

Data and Text Mining

**Part of
Jožef Stefan IPS Programme – ICT2**

2019 / 2020

Nada Lavrač

Jožef Stefan Institute
Ljubljana, Slovenia

2019/2020 Logistics: Course participants

Contacts: http://kt.ijs.si/petra_kralj/dmtm2.html

- Data Mining:
 - Nada Lavrač: nada.lavrac@ijs.si, Petra Kralj Novak: petra.kralj.novak@ijs.si, Martin Žnidaršič: martin.znidarsic@ijs.si
- Data preparation:
 - Bojan Cestnik: bojan.cestnik@temida.si
- Text mining
 - Dunja Mladenić: dunja.mladenic@ijs.si

Course Schedule – 2019/20

ICT3 and Statistics

Every Monday 15-17h, MPŠ

Mon. 21.10., not 28.10., 4.11.,
11.11, 18.11., 25.11., 2.12., 9.12.,
16.12., 23.12., 6.1., 13.1., 20.1.,
27.1.

with exceptions:

- to be communicated later

ICT2

Every Monday 17-19h, MPŠ

- Mon. 21.10., not 28.10., 4.11.,
11.11, 18.11., 25.11., 2.12., 9.12.,
16.12., 23.12., 6.1., 13.1., 20.1.,
27.1.

with following exceptions:

- Wed. 23.10., 17-19 Petra MPŠ
- Wed. 6.11., 17-19 Petra MPŠ
- Wed. 13.11., 17-19 Bojan MPŠ
- Wed. 20.11., 17-19 Bojan MPŠ
- Wed. 27.11., 17-19 Bojan MPŠ
- ... the rest to be communicated

Data and Text Mining: MSc Credits and Coursework for Data mining part

- Attending lectures
- Attending practical exercises
 - Theory exercises and hands-on (intro to Orange DM toolbox by dr. Petra Kralj Novak)
- **Written exam (40%)**
- **Seminar (60%):**
 - Data analysis of your own data (e.g., using Orange for questionnaire data analysis)
 - own initiatives are welcome ...

Data Mining: MSc Credits and coursework

Exam: Written exam (60 minutes) - Theory

Seminar: topic selection + results presentation

- One hour available for seminar topic discussion – one page written proposal defining the task and the selected dataset
- Deliver written report + electronic copy (4 pages in Information Society paper format, instructions on the web)
 - Report on data analysis of own data needs to follow the CRISP-DM methodology
 - Presentation of your seminar results (15 minutes each: 10 minutes presentation + 5 minutes discussion)

Data Mining: ICT2 Credits and Coursework

- 20 credits
 - 8 Nada Lavrač and Petra Kralj Novak
 - 4 Bojan Cestnik
 - 8 Dunja Mladenić

Course Outline

I. Introduction

- Data Mining and KDD process
- Introduction to Data Mining
- Data Mining platforms

II. Predictive DM Techniques

- Decision Tree learning
- Bayesian classifier
- Classification rule learning
- Classifier Evaluation

III. Regression

IV. Descriptive DM


- Predictive vs. descriptive induction
- Subgroup discovery
- Association rule learning
- Hierarchical clustering

V. Relational Data Mining

- RDM and Inductive Logic Programming
- Propositionalization
- Semantic data mining

VI. Advanced Topics

Part I. Introduction

- 
- Data Mining and the KDD process
 - Introduction to Data Mining
 - Data Mining platforms

Machine Learning and Data Mining

- **Machine Learning (ML)** – computer algorithms/machines that learn predictive models from class-labeled data
- **Data Mining (DM)** – extraction of useful information from data: discovering relationships and patterns that have not previously been known, and use of **ML** techniques applied to solving real-life data analysis problems
- **Knowledge discovery in databases (KDD)** – the process of knowledge discovery

Machine Learning and Data Mining

- **Machine Learning (ML)** – computer algorithms/machines that learn predictive models from class-labeled data
- **Data Mining (DM)** – extraction of useful information from data: discovering relationships and patterns that have not previously been known, and use of **ML** techniques applied to solving real-life data analysis problems
- **Knowledge Discovery in Databases (KDD)** – the process of knowledge discovery

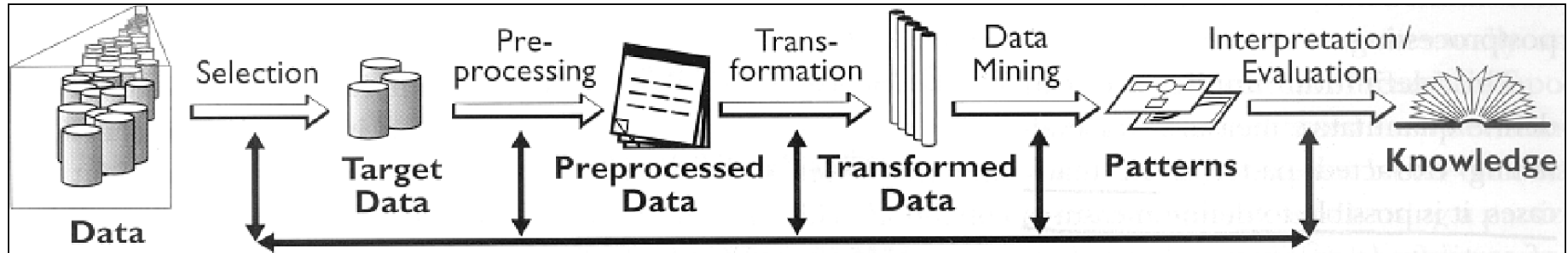
Data Mining and KDD

- Buzzword since 1996
- KDD is defined as “the process of identifying valid, novel, potentially useful and ultimately understandable models/patterns in data.” *
- Data Mining (DM) is the key step in the KDD process, performed by using data mining techniques for extracting models or interesting patterns from the data.

Usama M. Fayyad, Gregory Piatetsky-Shapiro, Pedhraic Smyth: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Comm ACM, Nov 96/Vol 39 No 11

KDD Process: CRISP-DM

KDD process of discovering useful knowledge from data

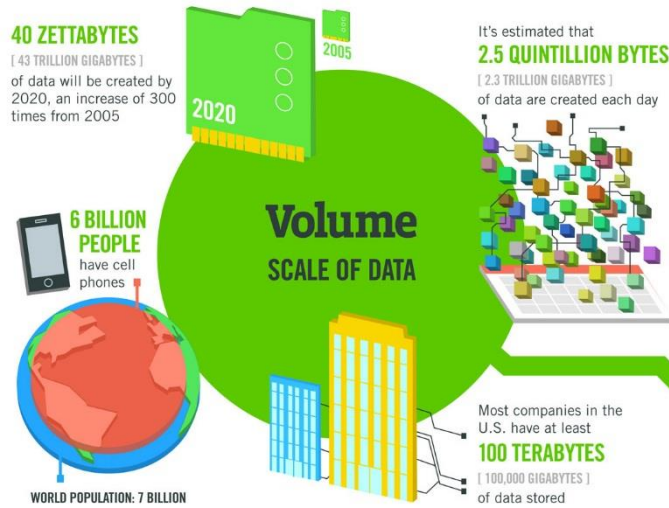


- KDD process involves several phases:
 - data preparation
 - data mining (machine learning, statistics)
 - evaluation and use of discovered patterns
- Data mining is the key step, but represents only 15%-25% of the entire KDD process

Big Data

- **Big Data** – Buzzword since 2008 (special issue of Nature on Big Data)
 - data and techniques for dealing with very large volumes of data, possibly dynamic data streams
 - requiring large data storage resources, special algorithms for parallel computing architectures.

The 4 Vs of Big Data



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT
are shared on Facebook every month



By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Variety
DIFFERENT FORMS OF DATA



The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session



Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

Velocity
ANALYSIS OF STREAMING DATA



By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** — almost 2.5 connections per person on earth



1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions



Poor data quality costs the US economy around **\$3.1 TRILLION A YEAR**



27% OF RESPONDENTS

Veracity
UNCERTAINTY OF DATA

in one survey were unsure of how much of their data was inaccurate

Data Science

- **Data Science** – buzzword since 2012 when Harvard Business Review called it "The Sexiest Job of the 21st Century"
 - an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to **data mining**.
 - used interchangeably with earlier concepts like business analytics, business intelligence, predictive modeling, and statistics.

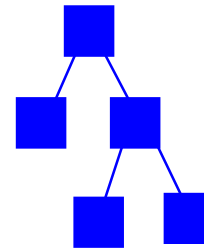
Data Mining in a Nutshell

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23
O24	56	hypermetrope	yes	normal	NONE

data

knowledge discovery
from data

Data Mining

model, patterns, ...

Given: transaction data table, relational database, text documents, Web pages

Find: a classification model, a set of interesting patterns

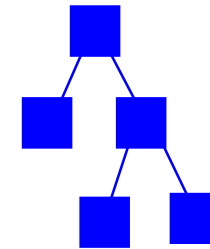
Data Mining in a Nutshell

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23
O24	56	hypermetrope	yes	normal	NONE

data

knowledge discovery
from data

Data Mining

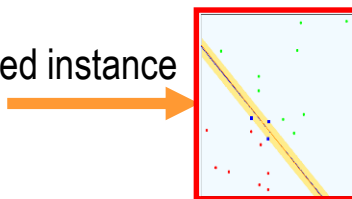


model, patterns, ...

Given: transaction data table, relational database, text documents, Web pages

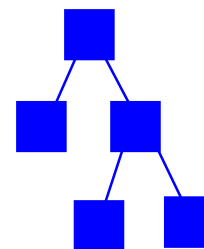
Find: a classification model, a set of interesting patterns

new unclassified instance



classified instance

black box classifier
no explanation



symbolic model
symbolic patterns

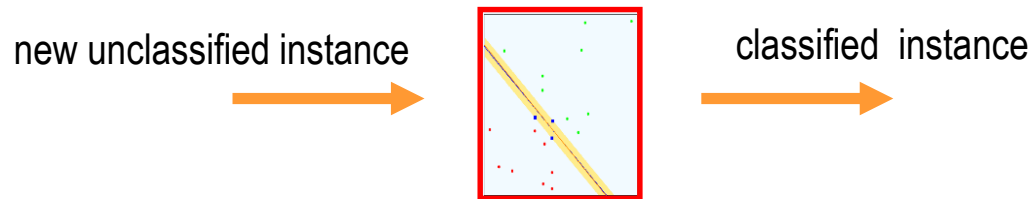
explanation



Why use black-box models

Given: the learned classification model
(e.g, a linear classifier, a deep neural network, ...)

Find: - the class label for a new unlabeled instance



Advantages:

- best classification results in image recognition and other complex classification tasks

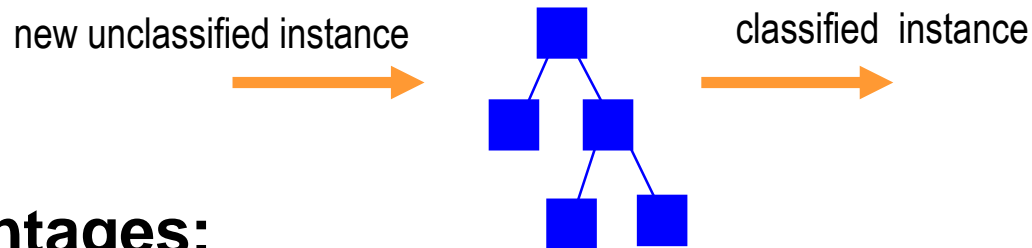
Drawbacks:

- poor interpretability of results
- can not be used for pattern analysis

Why learn and use symbolic models

Given: the learned classification model
(a decision tree or a set of rules)

Find: - the class label for a new unlabeled instance



Advantages:

- use the model for the explanation of classifications of new data instances
- use the discovered patterns for data exploration

Drawbacks:

- lower accuracy than deep NNs

Simplified example: Learning a classification model from contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23
O24	56	hypermetrope	yes	normal	NONE

Pattern discovery in Contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23
O24	56	hypermetrope	yes	normal	NONE

PATTERN

Rule:

IF

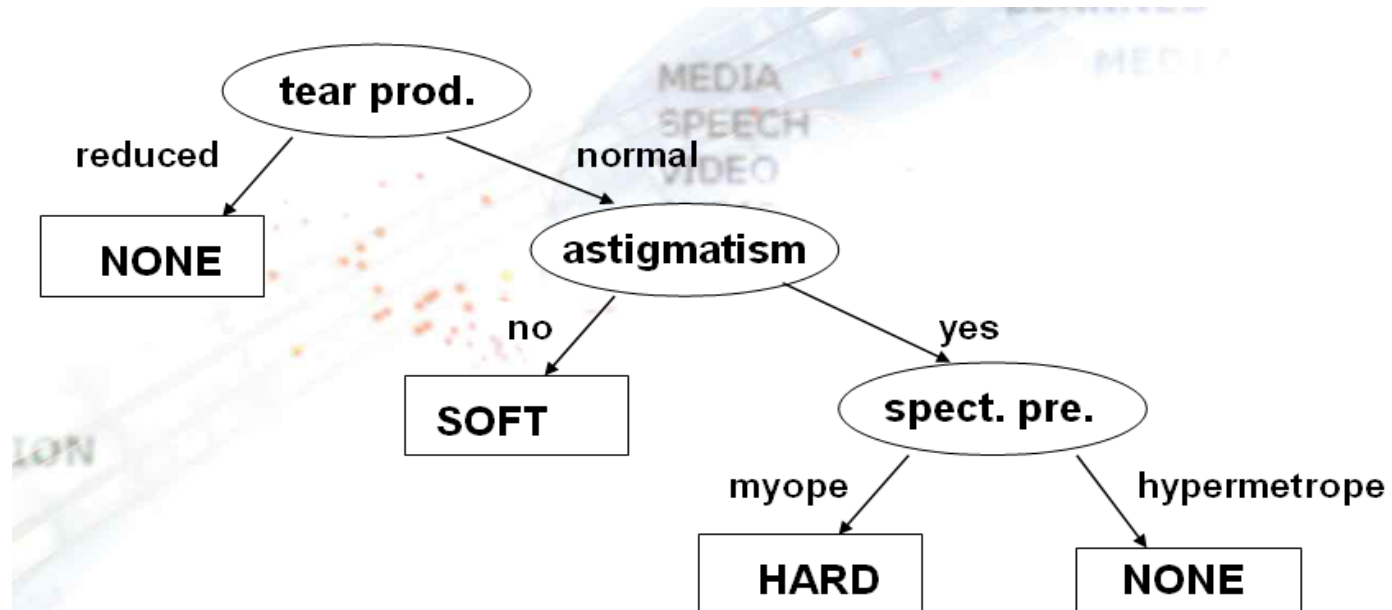
Tear prod. =
reduced

THEN

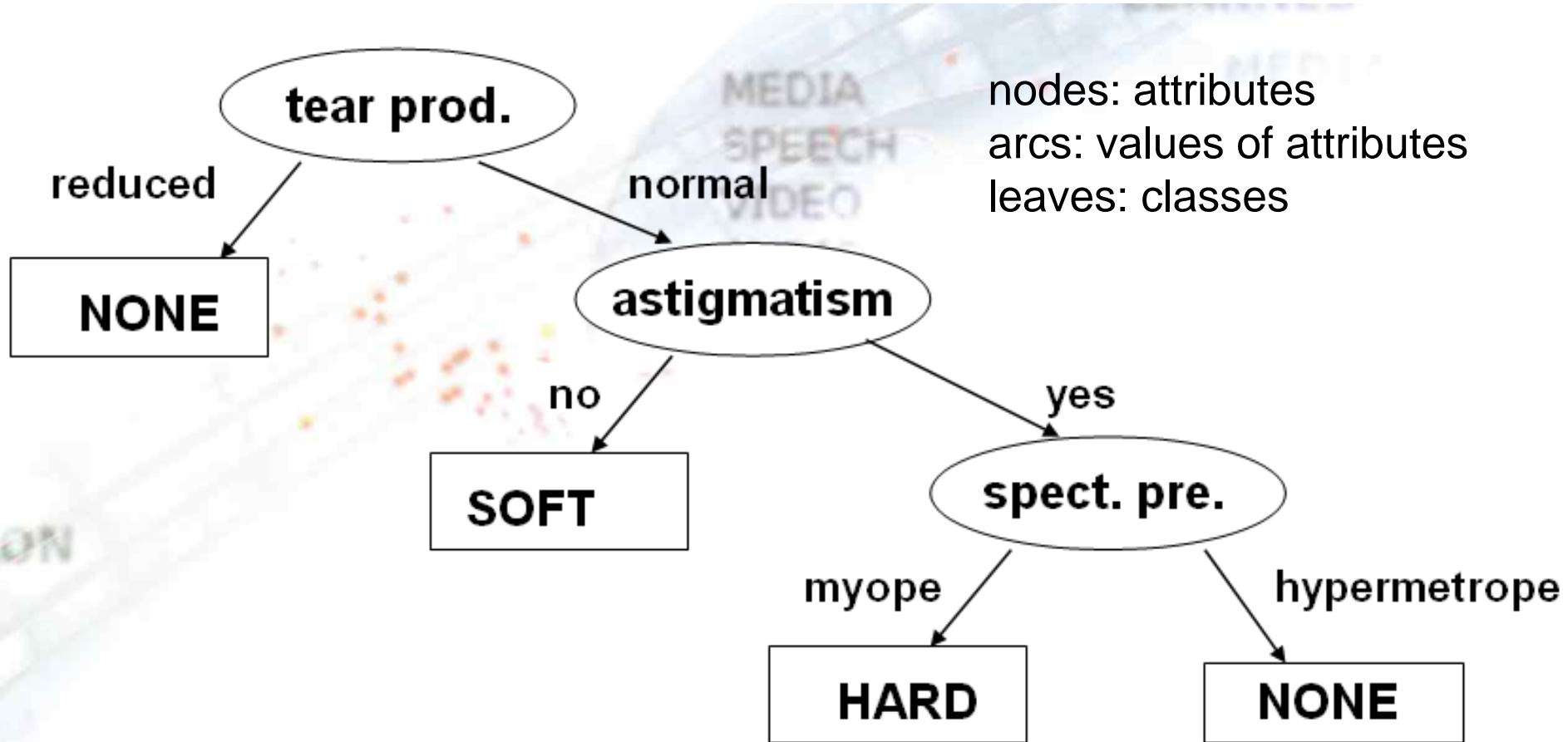
Lenses =
NONE

Learning a classification model from contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	young	myope	no	reduced	NONE
O2	young	myope	no	normal	SOFT
O3	young	myope	yes	reduced	NONE
O4	young	myope	yes	normal	HARD
O5	young	hypermetrope	no	reduced	NONE
O6-O13
O14	pre-presbyc	hypermetrope	no	normal	SOFT
O15	pre-presbyc	hypermetrope	yes	reduced	NONE
O16	pre-presbyc	hypermetrope	yes	normal	NONE
O17	presbyopic	myope	no	reduced	NONE
O18	presbyopic	myope	no	normal	NONE
O19-O23
O24	presbyopic	hypermetrope	yes	normal	NONE

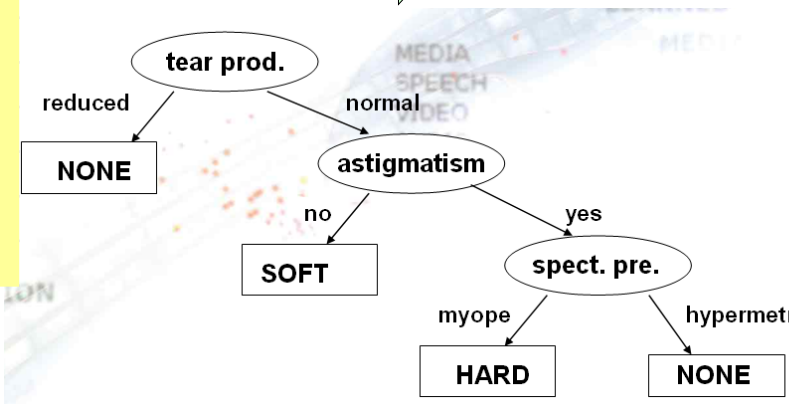


Decision tree classification model learned from contact lens data



Learning a classification model from contact lens data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23
O24	56	hypermetrope	yes	normal	NONE



lenses=NONE ← tear production=red

lenses=NONE ← tear production=normal AND astigmatism=yes AND spect. pre.=hypermetrope

lenses=SOFT ← tear production=normal AND astigmatism=no

lenses=HARD ← tear production=normal AND astigmatism=yes AND spect. pre.=myope

lenses=NONE ←

Classification rules model learned from contact lens data

lenses=NONE ← tear production=reduced

lenses=NONE ← tear production=normal AND
astigmatism=yes AND
spect. pre.=hypermetrope

lenses=SOFT ← tear production=normal AND
astigmatism=no

lenses=HARD ← tear production=normal AND
astigmatism=yes AND
spect. pre.=myope

lenses=NONE ←

Learning from Unlabeled Data

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NONE
O2	23	myope	no	normal	SOFT
O3	22	myope	yes	reduced	NONE
O4	27	myope	yes	normal	HARD
O5	19	hypermetrope	no	reduced	NONE
O6-O13
O14	35	hypermetrope	no	normal	SOFT
O15	43	hypermetrope	yes	reduced	NONE
O16	39	hypermetrope	yes	normal	NONE
O17	54	myope	no	reduced	NONE
O18	62	myope	no	normal	NONE
O19-O23
O24	56	hypermetrope	yes	normal	NONE

Unlabeled data - clustering: grouping of similar instances
 - association rule learning

Learning from Numeric Class Data

Person	Age	Spect. presc.	Astigm.	Tear prod.	LensPrice
O1	17	myope	no	reduced	0
O2	23	myope	no	normal	8
O3	22	myope	yes	reduced	0
O4	27	myope	yes	normal	5
O5	19	hypermetrope	no	reduced	0
O6-O13
O14	35	hypermetrope	no	normal	5
O15	43	hypermetrope	yes	reduced	0
O16	39	hypermetrope	yes	normal	0
O17	54	myope	no	reduced	0
O18	62	myope	no	normal	0
O19-O23
O24	56	hypermetrope	yes	normal	0

Numeric class values – regression analysis

Task reformulation: Binary Class Values

Person	Age	Spect. presc.	Astigm.	Tear prod.	Lenses
O1	17	myope	no	reduced	NO
O2	23	myope	no	normal	YES
O3	22	myope	yes	reduced	NO
O4	27	myope	yes	normal	YES
O5	19	hypermetrope	no	reduced	NO
O6-O13
O14	35	hypermetrope	no	normal	YES
O15	43	hypermetrope	yes	reduced	NO
O16	39	hypermetrope	yes	normal	NO
O17	54	myope	no	reduced	NO
O18	62	myope	no	normal	NO
O19-O23
O24	56	hypermetrope	yes	normal	NO

Binary classes (positive vs. negative examples of **Target class**)

- for Concept learning – classification and class description
- for Subgroup discovery – exploring patterns characterizing groups of instances of target class

Task reformulation: Binary Class and Feature Values

Person	Young	Myope	Astigm.	Reduced tea	Lenses
O1	1	1	0	1	NO
O2	1	1	0	0	YES
O3	1	1	1	1	NO
O4	1	1	1	0	YES
O5	1	0	0	1	NO
O6-O13
O14	0	0	0	0	YES
O15	0	0	1	1	NO
O16	0	0	1	0	NO
O17	0	1	0	1	NO
O18	0	1	0	0	NO
O19-O23
O24	0	0	1	0	NO

Binary features and class values

Data Mining, ML and Statistics

- All three areas have a long tradition of developing inductive techniques for data analysis.
 - reasoning from properties of a data sample to properties of a population
- **DM vs. ML - Viewpoint in this course:**
 - Data Mining is the application of Machine Learning techniques to hard real-life data analysis problems

Data Mining, ML and Statistics

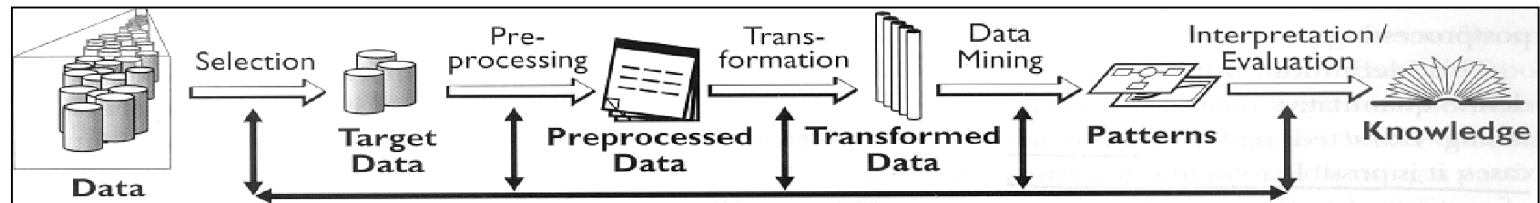
- All three areas have a long tradition of developing inductive techniques for data analysis.
 - reasoning from properties of a data sample to properties of a population
- **DM vs. Statistics:**
 - **Statistics**
 - Hypothesis testing when certain theoretical expectations about the data distribution, independence, random sampling, sample size, etc. are satisfied
 - Main approach: best fitting all the available data
 - **Data mining**
 - Automated construction of understandable patterns, and structured models
 - Main approach: structuring the data space, heuristic search for decision trees, rules, ... covering (parts of) the data space

First Generation Data Mining

- **First machine learning algorithms for**
 - Decision tree and rule learning in 1970s and early 1980s by Quinlan, Michalski et al., Breiman et al., ...
- **Characterized by**
 - Learning from data stored in a single data table
 - Relatively small set of instances and attributes
- **Lots of ML research followed in 1980s**
 - Numerous conferences ICML, ECML, ... and ML sessions at AI conferences IJCAI, ECAI, AAAI, ...
 - Extended set of learning tasks and algorithms addressed

Second Generation Data Mining

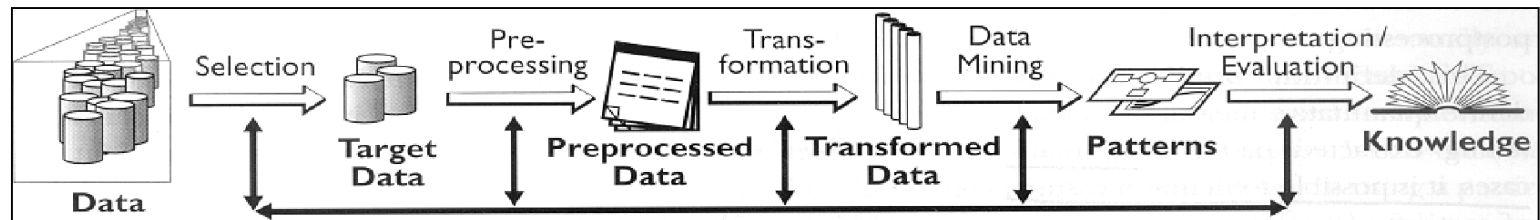
- **Developed since 1990s:**
 - Focused on data mining tasks characterized by large datasets described by large numbers of attributes
 - Industrial standard: CRISP-DM methodology (1997)



Second Generation Data Mining

- **Developed since 1990s:**

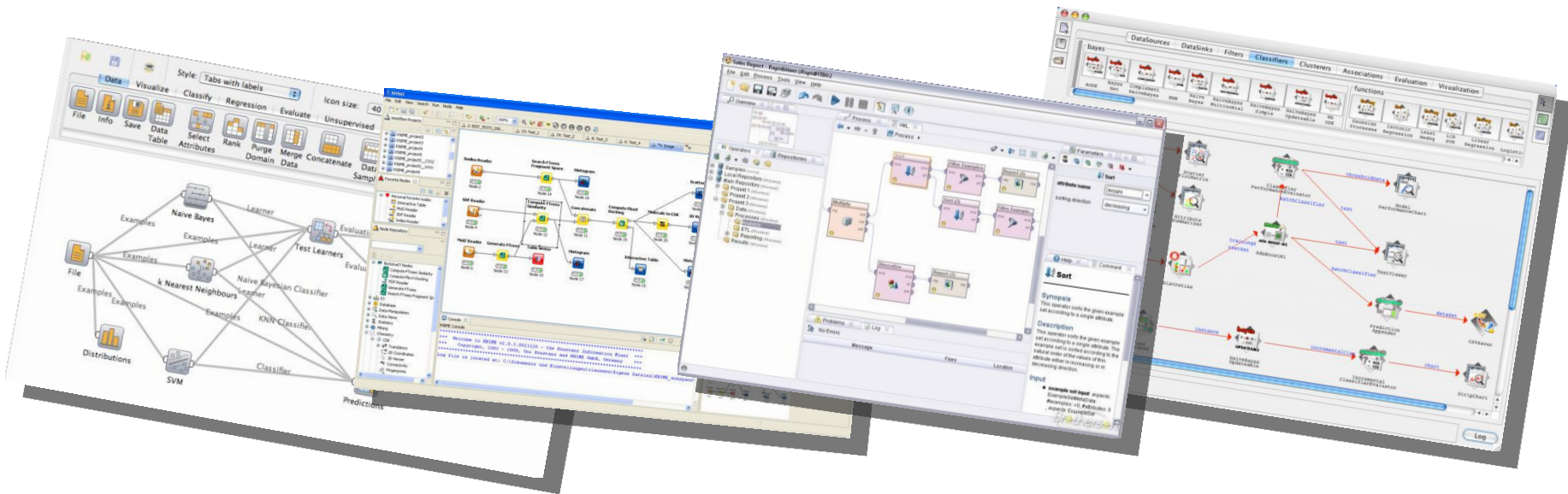
- Focused on data mining tasks characterized by large datasets described by large numbers of attributes
- Industrial standard: CRISP-DM methodology (1997)



- New conferences on practical aspects of data mining and knowledge discovery: KDD, PKDD, ...
- New learning tasks and efficient learning algorithms:
 - Learning predictive models: Bayesian network learning,, **relational data mining**, statistical relational learning, SVMs, ...
 - Learning descriptive patterns: association rule learning, **subgroup discovery**, ...

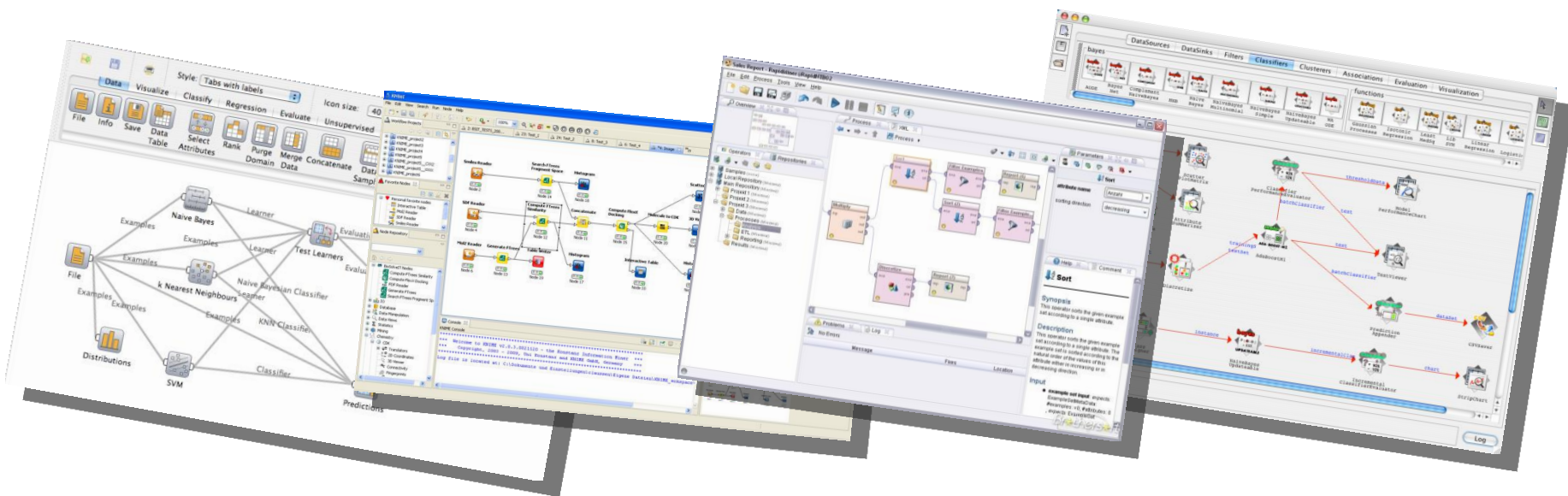
Second Generation Data Mining Platforms

Orange, WEKA, KNIME, RapidMiner, ...



Second Generation Data Mining Platforms

Orange, WEKA, KNIME, RapidMiner, ...

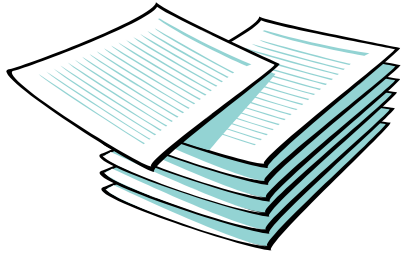


- include numerous data mining algorithms
- enable data and model visualization
- like Orange, Taverna, WEKA, KNIME, RapidMiner, also enable complex **workflow** construction

Third Generation Data Mining

- **Developed since 2010s:**
 - Focused on big data analytics
 - Addressing complex data mining tasks and scenarios
 - New conferences on data science and big data analytics; e.g., IEEE Big Data, Complex networks, ...
 - New learning tasks and efficient learning algorithms:
 - Analysis of dynamic data streams
 - Network analysis,
 - Text mining,
 - Semantic data analysis,
 - Analysis of heterogeneous information networks
 - Analysis of knowledge graphs ...

Bag-of-Words Data Transformation for Text mining



Step 1

BoW vector construction

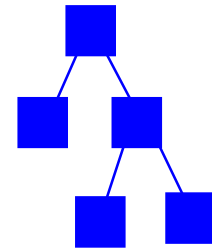
1. BoW features construction
2. Table of BoW vectors construction

Document	Word1	Word2	...	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23
d24	0	0	1	0	NO

Document	Word1	Word2	...	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23
d24	0	0	1	0	NO

Step 2

Data Mining



model, patterns, clusters,

...

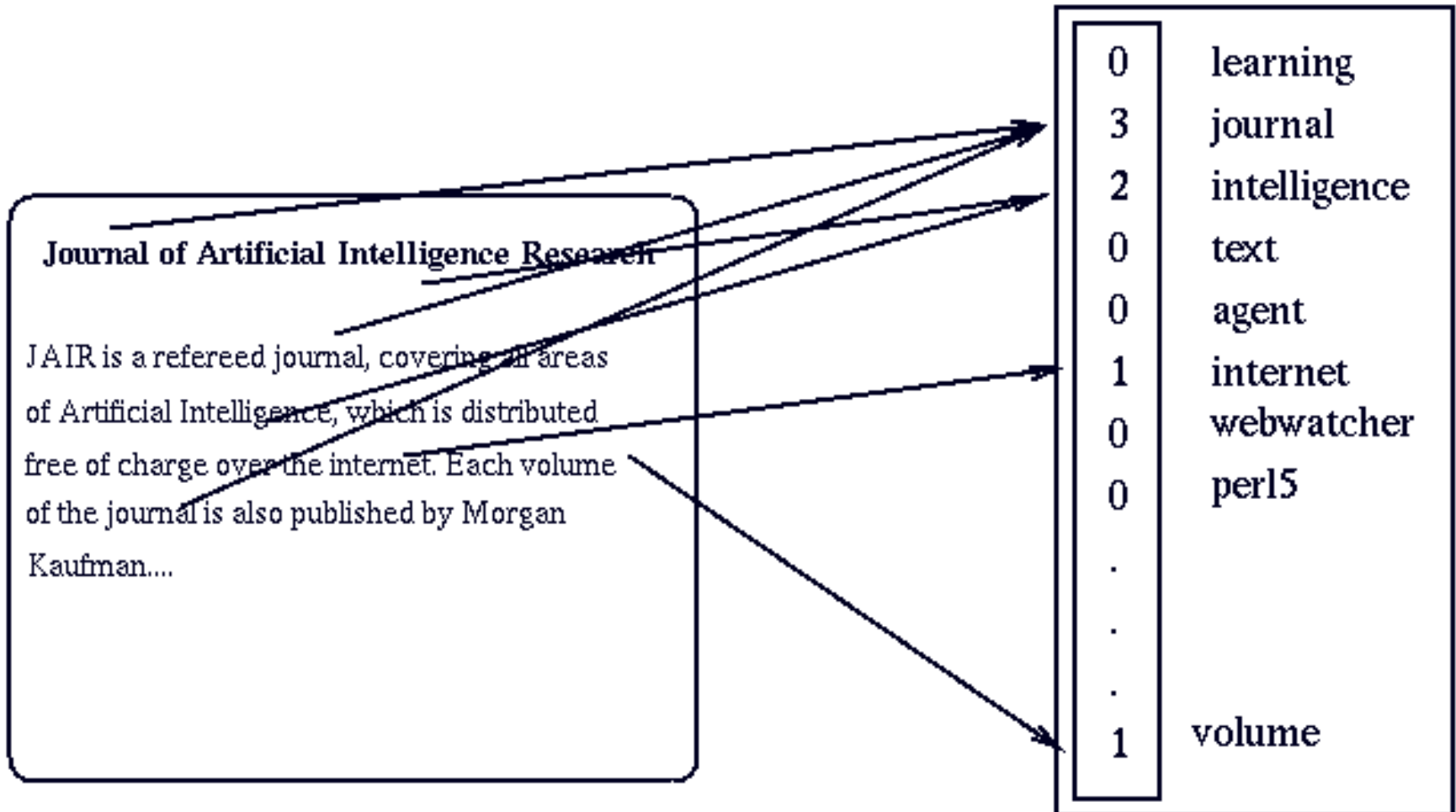
Text mining: Words/terms as binary features

Document	Word1	Word2	...	WordN	Class
d1	1	1	0	1	NO
d2	1	1	0	0	YES
d3	1	1	1	1	NO
d4	1	1	1	0	YES
d5	1	0	0	1	NO
d6-d13
d14	0	0	0	0	YES
d15	0	0	1	1	NO
d16	0	0	1	0	NO
d17	0	1	0	1	NO
d18	0	1	0	0	NO
d19-d23
d24	0	0	1	0	NO

Instances = documents

Words and terms = Binary features

Bag-of-Words document representation



Word weighting for BoW document representation

- In bag-of-words representation each word is represented as a separate variable having numeric weight.
- The most popular weighting schema is normalized word frequency TFIDF:

$$tfidf(w) = tf \cdot \log\left(\frac{N}{df(w)}\right)$$

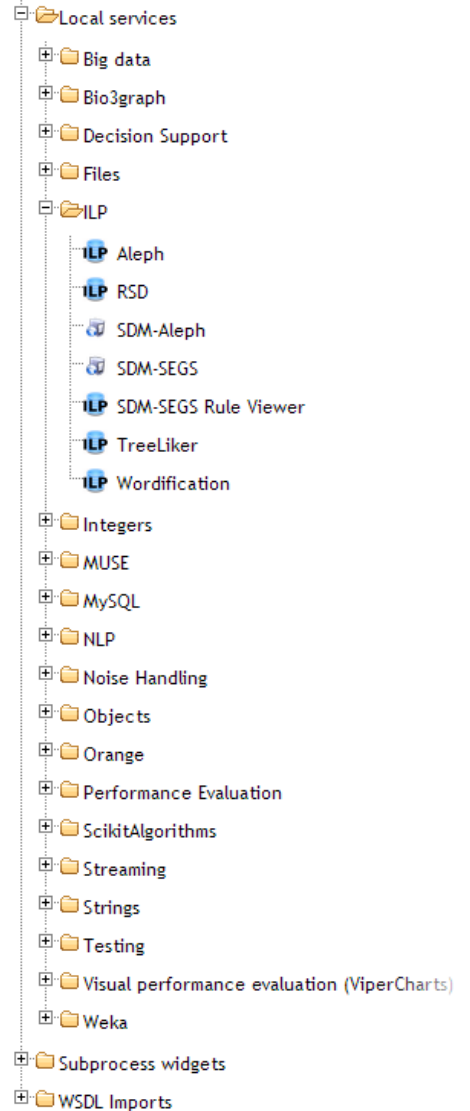
- $Tf(w)$ – term frequency (number of word occurrences in a document)
- $Df(w)$ – document frequency (number of documents containing the word)
- N – number of all documents
- $Tfidf(w)$ – relative importance of the word in the document

The word is more important if it appears several times in a target document

The word is more important if it appears in less documents

Third Generation Data Mining Platforms

- **Orange4WS** (Podpečan et al. 2009), **ClowdFlows** (Kranjc et al. 2012) and **TextFlows** (Perovšek et al. 2016)
 - are service oriented (DM algorithms as web services)
 - user-friendly HCI: canvas for workflow construction
 - include functionality of standard data mining platforms
 - WEKA algorithms, implemented as Web services
 - Include new functionality
 - relational data mining
 - semantic data mining
 - NLP processing and text mining
 - enable simplified construction of Web services from available algorithms
 - ClowdFlows and TextFlows run in a browser – enables data mining, workflow construction and sharing on the web

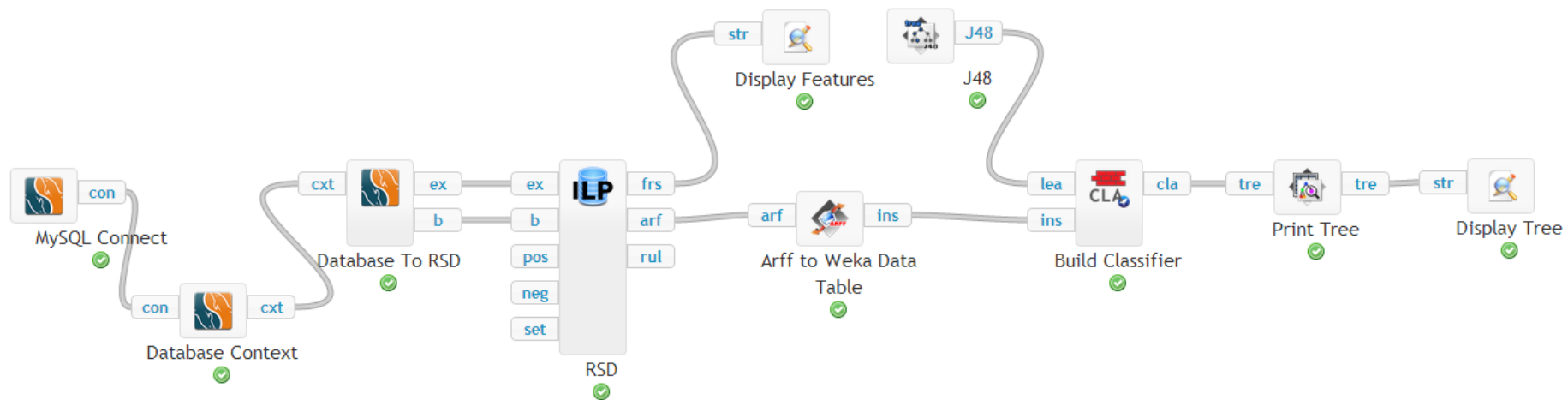


CloudFlows platform

- **Large algorithm repository**
 - Relational data mining
 - All Orange algorithms
 - WEKA algorithms as web services
 - Data and results visualization
 - Text analysis
 - Social network analysis
 - Analysis of big data streams
- **Large workflow repository**
 - Enables access to our technology heritage

CloudFlows platform

- Large repository of algorithms
- Large repository of workflows



Example workflow:

Propositionalization with RSD
available in CloudFlows at

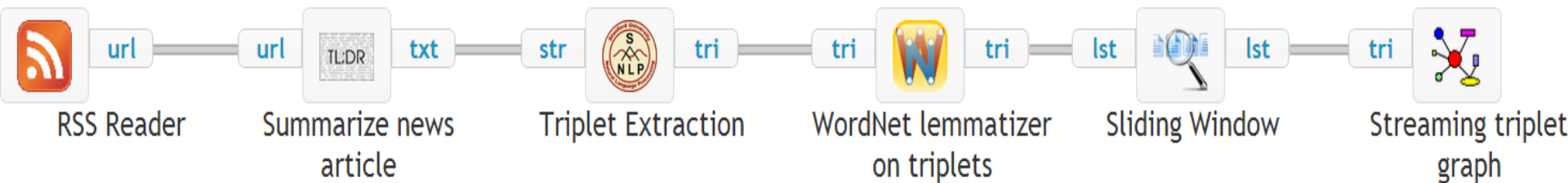
<http://clowdflows.org/workflow/611/>

TextFlows

- Motivation:
 - Develop an online text mining platform for composition, execution and sharing of text mining workflows
- TextFlows platform – fork of ClowdFlows.org:
 - Specialized on text mining
 - Web-based user interface
 - Visual programming
 - Big roster of existing workflow (mostly text mining) components
 - Cloud-based service-oriented architecture

“Big Data” Use Case

- Real-time analysis of big data streams
- Example: semantic graph construction from news streams. <http://clowdflows.org/workflow/1729/>.



- Example: news monitoring by graph visualization (graph of CNN RSS feeds)

<http://clowdflows.org/streams/data/31/1>



Part I: Summary

- KDD is the overall process of discovering useful knowledge in data
 - many steps including data preparation, cleaning, transformation, pre-processing
- Data Mining is the data analysis phase in KDD
 - DM takes only 15%-25% of the effort of the overall KDD process
 - employing techniques from machine learning and statistics
- Predictive and descriptive induction have different goals: classifier vs. pattern discovery
- Many application areas, many powerful tools available