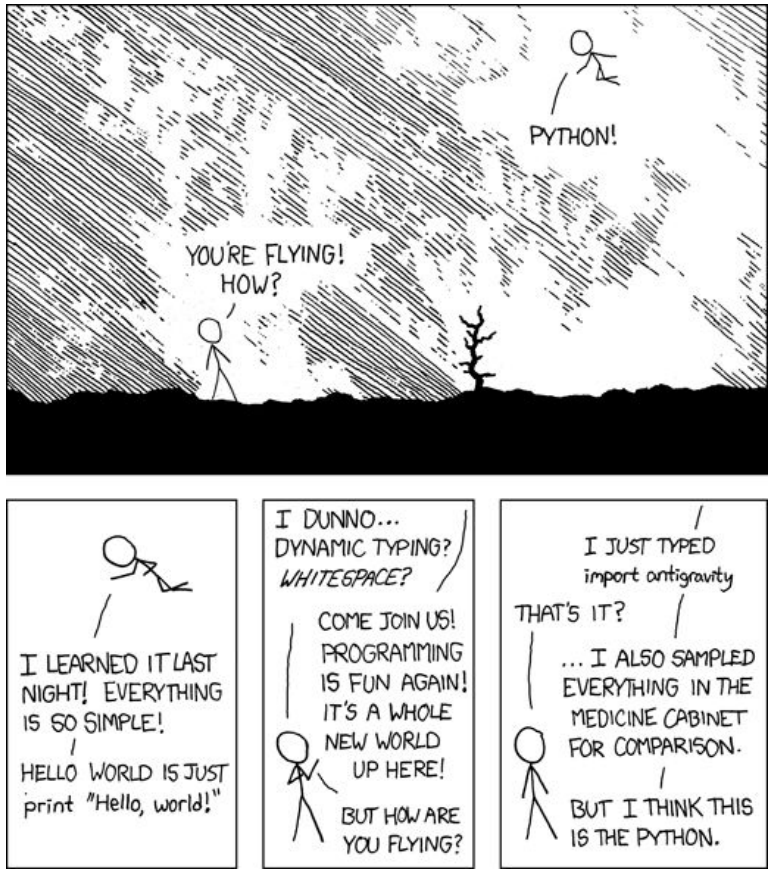
UMM
UNIVERSITÄTSMEDIZIN
MANNHEIM

Medizinische Fakultät Mannheim
der Universität Heidelberg

Universitätsklinikum Mannheim

Python for Big Data Analysis



Kim Hee
Graduate research assistant
at Heinrich-Lanz-Center (HLZ) for Digital Health

Tutorial organization

Time	Topic
09:00 - 09:30	Set the scene
09:30 - 10:45	Data analysis and Data visualization
10:45 - 11:15	Break
11:15 - 12:15	**Machine Learning**
12:15 - 13:00	Python at Scale (PySpark)

Agenda

- Machine Learning
- Decision tree and demo
- Clustering and demo

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

Arthur Samuel, 1959

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

Arthur Samuel, 1959

What is new in Machine Learning?

- Big data
  - A large volume of data
  - Distributed systems and databases
  - Distributed algorithms for machine learning
- Deep learning
  - Alexey Grigorevich Ivakhnenko published the first general, working learning algorithm for deep networks (Ivakhnenko and Lapa, 1965)
  - It is a learning method that involves more than one hidden layer of an Artificial Neural Networks (ANN)
  - It is a representation learning which is aspired to learn representations intrinsic in the data on its own, rather than manual feature engineering
  - Tools: Tensorflow, Keras, PyTorch and more (http://deeplearning.net/software_links)

Types of Machine Learning

- Supervised Learning
  - Learn **with** human guidance
  - It aims to predict a target value of unseen data from seen data by constructing a mathematical model from labeled data by means of correlating features in order to optimize an objective function
  - Examples: classification; regression
- Unsupervised Learning
  - Learn **without** human guidance
  - It infers a function to identify naturally occurring patterns (structure) in **unlabeled data**
  - Cluster algorithms have one or more hyper parameter that need to be set (e.g. the number of parameters k in k-means clustering); there are no known ways to set these to obtain optimal results (other than trial and error)
  - Examples: clustering; outlier detection

[Quiz] which of machine learning algorithms can be applied?

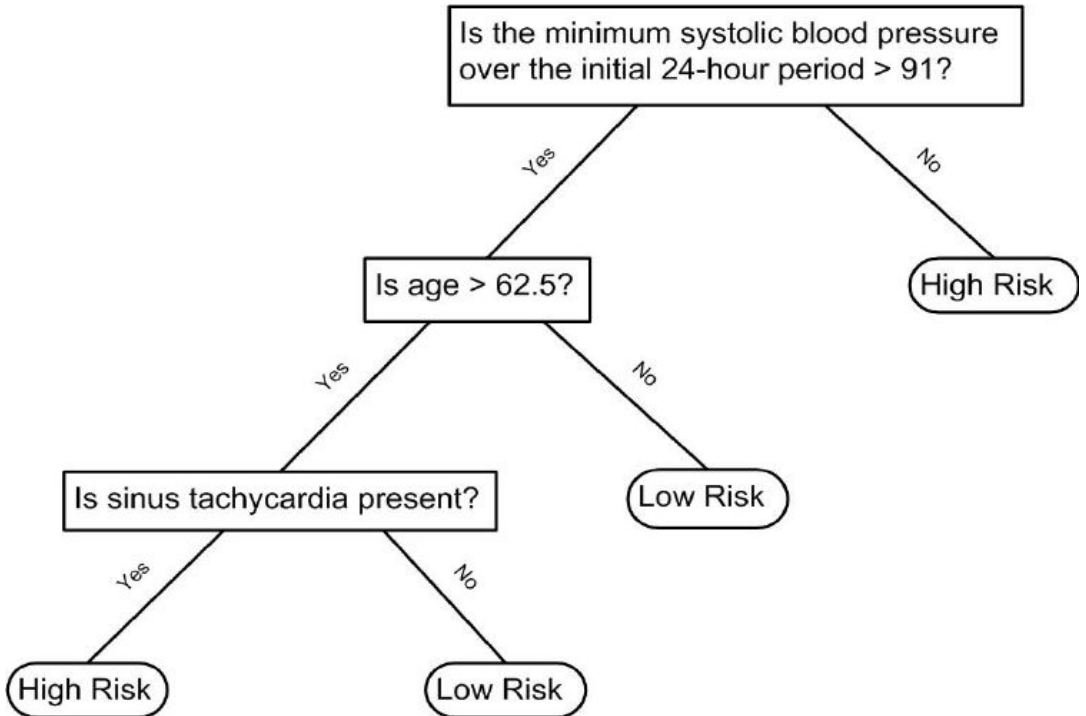
1. Type 2 diabetes **risk** forecasting
2. **Outlier detection** for patient monitoring and alerting
3. **Would this** patient readmission in 30 days?
4. **Segmentation** of brain tumor images
5. **How many** beds will be on demand today? / on a certain time?
6. **Mining** with rare disease cases

Agenda

- Machine Learning
- Decision tree and demo
- Clustering and demo

Supervised learning
Classification

- Decision Trees are a supervised learning method
- The model is built by learning decision rules inferred from the training data
- It is a tree of “Yes or No” questions, with questions of higher importance appearing earlier in the tree (e.g. feature importance is calculated based on information gain)
- Example tree that classifies the risk of sinus tachycardia



Supervised learning
Evaluation for classification algorithms

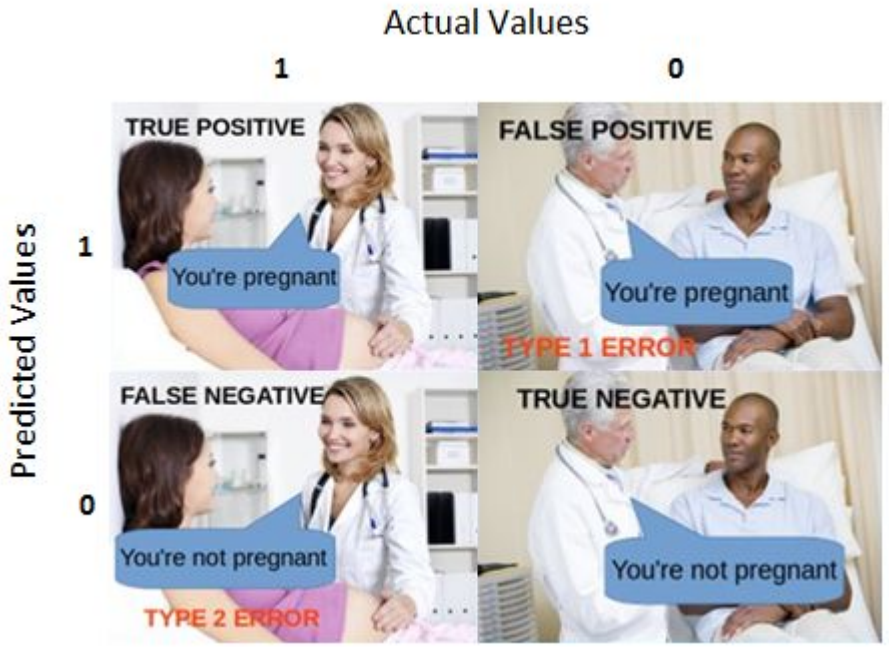
- A confusion matrix is a table that is used to describe the performance of a classification model (or “classifier”) on a set of test data.

		True condition	
		positive	negative
Predicted condition	positive	**True positive**	**False positive** Type I error
	negative	**False negative** Type II error	**True negative**

Precision

Recall/Sensitivity

Confusion matrix for a binary class problem (source: wikipedia)



[Demo] Decision Tree

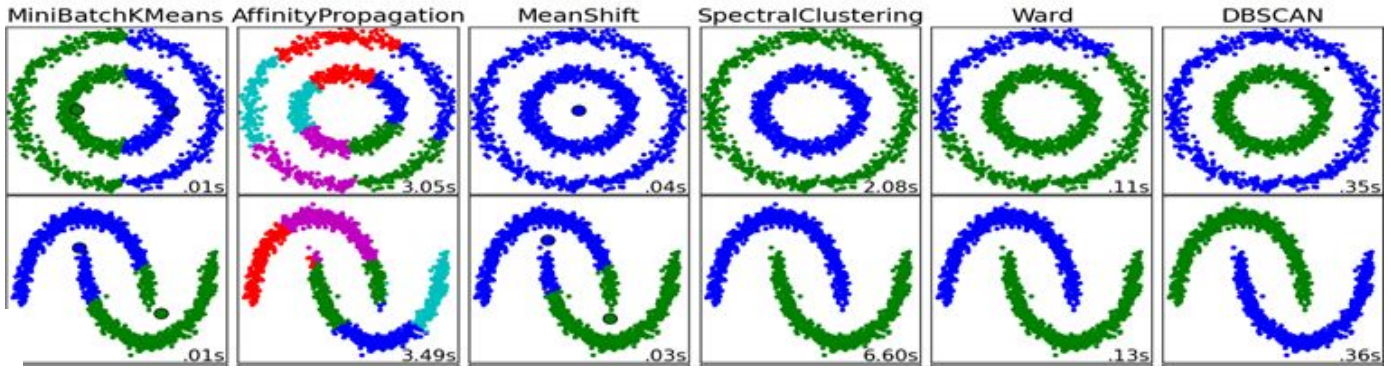
- Predict the onset of diabetes based on diagnostic measures

Agenda

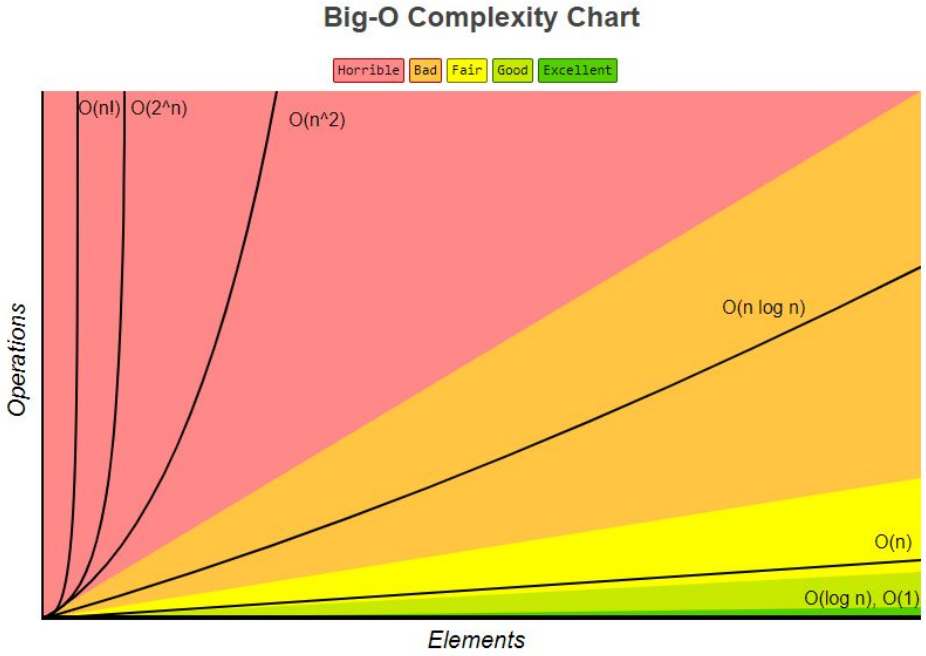
- Machine Learning
- Decision tree and demo
- Clustering and demo

Clustering

- Methods to identify possible groupings in a data-set Induced by the similarity function
- cluster is a set of data points that share a sub-set of the overall properties
- Clustering Algorithms
  - k-means clustering
  - Hierarchical clustering
  - Density based clustering
  - and more



ikit-learn documentation and it demonstrates how the different clustering algorithms worked on a same input data

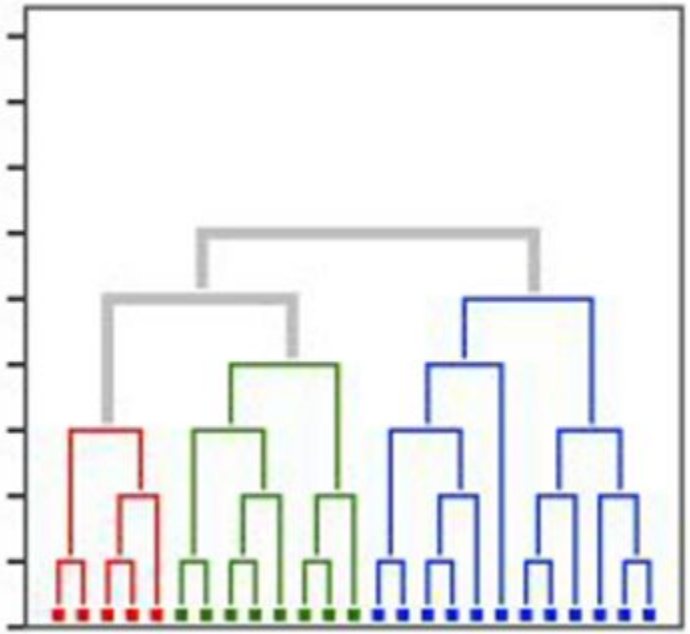


	KMeans	Hierarchical clustering	DBSCAN
time complexity	O(n^2)	O (n^3)	O (n^3)

- where n is the number of data points
- source: wikipedia

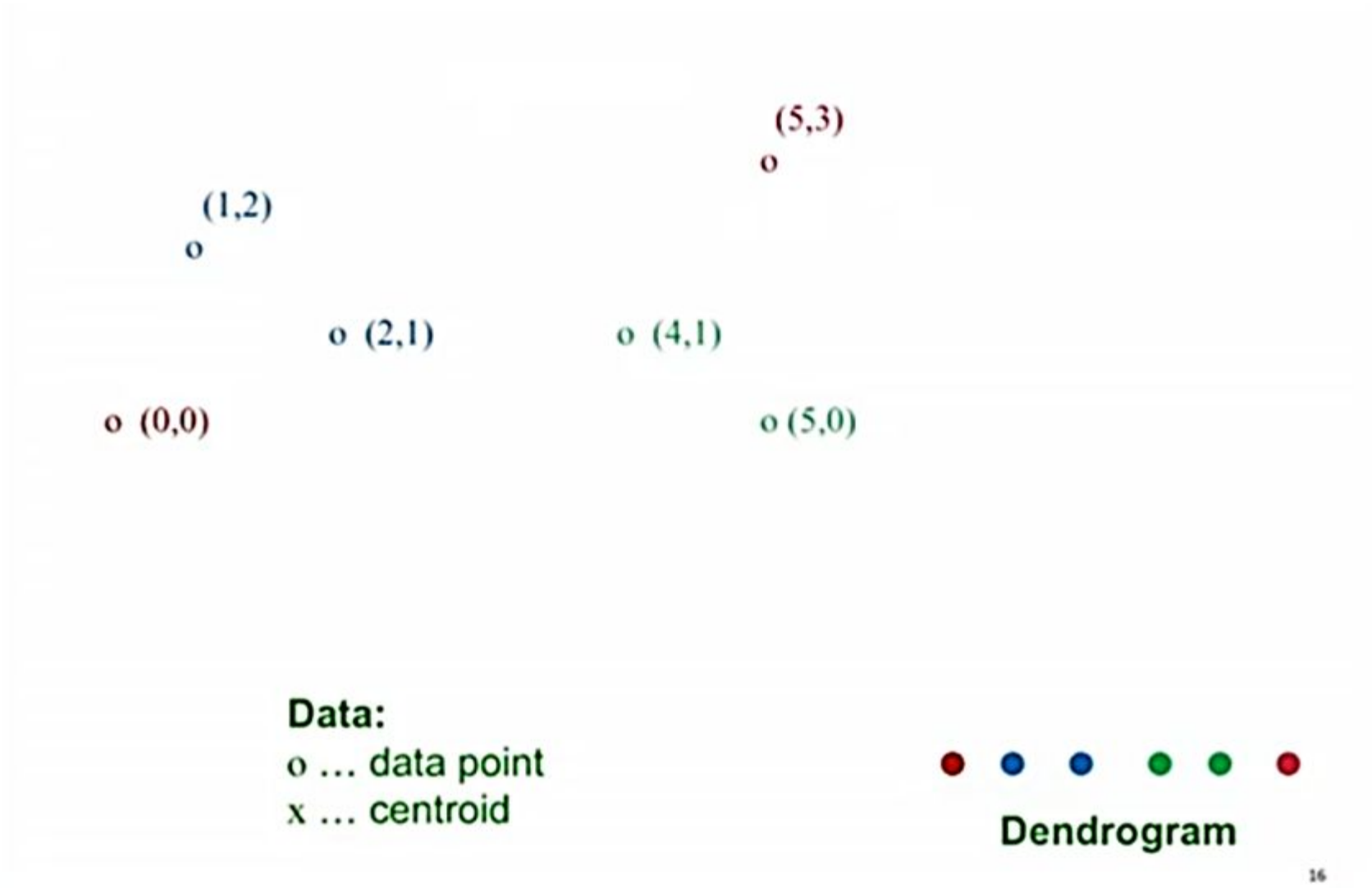
Hierarchical Clustering

- Easy to understand
- A dendrogram can be used to depict this hierarchical clustering process
- Bottom-up (=agglomerative) method: iterate over each data record
- Compare two data records using a similarity function, and join most similar pair



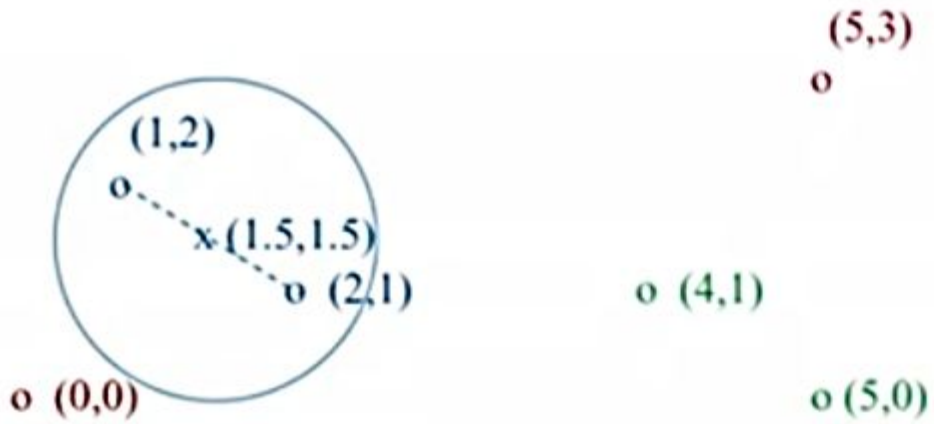
dendrogram

Hierarchical Clustering Demonstration (1/5)



Mining Massive Datasets - Stanford University

Hierarchical Clustering Demonstration (2/5)

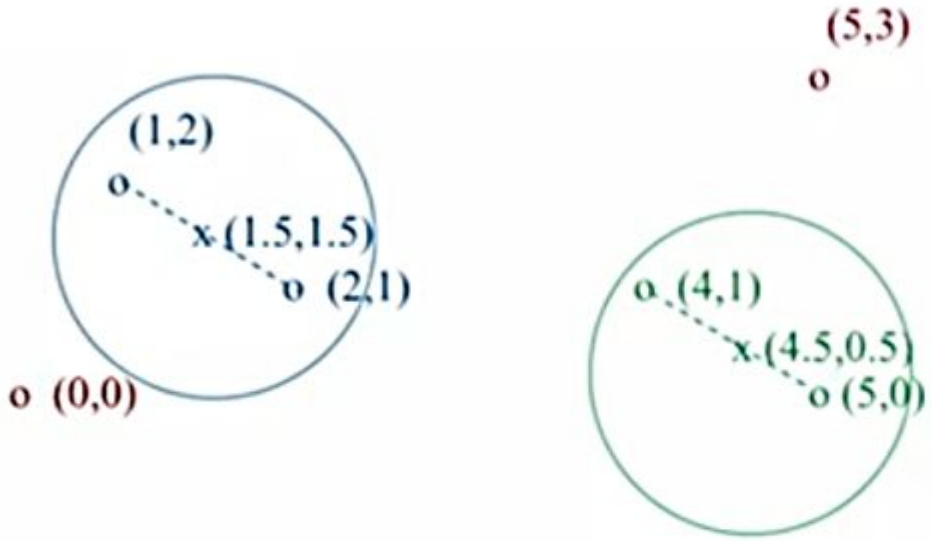


Data:
o … data point
x … centroid

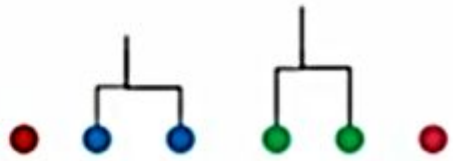
Dendrogram

16

Hierarchical Clustering Demonstration (3/5)

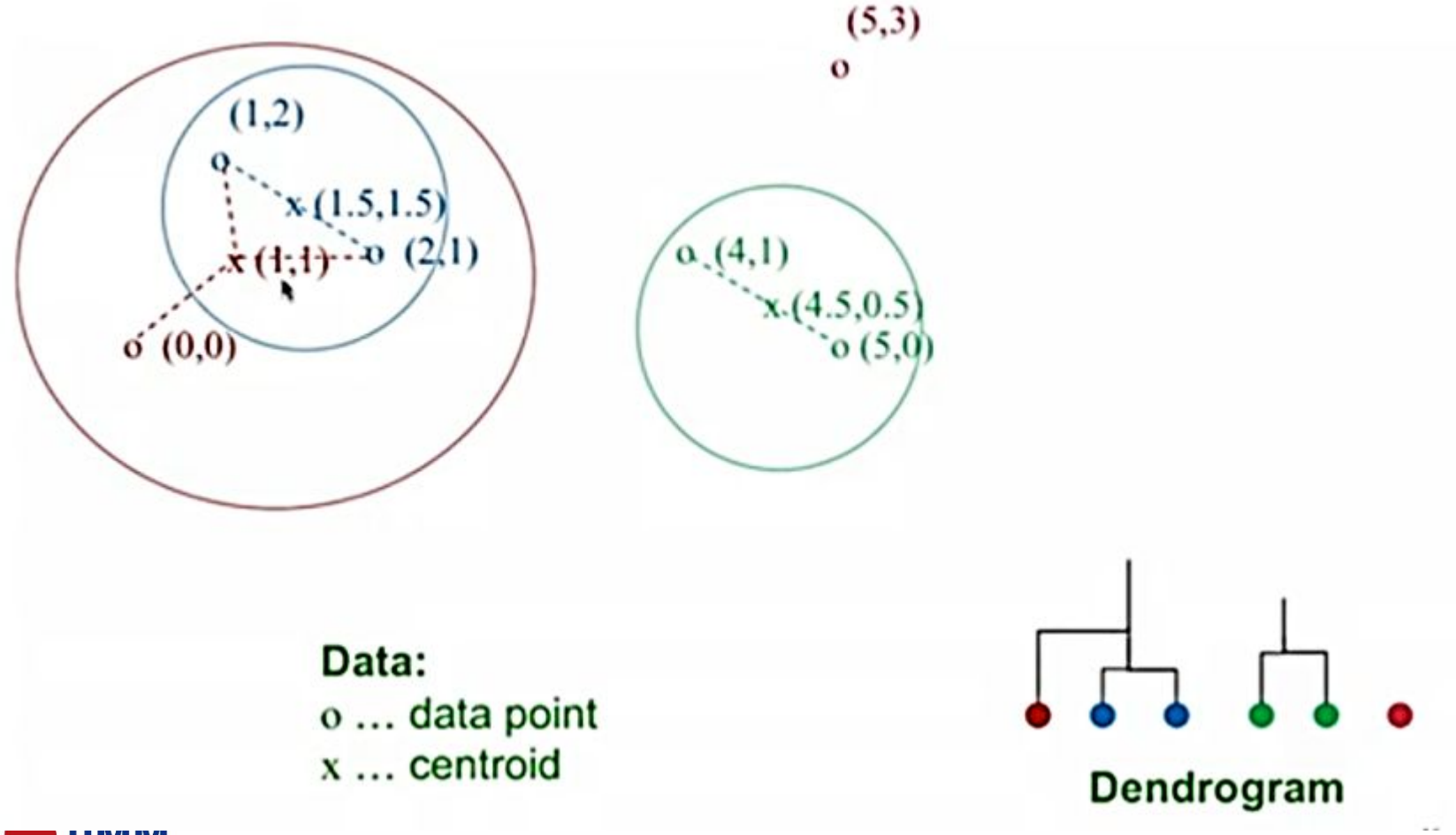


Data:
o … data point
x … centroid

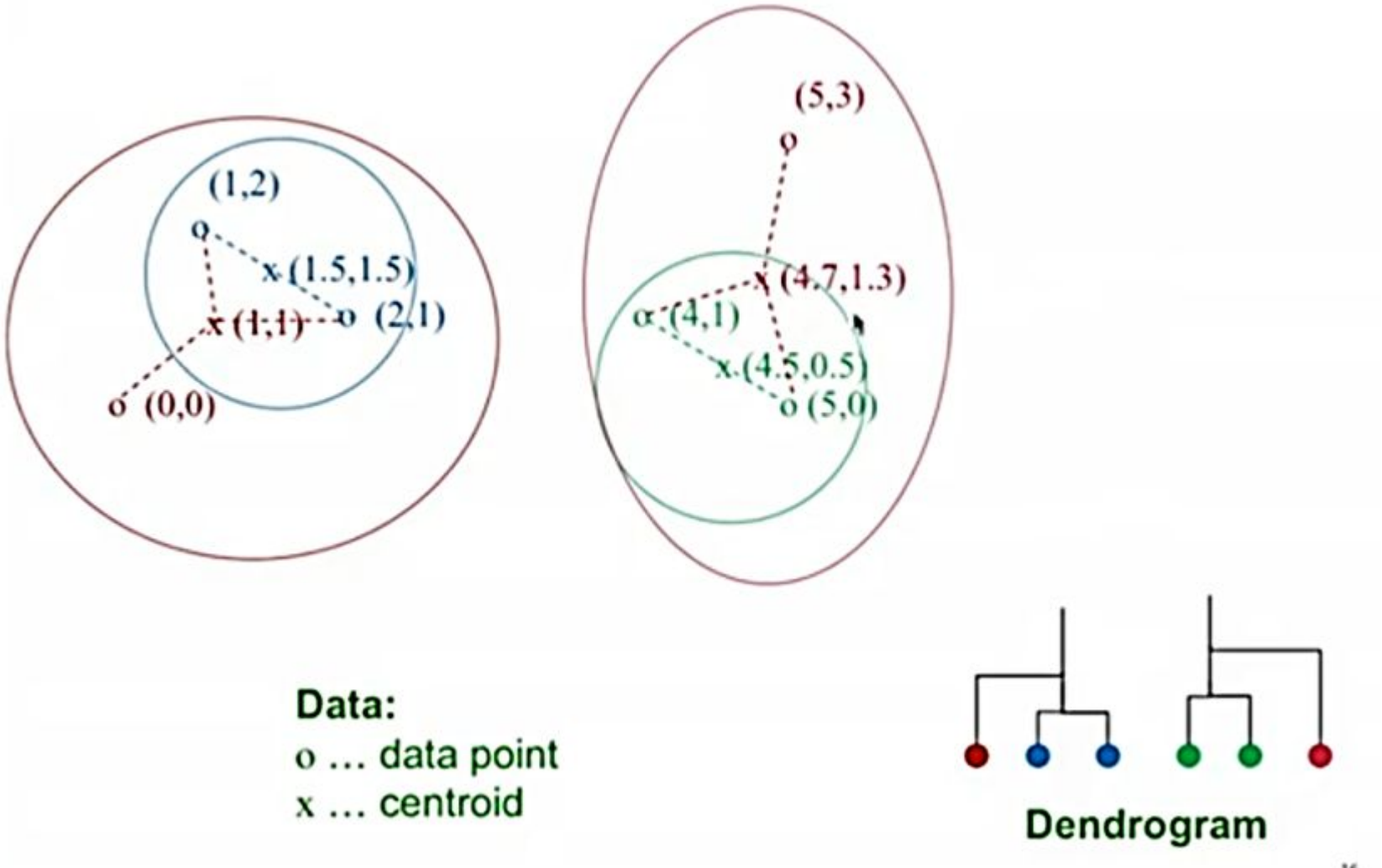


Dendrogram

Hierarchical Clustering Demonstration (4/5)



Hierarchical Clustering Demonstration (5/5)



k-means

- Simple and efficient; best known partitional clustering algorithm
- Maximize intracluster & minimize intercluster similarity
- **Sensitive to choice of seed clusters (initial centroids)**
- For every point in the dataset
  - Measure distance to each centroid
  - Assign point to centroid with lowest distance
- For every cluster
  - Calculate the mean position of all points
- Repeat until convergence criteria is met
  - No points change clusters
  - Centroid changes within a short distance
  - Fixed number of iterations

[Demo][Hands-on] Hierarchical Clustering

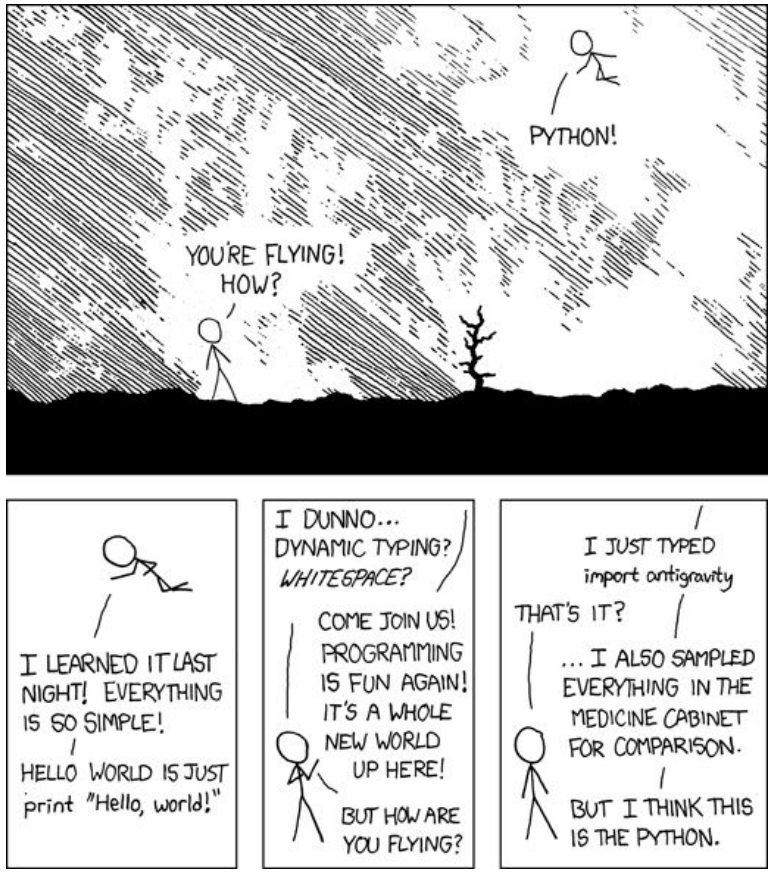
- Taxonomic classification of hemoglobin from different species

UMM
UNIVERSITÄTSMEDIZIN
MANNHEIM

Medizinische Fakultät Mannheim
der Universität Heidelberg

Universitätsklinikum Mannheim

THANK YOU



Kim Hee
Graduate research assistant
at Heinrich-Lanz-Center (HLZ) for Digital Health