Utrecht University

Faculty of Science
Department of Information and Computing Science
Master of Business Informatics

Seminar Medical Informatics

Data Science in Health Care

Verónica Burriel Coll

v.burriel@uu.nl

12 March 2019

Agenda for today

Data Science in Health Care

- Usefulness of data science in health care
- Benefits
- Characteristics of Big Data in health care
- Approaches to analyses
- Knowledge discovery and data mining

Assignment for next March 18th:
Data Science for Health Care's Workshop

Next weeks' schedule

Guest talk: Care for babies in the right place at the right time
by Devika Jagesar

Usefulness of Data Science in Health Care

Objectives:

- Create knowledge for learning
- Predict potential risks
- Address practical questions about benefits and cost

Large datasets of different data types

Identify trends, patterns and associations

Data categories	Examples of Collected Data
Web and social media	Facebook, Twitter, mHealth apps, health databases
Machine to machine	Uploads and readings from sensors and devices
Big transaction	Claims data and billing records
Biometric	Vital signs, medical imaging, fingerprints, genetics
Human generated	EHR, email, paper documents, data repositories

Benefits of Data Science in Health Care

- Inform health care delivery decisions based on complex information
- Advancement of science
- Improvements in health care, treatments and economics of healthcare
- Computerized clinical decision support systems

Characteristics of Big Data in Health Care

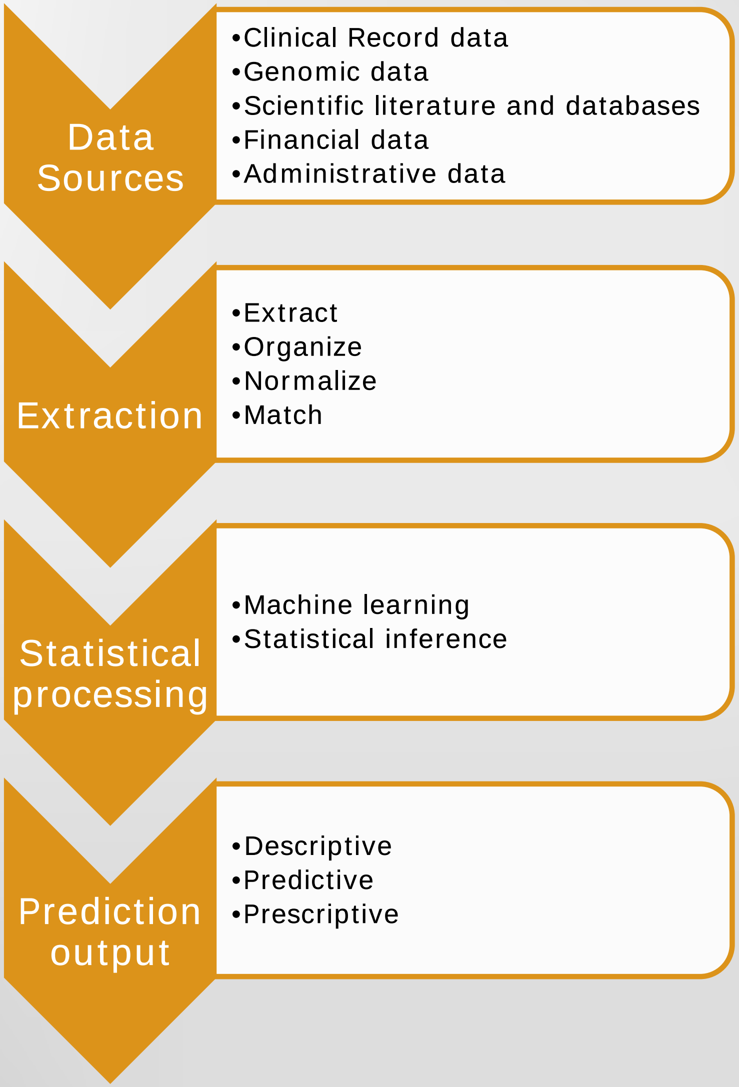
“The Five Vs”

- Volume : The sheer quantity of data generated and analyzed
- Velocity : The speed at which data are generated and change over time
- Variety : The data come from many different sources and in many different formats
- Veracity : The accuracy and completeness of the data (the “truth”)
- Value : The purpose of collecting, processing and analyzing data are to fill a need

Approaches of Data Analytics in Health Care

- **Descriptive (or Exploratory):** Prepares and analyzes retrospective data to identify patterns or trends
- **Predictive:** Development of analytic models that predict future trends based on retrospective or real-time data
  - **Regression**: Outcome or new observation
  - **Classification**: Category for a new outcome
  - **Clustering**: Grouping observations into similar groups
  - **Association rules**: Determining a new characteristic based on known characteristics of an observation
- **Prescriptive:** Use of models to evaluate and determine new ways of operating in a health system

The Analytics pipeline



Data Sources

- Clinical Record data
- Genomic data
- Scientific literature and databases
- Financial data
- Administrative data

Extraction

- Extract
- Organize
- Normalize
- Match

Statistical processing

- Machine learning
- Statistical inference

Prediction output

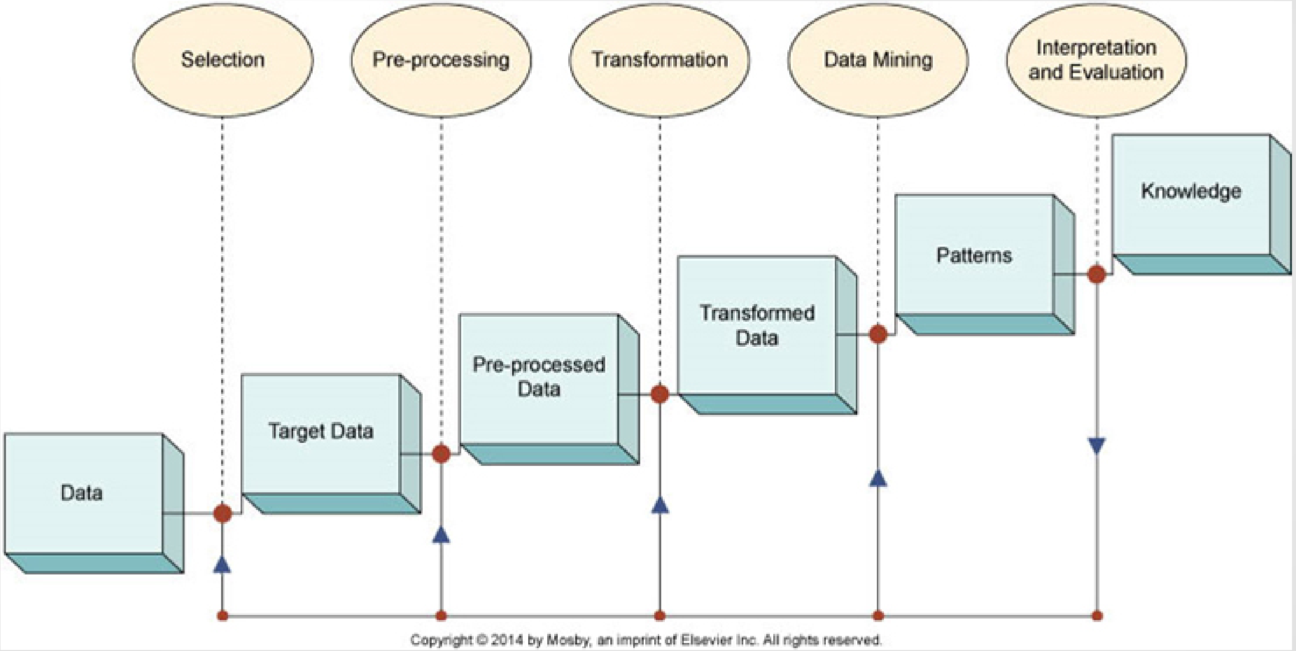
- Descriptive
- Predictive
- Prescriptive

Knowledge discovery and data mining

Advantages of KDDM models:

- Access and leverage valuable information contained in large repositories of clinical data
- Can be developed from large sample sizes or entire populations
- Models based on routinely collected data can be implemented in computerized systems to support decision making
- Can be induced directly from data, using machine learning methods, and often perform better than models manually developed by human experts

Steps of Knowledge Discovery and Data Mining Process



Copyright © 2014 by Mosby, an imprint of Elsevier Inc. All rights reserved.

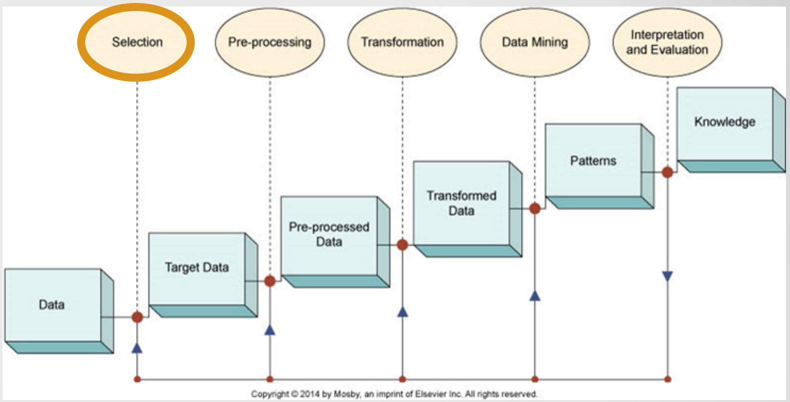
1. Retrieving a set of data for analysis
2. Preprocessing clinical data
3. Sampling, partitioning and transformation
4. Data mining
5. Interpretation and evaluation

Steps of Knowledge Discovery and Data Mining Process:

Retrieving a set of data for analysis

Data sources:

- Electronic Health Record
- Genomic data
- Scientific literature and databases
- Financial data
- Administrative data



Copyright © 2014 by Mosby, an imprint of Elsevier Inc. All rights reserved.

Define a subset of relevant data

- Work jointly with experts
- Sufficient but no overwhelming sample size

Understand the concepts represented in data

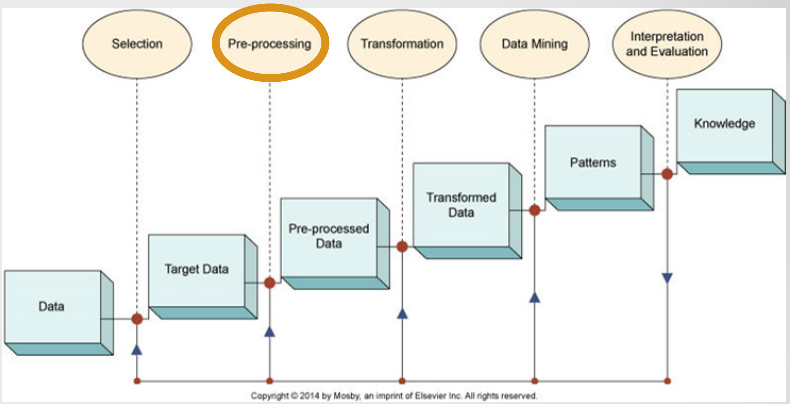
- Same concept with different words
- Terminologies and codes
- Data no structured or in free text

Collaboration to define effective queries

Steps of Knowledge Discovery and Data Mining Process:

Preprocessing clinical data

The majority of effort in the clinical KDDM process



Copyright © 2014 by Mosby, an imprint of Elsevier Inc. All rights reserved.

Different terminologies:
Data aggregated across time and across sites results in a dataset that represents similar concepts in multiple ways

Non-structured (text) data: Natural Language Processing

Structured (coded) data: Joint effort with clinicians.
May reveal conceptual gaps, absence of data, lack of quality data

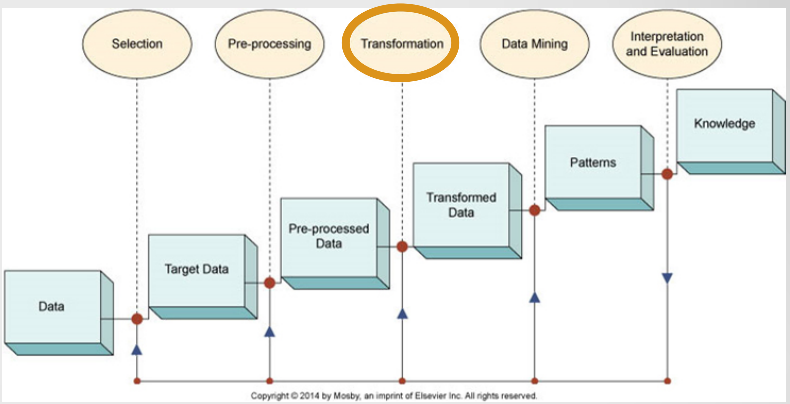
Steps of Knowledge Discovery and Data Mining Process:

Sampling, partitioning and transformation

Sampling: A smaller subset of the data is chosen for analysis

Partitioning: Assignment of individual records or rows in a dataset for a specific purpose: model development or model validation

Transformation: Migrate data from preprocessed files to the modeled environment

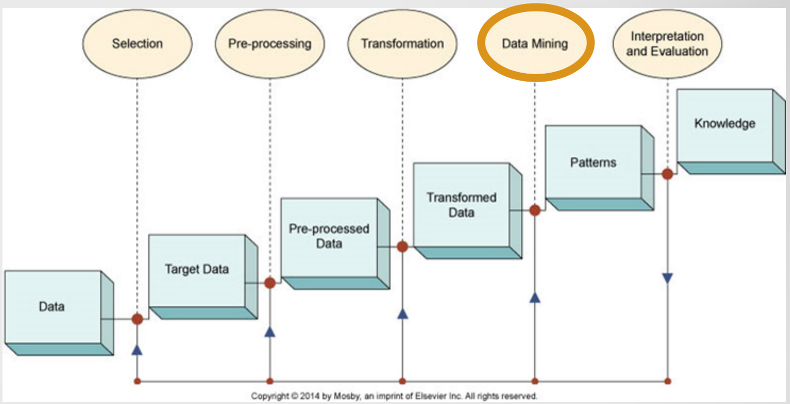


Copyright © 2014 by Mosby, an imprint of Elsevier Inc. All rights reserved.

Steps of Knowledge Discovery and Data Mining Process:

Data mining

Patterns are enumerated over a set of data



Copyright © 2014 by Mosby, an imprint of Elsevier Inc. All rights reserved.

Statistical approaches: Fit a model to the data

- Bayesian models
- Linear regression

Machine learning: Computer algorithms that learn to perform a task on the basis of examples

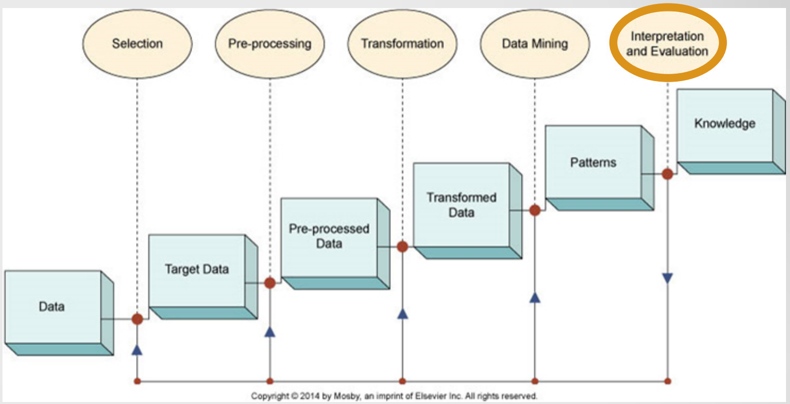
- Prediction or regression (predict a real value)
- Classification (predict class membership)

Steps of Knowledge Discovery and Data Mining Process:

Interpretation and evaluation

Comparing a model's predictions to actual values with a set of data for which the actual values are known

Use of validation set



Copyright © 2014 by Mosby, an imprint of Elsevier Inc. All rights reserved.

Prediction: Difference between real numbers

Classification : Classification matrix (true-false)

Food for further thought

Embracing big data - the future of healthcare
Willem Herter & Wouter Kroese | TEDxSaxionUniversity

1. They affirm that when a doctor prescribes you a treatment, he doesn’t have enough information to know if this treatment will work for you. What do you think about this affirmation? Do you think this it is an effectiveness problem of actual medicine?

2. Do you think that this problem can be solved using Data Science to analyze digital clinical data of patients? Could it be the solution to the problem?

3. How can you argue in opposite of their argumentation? Try to find any negative point.

4. Do you think they are realistic in wanting to analyze information from all these sources? Which are the possible threats to this?

Assignment for next March 18th: Data Science in Health Care Workshop

Assigned students:

- Leonardo Vida
- Jathin Nagesh
- Esmée van Vilsteren
- Noël Bainathsah
- Pratik Kushwaha

Each assigned student:

1. **Select a paper** of aprox. 8 pages about a Data Science in healthcare solution and send it to v.burriel@uu.nl **before Wednesday at 13.00.** During the afternoon all selected papers will be published on course's website.

2. **Prepare a presentation** of **7/8 minutes** about the paper and include some questions (at least 2) at the end of the presentation to challenge the audience and activate the discussion.

3. Join with the other assigned students and **prepare 1 or 2 group activities** to make during the last 30 minutes of the session. These activities should be related to the solutions presented.

Assignment for next March 18th:
Data Science in Health Care Workshop

Each no-assigned student:

1. **Read all the selected papers** and **prepare some questions or comments** (at least 2) per paper to discuss them after the presentation. Try to be critical and/or creative.

2. Send the questions/comments using this form **before Monday**
https://goo.gl/forms/K69vlahNqxzFe1ZG3

Next weeks' schedule

Week	Monday	Workshop	Tuesday	Lecture
12	March 18	Data Science solutions	March 19	Bioinformatics and Precision Medicine
	Wednesday *March 20, 5 PM*	**Paper mid-term submission**		
13	March 25	Bioinformatics solutions	March 26	**Multiple-choice quiz**
	Wednesday *March 27, 5 PM*	**Paper students assessment**		
14	April 1	**Mobile Apps presentations**	April 2	**Mobile Apps presentations**
15	*Friday* *April 12, 5 PM*	**Paper final submission**		

Submission details of mid-term version of your paper

Your mid-term version should include (at least):

- Introduction (incl. motivation)
- Problem statement (research about the disease/condition and needs detected)
- State of art (literature review of related Apps)
- Solution design (at least an overview to show how your App will be)

Paper should have **maximum 12 pages** and be in **Springer format**: You can find the templates in Word and Latex here: http://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines

Submission deadline: Wednesday March 20th at 5.00 PM

Submission procedure: Submission to **Peergrade** site. You will receive in the previous days an invitation to join Peergrade site.

Assessment of mid-term paper versions

On March 20: A paper will be assigned to you to make the peer review. **Be constructive!**

Students assessments submission deadline:
Wednesday March 27th at 5.00 PM in **Peergrade**.

After deadline on March 27: You can check your paper's assessment and feedback in Peergrade.

Use this exercise to get some ideas and improve the final version of your paper!

Grade

- 20%
  - 10% 2 students assessment of your paper
  - 10% your assessment of a paper

Multiple-choice quiz

Date: Tuesday March 26th 13.15
Place: HFG 611AB

Multiple choice questions

Knowledge acquired from lectures and workshop sessions will be evaluated

Grade

- Grade = 10%
- No retake opportunity if grade is lower than 4.0 or higher than 6.0
- At least 6.0 to pass the course (5.99 is not enough)

Mobile Apps presentations

Dates: **Monday April 1st 15.15 -17.00**
Tuesday April 2nd 13.15 -15.00

12 minutes presentations of your Mobile Apps for Health

Some questions from the audience

You will evaluate your mates using an assessment form

Grade

- 20% (10% students grade and 10% teacher grade)
- Attendance to both sessions is required to get the students grade
- Required to pass the course
- No second chance

Final paper version submission details

Your final paper should include (at least):

- Introduction (incl. motivation)
- Problem statement (research about the disease/condition and needs detected)
- State of art (literature review of related Apps)
- Solution design (including design of the App, technologies needed to create it and functionalities that solve the problem exposed in problem statement).
- Conclusions (including expected benefits of using this app)
- Future work

Paper should have **maximum 12 pages** and be in **Springer format.**

Submission deadline: **April 12th at 5.00 PM**
Submission procedure: Using the corresponding assignment in Blackboard.

Grade

- 30%
- No resubmit opportunity if grade is lower than 4.0 or higher than 5.5

Guest talk:

Care for babies in the right place at the right time: a data architecture to structure data flows to manage bed capacity in birth centres

by Devika Jagesar