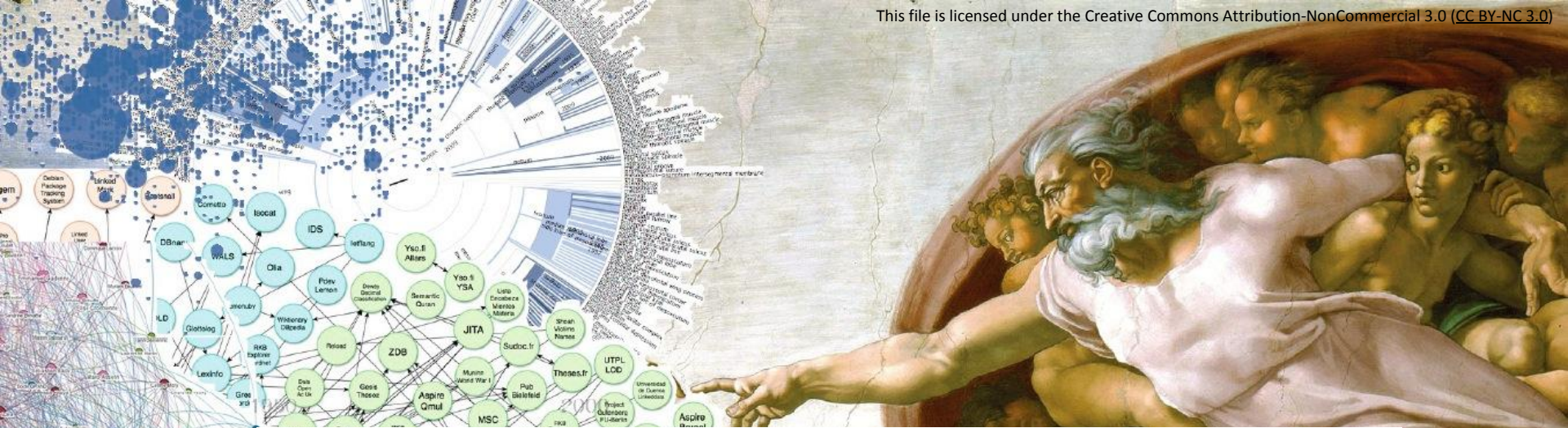




Information Service Engineering

Lecture 6 - Putting it All Together

6.3 Knowledge Mining and Information Extraction II



FIZ Karlsruhe

Leibniz Institute for Information Infrastructure



Hasso
Plattner
Institut



Karlsruher Institut für Technologie

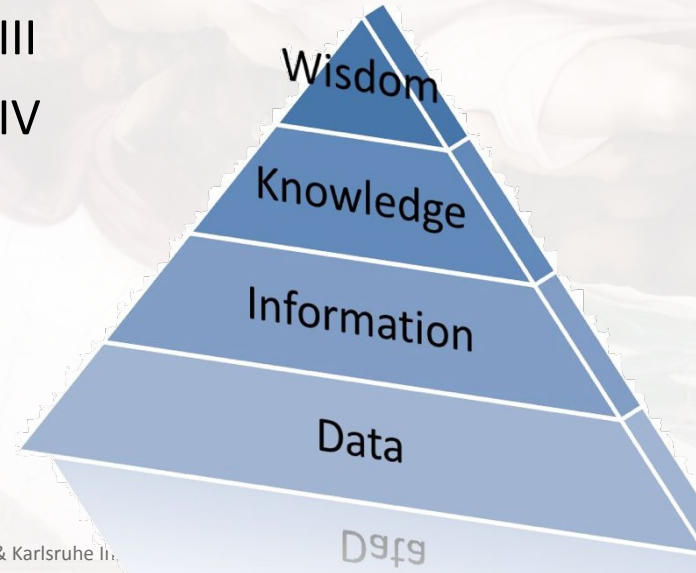
Prof. Dr. Harald Sack & Dr. Maria Koutraki
FIZ Karlsruhe

Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Spring 2018

Information Service Engineering

Lecture 6: Putting it All Together

- 6.1 Linked Data Programming
- 6.2 Knowledge Mining and Information Extraction
- 6.3 Knowledge Mining and Information Extraction II**
- 6.4 Knowledge Mining and Information Extraction III
- 6.5 Knowledge Mining and Information Extraction IV
- 6.6 Semantic Search
- 6.7 Exploratory Search



Knowledge Mining and Knowledge Discovery

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

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 postprocessing.

Knowledge Mining and Knowledge Discovery

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

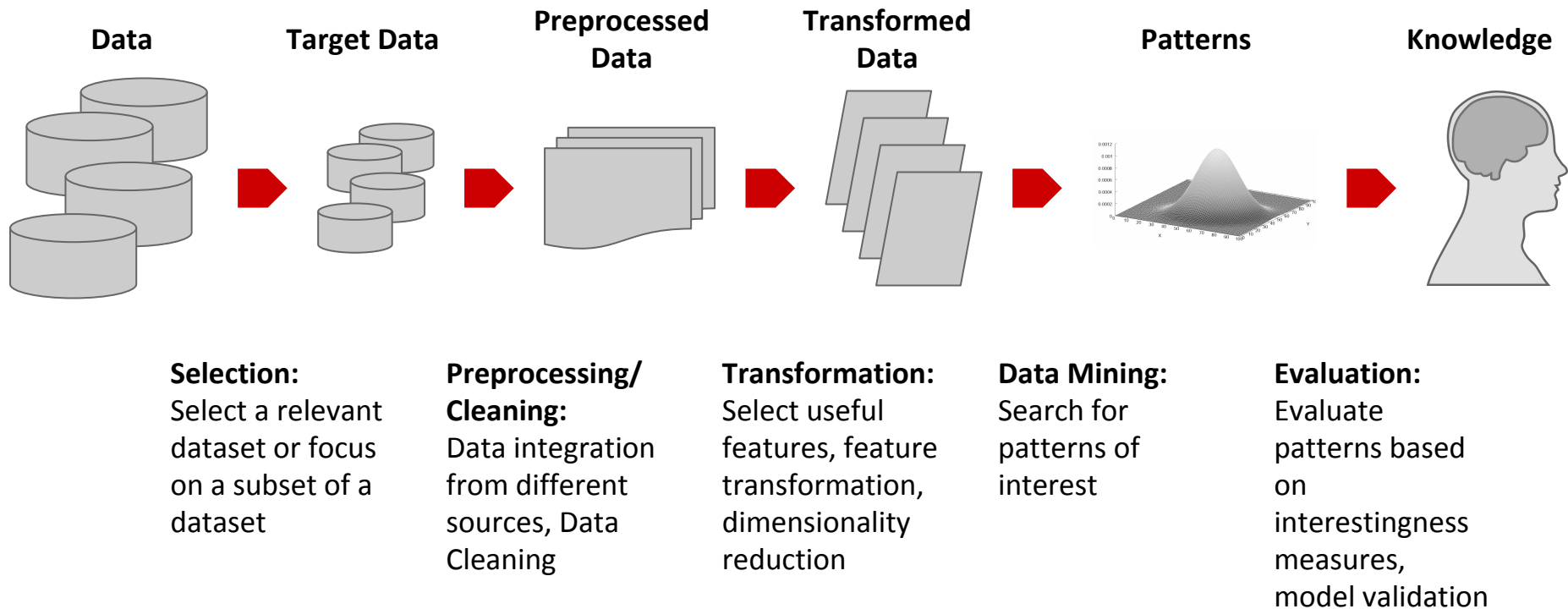
Knowledge Discovery [in Databases] (KDD) is the nontrivial process of identifying **valid, novel, potentially useful, and ultimately understandable patterns** in (massive) data sources.

(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

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II



Data Cleaning

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

- **“Dirty” Data:**
 - Dummy values, absence of data, contradicting data, etc.
- **Steps in Data Cleaning**
 - **Parsing:** locates and identifies individual data elements in raw data
 - **Correcting:** corrects parsed individual data components using sophisticated data algorithms
 - **Normalization:** applies conversion routines to transform data into standard formats
 - **Matching:** searching and matching records within and across data based on predefined rules
 - **Consolidating:** merges data into one representation



Knowledge Mining Functionality

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

- **Characterization:** summarizing general features of objects in a target class (concept description)
- **Discrimination:** comparing general features of objects between a target class and a contrasting class (concept comparison)
- **Association:** studying the frequency of items occurring together
- **Prediction:** predicting some unknown or missing attribute values
- **Classification:** organizing data in given classes based on attribute values (supervised)
- **Clustering:** organizing data in classes based on attribute values (unsupervised)
- **Outlier analysis:** identifying and explaining exceptions (surprises)
- **Time-series analysis:** analyzing trends and deviations

Data Analysis

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

- **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

Hands-On Data Analysis Example Task

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II



Hands-On Data Analysis Example Task

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

1. Examine available Data

Look Up a sample of the data you want to collect

Wikidata

Item | Discussion | Read | View history | Search Wikidata

Ronaldinho (Q39444)

Brazilian association football player

Ronaldo de Assis Moreira | Ronaldinho Gaúcho | Ronaldinho Gaúcho

In more languages Configure:

Language	Label	Description	Also known as
English	Ronaldinho	Brazilian association football player	Ronaldo de Assis Moreira Ronaldinho Gaúcho Ronaldinho Gaúcho
German	Ronaldinho	brasilianischer Fußballspieler	Ronaldo de Assis Moreira Ronaldinho Gaúcho Ronaldinho Gaúcho
French	Ronaldinho	joueur de football brésilien	Ronaldinho Ronaldinho de Assis Moreira Ronaldinho Ronaldo de Assis Moreira Ronaldinho Ronaldinho Gaúcho
Bavarian	No label defined	No description defined	

All entered languages

Statements

instance of human edit

» 1 reference

+ add value

Image

edit

[Wikidata sample page](#)

Hands-On Data Analysis Example Task

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

1. Examine available Data

Look Up a sample of the data you want to collect

statement

member of sports team	FC Barcelona start time 2002 end time 2009 number of matches played 145±0 number of points/goals/set scored 70±0 ▶ 1 reference edit
	Grêmio Foot-Ball Porto Alegrense start time 1998 end time 2001 number of matches played 52±0 number of points/goals/set scored 21±0 ▶ 1 reference edit
	A.C. Milan start time 2010 end time 2011 ▶ 1 reference edit

<https://www.wikidata.org/wiki/Q39444>

Wikidata Recap

object
=
subject/context
for statement

Access via **different namespaces for properties:**

- **wdt:** connects an item to a value
wd:Q39444 wdt:P54 ?team .

member of sports team

FC Barcelona

start time	2002
end time	2009
number of matches played	145±0
number of points/goals/set scored	70±0
▶ 1 reference	

Grêmio Foot-Ball Porto Alegre

start time	1998
end time	2001
number of matches played	52±0
number of points/goals/set scored	21±0
▶ 1 reference	

A.C. Milan

start time	2010
end time	2011
▶ 1 reference	

Clube Atlético Mineiro

start time	2012
------------	------

<https://www.wikidata.org/wiki/Q39444>

Wikidata Recap

Access via **different namespaces for properties:**

- **wdt:** connects an item to a value
wd:Q39444 wdt:P54 ?team .
- **p:** connects a subject to a statement
wd:Q39444 p:P54 ?team_statement .

member of sports team	FC Barcelona
start time	2002
end time	2009
number of matches played	145±0
number of points/goals/set scored	70±0
	▶ 1 reference

	Grêmio Foot-Ball Porto Alegre
start time	1998
end time	2001
number of matches played	52±0
number of points/goals/set scored	21±0
	▶ 1 reference

	A.C. Milan
start time	2010
end time	2011
	▶ 1 reference

	Clube Atlético Mineiro
start time	2012

<https://www.wikidata.org/wiki/Q39444>

statement

Wikidata Recap

Access via **different namespaces for properties:**

- **wdt:** connects an item to a value
wd:Q39444 wdt:P54 ?team .
- **p:** connects a subject to a statement
wd:Q39444 p:P54 ?team_statement .
- **pq:** connects statement to qualifier value
?team_statement pq:1351 ?statement_value

member of sports team

FC Barcelona	
start time	2002
end time	2009
number of matches played	145±0
number of points/goals/set scored	70±0
▶ 1 reference	

Grêmio Foot-Ball Porto Alegrense	
start time	1998
end time	2001
number of matches played	52±0
number of points/goals/set scored	21±0
▶ 1 reference	

A.C. Milan	
start time	2010
end time	2011
▶ 1 reference	

Clube Atlético Mineiro	
start time	2012

property and
object/value of
statement

<https://www.wikidata.org/wiki/Q39444>

Hands-On Data Analysis Example Task

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

2. Get Data

Compose a SPARQL query to analyse Football data available in Wikidata

```
SELECT (SUM(?goals) as ?total_goals) (SAMPLE(?height) as ?height)
(SAMPLE(?weight) as ?weight) (SAMPLE(?posLabel) as ?pos)
  (SAMPLE(xsd:date(?birthdate)) as ?bday)
WHERE {
  ?s wdt:P106 wd:Q937857 ;
    p:P54 ?team_statement ;
    wdt:P2048 ?height ;
    wdt:P2067 ?weight ;
    wdt:P413 ?pos ;
    wdt:P569 ?birthdate .
  ?team_statement pq:P1351 ?goals .
  ?pos rdfs:label ?posLabel FILTER (lang(?posLabel)="en") .
} GROUP BY ?s
```

[SPARQL query](#)

Hands-On Data Analysis Example Task

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

2. Get Data

Compose a SPARQL query to analyse Football data available in Wikidata

The screenshot shows the Wikidata Query interface. The query is as follows:

```
1 SELECT (SUM(?goals) as ?total_goals) (SAMPLE(?height) as ?height) (SAMPLE(?weight) as ?weight) (SAMPLE(?posLabel) as ?pos)
2 (SAMPLE(?bday) as ?bday)
3 WHERE {
4   ?s wdt:P106 wdt:Q937857 ;
5   p:P54 ?team_statement ;
6   wdt:P2048 ?height ;
7   wdt:P2047 ?weight ;
8   wdt:P413 ?pos ;
9   wdt:P569 ?bday .
10  ?team_statement pq:P1351 ?goals .
11  ?pos rdfs:label ?posLabel FILTER (lang(?posLabel)="en") .
12 } GROUP BY ?s
```

The results table shows 13 rows of data:

total_goals	height	weight	pos	bday
25	174	71	midfielder	1988-06-29
8	183	76	midfielder	1982-11-26
62	172	69	midfielder	1986-09-28
29	180	76	midfielder	1989-04-16
1	170	67	midfielder	1988-12-06
15	176	79	midfielder	1993-10-11
25	183	74	midfielder	1985-09-27
12	182	77	midfielder	1986-10-21
9	169	64	midfielder	1992-02-10
12	185	80	midfielder	1994-12-12
70	174	64	midfielder	1982-10-10
13	180	76	midfielder	1987-04-27

[SPARQL query](#)

Hands-On Data Analysis Example Task

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

2. Get Data Save data...

[Data in Google Doc Spreadsheet](#)

6.3 - Soccer Stats Example

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

	A	B	C	D	E	F	G	H
1	total_goals	height	weight	pos	bday			
2	8	183	76	midfielder	1982-11-26			
3	62	172	69	midfielder	1986-09-28			
4	29	180	76	midfielder	1989-04-16			
5	25	174	71	midfielder	1988-06-29			
6	17	170	64	midfielder	1986-12-07			
7	5	183	78	midfielder	1989-01-16			
8	7	180	60	midfielder	1987-12-06			
9	19	177	71	midfielder	1988-01-12			
10	6	170	65	midfielder	1987-06-11			
11	22	170	70	midfielder	1975-02-08			
12	4	178	70	midfielder	1979-04-18			
13	14	169	68	midfielder	1981-09-12			
14	1	170	67	midfielder	1988-12-06			
15	15	176	79	midfielder	1993-10-11			
16	25	183	74	midfielder	1985-09-27			
17	26	176	65	midfielder	1991-02-16			
18	43	179	67	midfielder	1991-09-09			
19	574	180	82	midfielder	1976-09-27			
20	14	179	70	midfielder	1990-06-17			
21	23	169	68	midfielder	1988-03-26			
22	104	180	73	midfielder	1973-09-29			
23	70	174	64	midfielder	1982-10-10			
24	45	184	72	goalkeeper	1983-09-20			
25	0	184	78	goalkeeper	1980-08-05			
26	0	183	79	goalkeeper	1983-03-30			
27	22	162	61	forward	1986-03-23			
28	90	181	69	forward	1980-09-18			
29	0	191	80	goalkeeper	1987-01-10			

Hands-On Data Analysis Example Task

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

3. CleanUp Data

This might require
Several rounds...

[Data in Google Doc Spreadsheet](#)

6.3 - Soccer Stats Example

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

	A	B	C	D	E	F	G	H	I
1	total_goals	height	weight	year	pos	bday			
2	1859	180	74	1943	forward	1943-06-11			
3	1457	178	89	1982	fly-half	1982-03-05			
4	1358	172	75	1940	midfielder	1940-10-23			
5	1318	190	84	1970	midfielder	1970-08-05			
6	1142	177	78	1977	forward	1977-06-08			
7	946	187	83.5	1985	forward	1985-02-05			
8	902	170	77	1977	winger	1977-02-26			
9	849	164	57	1975	forward	1975-02-08			
10	831	184	80	1955	goalkeeper	1955-05-16			
11	691	181	84	1976	goalkeeper	1976-10-21			
12	646	180	68	1947	midfielder	1947-04-25			
13	628	192	80	1954	midfielder	1954-02-19			
14	612	165	68	1960	midfielder	1960-10-30			
15	611	172	72	1907	forward	1907-09-09			
16	594	186	73	1972	midfielder	1972-04-19			
17	592	172	72	1927	forward	1927-04-02			
18	586	170	72	1933	midfielder	1933-10-28			
19	581	171	69	1973	forward	1973-08-11			
20	574	180	82	1976	midfielder	1976-09-27			
21	562	188	85	1982	goalkeeper	1982-02-13			
22	560	182	82	1963	defender	1963-03-21			
23	557	176	68	1945	forward	1945-11-03			
24	556	183	72	1976	centre-back	1976-09-29			
25	552	185	85	1972	goalkeeper	1972-12-14			
26	548	183	78	1969	midfielder	1969-09-29			
27	525	181	83	1985	midfielder	1985-06-04			
28	522	181	76	1977	forward	1977-10-25			
29	497	173	75	1913	forward	1913-09-06			

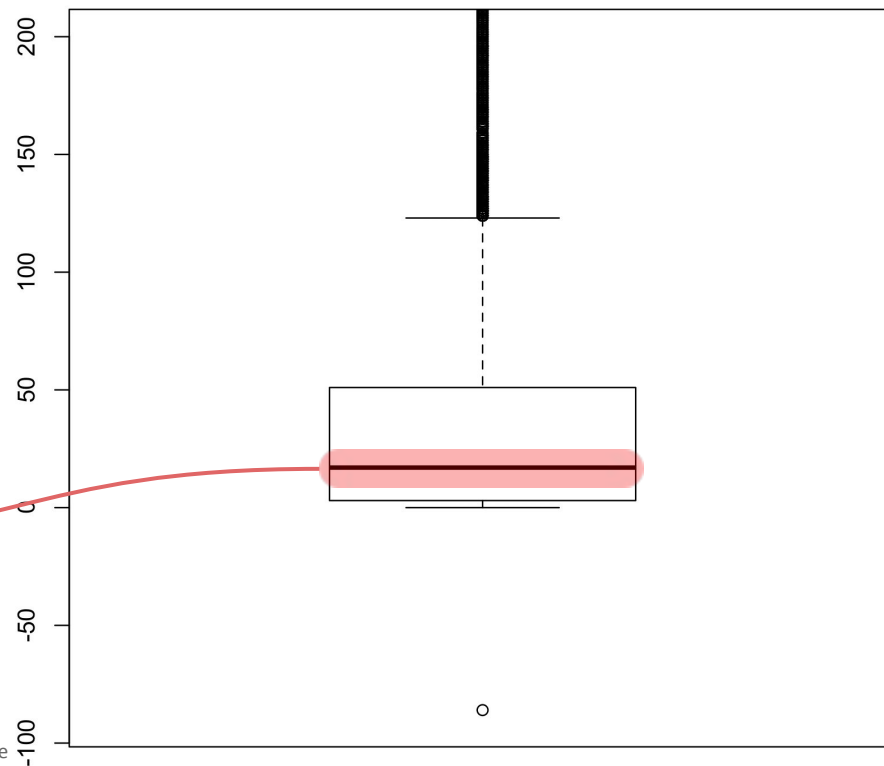
+ Wikidata CleanedUp Wikidata original DBpedia

Excursion: Learn how to read Boxplots

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

X0

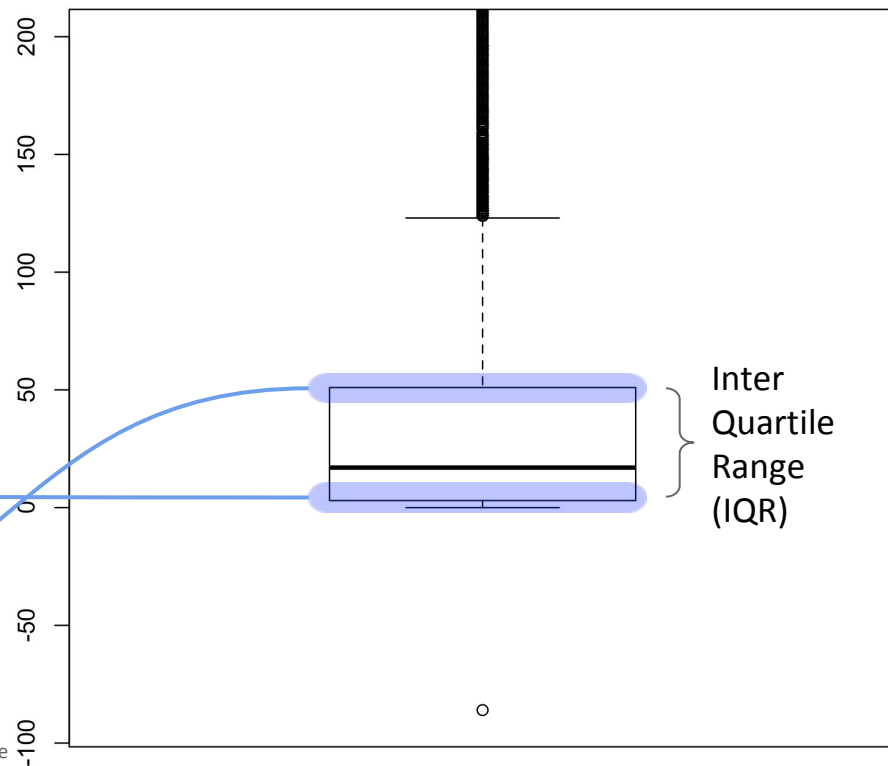
Min.	: -86.00
1st Qu.:	3.00
Median :	17.00
Mean	: 38.85
3rd Qu.:	51.00
Max.	:1329.00



Excursion: Learn how to read Boxplots

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

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



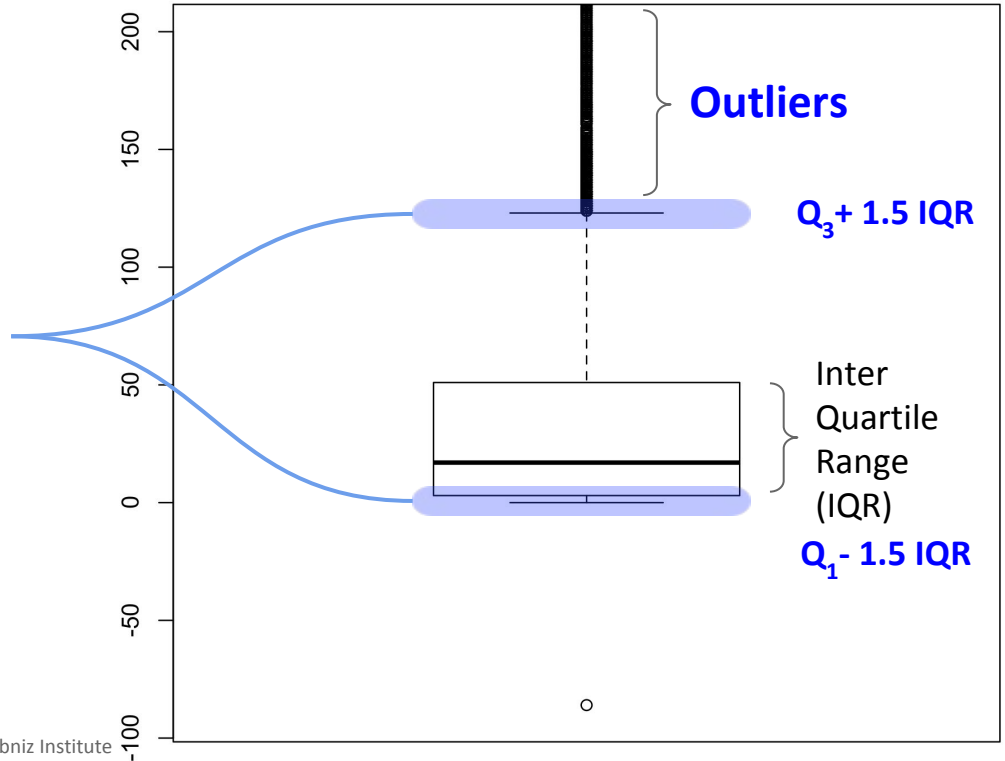
Excursion: Learn how to read Boxplots

Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

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

X0

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

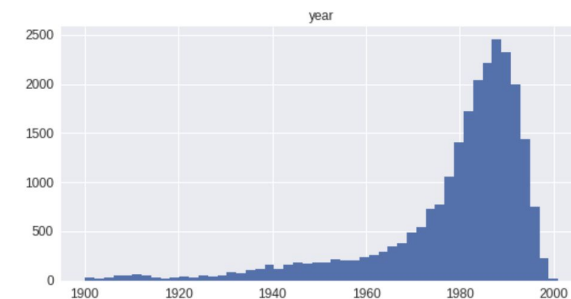
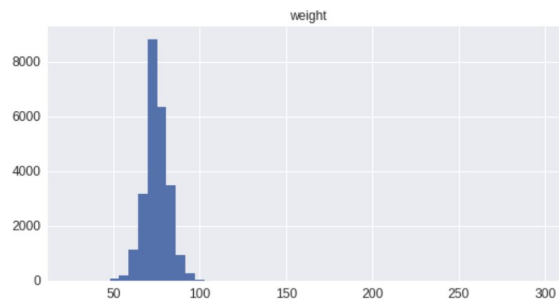
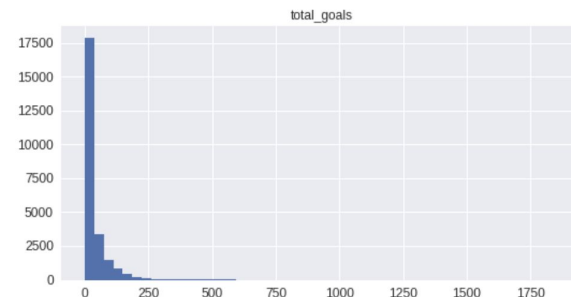
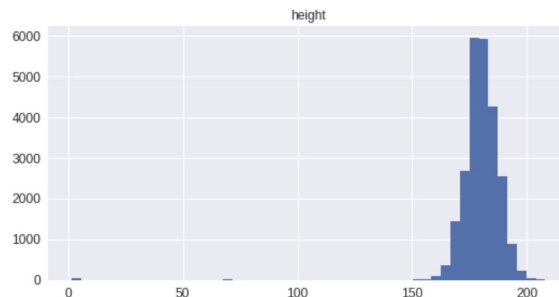


Hands-On Data Analysis Example Task

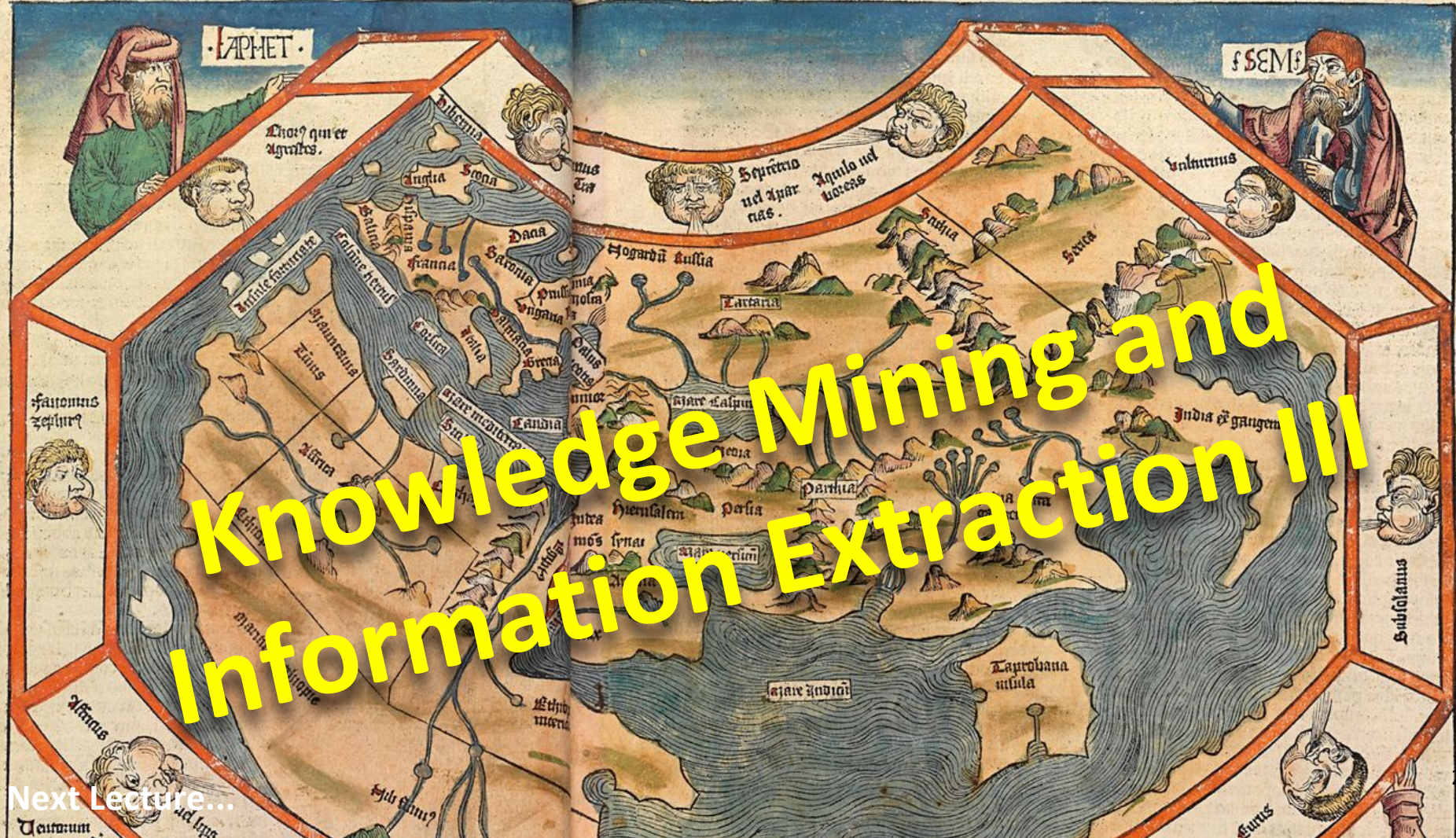
Information Service Engineering - Lecture 6 Putting it All Together - 6.3 Knowledge Mining and Information Extraction II

4. Analyse the Data

E.g. via python



[Data in Google Collab Notebook](#)



Next Lecture...

Picture References:

- P.6: Nick Youngson - link to - <http://nyphotographic.com/> , [CC BY-SA 3.0](#) Alpha Stock Images - link to - <http://alphastockimages.com/http://www.creative-commons-images.com/cleaning/cleaning04.html>
- P.9: A football match between Dresden and Berlin played at Jahre, Germany. 1892, [https://commons.wikimedia.org/wiki/File:Fu%C3%9Fballgeschichte_\(1892\).jpg](https://commons.wikimedia.org/wiki/File:Fu%C3%9Fballgeschichte_(1892).jpg)
- P.23: Hartman Schedel, Nuremberg Chronicle, 1493, https://commons.wikimedia.org/wiki/File:Nuremberg_chronicles_-_f_13b.png