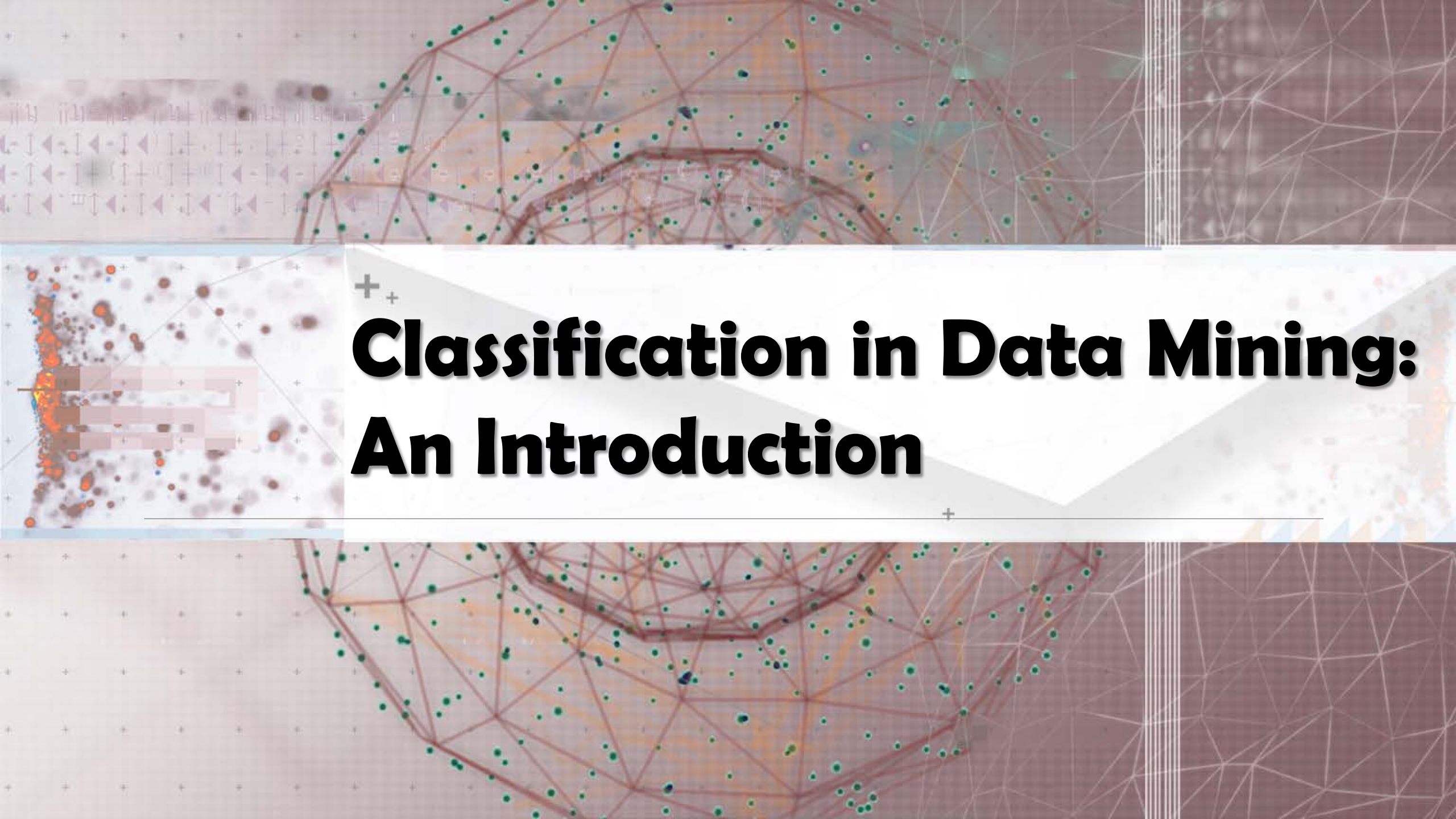
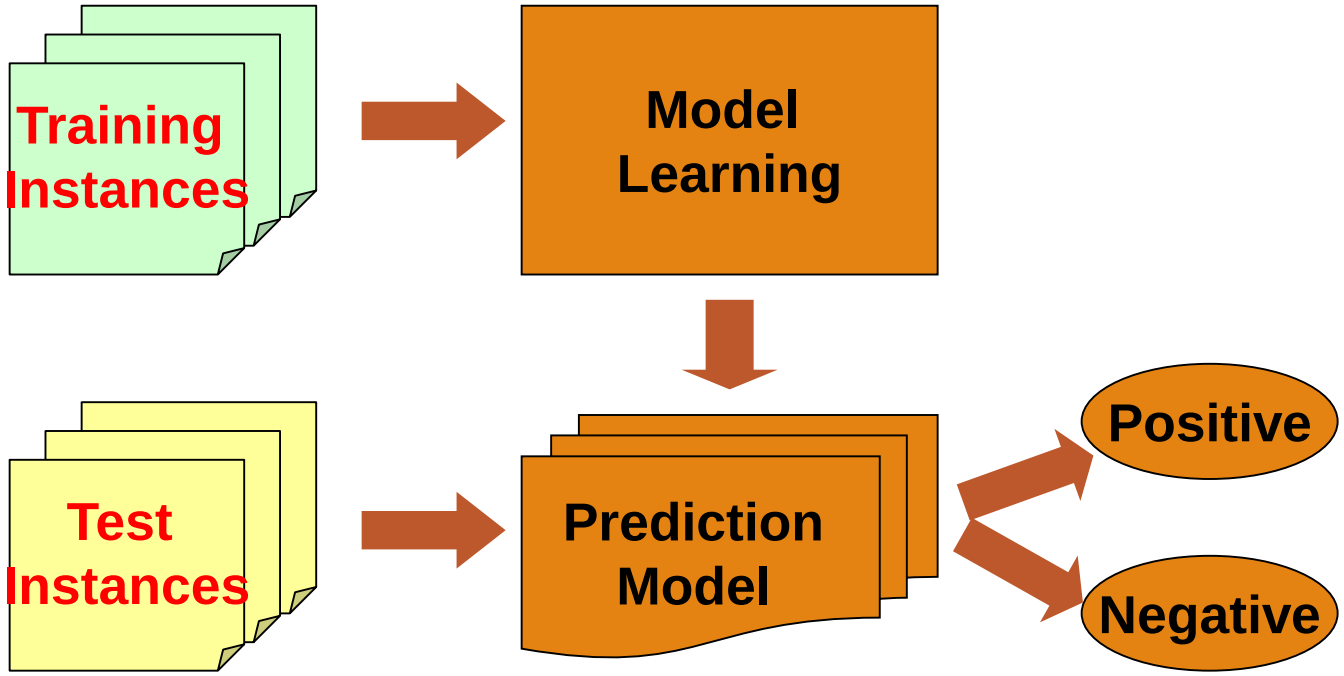


Classification in Data Mining: An Introduction

Supervised vs. Unsupervised Learning (1)

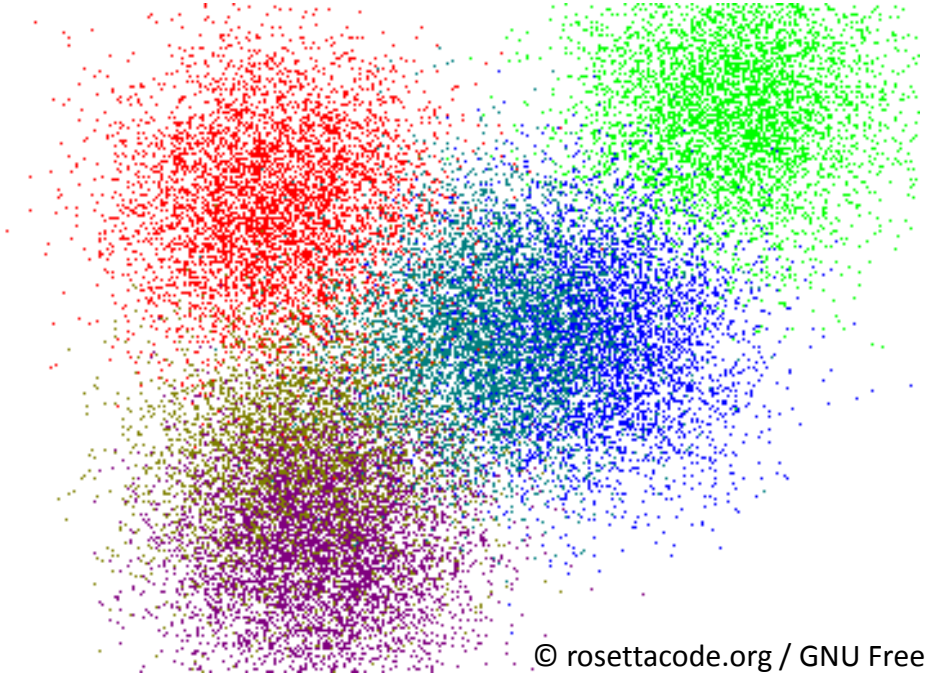
- Supervised learning (classification)
  - Supervision: The training data, such as observations or measurements, are accompanied by **labels** indicating the classes to which they belong
  - New data is classified based on the models built from the training set

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



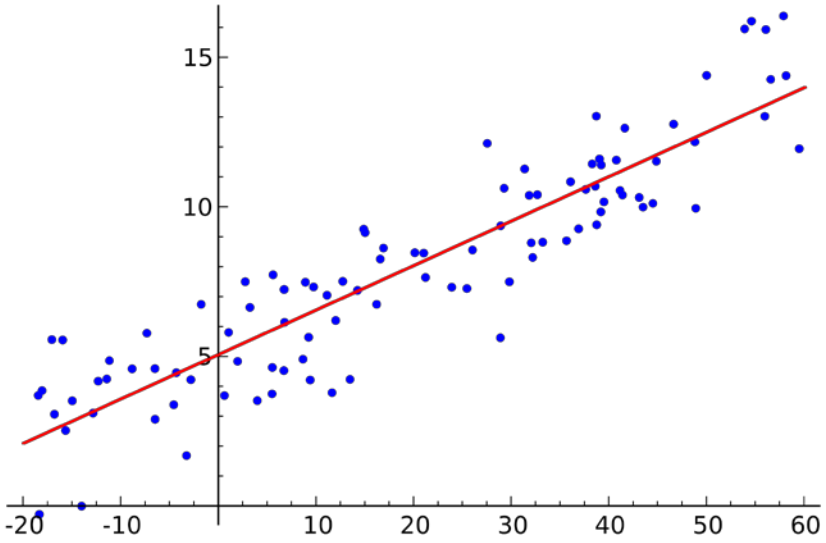
Supervised vs. Unsupervised Learning (2)

- Unsupervised learning (clustering)
  - The class labels of training data are unknown
  - Given a set of observations or measurements, establish the possible existence of classes or clusters in the data



Prediction Problems: Classification vs. Numeric Prediction

- Classification
  - Predict categorical class labels (discrete or nominal)
  - Construct a model based on the training set and the **class labels** (the values in a classifying attribute) and use it in classifying new data
- Numeric prediction
  - Model continuous-valued functions (i.e., predict unknown or missing values)
- Typical applications of classification
  - Credit/loan approval
  - Medical diagnosis: If a tumor is cancerous or benign
  - Fraud detection: If a transaction is fraudulent
  - Web page categorization: Which category it is

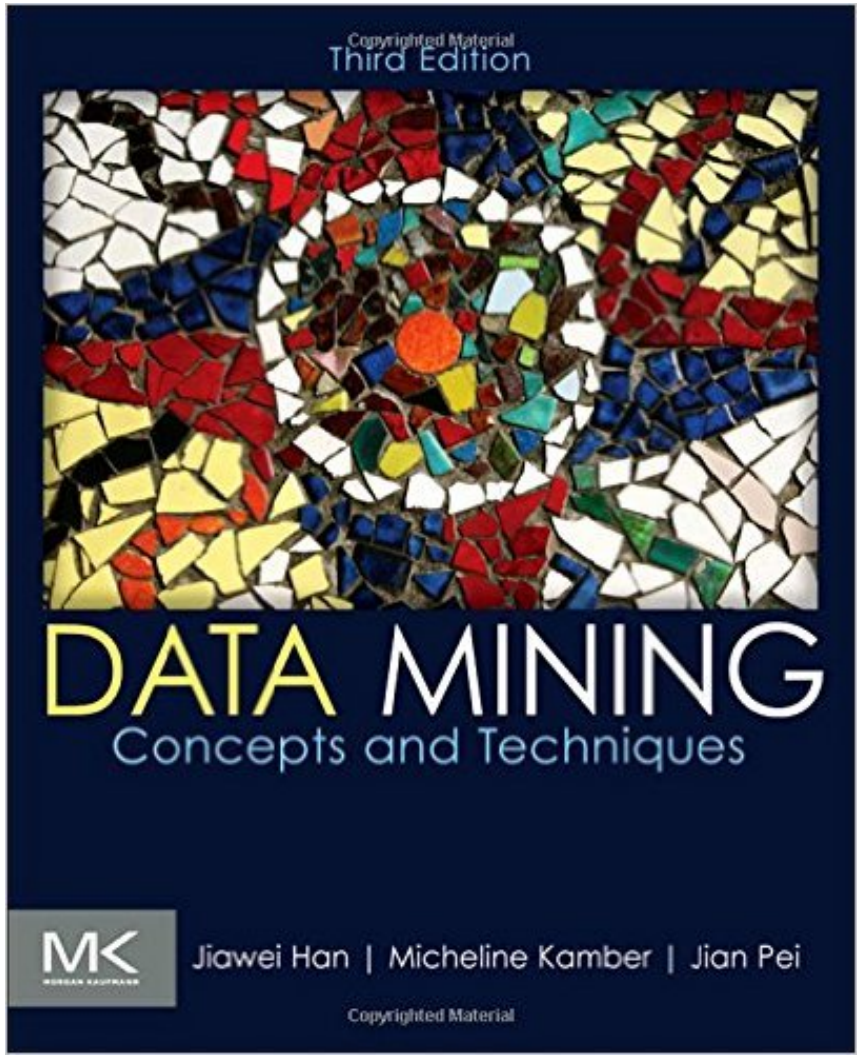


Classification—Model Construction, Validation and Testing

- **Model construction**
  - Each sample is assumed to belong to a predefined class (shown by the **class label**)
  - The set of samples used for model construction is the **training set**
  - **Model**: Represented as decision trees, rules, mathematical formulas, or other forms
- **Model validation and testing**:
  - **Test:** Estimate accuracy of the model
    - The known label of test sample is compared with the classified result from the model
    - *Accuracy:* % of test set samples that are correctly classified by the model
    - Test set is independent of training set
  - **Validation**: If *the test set* is used to select or refine models, it is called **validation** (or development) **(test) set**
- **Model deployment:** If the accuracy is acceptable, use the model to classify new data

Major Reference Readings for the Course

- **Textbook**
  - Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.
  - **Chapters most related to the course**
    - Chapter 8: Classification: Basic Concepts
    - Chapter 9: Classification: Advanced Methods
- Other references will be listed at the end of each lecture video



Course Structure

- Lesson 0: Classification in Data Mining: An Introduction
- Lesson 1: Decision Tree Induction
- Lesson 2: Bayes Classifier and Bayesian Networks
- Lesson 3: Model Evaluation, Selection, and Improvements
- Lesson 4: Linear Classifier and Support Vector Machines
- Lesson 5: Neural Networks and Deep Learning
- Lesson 6: Pattern-Based Classification and K-Nearest Neighbors Algorithm

Course General Information

- Instructor:

  Jiawei Han, Abel Bliss Professor

  Department of Computer Science

  University of Illinois at Urbana-Champaign

- Teaching assistants
- Course prerequisite:

  Familiarity with basic data structures and algorithms

- Course assessments
  - In-video questions
  - Lesson quizzes
  - Programming assignments
  - Exam

Recommended Readings

- Aggarwal, C. C. (2015). *Data mining: The textbook.* New York, NY: Springer.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). Hoboken, NJ: John Wiley.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.
- Mitchell, T. M. (1997). *Machine Learning*. Columbus, OH: McGraw Hill.
- Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2013). *Introduction to data mining* (2nd ed.). Boston, MA: Addison-Wesley.
- Weiss, S. M. & Kulikowski, C. A. (1991). *Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Burlington, MA: Morgan Kaufmann.
- Witten, I. H. & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). Burlington, MA: Morgan Kaufmann.
- Zaki, M. J. & Meira Jr., W. (2014). *Data mining and analysis: Fundamental concepts and algorithms*. Cambridge, UK: Cambridge University Press.

References

- Morgan Kaufmann. (2011). *Data mining: Concepts and techniques (3rd ed.) book cover* [Online image]. Retrieved Feb 16, 2018 from https://www.elsevier.com/books/data-mining-concepts-and-techniques/han/978-0-12-381479-1
- rosettacode.org. (2018). *Cluster diagram* [Online image]. Retrieved Feb 16, 2018 from https://goo.gl/g7KTCQ

-