Machine Learning for Biology

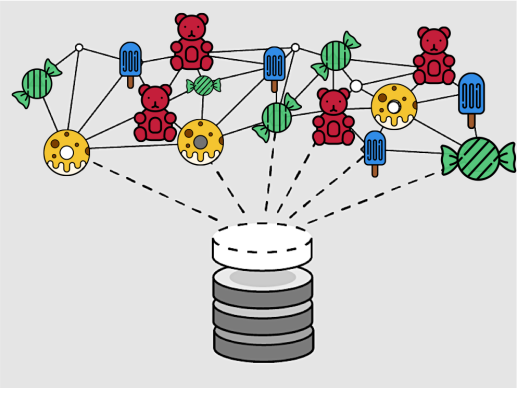
Workshop, D-BSSE Retreat

Anja Gumpinger Catherine Jutzeler Bastian Rieck Caroline Weis

28 June 2019

Machine leaning and candy

A good fit



What is machine learning?

Group work 1

- What is machine learning?
- Which machine learning techniques do you already know?

Instructions: please discuss the questions above within your group. Use the provided pens and sheets to write down your answers. ($\approx$ 5 min)

Machine learning (ML)

"Machine learning is the field of study that gives the computer the *ability to learn without being explicitly programmed*."

Arthur Samuel, Computer Scientist, 1959

Machine learning (ML)

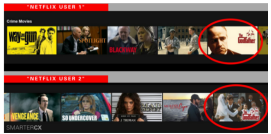
"Machine learning is the field of study that gives the computer the *ability to learn without being explicitly programmed*."

Arthur Samuel, Computer Scientist, 1959

"A computer program is said to *learn from Experience E* with respect to some *class of Task T* and some *performance measure P*, if its performance on T, as measured by P, improves with *Experience E*."

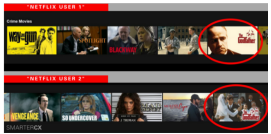
Tom Mitchell, Computer Scientist, 1997

Everyday applications of machine learning

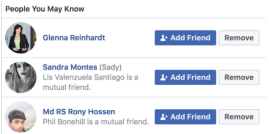


Personalised entertainment recommendations

Everyday applications of machine learning

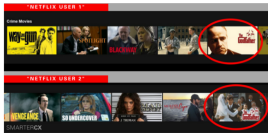


Personalised entertainment recommendations

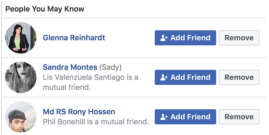


Personalised contact recommendations

Everyday applications of machine learning



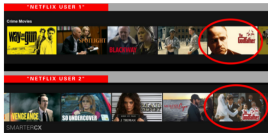
Personalised entertainment recommendations



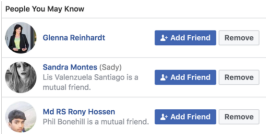
Personalised contact recommendations

Music recognition

Everyday applications of machine learning



Personalised entertainment recommendations

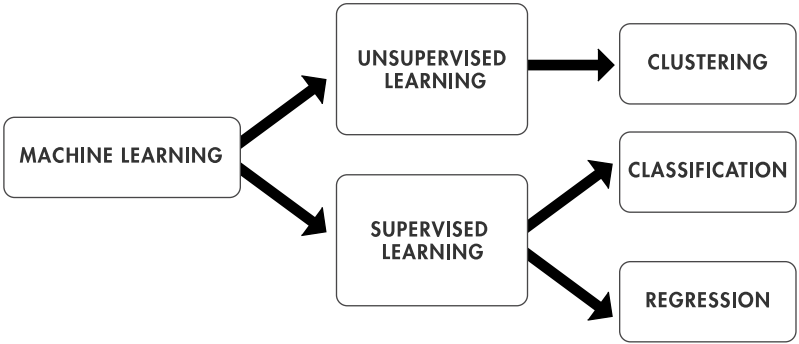


Personalised contact recommendations

Music recognition

Traffic prediction

A bestiary for machine learning


MACHINE LEARNING
UNSUPERVISED LEARNING
SUPERVISED LEARNING
CLUSTERING
CLASSIFICATION
REGRESSION

Clustering methods

What is clustering?

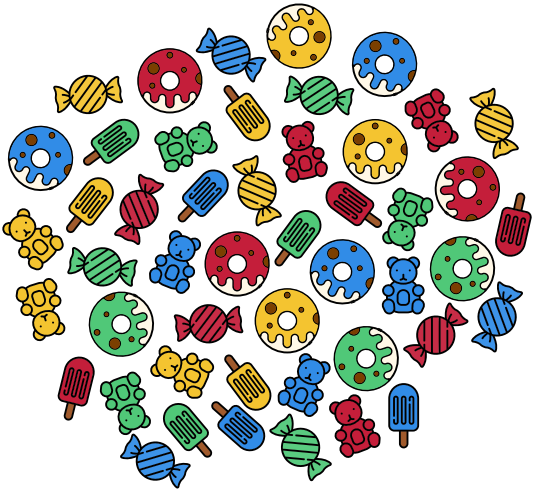
Problem definition

Given a set of n objects, how can we group them into k clusters, such that all objects in one cluster are more *similar* to each other than to the objects in the remaining clusters?

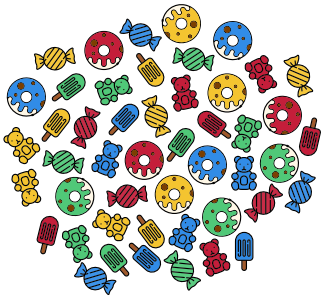
- *Unsupervised* technique: labels are not known
- Some parameter choices:
  - How many clusters?
  - How to measure similarity?

Group work 2

Candy mountain clustering



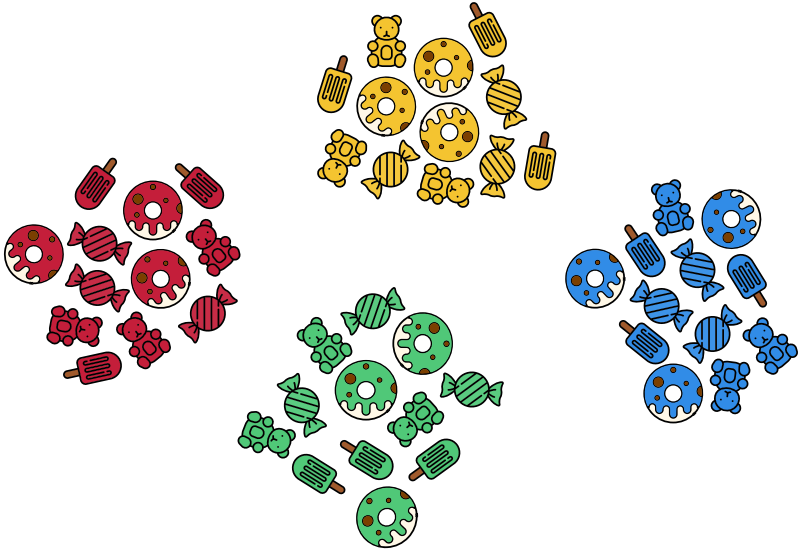
Some constraints



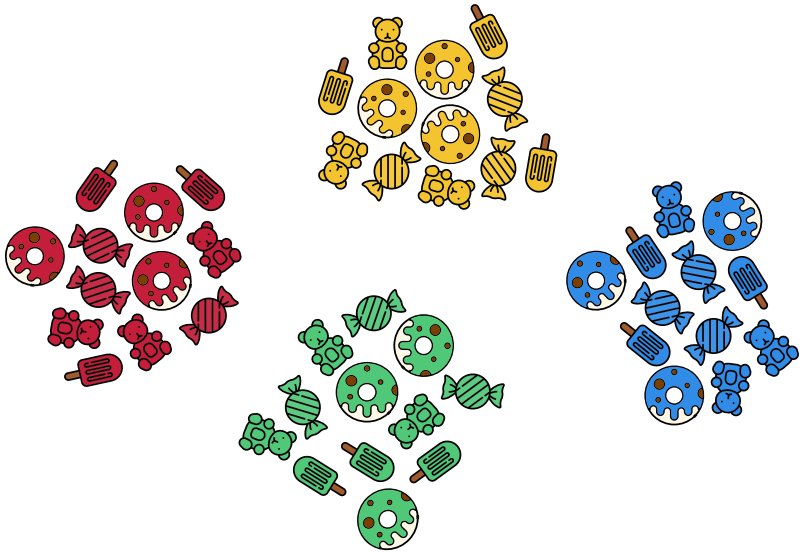
- There are $n = 48$ objects
- Let us assume we want $k = 4$ clusters
- There are $S(48, 4) = 330116014368723828972353170 1$ ways of doing so: three octillion, three hundred one septillion, one hundred sixty sextillion, one hundred forty three quintillion, six hundred eighty seven quadrillion, two hundred thirty eight trillion, two hundred eighty nine billion, seven hundred twenty three million, five hundred thirty one thousand, seven hundred one

There is no *best* clustering.

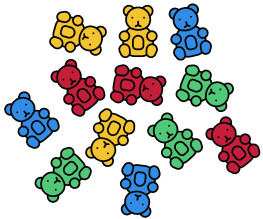
?

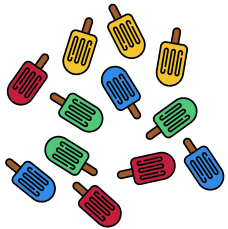


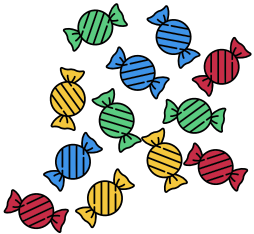
Clustering by colour ($k = 4$)

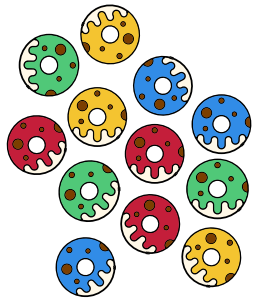


?

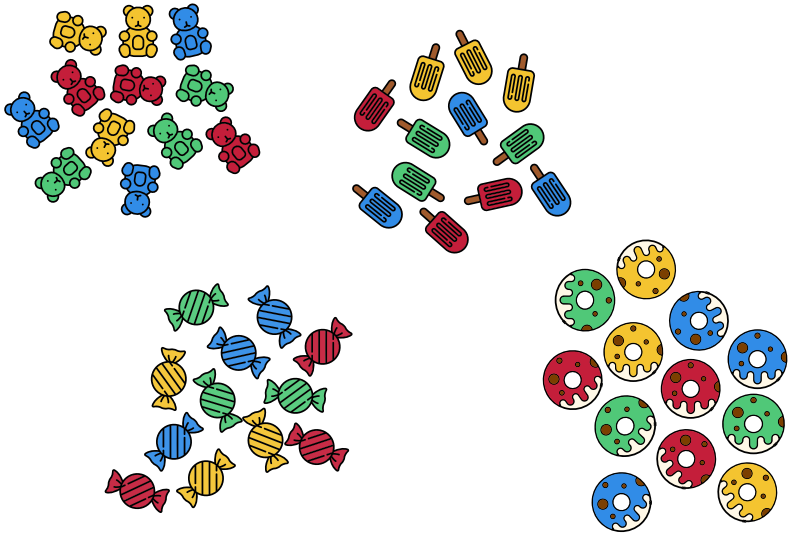




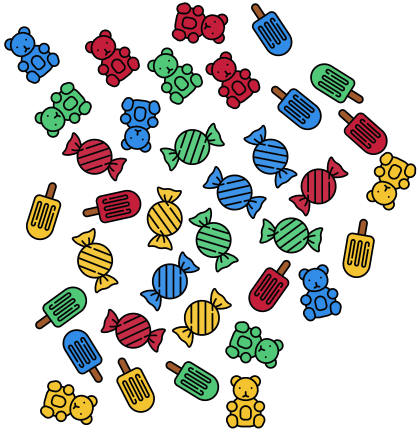


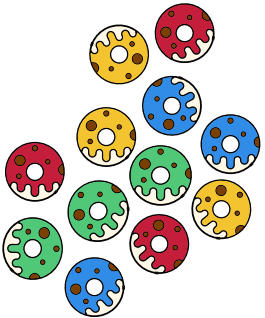


Clustering by type ($k = 4$)

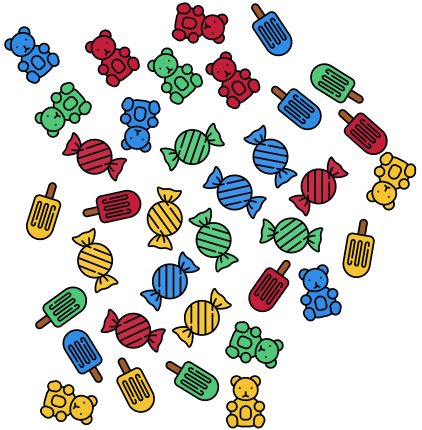


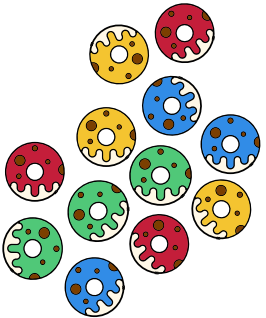
?



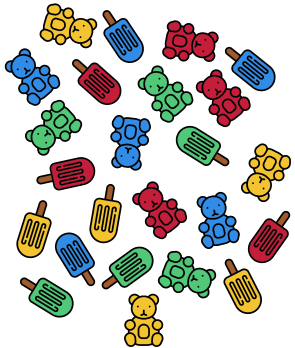


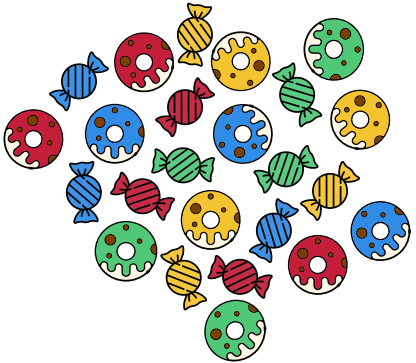
Clustering by topology ($k = 2$)



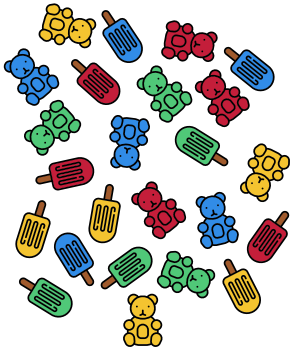


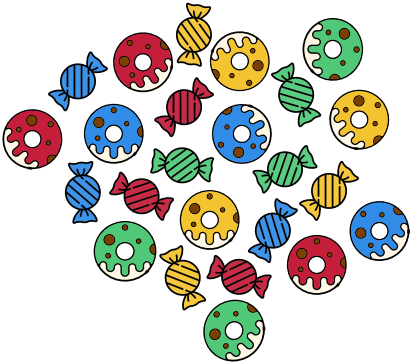
?





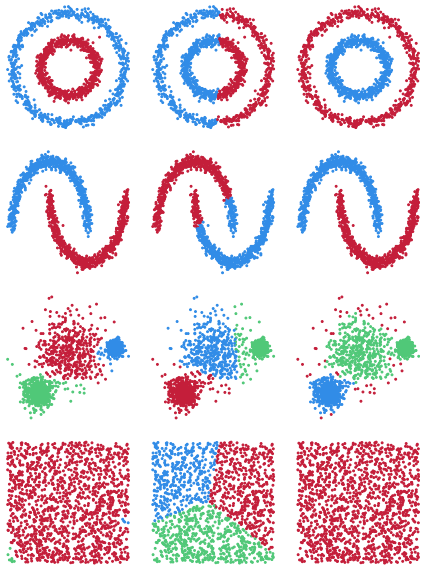
Clustering by ingredients ($k = 2$)





A selection of clustering methods

- Hierarchical clustering (agglomerative, divisive, …)
- k-means
- Density-based clustering (DBSCAN, DENCLUE)


Agglomerative k-means
DBSCAN

Similarity measures

Which metric should I use for…?	
Continuous data (temperature, length, …)	Euclidean distance, Mahalanobis distance
Categorical data (blood type, sex, …)	Hamming distance
Images	Euclidean distance (?)
Graphs	Weisfeiler–Lehman graph kernels
Time series	Dynamic time warping (DTW)

How to choose k?

- β-*CV*: ratio of mean intracluster distance to mean intercluster distance
- *C index*: ratio of intracluster distances to sum of the largest distances between points
- *Dunn index*: ratio of smallest distance between points in different clusters to largest diameter of a cluster
- *Silhouette score*: ensures that clusters are *cohesive* while also being *separate*

All of these measures have their shortcomings, though, and can be "tricked". Ideally, ground truth information is available to verify a clustering.

Classification methods

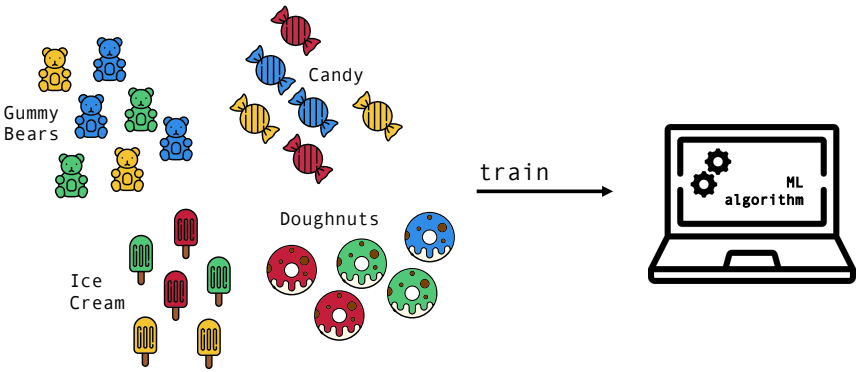
Classification

Problem:
Identify a new object based on previously seen objects of similar type.

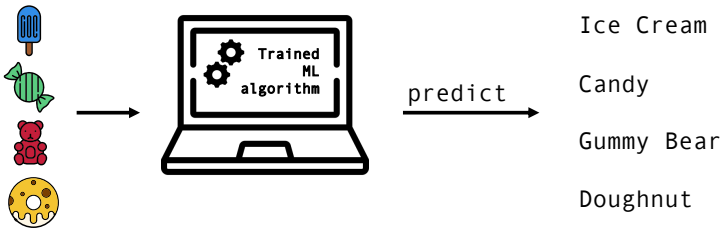
Properties:

- instance of *supervised* machine learning, i.e. requires *labels* for data points
- requires *training*, i.e. presentation of examples to an algorithm
- *prediction*, i.e. the prediction of the label for previously unseen data points

Classification - Concept: Training



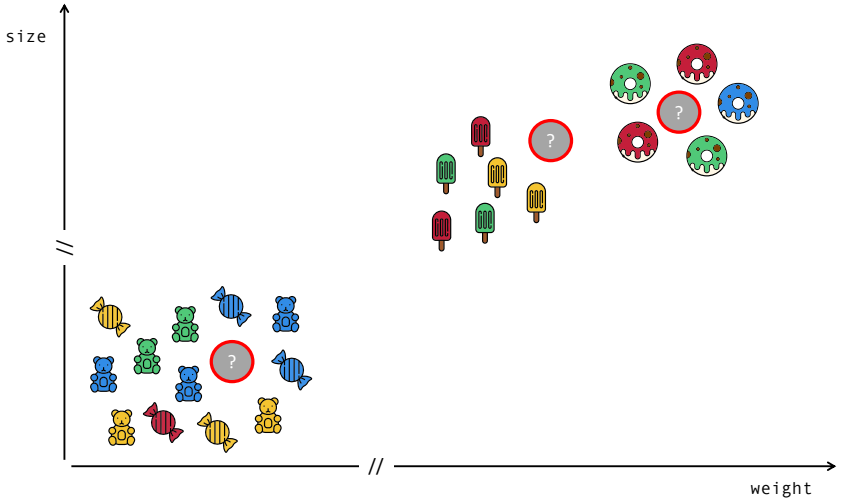
Classification - Concept: Prediction



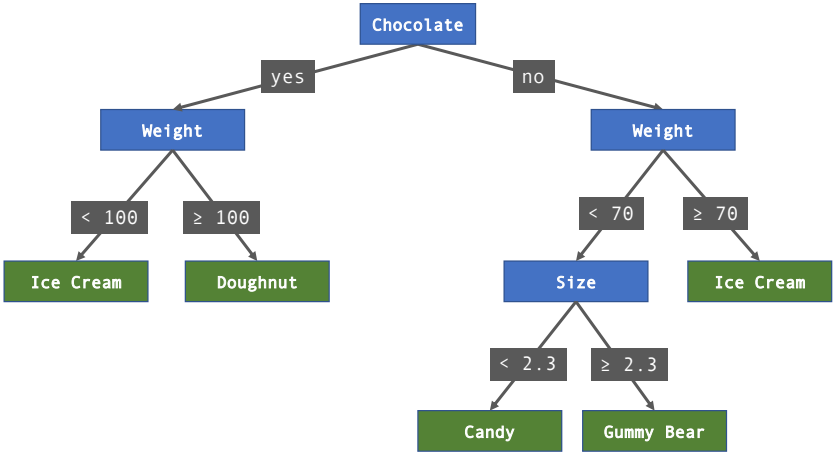
Classification - Data Representation

	color	size [cm]	weight [g]	chocolate	label
	red	2.02	1.99	0	Gummy Bear
	green	1.98	2.10	0	Gummy Bear
	yellow	14.20	56.15	1	Ice Cream
	blue	2.45	4.30	0	Candy
	green	2.49	4.17	0	Candy
		⋮			⋮
	blue	14.20	56.15	1	Ice Cream
	yellow	16.80	110.30	1	Doughnut

Classification: Nearest Neighbor Classifiers



Classification: Decision Tree Example



Famous Classification Methods

- Logistic Regression
- k-nearest Neighbor Classification (*k*-NN)
- Naïve Bayes
- Support Vector Machines (SVM)
- Decision Trees and Random Forests
- Neural Networks

Classification vs. Regression

- Very similar concepts and methods
- Instead of predicting a discrete *class label* for each sample, predict a measurable *continuous quantity*

Candy Regression

The target to predict could become e.g.

- sugar content of a sweet
- calories
- ...

Common myths in machine learning

Myths in machine learning

1. Machine learning does not need a clear objective.
2. Machine learning does not need a clear hypothesis.
3. Machine learning can extract the *right* signal from your data.

Machine learning needs clear objectives

Prior to applying any machine learning algorithm, you should have a clear objective in mind. Do you want to cluster data to learn about (unknown) structures? Do you need to classify *unseen* data in order to save valuable working time?

Question: What would an "ideal" algorithm do for *your* project?

Machine learning needs a hypothesis

Decide beforehand what you are looking for in your data. Do you hypothesise that a certain biomarker might exist in your data that is correlated with a phenotype? Do you hypothesise that there are subgroups of subjects of your study that give rise to similar behaviour?

Take-away: The results of a machine learning algorithm should be able to surprise you (because a hypothesis has been disproved).

Machine learning cannot extract something from nothing

If your data are very "noisy", the real signal might be hidden—and in the worst case, an algorithm might be unable to find it. Larger data sets might lead to better performance, but only if the data are properly curated.

Take-away: Be aware of inconsistencies in your data.

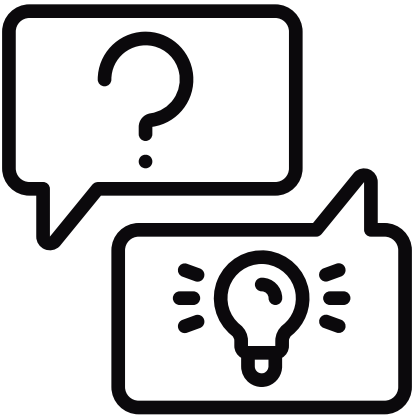
Your own projects

Group Work 3: Discussion of individual projects (45min)

Instructions:

- What does your data look like (time series, images, genetic data, ...)?
- How large is the data set (number of samples/features)?
- Are there any pitfalls (high noise, missing data, incorrect measurements, ...)?
- What would you like to learn? What is the question you want to answer?
- Are you dealing with a clustering/classification/regression problem? What is your target variable?
- Can your question be answered with your current data set? If not, how would the data set have to be changed to answer the question (larger data set? different experiments, ...)?

Question-Answer-Session



Machine Learning Resources

Online Resources:

- python's `scikit-learn` package plus documentation (hands-on ML techniques with examples)
- ML-lecture by A. Ng: https://www.coursera.org/learn/machine-learning
- `https://shapeofdata.wordpress.com`

Textbooks:

- *Elements of Statistical Learning* by T. Hastie, R. Tibshirani, J. Friedmann
- *Machine Learning: a Probabilistic Perspective* by K. Murphy
- *Understanding Machine Learning: From Theory to Algorithms* by S. Shalev-Shwartz, S. Ben-David
- *Data Mining and Analysis: Fundamental Concepts and Algorithms* by M.J. Zaki and W. Meira Jr.
- *Data Mining: The Textbook* by C.C. Aggarwal

…and of course: the Borgwardt-lab ;)