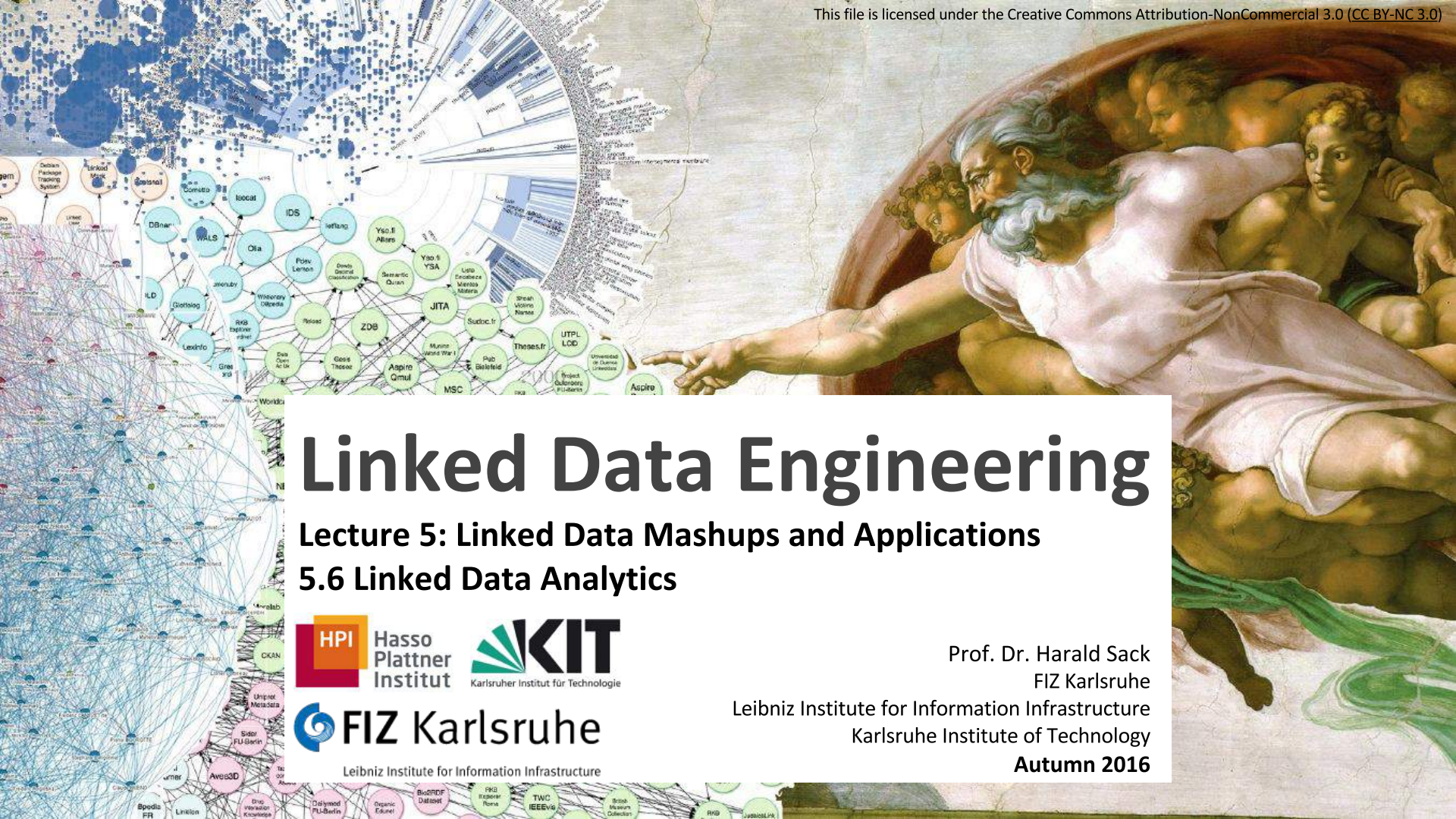




Linked Data Engineering

Lecture 5: Linked Data Mashups and Applications

5.6 Linked Data Analytics



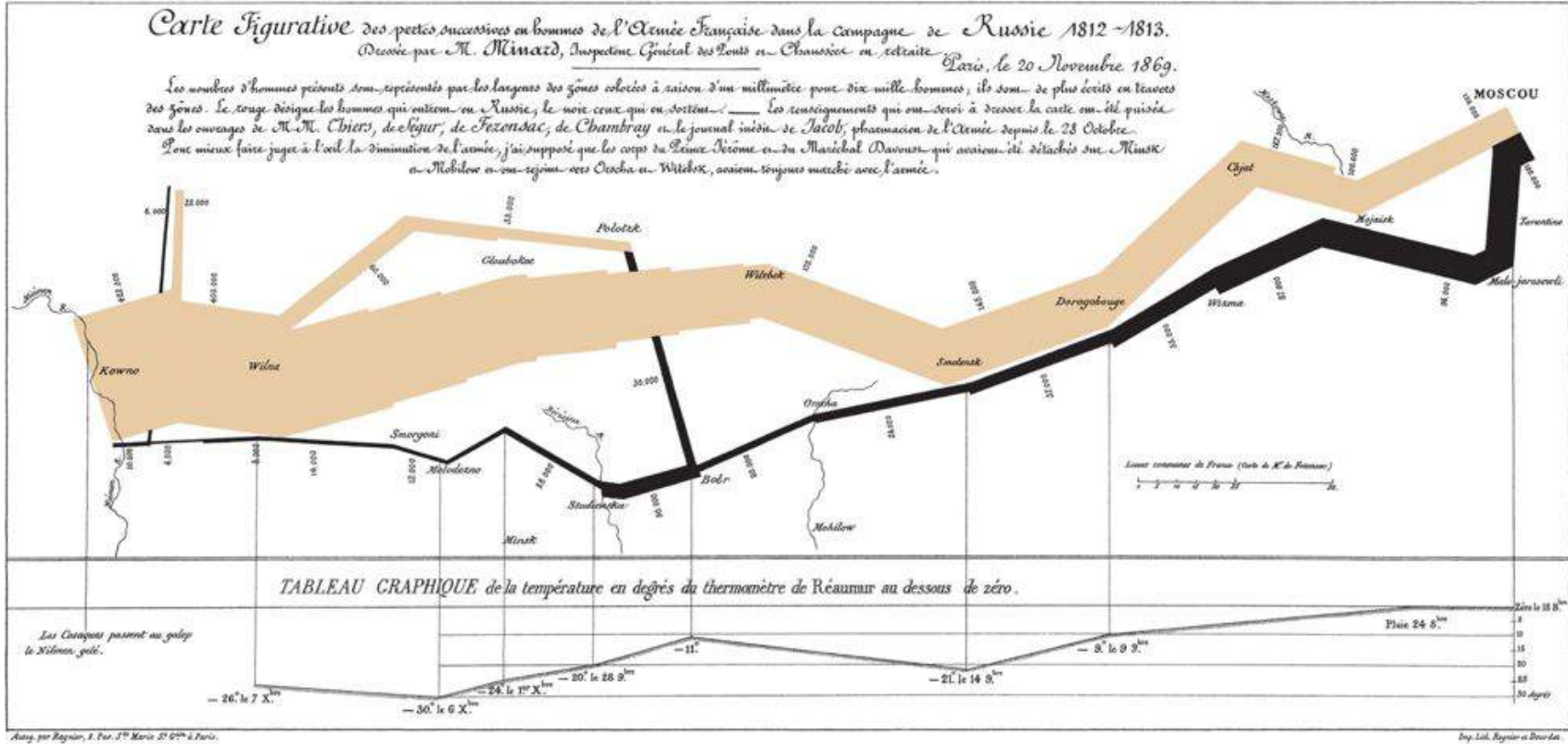
Leibniz Institute for Information Infrastructure

Prof. Dr. Harald Sack
FIZ Karlsruhe

Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology

Autumn 2016

A Picture tells more than a 1000 Words...

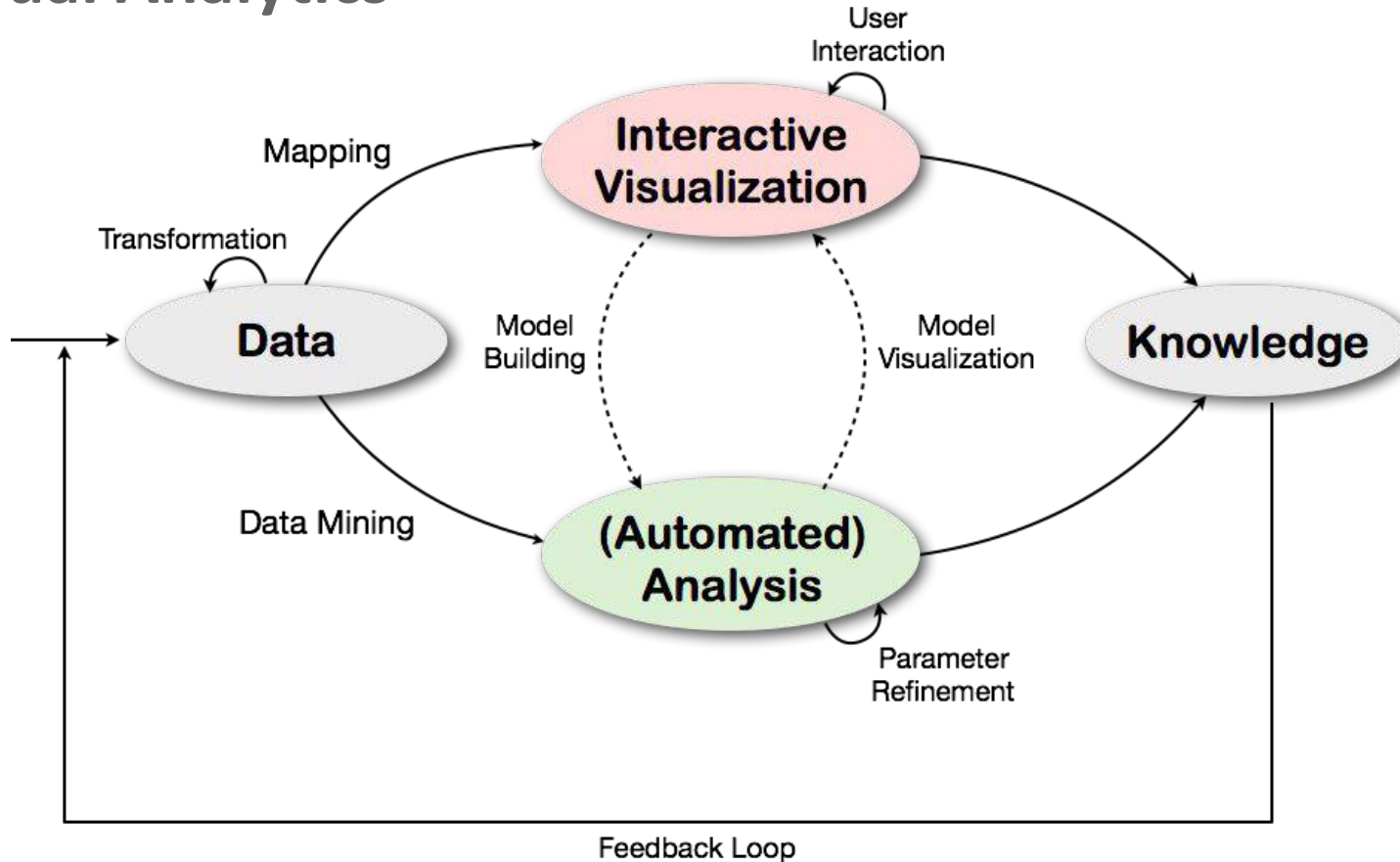


Charles Joseph Minard (1781-1870) - Napoleon's Russian Campaign

Data Analysis

- **Data Analysis** is a fundamental iterative process:
 1. Formulation and execution of a query
 2. Analysis of the results
 3. Formulation of a consecutive query based on the achieved results
- **Goals of Data Analysis:**
 - maximize understanding of analyzed data
 - uncover hidden structures/patterns
 - extraction of important variables
 - detection of anomalies and outliers
 - testing of hypotheses
 - development of a simple model

Visual Analytics



Interactive Data Analysis with Linked Data

- Available Toolset:
 - Data: **DBpedia SPARQL endpoint**
- Query language:
 - **SPARQL**
- simple statistics and visualization:
 - **R**
<https://www.r-project.org/>



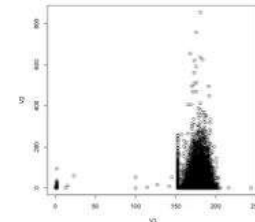
Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?type COUNT(DISTINCT ?person)
WHERE {
  ?person rdf:type dbpedia-owl:SoccerPlayer .
  ?person rdf:type ?type FILTER (regex(?type,"player|football|people|soccer|expatriate|emigrant|citizen","i"))
}
GROUP BY ?type
HAVING COUNT(DISTINCT ?person)<20000
ORDER BY DESC(COUNT(DISTINCT ?person))
```



Example from DBpedia:

- How Do I determine the **importance of an entity** within the Linked Data Graph

(1) InDegree

- simple **Link Popularity**
- can be computed via SPARQL

(2) PageRank

- more sophisticated graph measure for the **importance** of a linked resource
- available at public DBpedia SPARQL endpoint from separate graph:

http://people.aifb.kit.edu/ath/#DBpedia_PageRank

Linked Data Analytics

- **Page Rank at DBpedia**
 - First example: Analyze the the importance of authors

```
PREFIX vrank:<http://purl.org/voc/vrank#>

SELECT ?author ?rank
FROM <http://dbpedia.org>
FROM <http://people.aifb.kit.edu/ath/#DBpedia_PageRank>
WHERE {
    ?author rdf:type dbo:Writer ;
            vrank:hasRank/vrank:rankValue ?rank .
} ORDER by DESC(?rank)
```

Linked Data Analytics

- now save your result locally as **csv** for further analysis
- -> authors.csv

[Link to the SPARQL query](#)

author	rank
http://dbpedia.org/resource/William_Shakespeare	376.536
http://dbpedia.org/resource/Charles_Dickens	138.343
http://dbpedia.org/resource/Johann_Wolfgang_von_Goethe	123.192
http://dbpedia.org/resource/J._R._R._Tolkien	119.514
http://dbpedia.org/resource/Roger_Ebert	111.78
http://dbpedia.org/resource/Virgil	107.56
http://dbpedia.org/resource/Dante_Alighieri	105.222
http://dbpedia.org/resource/Plutarch	104.906
http://dbpedia.org/resource/Voltaire	102.866
http://dbpedia.org/resource/Robert_Christgau	98.7989
http://dbpedia.org/resource/Mark_Twain	95.103
http://dbpedia.org/resource/Tacitus	93.9836
http://dbpedia.org/resource/Edgar_Allan_Poe	84.0183
http://dbpedia.org/resource/John_Milton	83.8589
http://dbpedia.org/resource/George_Orwell	78.8637
http://dbpedia.org/resource/Ovid	77.4266
http://dbpedia.org/resource/Leo_Tolstoy	75.5519
http://dbpedia.org/resource/Walter_Scott	74.7785
http://dbpedia.org/resource/T._S._Eliot	73.4492
http://dbpedia.org/resource/Ernest_Hemingway	73.1249
http://dbpedia.org/resource/H._G._Wells	72.2182
http://dbpedia.org/resource/George_Bernard_Shaw	70.9588
http://dbpedia.org/resource/Livy	67.8619
http://dbpedia.org/resource/Samuel_Johnson	67.7929
http://dbpedia.org/resource/Oscar_Wilde	67.7591
http://dbpedia.org/resource/Rudyard_Kipling	67.51
http://dbpedia.org/resource/Lord_Byron	66.9814
http://dbpedia.org/resource/Stephen_King	64.6965
http://dbpedia.org/resource/C._S._Lewis	64.1277
http://dbpedia.org/resource/Arthur_Conan_Doyle	61.2841
http://dbpedia.org/resource/Isaac_Asimov	60.2163

Linked Data Analytics

- First example: Analyze the importance of authors
- Analysis via R

- read data:

```
authors <- read.csv("authors.csv", header=TRUE)
```

- summarize data:

```
summary(authors)
```

```
> authors <- read.csv("authors.csv", header=TRUE)
> summary(authors)
      rank
Min.   : 0.5819
1st Qu.: 0.7708
Median : 1.1333
Mean   : 2.7408
3rd Qu.: 2.1398
Max.   :376.5360
> █
```



Linked Data Analytics

- First example: Determine the most important authors
- Analysis via R
 - analyze data via boxplot:

```
boxplot(authors)
```

```
> authors <- read.csv("authors.csv", header=TRUE)
```

```
> summary(authors)
```

```
rank
```

```
Min. : 0.5819
```

```
1st Qu.: 0.7708
```

```
Median : 1.1333
```

```
Mean : 2.7408
```

```
3rd Qu.: 2.1398
```

```
Max. : 376.5360
```

```
>
```

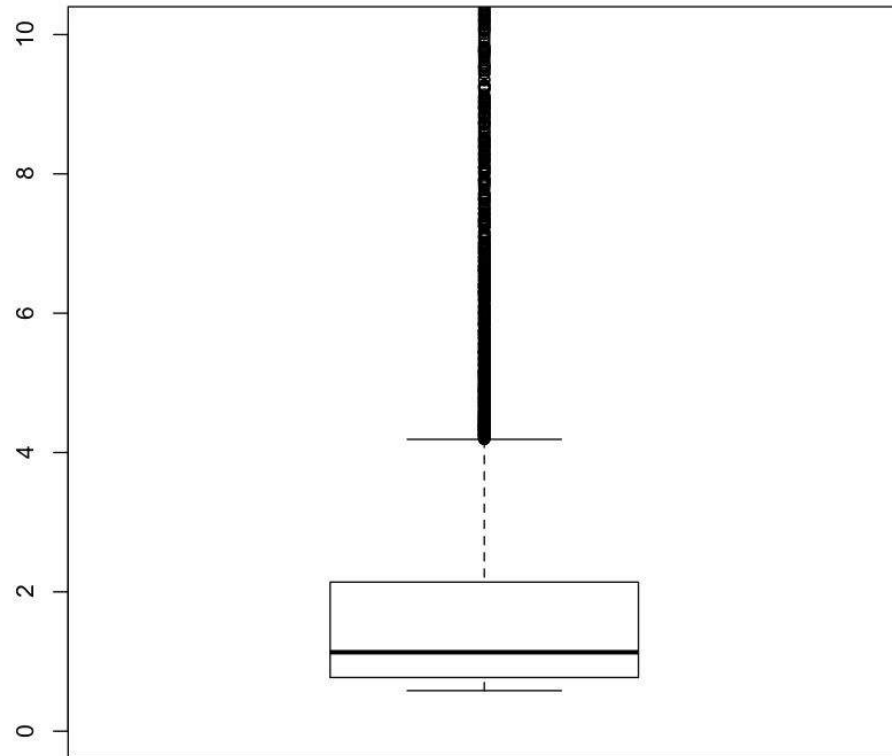
```
> boxplot(authors)
```

```
> boxplot(authors, ylim=c(0,100))
```

```
> boxplot(authors, ylim=c(0,20))
```

```
> boxplot(authors, ylim=c(0,10))
```

```
>
```



Linked Data Analytics

- First example: Determine the most important authors
- Analysis via R
 - analyze data via boxplot:

```
boxplot(authors)
```

```
> authors <- read.csv("authors.csv", header=TRUE)
> summary(authors)
```

rank

Min. : 0.5819

1st Qu.: 0.7708

Median : 1.1333

Mean : 2.7408

3rd Qu.: 2.1398

Max. : 376.5360

```
>
```

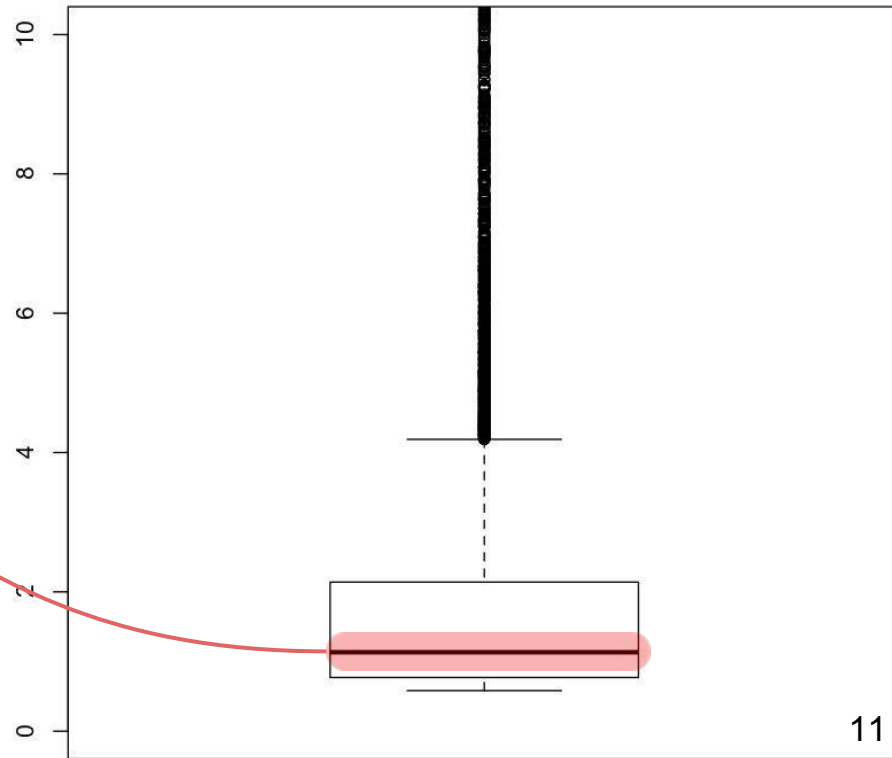
```
> boxplot(authors)
```

```
> boxplot(authors, ylim=c(0,100))
```

```
> boxplot(authors, ylim=c(0,20))
```

```
> boxplot(authors, ylim=c(0,10))
```

```
>
```



Linked Data Analytics

- First example: Determine the most important authors
- Analysis via R
 - analyze data via boxplot:

```
boxplot(authors)
```

```
> authors <- read.csv("authors.csv", header=TRUE)
```

```
> summary(authors)
```

rank

Min. : 0.5819

1st Qu.: 0.7708

Median : 1.1333

Mean : 2.7408

3rd Qu.: 2.1398

Max. : 376.5360

```
>
```

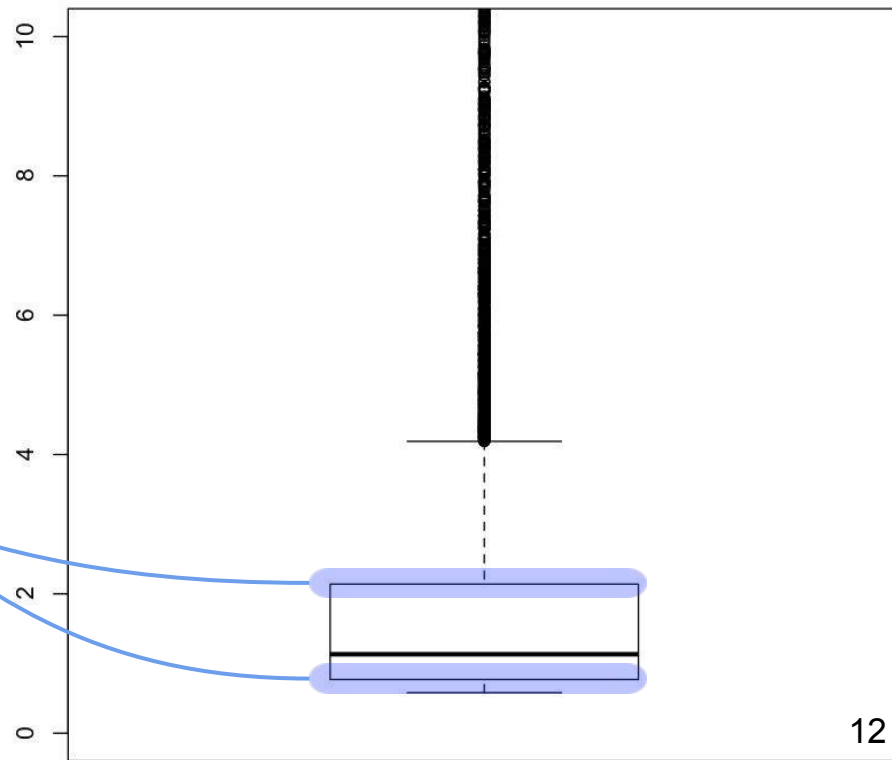
```
> boxplot(authors)
```

```
> boxplot(authors, ylim=c(0,100))
```

```
> boxplot(authors, ylim=c(0,20))
```

```
> boxplot(authors, ylim=c(0,10))
```

```
>
```



Linked Data Analytics

- First example: Determine the most important authors
- Analysis via R
 - analyze data via boxplot:

```
boxplot(authors)
```

```
> authors <- read.csv("authors.csv", header=TRUE)
> summary(authors)
```

```
rank
```

```
Min.   : 0.5819
1st Qu.: 0.7708
Median : 1.1333
Mean   : 2.7408
3rd Qu.: 2.1398
Max.   :376.5360
```

```
>
```

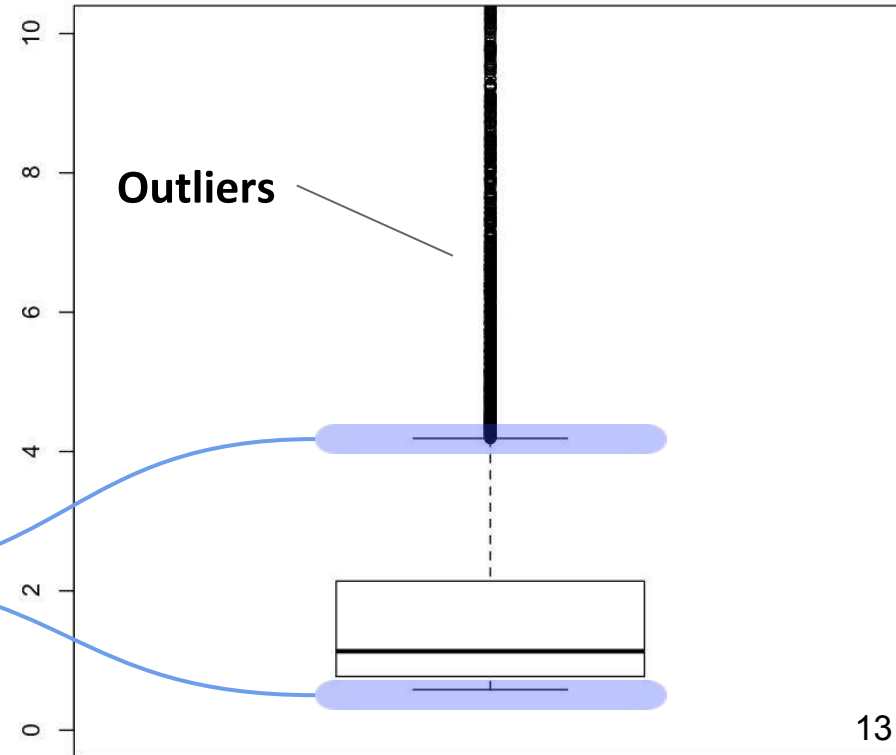
```
> boxplot(authors)
```

Whiskers: $IQR \times 1.5 = (Q_3 - Q_1) \times 1.5$

```
> boxplot(authors, ylim=c(0,20))
```

```
> boxplot(authors, ylim=c(0,10))
```

```
>
```



- 2nd example: include **the number of books for each author**

```
PREFIX vrank:<http://purl.org/voc/vrank#>

SELECT ?author SAMPLE(?rank) as ?arank COUNT(?book) as ?bcount
FROM <http://dbpedia.org>
FROM <http://people.aifb.kit.edu/ath/#DBpedia_PageRank>
WHERE {
  ?author rdf:type dbo:Writer ;
          vrank:hasRank/vrank:rankValue ?rank .
  ?book dbo:author ?author .
} GROUP BY ?author
ORDER BY DESC(?arank)
```

Linked Data Analytics

- now save your result locally as **csv** for further analysis
- -> authors2.csv

[Link to the SPARQL query](#)

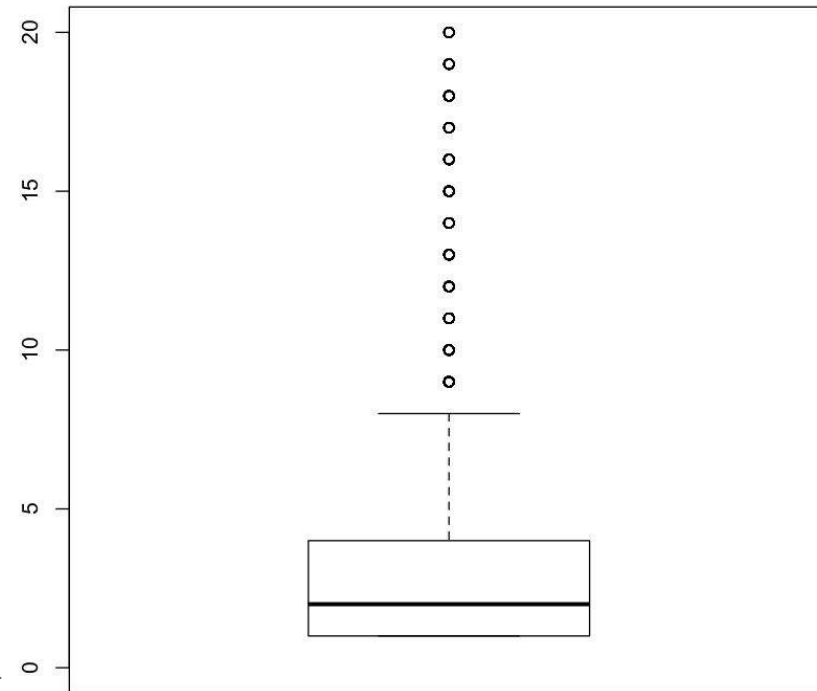
author	arank	bcount
http://dbpedia.org/resource/William Shakespeare	376.536	23
http://dbpedia.org/resource/Charles Dickens	138.343	41
http://dbpedia.org/resource/Johann Wolfgang von Goethe	123.192	7
http://dbpedia.org/resource/J. R. R. Tolkien	119.514	31
http://dbpedia.org/resource/Plutarch	104.906	1
http://dbpedia.org/resource/Voltaire	102.866	7
http://dbpedia.org/resource/Robert Christgau	98.7989	1
http://dbpedia.org/resource/Mark Twain	95.103	40
http://dbpedia.org/resource/Edgar Allan Poe	84.0183	44
http://dbpedia.org/resource/John Milton	83.8589	2
http://dbpedia.org/resource/George Orwell	78.8637	13
http://dbpedia.org/resource/Ovid	77.4266	1
http://dbpedia.org/resource/Leo Tolstoy	75.5519	24
http://dbpedia.org/resource/Walter Scott	74.7785	31
http://dbpedia.org/resource/T. S. Eliot	73.4492	4
http://dbpedia.org/resource/Ernest Hemingway	73.1249	25
http://dbpedia.org/resource/H. G. Wells	72.2182	88
http://dbpedia.org/resource/George Bernard Shaw	70.9588	61
http://dbpedia.org/resource/Samuel Johnson	67.7929	2
http://dbpedia.org/resource/Oscar Wilde	67.7591	12
http://dbpedia.org/resource/Rudyard Kipling	67.51	18
http://dbpedia.org/resource/Lord Byron	66.9814	6
http://dbpedia.org/resource/Stephen King	64.6965	206
http://dbpedia.org/resource/C. S. Lewis	64.1277	27
http://dbpedia.org/resource/Arthur Conan Doyle	61.2841	93
http://dbpedia.org/resource/Isaac Asimov	60.2163	205
http://dbpedia.org/resource/Victor Hugo	58.8837	15
http://dbpedia.org/resource/Lewis Carroll	56.6603	9
http://dbpedia.org/resource/Fyodor Dostoyevsky	55.6083	20
http://dbpedia.org/resource/Samuel Taylor Coleridge	52.9977	1
http://dbpedia.org/resource/Bertolt Brecht	51.6228	16

- 2nd example: include **the number of books for each author**

```
> authors2 <- read.csv("authors2.csv", header=TRUE)
> summary(authors2)
```

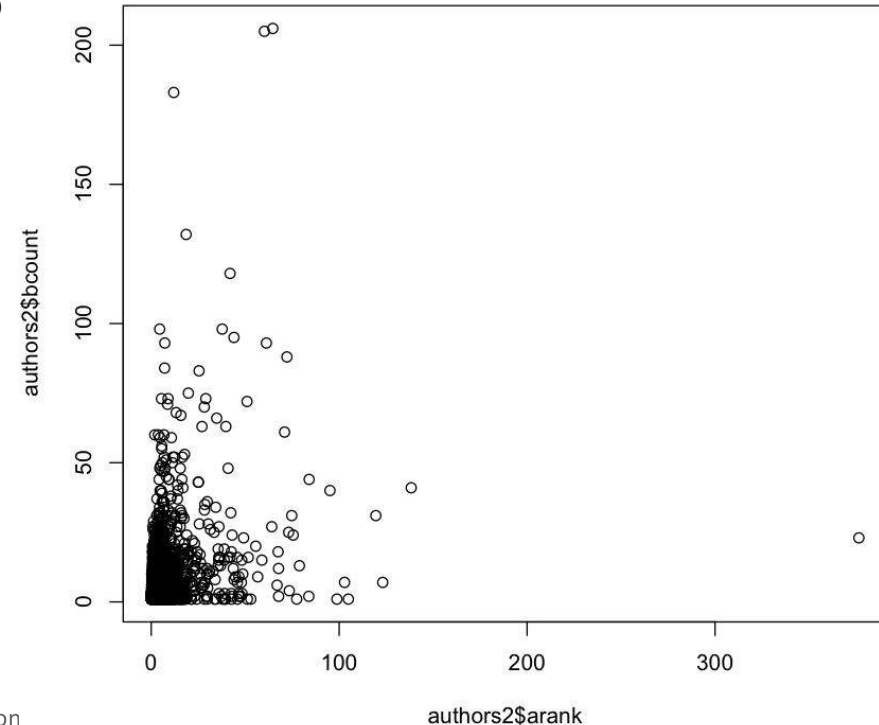
arank	bcount
Min. : 0.1548	Min. : 1.000
1st Qu.: 0.4566	1st Qu.: 1.000
Median : 0.9826	Median : 2.000
Mean : 3.0980	Mean : 4.238
3rd Qu.: 2.4136	3rd Qu.: 4.000
Max. : 376.5360	Max. : 206.000

```
> boxplot(authors2$bcount)
> boxplot(authors2$bcount,ylim=c(0,100))
> boxplot(authors2$bcount,ylim=c(0,20))
>
```

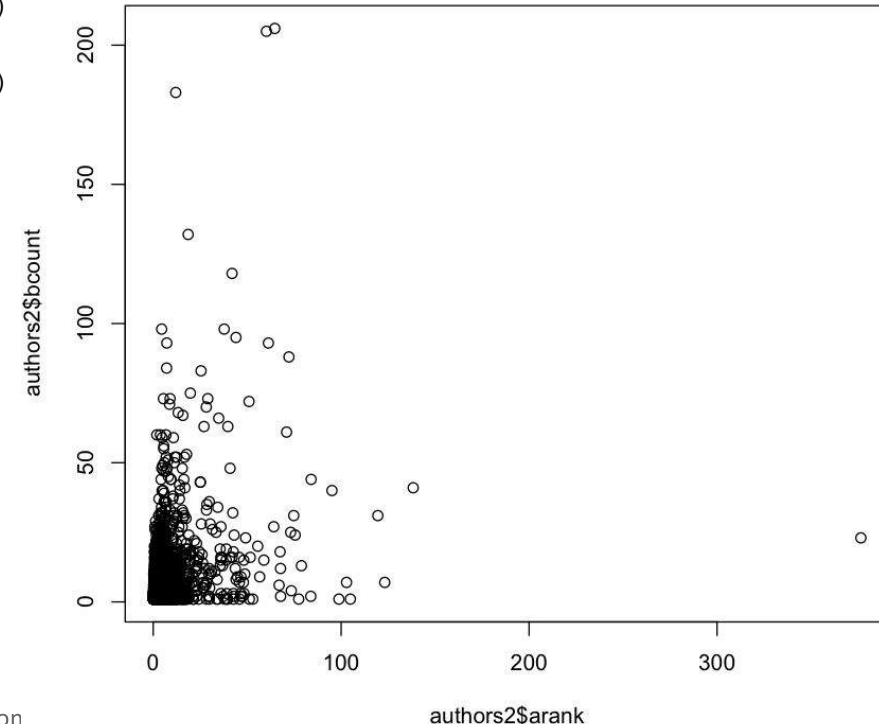


Linked Data Analytics

- 2nd example: include **the number of books for each author**
 - **Hypothesis:** *“important authors have written many books”*
 - `plot(authors2$arank, authors2$bcount)`

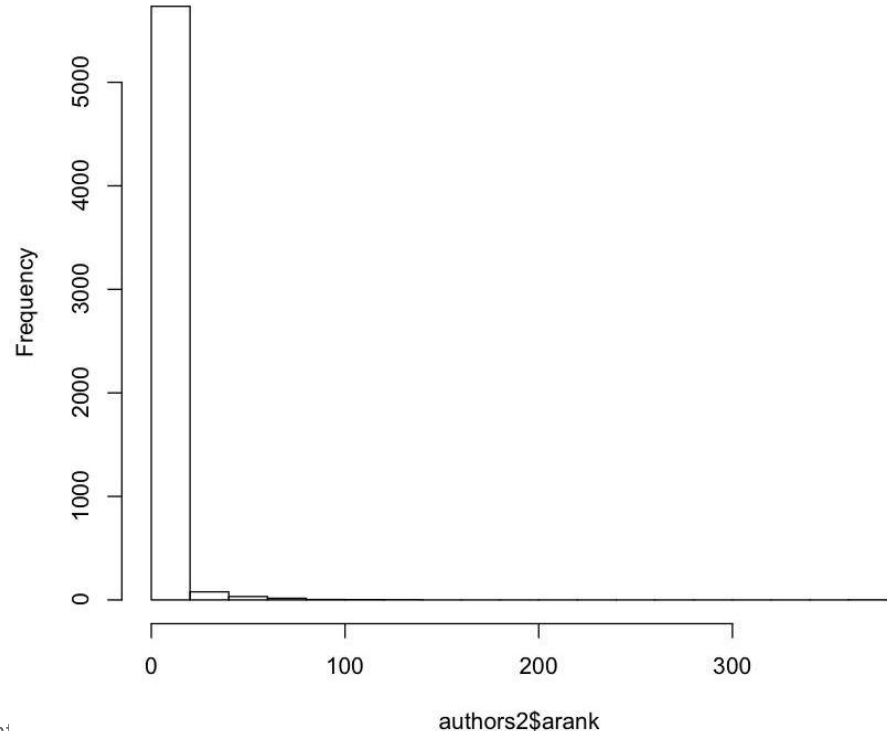


- 2nd example: include **the number of books for each author**
 - **Hypothesis:** *“important authors have written many books”*
 - `plot(authors2$arank, authors2$bcount)`
 - `cor(authors2$arank, authors2$bcount)`
`[1] 0.3560807`

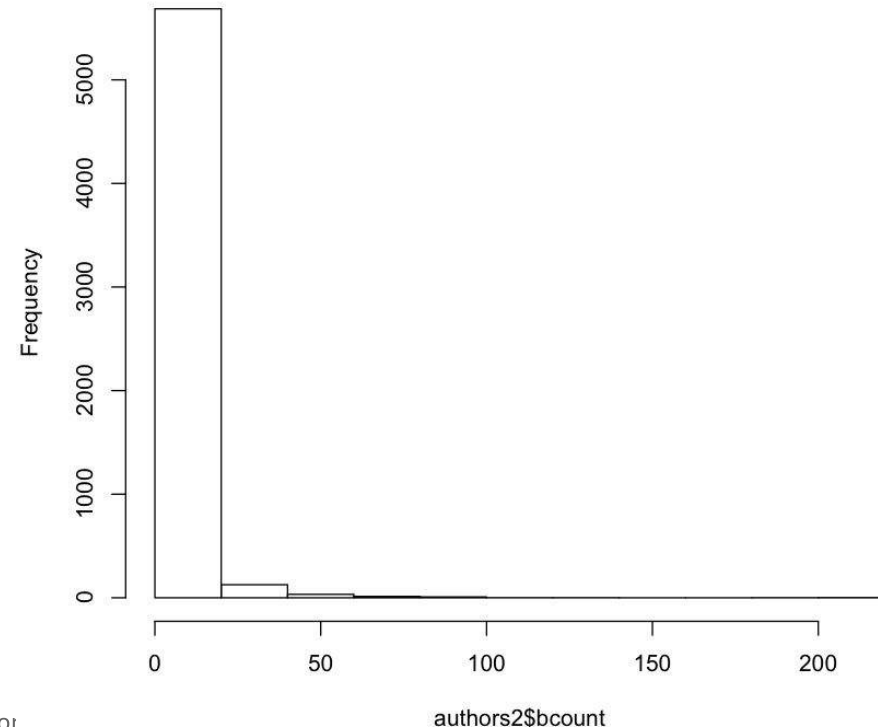


Linked Data Analytics

- 2nd example: include **the number of books for each author**
 - **Question:** *“How are importance and number of books distributed?”*
 - `hist(authors2$arank)`



- 2nd example: include **the number of books for each author**
 - **Question:** *“How are author importance and number of books distributed?”*
 - `hist(authors2$arank)`
 - `hist(authors2$bcount)`



- Linked Data Analytics is an **iterative** and **exploratory** process
- **Visualization** is an important tool for data analysis
- Data analysis enables the **discovery of previously unknown relations**



Lecture 6 - Advanced Linked Data Based Applications - OpenHPI - Course Linked Data Engineering