates who were born in that state and conduct reasearch there.
This allows us to understand how many Nobel Prize winners conducted research in their home country or abroad in order to earn the recognition.

```
PREFIX : <http://www.semanticweb.org/a3d/ontologies/2024/10/nobelOntology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?state ?bornIn ?bornAndWorkIn WHERE {
  {
        SELECT ?state (COUNT(DISTINCT ?laureate) AS ?bornIn) WHERE {
          ?laureate a :Laureate ;
              :bornIn ?city .
          ?city :locatedIn ?country .
          ?country foaf:name ?state .
        } GROUP BY ?state
    }

    {
        SELECT ?state (COUNT(distinct ?laureate) AS ?bornAndWorkIn) WHERE {
            ?laureate a :Laureate ;
                :worksFor ?organization ;
                :bornIn ?birthCity .
            ?organization :basedIn ?orgCity .
            ?orgCity :locatedIn ?country .
            ?birthCity :locatedIn ?country .
            ?country foaf:name ?state
        } GROUP BY ?state
    }
} ORDER BY DESC (?bornIn)
```

Figure 5 displays the data returned by the query, while table 7 reports the five states with the lowest percentage of laureates active in research in their home country.
As expected, the US is the state with the most laureates and also with the highest percentage of laureates active in their home country (88%), thanks to the large fundings allocated in R&D, as seen before.
An interesting fact is that Japan, despite being the 7th country in terms of Nobel Prizes won, manages to retain nearly 67% of its laureates within its borders. On the other hand, Russia, despite being the 6th country in terms of Nobel Prizes won, has a retention rate of only 35%.

State	Laureates	Laureates active in their home country	Percentage
Canada	19	2	10,5%
China	11	2	18,2%
Italy	19	4	21%
Austria	17	4	23,5%
Australia	11	3	27,3%

Table 7: How many laureates are active in research in their home country.

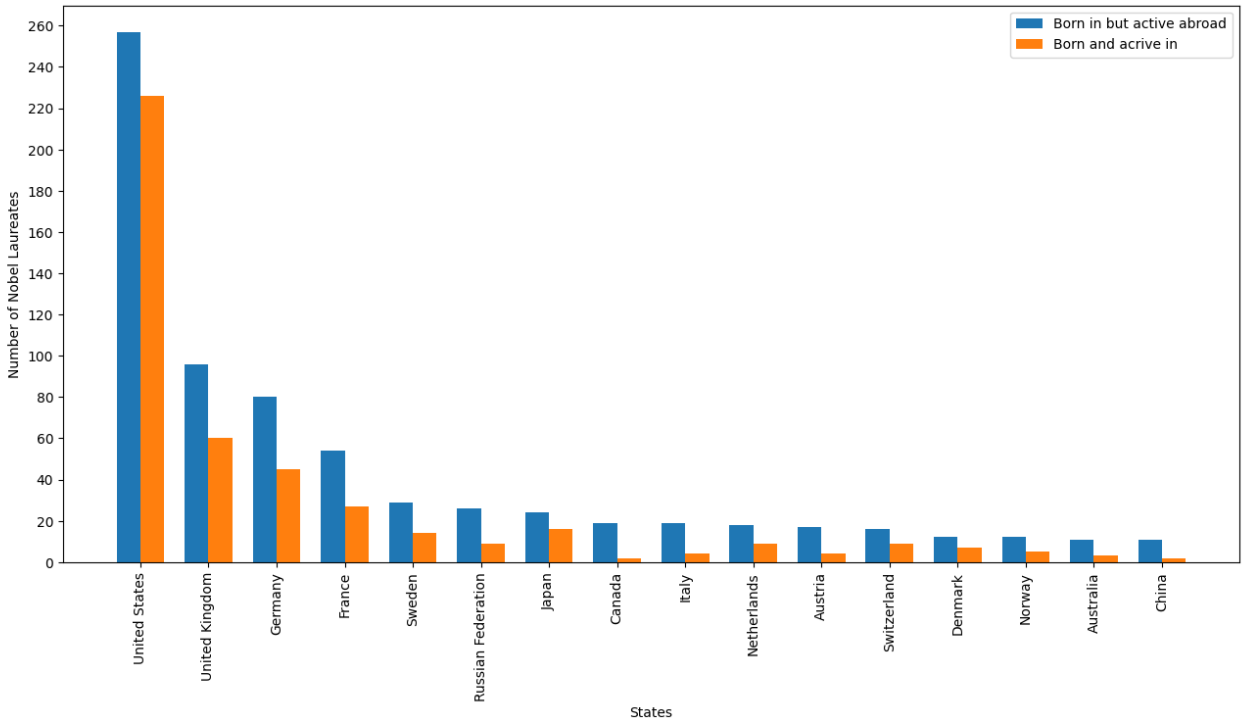


Figure 5: The plot shows for each state two columns. The blue one indicates the number of laureates born in that state but active abroad. The orange one indicates the number of laureates born and active in that state.