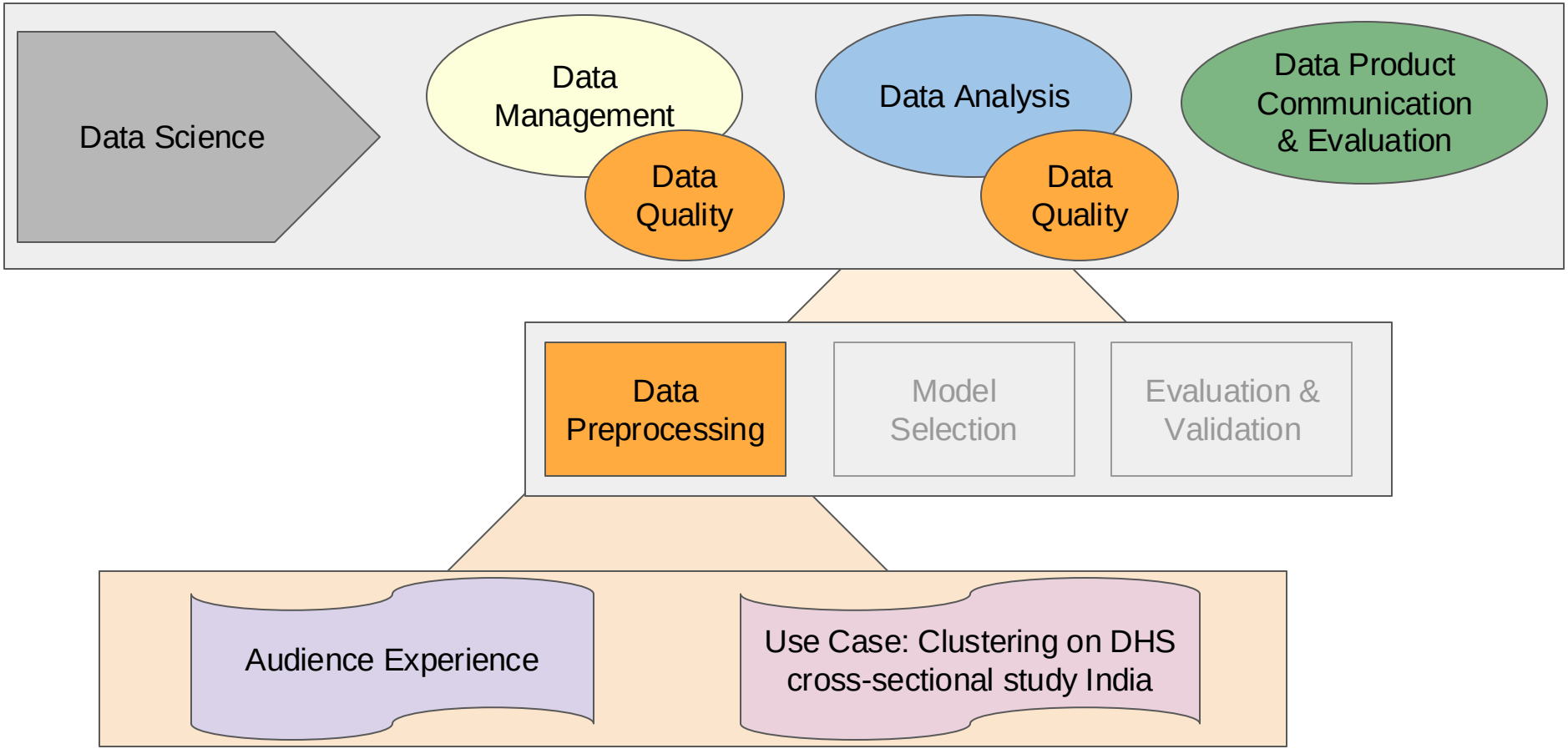
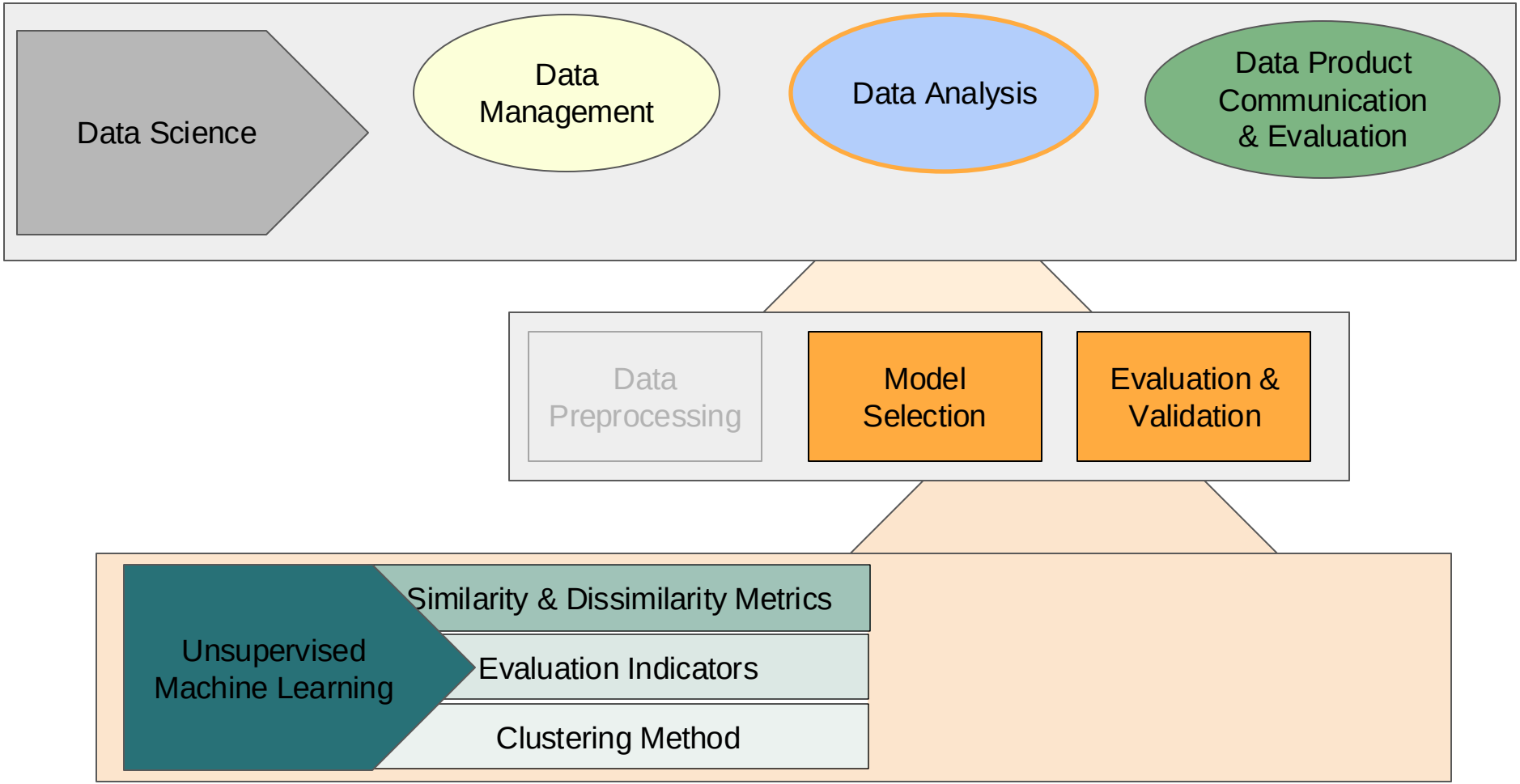
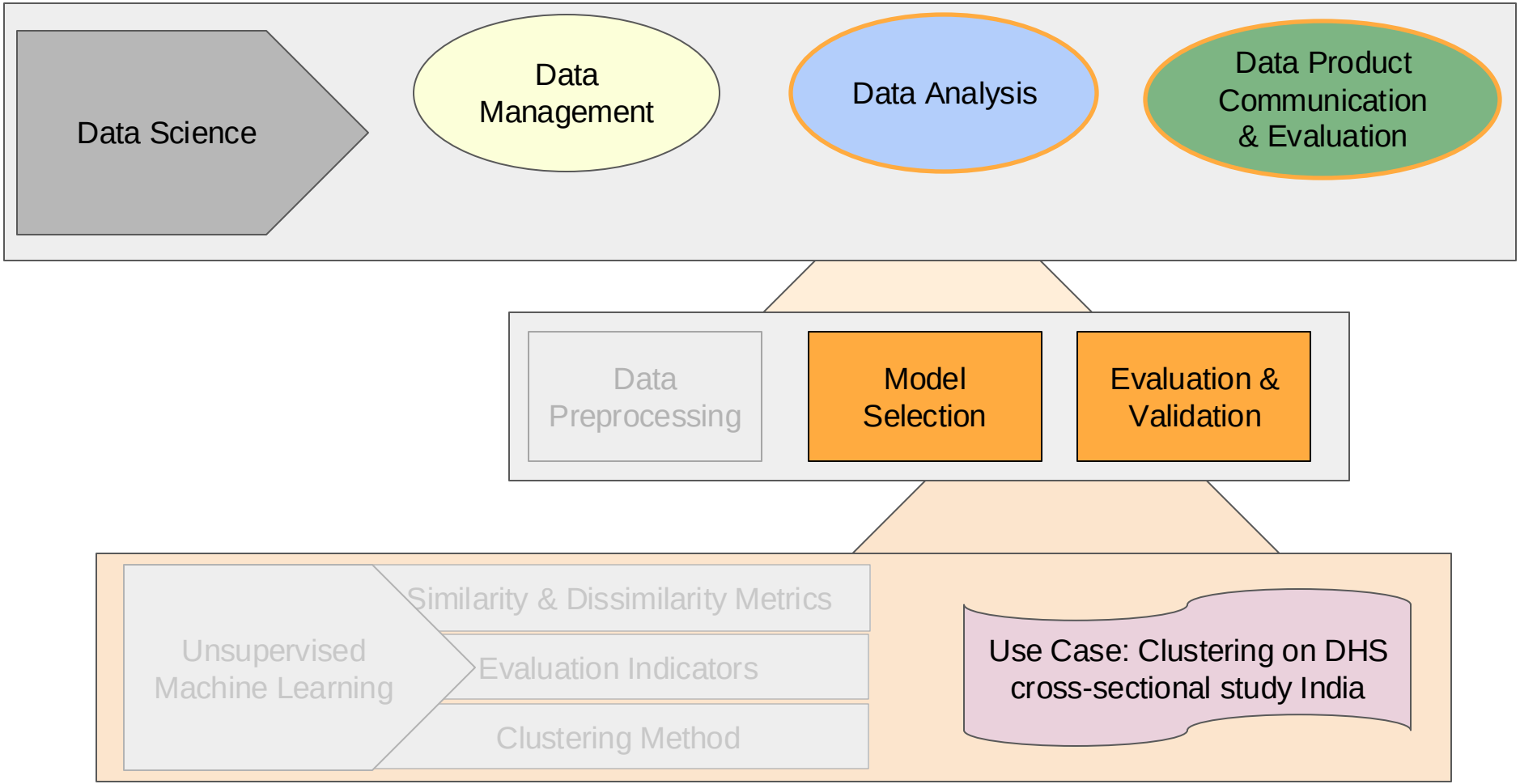
Introduction to unsupervised Machine Learning

Anna-Katharina Nitschke
& Carlos Brandl, Carola Behr, Fabian Egersdörfer and Prof. Matthias Weidemüller
Physikalisches Institut, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 226, 69120 Heidelberg, Germany

in collaboration with Prof. Till Bärnighausen
Global Health Institut Heidelberg


Data Science
Data Management
Data Quality
Data Analysis
Data Quality
Data Product Communication & Evaluation
Data Preprocessing
Model Selection
Evaluation & Validation
Audience Experience
Use Case: Clustering on DHS cross-sectional study India


Data Science
Data Management
Data Analysis
Data Product Communication & Evaluation
Data Preprocessing
Model Selection
Evaluation & Validation
Unsupervised Machine Learning
Similarity & Dissimilarity Metrics
Evaluation Indicators
Clustering Method


Data Science
Data Management
Data Analysis
Data Product Communication & Evaluation
Data Preprocessing
Model Selection
Evaluation & Validation
Unsupervised Machine Learning
Similarity & Dissimilarity Metrics
Evaluation Indicators
Clustering Method
Use Case: Clustering on DHS cross-sectional study India

Model Selection – Data Specification

Scales of Mathematical Modeling in Health Care

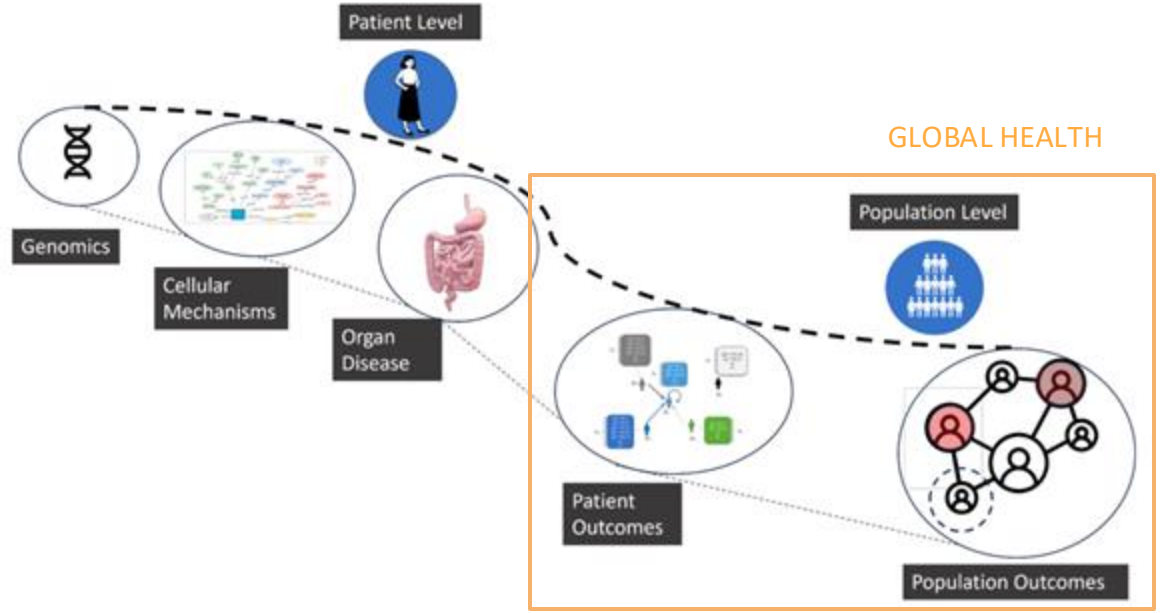


Figure 6 - Graph models provide holistic representations, for which the scale of the system to be mathematically modeled can range from fine to superficial. Modeling scales can include information levels like genomics, cellular mechanisms, organ disease, patient outcomes and population outcomes. The topology of the graph model for cellular mechanisms (sub graph left) has been extracted from Theodopolous et al. (Theodoropoulos et al. 2023).

Nitschke et al.; 2024; preprint From Knowledge Graphs to Digital Twins: Perspectives on Modeling Patient Outcomes for Healthcare Quality Assessment

Model Selection – Research Question Specification

The distinctions highlight different phases and goals in data analysis, helping guide the choice of techniques depending on whether the focus is on understanding the data, confirming a hypothesis, or predicting future outcomes:

Exploratory Analysis vs. Explanatory Analysis → reflects the stage and purpose of **understanding and theory testing**

1. **Exploratory Analysis**:

- no predefined models or hypotheses
- The goal is to explore the data, uncover patterns, and understand its general structure without any initial assumptions.
- Techniques like clustering, principal component analysis, multidimensional scaling

2. **Explanatory Analysis**:

- test specific hypotheses or theories
- The goal is to understand the relationships between variables and validating the theoretical constructs
- Techniques like Regression analysis, structural equation modeling, causal inference methods.

Anil K. Jain; 2009; Data clustering: 50 years beyond K-means; Pattern Recognition Letters

Model Selection – Model Aim Specification

The distinctions highlight different phases and goals in data analysis, helping guide the choice of techniques depending on whether the focus is on understanding the data, confirming a hypothesis, or predicting future outcomes:

Confirmatory Modeling vs. Predictive Modeling → reflects the goal of hypothesis validation versus forecasting

1. **Predictive Modeling**:

- The focus is on using the data to make predictions about new or future observations (i.e. anticipating future outcomes based on patterns learned from the data).
- The goal is accurate prediction rather than testing a theoretical hypothesis.
- Techniques like Machine learning models (e.g., decision trees, neural networks), time series forecasting, predictive analytics.

2. **Confirmatory (or Causal) Modeling**:

- To confirm the validity of a hypothesis or model. The investigator has a model or set of assumptions that they want to validate using the data.
- The goal is to assess the accuracy of a model or hypothesis given existing data
- Techniques like linear regression, ANOVA, confirmatory factor analysis

Anil K. Jain; 2009; Data clustering: 50 years beyond K-means; Pattern Recognition Letters

Pattern Recognition

„Pattern recognition is the process of identifying and classifying patterns or regularities in data.“

→ part of predictive modeling: Given some training data, we want to predict the behavior of the unseen test data. This task is also referred to as learning.

Learning problems are:

CLUSTERING

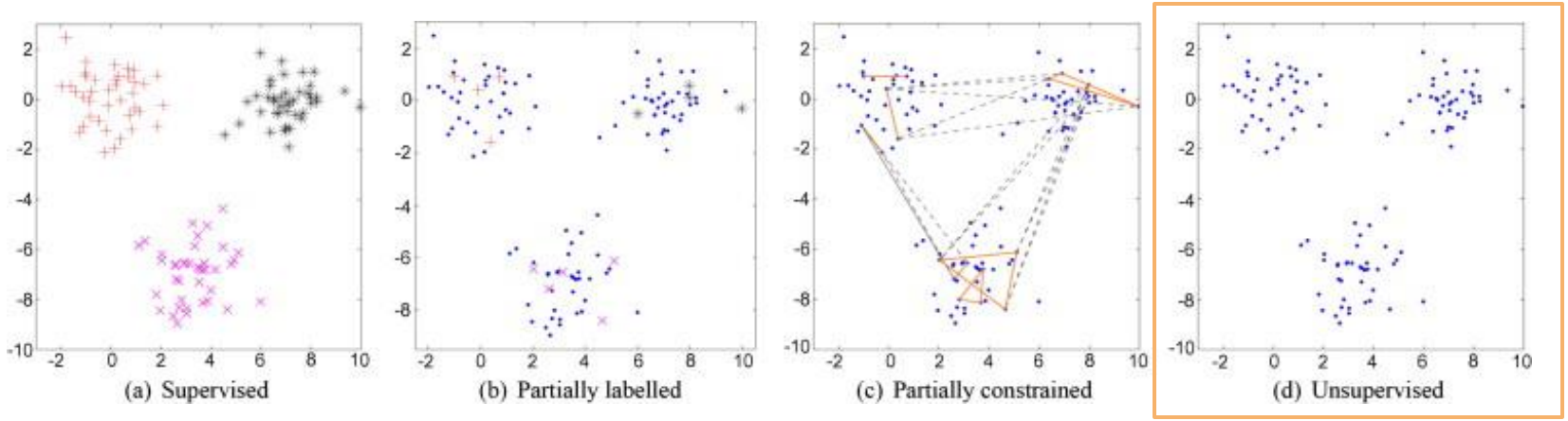


Fig. 1. Learning problems: dots correspond to points without any labels. Points with labels are denoted by plus signs, asterisks, and crosses. In (c), the must-link and cannot-link constraints are denoted by solid and dashed lines, respectively (figure taken from Lange et al. (2005).

Anil K. Jain; 2009; Data clustering: 50 years beyond K-means; Pattern Recognition Letters

Clustering

“Clustering is a learning task that aims to decompose a given set of observations into subgroups (clusters) based on data similarity, such that observations in the same cluster are more closely related to each other than observations in different clusters. It is an unsupervised learning task, since it identifies structures in unlabeled datasets, and a classification task, since it can give a label to observations according to the cluster they are assigned to.“

Calvin J. Chiew, 2020; *Leveraging Data Science for Global Health;* Chapter 8: Applied Statistical Learning in Python https://link.springer.com/book/10.1007/978-3-030-47994-7

Application of clustering

Data clustering has been used for the following three main purposes:

1. **Underlying structure:** to gain insight into data, generate hypotheses, detect anomalies
2. **Natural classification:** to identify the degree of similarity among forms or organisms (phylogenetic relationship).
3. **Compression:** as a method for organizing the data and summarizing it through cluster prototypes.

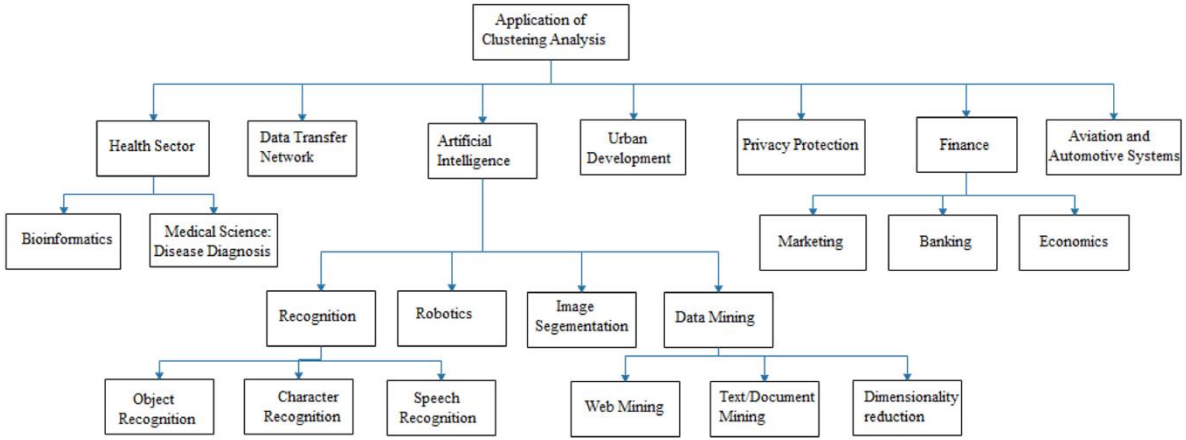
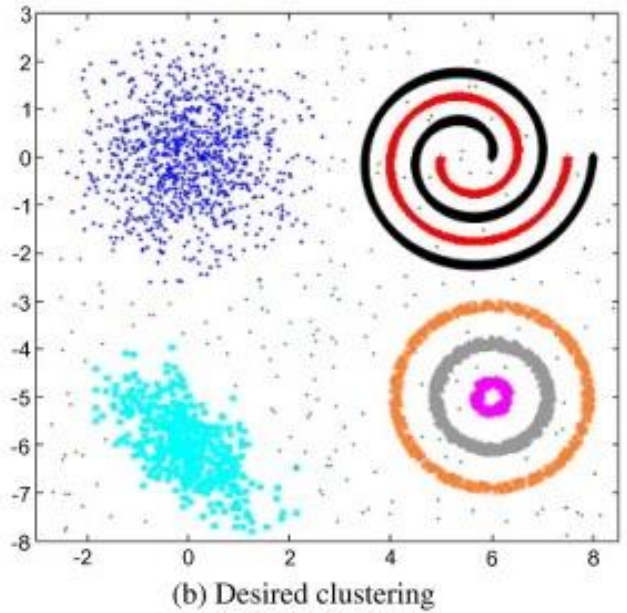


Fig. 10. Taxonomy of clustering algorithm application.

Absalom E. Ezugwu et al.; 2022; A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects https://doi.org/10.1016/j.engappai.2022.104743

Calvin J. Chiew, 2020; *Leveraging Data Science for Global Health;* Chapter 8: Applied Statistical Learning in Python https://link.springer.com/book/10.1007/978-3-030-47994-7

Diversity of Clusters



(b) Desired clustering

An ideal cluster :
„ A cluster is a set of points that is compact and isolated."

In reality:
„A cluster is a subjective entity that is in the eye of the beholder and whose significance and interpretation requires domain knowledge.“

But, while humans are excellent cluster seekers in two and possibly three dimensions, we need automatic algorithms for high-dimensional data.

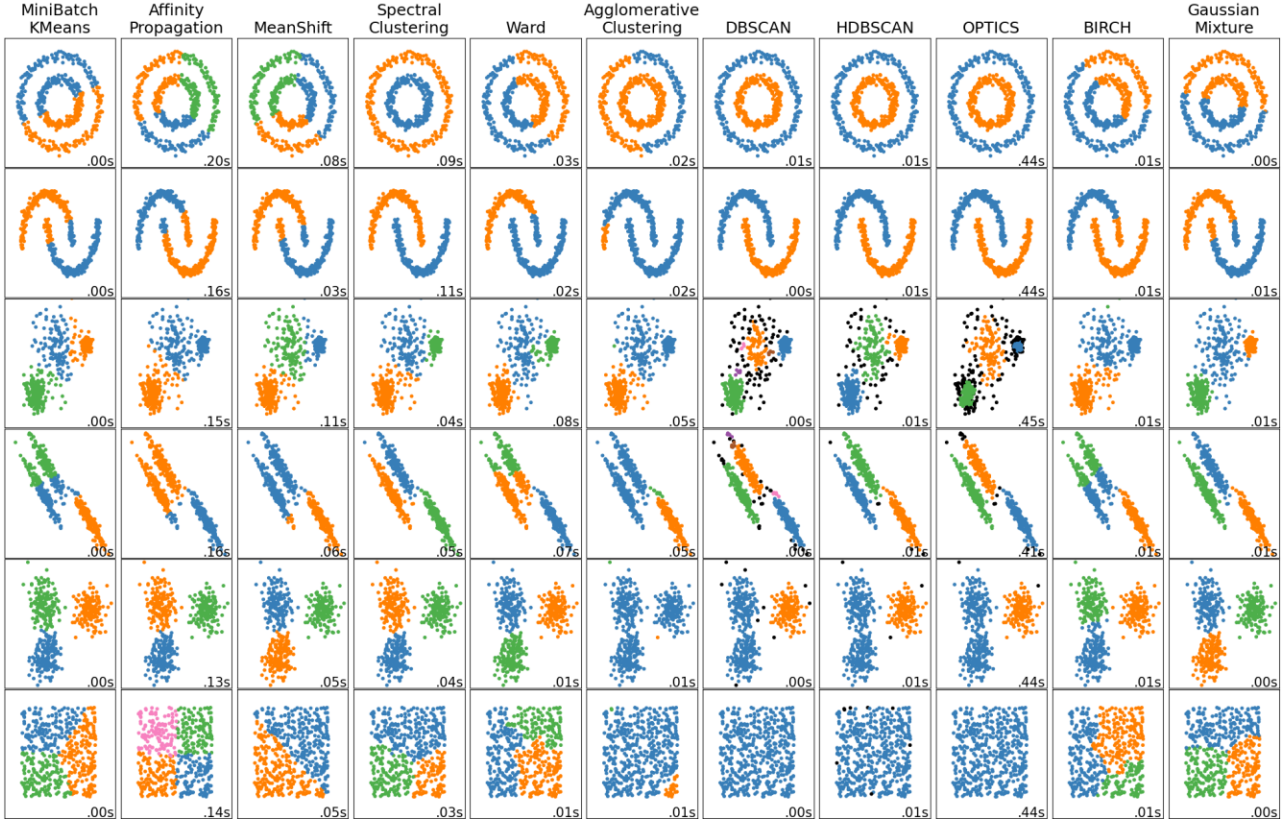
Challenge one: Although these clusters are apparent to a data analyst, none of the available clustering algorithms can detect all these clusters. → importance of method selection

Challenge two: Number of clusters not known

- seven clusters differ in terms of their shape, size, and density.
- presence of noise in the data makes the detection of the clusters even more difficult.

Anil K. Jain; 2009; Data clustering: 50 years beyond K-means; Pattern Recognition Letters

Performance of Several Clustering Algorithms



https://scikit-learn.org/stable/modules/clustering.html

Similarity Metrics for Clustering

- **similarity** is fundamental to the definition of a cluster and measures the similarity between two instances drawn from the same feature space
  → qualitative features
- ***dissimilarity*** (distance measure) = opposite measure to similarity, which is often calculated instead
  → quantitative features
- Metrics need to be chosen carefully, depending on:
  - Feature types and heterogeneity
  - Feature scales (normalization)
  - Feature correlations (Mahalanobis distance)
  - effect of surrounding or neighboring points = ***context***
  - additional domain information = **concept**

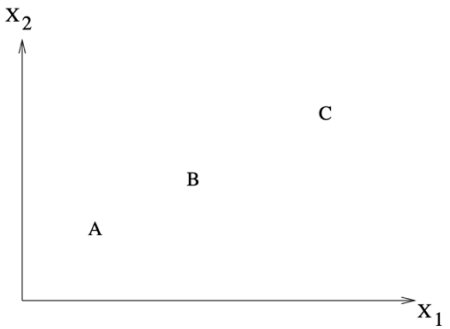


Figure 4. A and B are more similar than A and C.

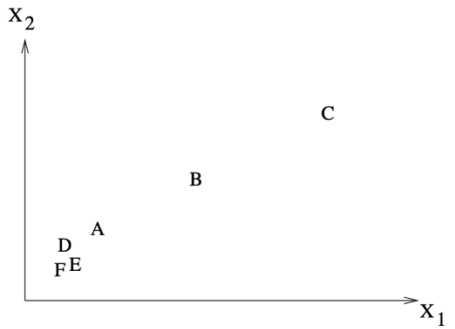


Figure 5. After a change in context, B and C are more similar than B and A.

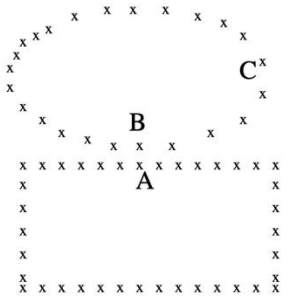


Figure 6. Conceptual similarity between points .

A.K. Jain et al.; 1999; **Data Clustering: A Review;** https://dl.acm.org/doi/abs/10.1145/331499.331504

Similarity Metrics for Clustering

- **similarity** is fundamental to the definition of a cluster and measures the similarity between two instances drawn from the same feature space
  → qualitative features
- ***dissimilarity*** (distance measure) = opposite measure to similarity, which is often calculated instead
  → quantitative features

The most popular metric for continuous features is the **Euclidean distance**

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2\right)^{1/2}$$

$$= \|\mathbf{x}_i - \mathbf{x}_j\|_2,$$

which is a special case of the **Minkowski metric:**

$$d_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{d} |x_{i,k} - x_{j,k}|^p\right)^{1/p}$$

$$= \|\mathbf{x}_i - \mathbf{x}_j\|_p.$$

A.K. Jain et al.; 1999; **Data Clustering: A Review;** https://dl.acm.org/doi/abs/10.1145/331499.331504

Similarity Metrics for Clustering

- **similarity** is fundamental to the definition of a cluster and measures the similarity between two instances drawn from the same feature space
  → qualitative features
- ***dissimilarity*** (distance measure) = opposite measure to similarity, which is often calculated instead
  → quantitative features

Xu and Wunsch II 2005; Murtagh and Contreras 2012, 2017

Properties of distance/ dissimilarity metric:

(1) Symmetric, $D(x_i, x_j) = D(x_j, x_i)$;

(2) Positivity, $D(x_i, x_j) \geq 0$ for all x_i and x_j;

(3) Triangle inequality, $D(x_i, x_j) \leq D(x_i, x_k) + D(x_k, x_j)$ for all x_i, x_k and x_j;

(4) $D(x_i, x_j) = 0$ iff $x_i = x_j$.

A.K. Jain et al.; 1999; **Data Clustering: A Review;** https://dl.acm.org/doi/abs/10.1145/331499.331504

Xingcheng Ran, Yue Xi, Yonggang Lu, Xiangwen Wang & Zhenyu Lu ; 2023; **Comprehensive survey on hierarchical clustering algorithms and the recent developments; https://link.springer.com/article/10.1007/s10462-022-10366-3**

Similarity Metrics for Clustering

- **similarity** is fundamental to the definition of a cluster and measures the similarity between two instances drawn from the same feature space
  → qualitative features
- ***dissimilarity*** (distance measure) = opposite measure to similarity, which is often calculated instead
  → quantitative features

Xu and Wunsch II 2005; Murtagh and Contreras 2012, 2017

Properties of similarity metrics:

(1) Symmetric, $S(x_i, x_j) = S(x_j, x_i)$;

(2) Positivity, $0 \leq S(x_i, x_j) \leq 1$ for all x_i and x_j;

(3) $S(x_i, x_j)S(x_j, x_k) \leq [S(x_i, x_j) + S(x_j, x_k)]S(x_i, x_k)$ for all x_i, x_k and x_j;

(4) $S(x_i, x_j) = 1$ iff $x_i = x_j$.

A.K. Jain et al.; 1999; **Data Clustering: A Review;** https://dl.acm.org/doi/abs/10.1145/331499.331504

Xingcheng Ran, Yue Xi, Yonggang Lu, Xiangwen Wang & Zhenyu Lu ; 2023; **Comprehensive survey on hierarchical clustering algorithms and the recent developments; https://link.springer.com/article/10.1007/s10462-022-10366-3**

Evaluation Indicators for Clustering

- Purpose: test the validity of the clustering algorithm
- two categories:
  a) **internal evaluation indicators** (measure performance upon data set) → Table 3

Table 3 Evaluation indicators

Name	Formula or measure method	Explanation
Davies–Bouldin indicator	$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right)$	K stands for the number of clusters, C_x is the center of cluster x, σ_x is the average distance between any data in cluster x and C_x, $d(c_i, c_j)$ is the distance between c_i and c_j
Dunn indicator	$D = \min_{1\le i\le n}\left\{\min_{1\le j\le n, i\neq j}\left\{\frac{d(i,j)}{\max_{1\le k\le n} d'(k)}\right\}\right\}$	1. Mainly for the data that has even density and distribution 2. $d(c_i, c_j)$ is the distance between c_i and c_j, $d'(k)$ stands for the distance in cluster k
Silhouette coefficient	Evaluate the clustering result based on the average distance between a data point and other data points in the same cluster and average distance among different clusters	

An ideal cluster :

„ A cluster is a set of points that is compact and isolated."

Properties that we want to evaluate for measuring clustering performance:

(1) Minimal distances within one cluster
= compactness

(2) Maximal distance across clusters
= separation

Dongkuan Xu1,2 · Yingjie Tian; 2015; **A Comprehensive Survey of Clustering Algorithms**;
https://link.springer.com/article/10.1007/s40745-015-0040-1?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

Evaluation Indicators for Clustering

- Purpose: test the validity of the clustering algorithm
- two categories:
  a) **internal evaluation indicators** (measure performance upon data set) → Table 3
  b) **external evaluation indicators** (compare against externally given clustering; labels) → Table 4
  = measuring accordance to other clustering result

Table 4 Evaluation indicators

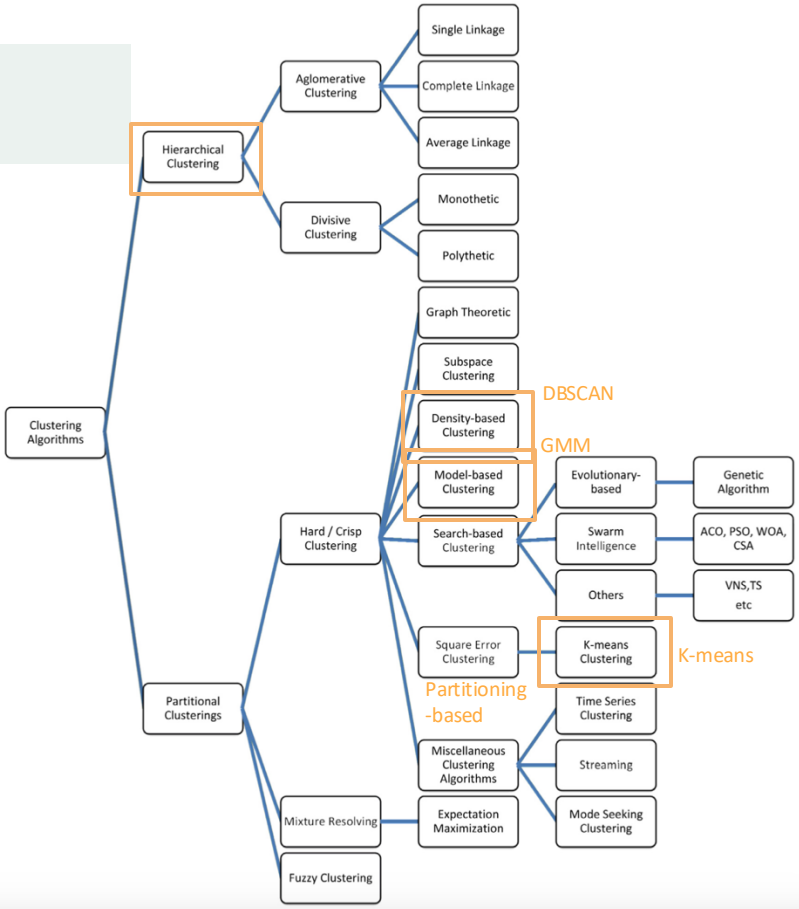
Name	Formula or measure method	Explanation
Rand indicator	$RI = \frac{TP+TN}{TP+FP+FN+TN}$	1. TP is the number of true positives 2. TN is the number of true negatives 3. FP is the number of false positives 4. FN is the number of false negatives
F indicator	$F_\beta = \frac{\left(\beta^2+1\right)\cdot P\cdot R}{\beta^2\cdot P+R}$	1. $P = \frac{TP}{TP+FP}$ stands for the accuracy, $R = \frac{TP}{TP+FN}$ stands for the recall rate 2. TP, TN, FP, and FN are defined as before
Jaccard indicator	$J(A,B) = \frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert} = \frac{TP}{TP+FP+FN}$	1. Measure the similarity of two sets 2. |X| Stands for the number of elements of set X 3. TP, TN, FP, and FN are defined as before
Fowlkes–Mallows indicator	$FM = \sqrt{\frac{TP}{TP+FP}\cdot\frac{TP}{TP+FN}}$	TP, TN, FP, and FN are defined as before
Mutual information	To measure, based on information theory, how much information is shared by two clusters, between which the nonlinear correlation can be detected	
Confusion matrix	To figure out the difference between a cluster and a gold-standard cluster	

Dongkuan Xu1,2 · Yingjie Tian; 2015; **A Comprehensive Survey of Clustering Algorithms**;
https://link.springer.com/article/10.1007/s40745-015-0040-1?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

Clustering Algorithms

Traditional algorithms

Category	Typical algorithm
Clustering algorithm based on partition	K-means, K-medoids, PAM, CLARA, CLARANS
Clustering algorithm based on hierarchy	BIRCH, CURE, ROCK, Chameleon
Clustering algorithm based on fuzzy theory	FCM, FCS, MM
Clustering algorithm based on distribution	DBCLASD, GMM
Clustering algorithm based on density	DBSCAN, OPTICS, Mean-shift
Clustering algorithm based on graph theory	CLICK, MST
Clustering algorithm based on grid	STING, CLIQUE
Clustering algorithm based on fractal theory	FC
Clustering algorithm based on model	COBWEB, GMM, SOM, ART



Dongkuan Xu1,2 · Yingjie Tian; 2015; **A Comprehensive Survey of Clustering Algorithms**; https://link.springer.com/article/10.1007/s40745-015-0040-1?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

Absalom E. Ezugwu et al.; 2022; A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects https://doi.org/10.1016/j.engappai.2022.104743

Clustering Algorithms

1. **hierarchical clustering**:
   „… partitions data objects into clusters in a hierarchical form through merging (agglomerative approach) or splitting (divisive approach) data objects. The corresponding cluster hierarchy generated as output is called dendrogram."
   a. bottom-up approach (agglomerative method): individual data objects are merged iteratively based on their similarity
   b. top-down approach (divisive method): the initial dataset is taken as a single cluster and broken down iteratively using data object similarity until each data object forms a single cluster or a set criterion is met.

2. **partitional clustering approach**:
   „… identifies a single partition of the initial dataset, through a heuristic approach that is optimizing a criterion function defined globally on all the data objects in the set or locally on the subset of the data objects.“
   a. partitioning-based: **K-Means**
   b. density-based: **DBSCAN**
   c. distribution-based/ model-based: **GMM**

Abiodun M. Ikotun et al.; 2022K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data
https://www.sciencedirect.com/science/article/pii/S0020025522014633

Clustering Algorithms based on Hierarchy

- Goal = construct the hierarchical relationship among data in order to cluster
- Suppose that each data point stands for an individual cluster in the beginning, and then, the most neighboring two clusters are merged into a new cluster until there is only one cluster left. Or reverse…
- Advantages: suitable for arbitrary data shape and type, relatively high scalability
- Disadvantages: high in time complexity, the number of clusters needs to be preset

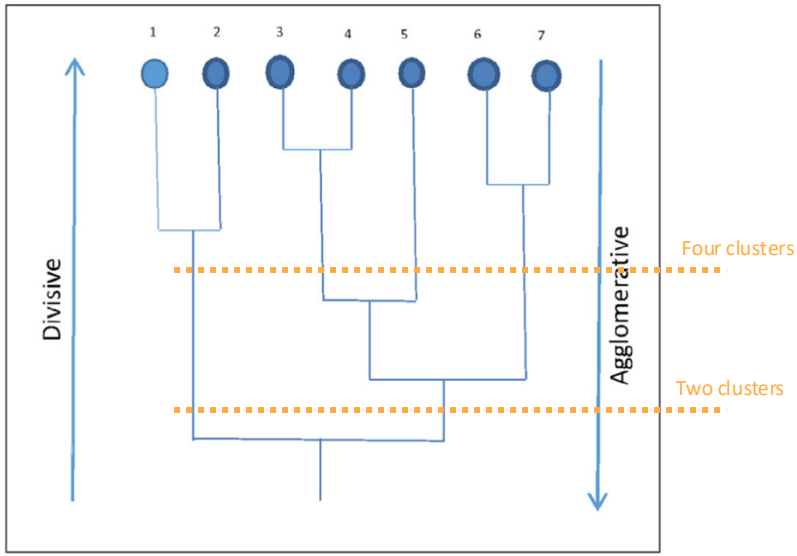
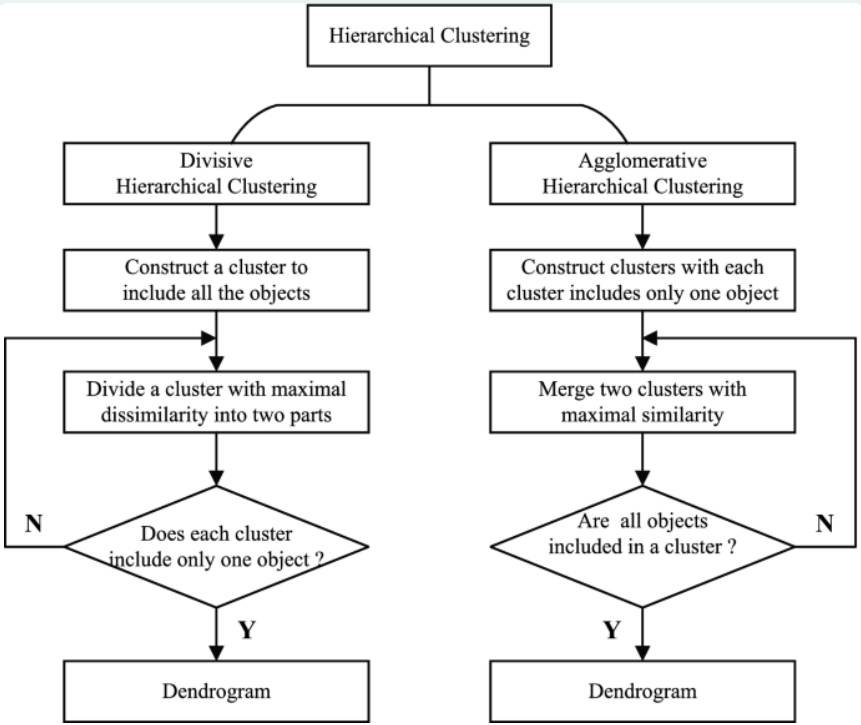


Fig. 2. A dendrogram representation for hierarchical clustering of data objects 1, 2, 3, 4, 5, 6, 7..

Dongkuan Xu1,2 · Yingjie Tian; 2015; **A Comprehensive Survey of Clustering Algorithms**;
https://link.springer.com/article/10.1007/s40745-015-0040-1?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

Clustering Algorithms based on Hierarchy

- Goal = construct the hierarchical relationship among data in order to cluster
- Suppose that each data point stands for an individual cluster in the beginning, and then, the most neighboring two clusters are merged into a new cluster until there is only one cluster left. Or reverse…
- Advantages: suitable for arbitrary data shape and type, relatively high scalability
- Disadvantages: high in time complexity, the number of clusters needs to be preset



Dongkuan Xu1,2 · Yingjie Tian; 2015; **A Comprehensive Survey of Clustering Algorithms**; https://link.springer.com/article/10.1007/s40745-015-0040-1?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

Xingcheng Ran, Yue Xi, Yonggang Lu, Xiangwen Wang & Zhenyu Lu ; 2023; **Comprehensive survey on hierarchical clustering algorithms and the recent developments; https://link.springer.com/article/10.1007/s10462-022-10366-3**

Clustering Algorithms based on Partitioning – K-Means

- Goal = regard the center of data points as the center of the corresponding cluster, by iterative computation until some criteria for convergence is met
- Advantages: relatively low time complexity and high computing efficiency
- Disadvantages: not suitable for non-convex data, relatively sensitive to the outliers, easily drawn into local optimal, the number of clusters needed to be preset, and the clustering result sensitive to the number of clusters

One of the most popular and simple clustering algorithms, K-means, was first published in 1955.

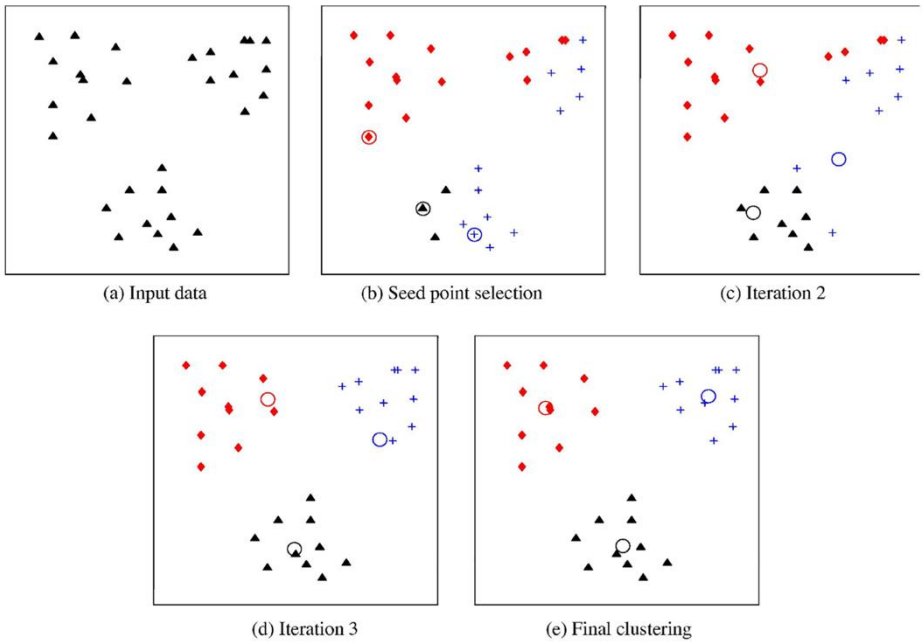


Illustration of K-means algorithm. (a) Two-dimensional input data with three clusters; (b) three seed points selected as cluster centers and initial assignment of the data points to clusters; (c) and (d) intermediate iterations updating cluster labels and their centers; (e) final clustering obtained by K- means algorithm at convergence.

Dongkuan Xu1,2 · **Yingjie Tian**; 2015; **A Comprehensive Survey of Clustering Algorithms**; https://link.springer.com/article/10.1007/s40745-015-0040-1?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

Abiodun M. Ikotun et al.; 2022K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data https://www.sciencedirect.com/science/article/pii/S0020025522014633

Clustering Algorithms based on Partitioning – K-Means

- Goal = regard the center of data points as the center of the corresponding cluster, by iterative computation until some criteria for convergence is met
- Advantages: relatively low time complexity and high computing efficiency
- Disadvantages: not suitable for non-convex data, relatively sensitive to the outliers, easily drawn into local optimal, the number of clusters needed to be preset, and the clustering result sensitive to the number of clusters

10.1.1 K-means Clustering Algorithm

Consider a (training) dataset composed of N observations:

$$x_1, x_2, ..., x_N$$

Initialize K centroids $\mu_1, \mu_2, ..., \mu_K$ randomly.
Repeat until convergence:

1. Cluster assignment
   Assign each x_i to the nearest cluster. For every i do:

$$\underset{j}{argmin} \left\| x_i - \mu_j \right\|^2,$$

   where $j = 1, 2, ..., K$.
2. Cluster updating
   Update the cluster centroids μ_j. For every j do:

$$\mu_j = \frac{1}{N_j}\left[x_1^j + x_2^j + ... + x_{N_j}^j\right],$$

   where N_j is the number of observations assigned to cluster $j, k = 1, 2, ..., N_j$, and x_k^j represents observation k assigned to cluster j. Each new centroid corresponds to the mean of the observations assigned in the previous step.

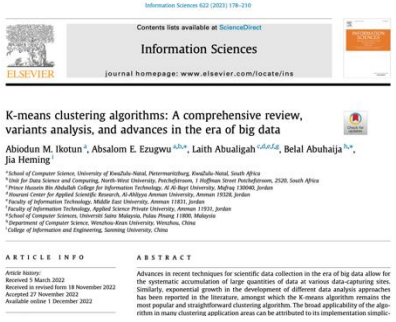
Dongkuan Xu1,2 · Yingjie Tian; 2015; **A Comprehensive Survey of Clustering Algorithms**; https://link.springer.com/article/10.1007/s40745-015-0040-1?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

Calvin J. Chiew, 2020; *Leveraging Data Science for Global Health;* Chapter 8: Applied Statistical Learning in Python https://link.springer.com/book/10.1007/978-3-030-47994-7

Clustering Algorithms based on Partitioning – K-Means

Due to the wide acceptability and use of the K-means clustering algorithm and its easy implementation and simplicity, several algorithm variants have been proposed in the literature to address the known limitations of the standard algorithm.

- K-means algorithm implementation variants
- K-means algorithm design variants



Information Sciences 622 (2023) 178–210

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data

Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming

ARTICLE INFO

ABSTRACT

Advances in recent techniques for scientific data collection in the era of big data allow for the systematic accumulation of large quantities of data at various data-capturing sites. Similarly, exponential growth in the development of different data analysis approaches has been reported in the literature, amongst which the K-means algorithm remains the most popular and straightforward clustering algorithm. The broad applicability of the algorithm in many clustering application areas can be attributed to its implementation simplic-

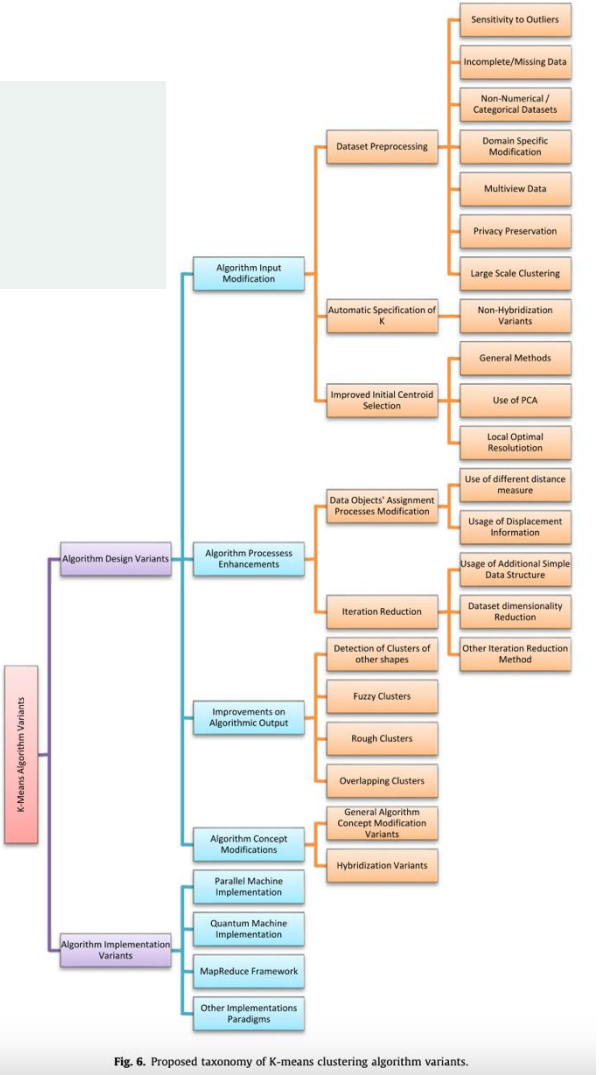


Fig. 6. Proposed taxonomy of K-means clustering algorithm variants.

Abiodun M. Ikotun et al.; 2022K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data
https://www.sciencedirect.com/science/article/pii/S0020025522014633

Clustering Algorithms based on Distribution – GMM

- basic idea: data, generated from the same distribution, belongs to the same cluster if there exist several distributions in the original data.
- The core idea of GMM (Gaussian Mixture Model): There exist several Gaussian distributions from which the original data is generated and the data, obeying the same independent Gaussian distribution, is considered to belong to the same cluster.
- Advantages: more realistic to give the probability of belonging, high scalability by changing the distribution, number of clusters and so on, and supported by the well developed statistical science
- Disadvantages: the premise not completely correct, involved in many parameters which have a strong influence on the clustering result and relatively high time complexity

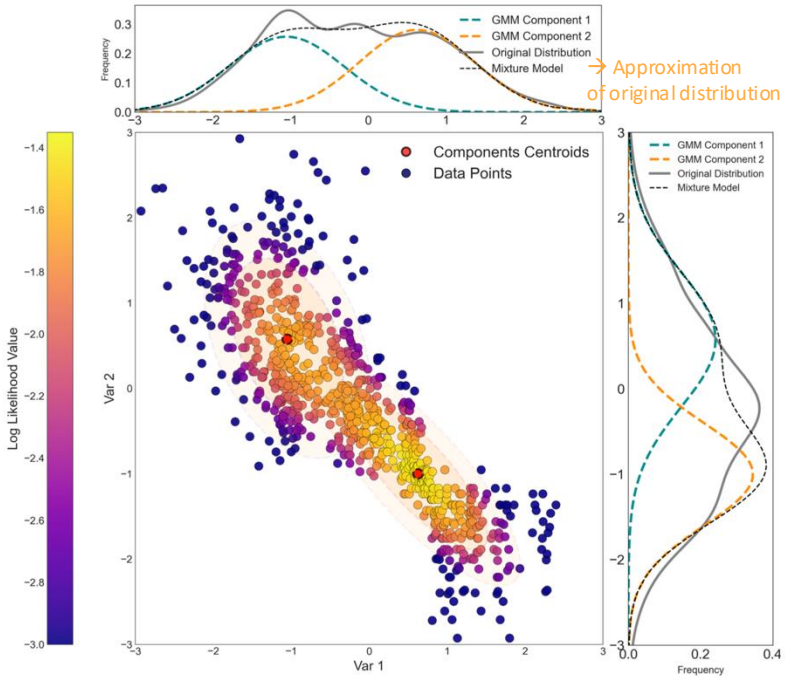


Figure 2: Gaussian Mixture of 2 components fitting bivariate distributions, with respective probability distributions in shared axes.

Dongkuan Xu1,2 · Yingjie Tian; 2015; **A Comprehensive Survey of Clustering Algorithms**; https://link.springer.com/article/10.1007/s40745-015-0040-1?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

Caio Cardoso Gomes1, Jeff Boisvert2, and Clayton V. Deutsch; 2022; Gaussian Mixture Models

Clustering Algorithms based on Distribution – GMM

for a given total number of components *K*, the observations $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ are realizations from the mixture model with density:

$$f(x; \nu) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}; \boldsymbol{\nu}_k)$$

where $f_k(\boldsymbol{x}; \nu_k)$ represents the density of the *k*th group parametrized by ν_k and π_k represents the mixing proportion of the *k*th group.

GMM: each $f_k(\boldsymbol{x};\nu_k)$ is taken to be the multivariate **Gaussian density** $\phi(\boldsymbol{x};\mu k,k)$ with mean μ_k and covariance matrix Σ_k (spread, correlation and orientation of components). Estimation is via the **Expectation-Maximization** (EM) algorithm.

(i) *Initialization*. Obtain starting values $\{(\boldsymbol{\Sigma}_k^{\circ}, \boldsymbol{\mu}_k^{\circ}, \pi_k^{\circ}); k = 1, 2, \ldots, K\}$.

(ii) *E-step updates*. For $k = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, n$, calculate the posterior probability that the *i*th observation arises from the *k*th group:

$$\pi_{ik}^{\circ} = \frac{\pi_k^{\circ} \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l^{\circ} \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}. \quad (3)$$

.

(iii) *M-step updates*. For $k = 1, 2, \ldots, K$, obtain updates:

$$\pi_k^{\circ} = \frac{\sum_{i=1}^{n} \pi_{ik}^{\circ}}{\sum_{k=1}^{K} \sum_{i=1}^{n} \pi_{ik}^{\circ}}, \quad (4)$$

$$\boldsymbol{\mu}_k^{\circ} = \frac{\sum_{i=1}^{n} \pi_{ik}^{\circ} \boldsymbol{x}_i}{\sum_{i=1}^{n} \pi_{ik}^{\circ}}, \text{ and} \quad (5)$$

$$\boldsymbol{\Sigma}_k^{\circ} = \frac{\sum_{i=1}^{n} \pi_{ik}^{\circ} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\circ})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\circ})^T}{\sum_{i=1}^{n} \pi_{ik}^{\circ}}. \quad (6)$$

(iv) Alternate between the E- and M-steps until numerical convergence.

Souradeep Chattopadhyay and Ranjan Maitra; 2017; Gaussian-mixture-model-based cluster analysis finds five kinds of gamma-ray bursts in the BATSE catalogue; doi:10.1093/mnras/stx1024

Clustering Algorithms based on Distribution – GMM

- mixture model:

$$p(X_i|\theta) = \sum_k \pi_k p(X_i|\beta_k)$$

  with $\sum_k \pi_k = 1, \pi_k > 0 \ \ \forall k$ for the mixture coefficients.

- gaussian mixture model: $p_k(X_i|\beta_k) = \mathcal{N}(X_i|\mu_k, S_k)$
- Data log likelihood: $L(\theta) = \sum_{i=1}^{N} \log p(X_i|\theta) = \sum_i \log(\sum_k \pi_k p_k(X_i|\beta_k)) \rightarrow \max$
-

$$\frac{\partial L}{\partial \beta_k} = \sum_i \gamma_{ik} \frac{\partial}{\partial \beta_k} \log p_k(X_i|\beta_k) =^{!} 0$$

  $\Rightarrow$ with γ_{ik} fixed, this is a weighted ML problem for the kth mixture component

  $\Rightarrow$ alternating optimization: Solve $\sum_i \gamma_{ik}(\theta^{(t-1)}) \frac{\partial}{\partial \beta_k} \log p_k(X_i|\beta_k)$ to get $\beta_k^{(t)}$

Clustering Algorithms based on Distribution – GMM

$$\frac{\partial(L + \lambda(\sum_k \pi_k - 1))}{\partial \pi_k} =^! 0 \qquad \Rightarrow \pi_k^{(t)} = \frac{1}{N} \sum_i \gamma_{ik}(\theta^{(t-1)})$$

- algorithm (for Gaussian mixture models)
  - choose number of mixture components C (difficult! see k-means)
  - define initial guess $\theta^{(0)}$ (random, k-means++ initialization)
  - repeat until convergence $t = 1, ..., T$
    - compute responsibilities γ_{ik}

$$\gamma_{ik} = p(Y_i = k | X_i, \theta^{(t-1)}) = \frac{\pi_k^{(t-1)} \mathcal{N}\left(X_i | \mu_k^{(t-1)}, s_k^{(t-1)}\right)}{\sum_{k'} \pi^{(t-1)} \mathcal{N}\left(X_i | \mu_{k'}^{(t-1)}, s_k^{(t-1)}\right)}$$

    - E-step (expectation step)

$$\pi_k^{(t)} = \frac{1}{N} \sum_i \gamma_{ik}$$

    - M-step (maximization step): determine the weighted ML solution for each mixture component individually

$$\mu_k^{(t)} = \frac{\sum_i \gamma_{ik} X_i}{\sum_i \gamma_{ik}}, \quad s_k^{(t)} = \frac{\sum_i \gamma_{ik} (X_i - \mu_k^{(t)})^\top (X_i - \mu_k^{(t)})}{\sum_i \gamma_{ik}}$$

Clustering Algorithms based on Density – DBSCAN

- basic idea: data which is in the region with high density of the data space is considered to belong to the same cluster.
- DBSCAN (density-based spatial clustering of applications with noise) is the most well known density-based clustering algorithm
- Advantages: high efficiency and suitable for data with arbitrary shape
- Disadvantages: resulting in a clustering result with low quality when the density of data space isn't even, a memory with big size needed when the data volume is big, and the clustering result highly sensitive to the radius of the neighborhood and the minimum number of points in a neighborhood

Algorithm 1 DBSCAN

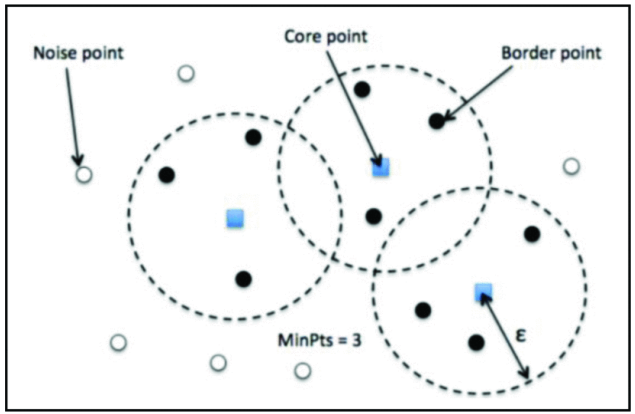
Require: dataset D, adjacent area radius EPS, minimum number of points in the neighborhood $MinPts$
Ensure: Clusters C

```
Set all points as unmarked
Do
Choose one unmarked points X randomly
Mark X as marked
if the number of points in adjacent area radius EPS of X ≥ MinPts then
    Produce a new cluster C and put P into C
    Let P be the set of points within EPS of X neighborhood radius
    for P for each pointx do
        if x is unmarked then
            Set X as marked
            if the number of points in X adjacent area radius EPS >MinPts then
                put them all to X
                if X does not belong to any cluster then
                    Put X to C
                end if
            end if
        end if
    end for
    Output C
else
    Marks X as noise points
end if
repeat
    Above procedure
until there are no unmarked points
```

Dongkuan Xu1,2 · **Yingjie Tian**; 2015; **A Comprehensive Survey of Clustering Algorithms**; https://link.springer.com/article/10.1007/s40745-015-0040-1?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot

https://towardsdatascience.com/dbscan-explained-in-5-minutes-133f6a9766e4

Clustering Algorithms based on Density – DBSCAN



Hyper Parameters:

•eps(ε): - This is the radius around the point which is checked for other points in its proximity. It is used to determine the density of region. If large eps value is chosen, there will be low density clusters.

•minPts: - This is the number of points present in the surrounding (eps distance from the point) to form a cluster.

Point Classification:

1.Core — If different clusters are made this is the point from which there are at least k number of points in r distance of radius.

2.Border — This can be any point that has one or more core points within r distance of radius.

3.Noise – Any point which is neither a core and a border and has at least k number of points in r distance of radius.

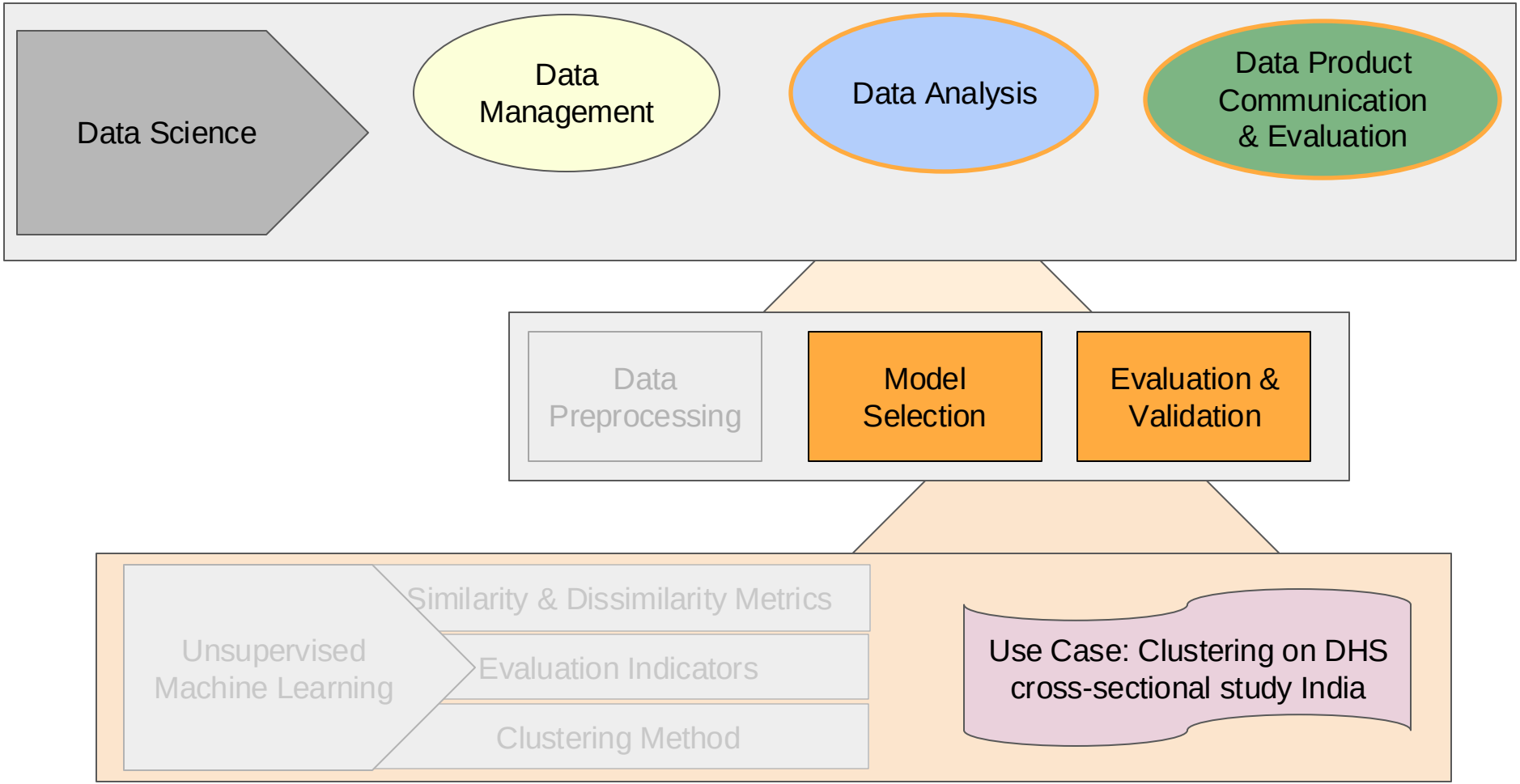
Harsh Vardhan Singh; Ashwin Girdhar; Sonika Dahiya; 2022; A **Literature survey based on DBSCAN algorithms**

There is no best Clustering Algorithm

„Each clustering algorithm imposes a structure on the data either explicitly or implicitly. When there is a good match between the model and the data, good partitions are obtained. Since the structure of the data is not known a priori, one needs to try competing and diverse approaches to determine an appropriate algorithm for the clustering task at hand. “

→ impossibility theorem (Kleinberg, 2002): no single clustering algorithm simultaneously satisfies a set of basic axioms of data clustering

Anil K. Jain; 2009; Data clustering: 50 years beyond K-means; Pattern Recognition Letters


Data Science
Data Management
Data Analysis
Data Product Communication & Evaluation
Data Preprocessing
Model Selection
Evaluation & Validation
Unsupervised Machine Learning
Similarity & Dissimilarity Metrics
Evaluation Indicators
Clustering Method
Use Case: Clustering on DHS cross-sectional study India

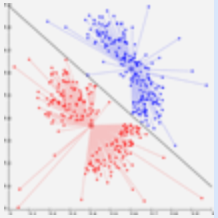
Objective: Address population health provision (i.e. health care delivery) for chronic non-communicable disease through identification of subpopulations with distinct prevalence patterns and their identifiers.

The DHS Program
Demographic and Health Surveys

DHS Program data are collected so they can be used to guide programs and policies to improve health and well-being.
NFHS-5 India: 625.000 women; 80.000 men

= **Cross-Sectional Study**
(data collection at one time point across study population with stratified sampling)

Clustering



Clustering is a procedure of organizing the objects in groups whose member exhibits some kind of similarity.

[3] https://doi.org/10.1145/2393216.2393327

Clusters are constructed though the analysis of the distribution of instance (person) within a parameter space.

disease markers
biological parameters
numerical
exposure markers
socio-economic factors
behavioural information
categorical

Data Preprocessing

Data Source
Analysis Model
Data Type

All shared results on the following slides are preliminary versions of our data analysis. There will be a publication following soon!

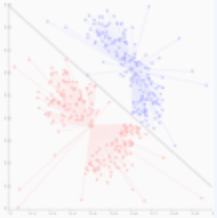
The DHS Program
Demographic and Health Surveys

DHS Program data are collected so they can be used to guide programs and policies to improve health and well-being.
NFHS-5 India: 625.000 women; 80.000 men

= **Cross-Sectional Study**
(data collection at one time point across study population with stratified sampling)

Data Source

Clustering



Clustering is a procedure of organizing the objects in groups whose member exhibits some kind of similarity.

[3] https://doi.org/10.1145/2393216.2393327

Clusters are constructed though the analysis of the distribution of instance (person) within a parameter space.

Analysis Model

disease markers

biological parameters

numerical

exposure markers

socio-economic factors

behavioural information

categorical

Data Type

Data Preprocessing

Using Clusters for Predictions

We have seen how clustering can be used to stratify patients, but not how it can be used to predict outcomes.
→ Predictive models can be informed by the information provided by clustering

a. By using identified clustering models as classification models

b. By using supervised ML to train upon labels assign by clustering model (using same parameter or even different ones)

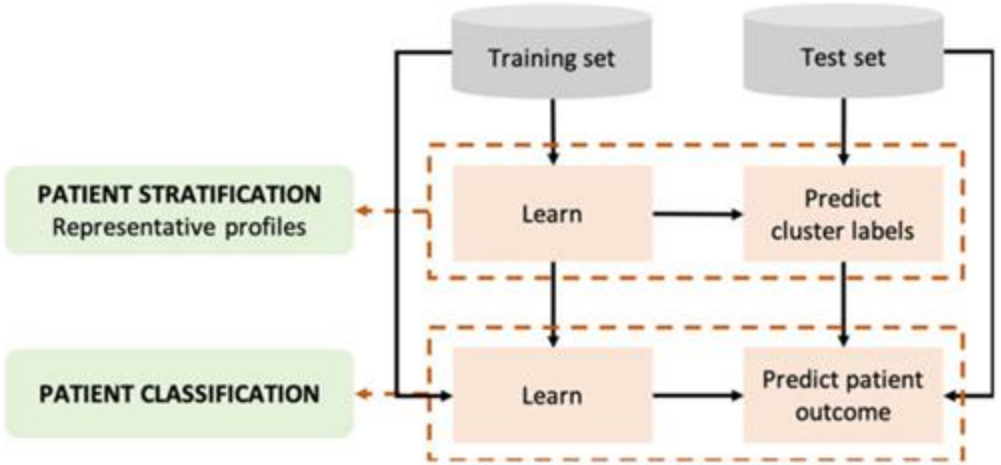
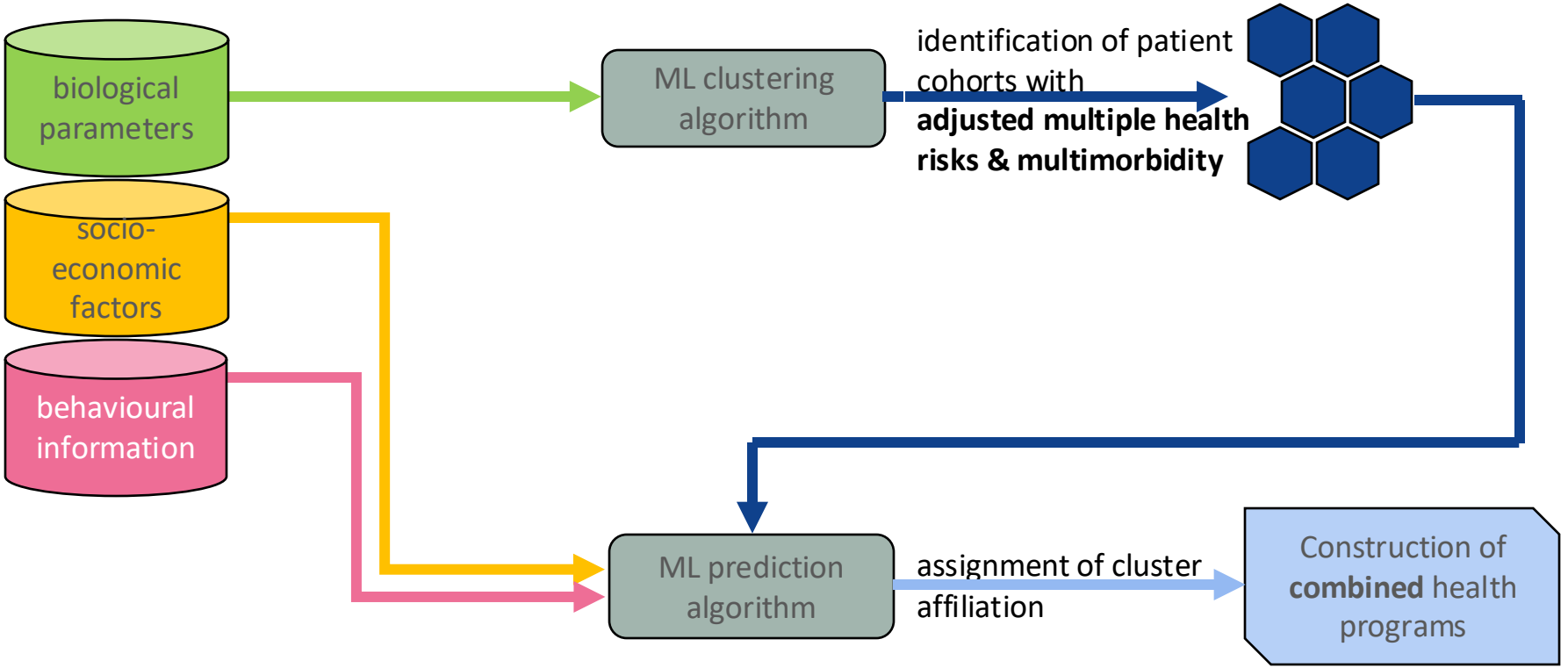


Fig. 10.1 Schematic representation of the machine learning steps

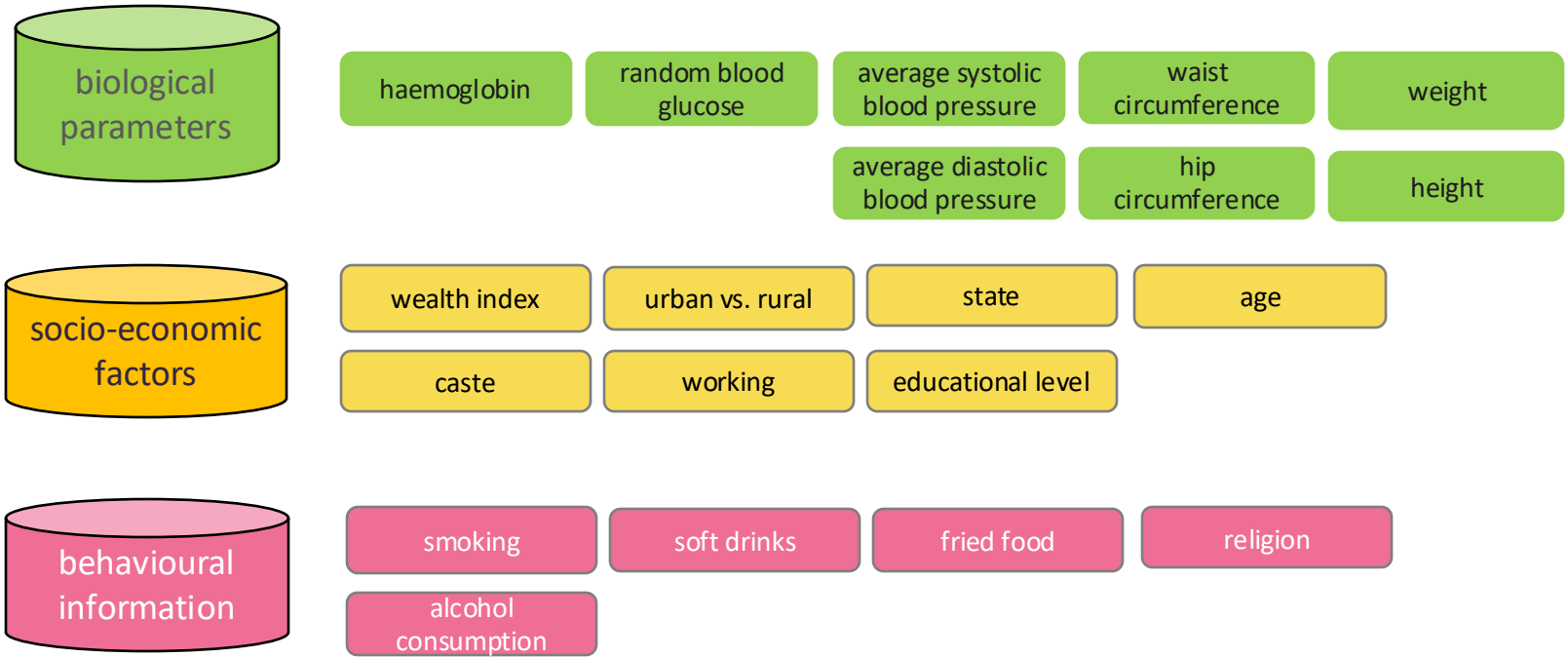
Calvin J. Chiew, 2020; *Leveraging Data Science for Global Health;* Chapter 8: Applied Statistical Learning in Python https://link.springer.com/book/10.1007/978-3-030-47994-7

Full Data Analysis



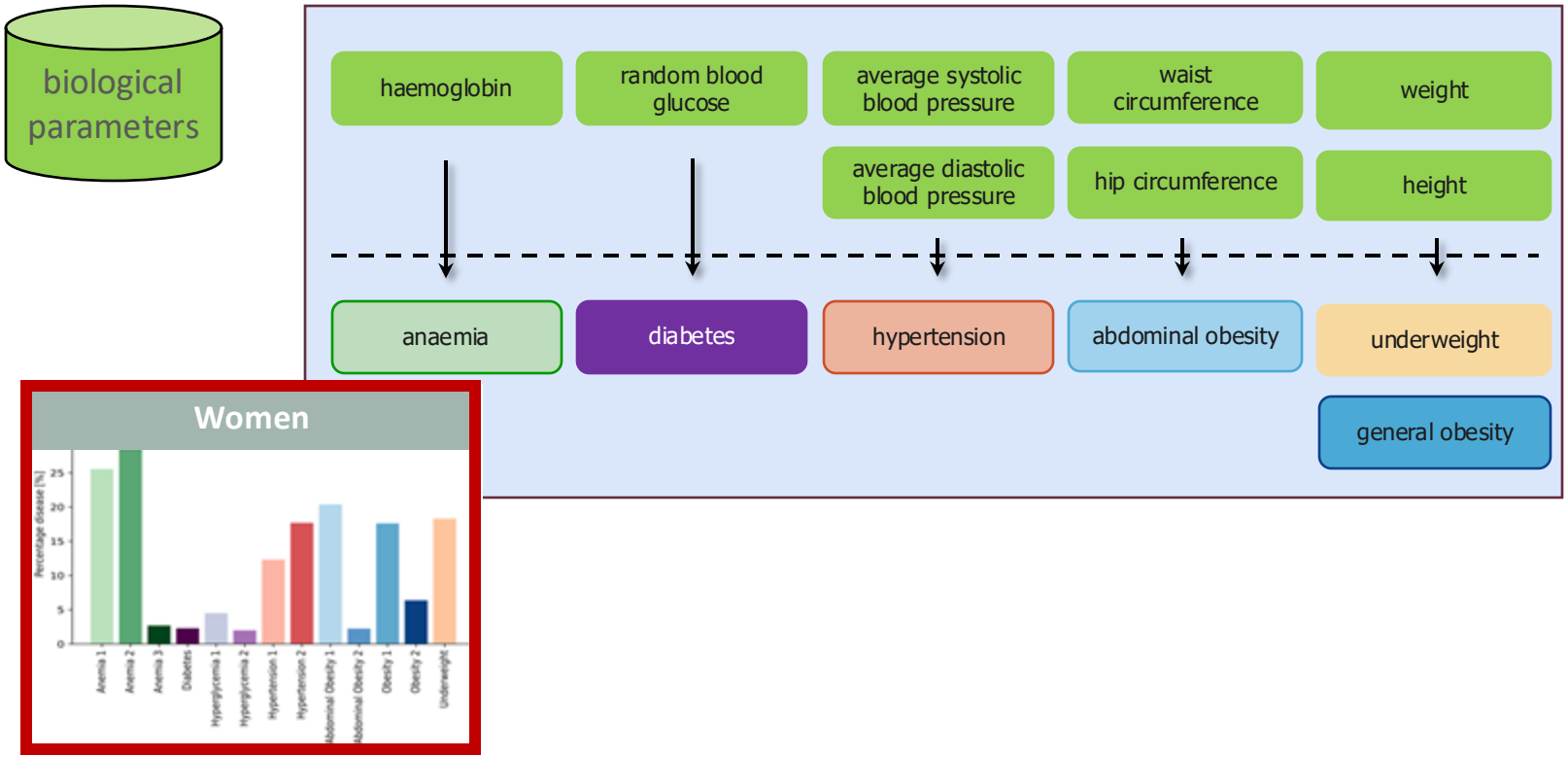
Nitschke, Brandl, Bärnighausen et al.; 2024; preprint Machine learning-based Clustering of Chronic non-communicable Multimorbidity in India: A National Representative Study including 718811 Individuals

Variable Selection



Nitschke, Brandl, Bärnighausen et al.; 2024; preprint Machine learning-based Clustering of Chronic non-communicable Multimorbidity in India: A National Representative Study including 718811 Individuals

Variable Selection



Nitschke, Brandl, Bärnighausen et al.; 2024; preprint Machine learning-based Clustering of Chronic non-communicable Multimorbidity in India: A National Representative Study including 718811 Individuals

Method Selection & Results Validation

Data Analysis

Data-driven | Optimization

1

Change Data

2

Change Model

3

Performance Scores

4

Change Metric

Method Selection & Results Validation

1

Change Data

Bootstrapping

Hold out validation - Separate data into training (and validation) and test data set:

- training set is used to train/build the learning algorithm
- validation (or development) set is used to tune parameters select features, and make other decisions regarding the learning algorithm
- test set is used to evaluate the performance of the algorithm

Cluster stability is measured as the amount of variation in the clustering solution over different subsamples drawn from the input data.

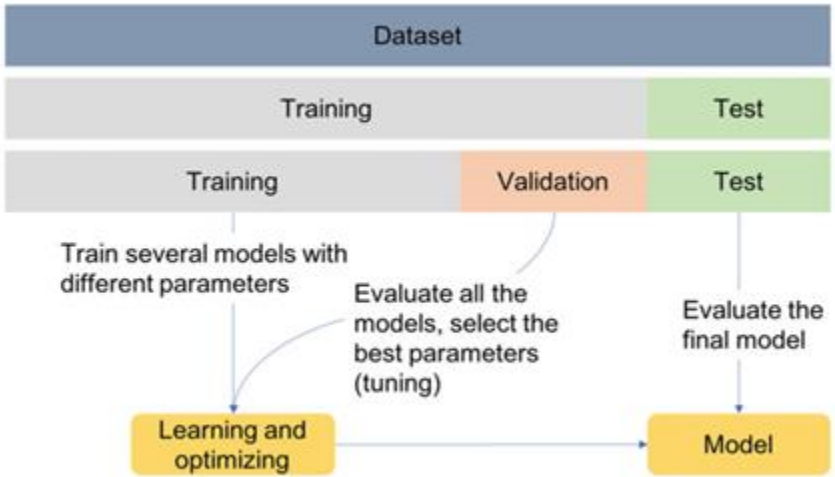


Fig. 9.1 Illustrative scheme of the holdout validation method

Calvin J. Chiew, 2020; *Leveraging Data Science for Global Health;* Chapter 8: Applied Statistical Learning in Python https://link.springer.com/book/10.1007/978-3-030-47994-7

Method Selection & Results Validation

1

Change Data

Bootstrapping

Cross-validation is a resampling method that can be used to tune parameters of a model.

- In k-fold CV, we split the training data into k folds, take one fold to validate and remaining k-1 folds to train.
- Then calculate the chosen performance metric, repeat k times and average the result.

Cluster stability is measured as the amount of variation in the clustering solution over different subsamples drawn from the input data.

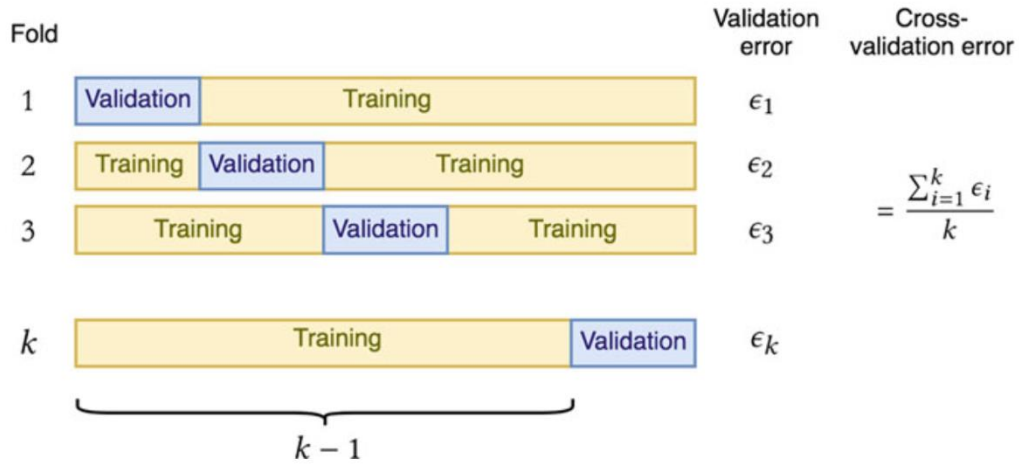


Fig. 12.2 K-fold cross-validation

Calvin J. Chiew, 2020; *Leveraging Data Science for Global Health;* Chapter 8: Applied Statistical Learning in Python https://link.springer.com/book/10.1007/978-3-030-47994-7

Method Selection & Results Validation

1

Change Data

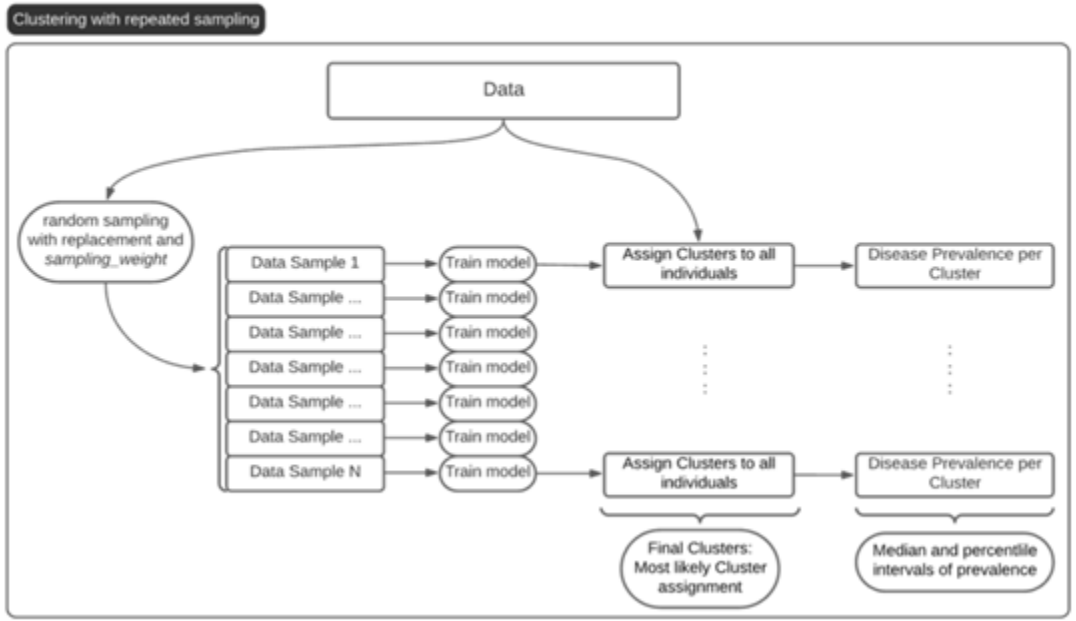
Bootstrapping

Our Approach

- Use repeated weighted sampling (using sampling weights from survey)
- Calculate median and percentile intervals

Cluster stability is measured as the amount of variation in the clustering solution over different subsamples drawn from the input data.

Supplemental Figure 5: Diagram of Clustering Validation process / repeated sampling

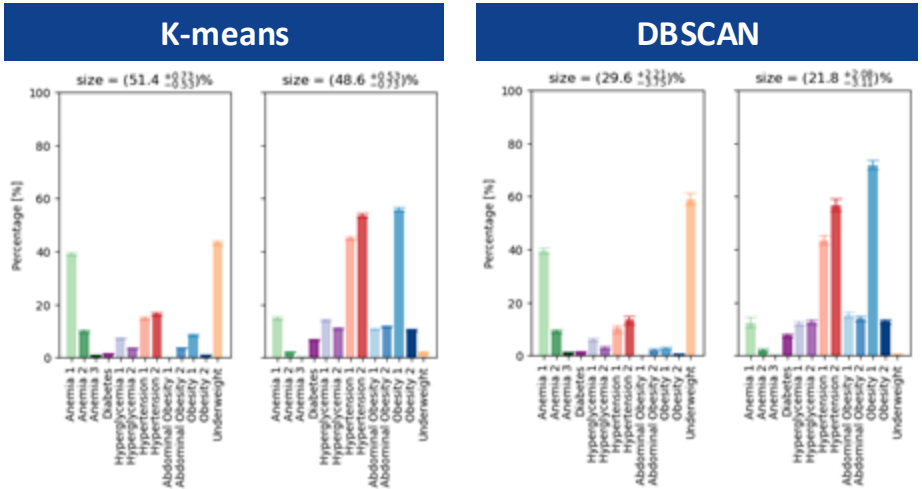


Method Selection & Results Validation

2

Change Model

K-means
Gaussian Mixture Model
Agglomerative Clustering
HDBSCAN



External validation: compare results across different models

a.) visual comparison
b.) External Validation Indicator Metrics (Rand Index)

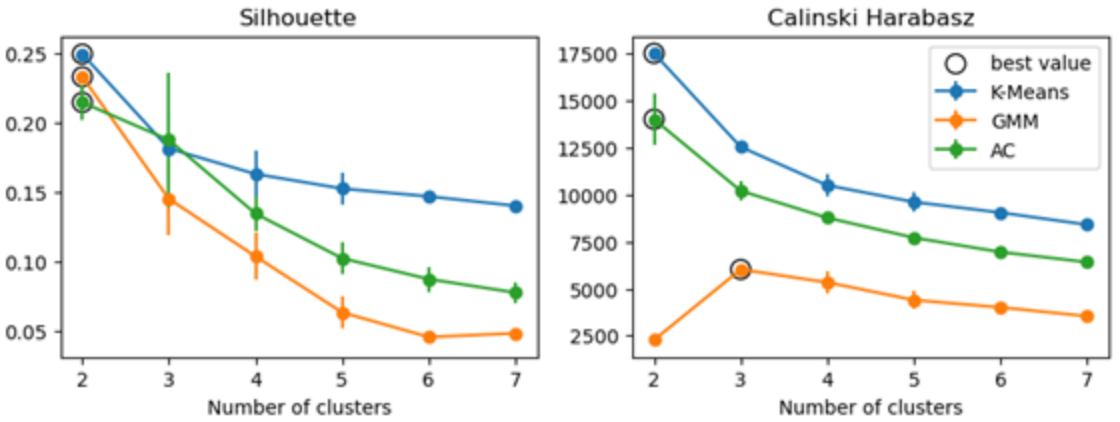
Method Selection & Results Validation

3

Performance Scores

Internal Validation Indicator Metrics (Silhouette; Calinski Harabasz)

compare results across
a.) different models
b.) different numbers of clusters



Method Selection & Results Validation

4

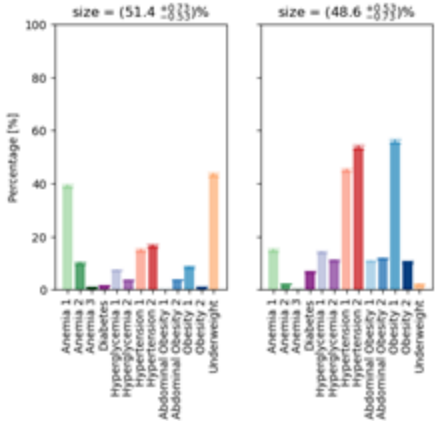
Change Metric

Canberra
Euclidean

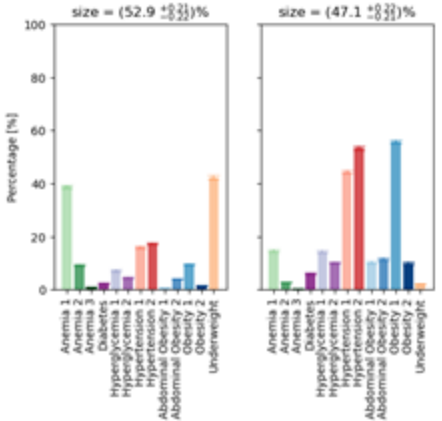
External validation: compare results across different metrics

a.) visual comparison
b.) External Validation Indicator Metrics (Rand Index)

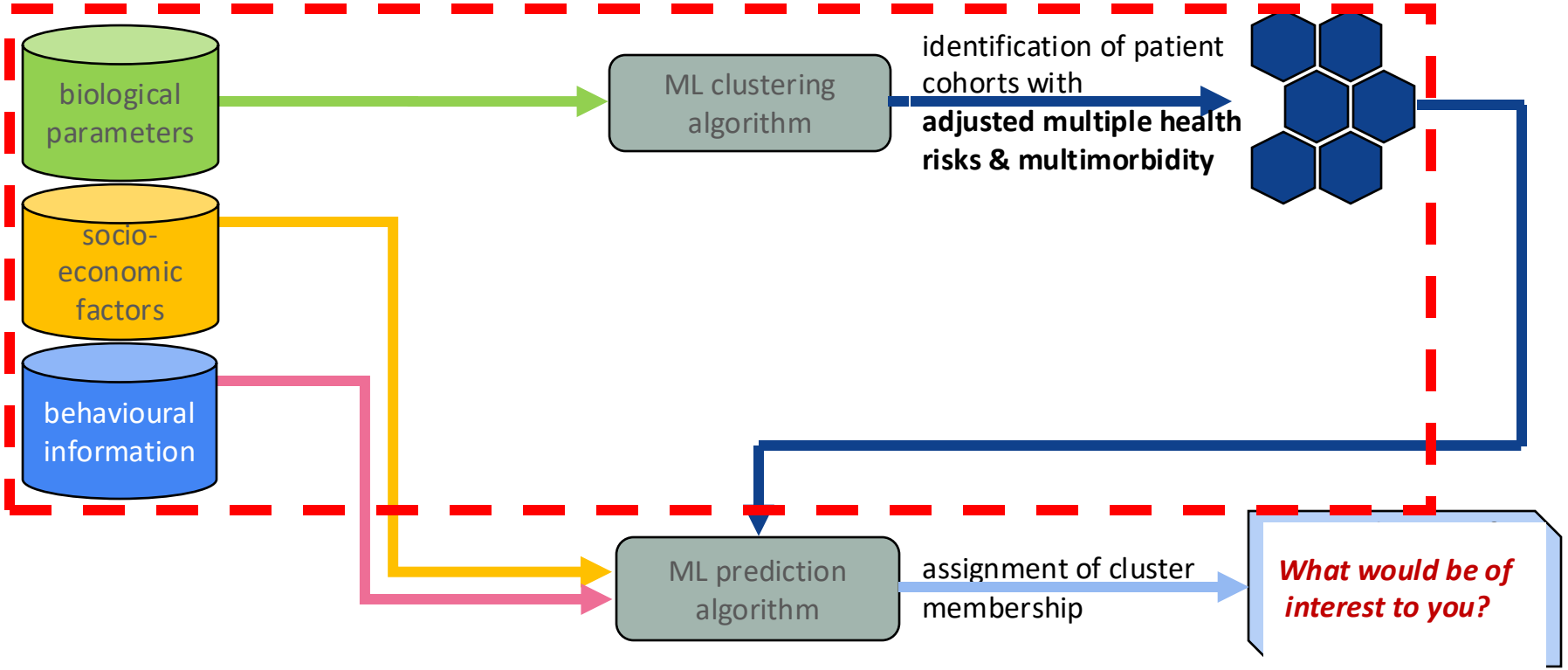
K-means: Euclidean



K-means: Canberra



Full Data Analysis



Data Product Communication & Evaluation

Data Product Communication & Evaluation

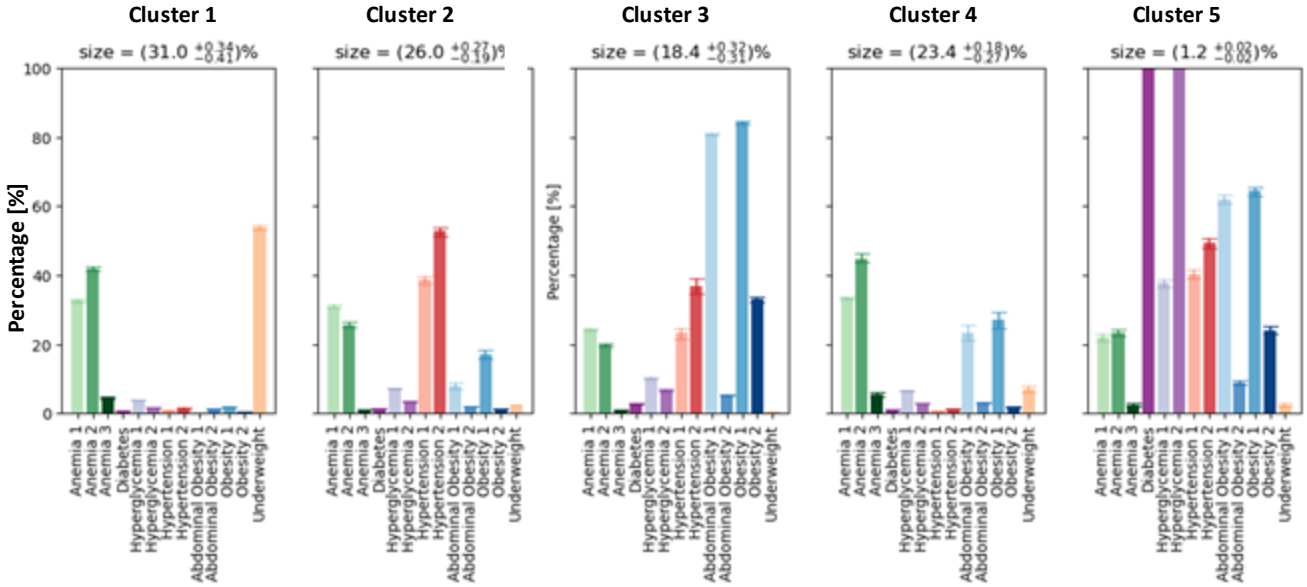
Domain-knowledge driven | Plausibility

A	B	C	D	E	F
Visualisation individual disease prevalence	Visualisation individual disease patterns	Visualisation correlations between disease	Visualisation Multi-Morbidity	Visualisation cluster distribution in country	Visualisation of other parameters within clusters

Data Product Communication & Evaluation

A

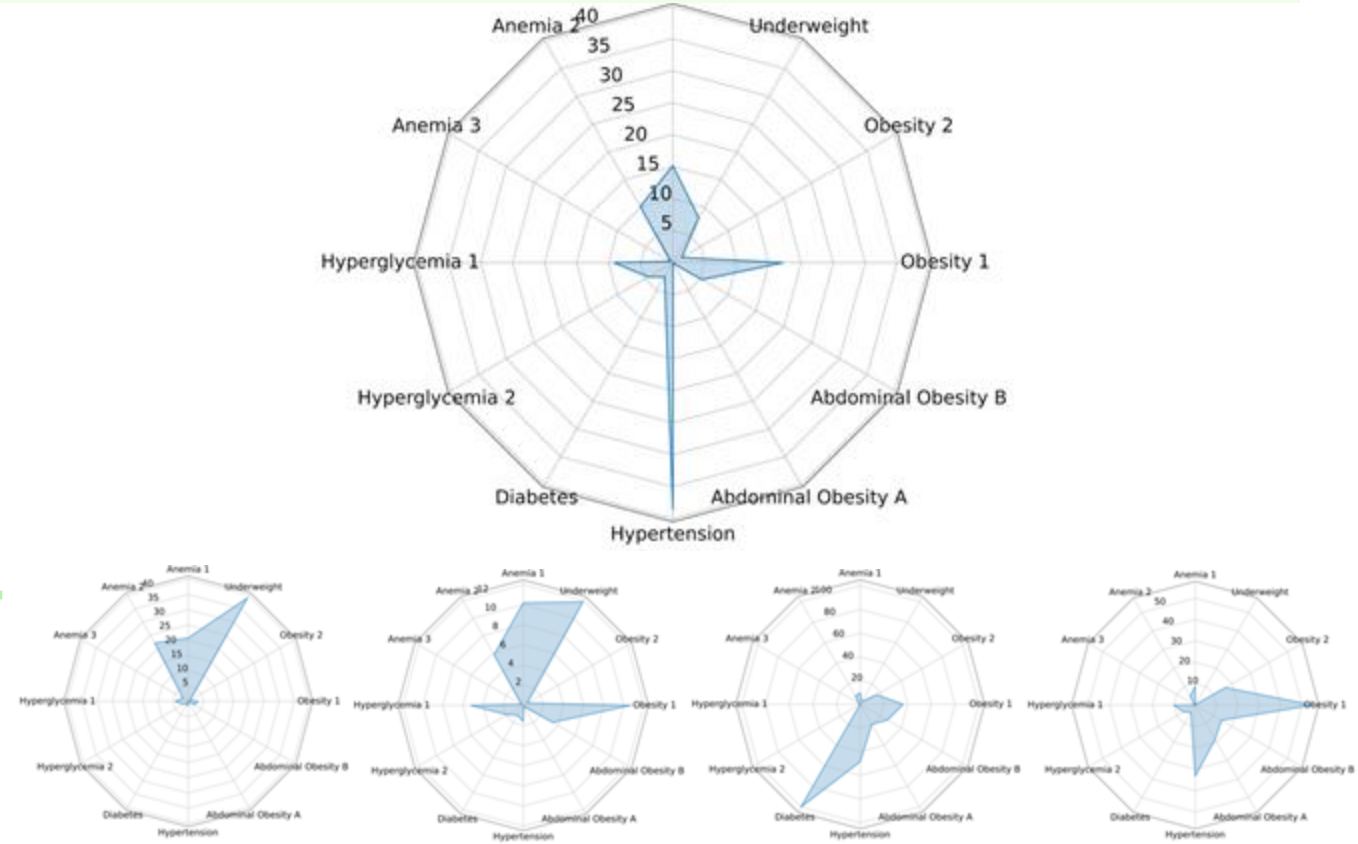
Visualisation individual disease prevalence



Data Product Communication & Evaluation

B

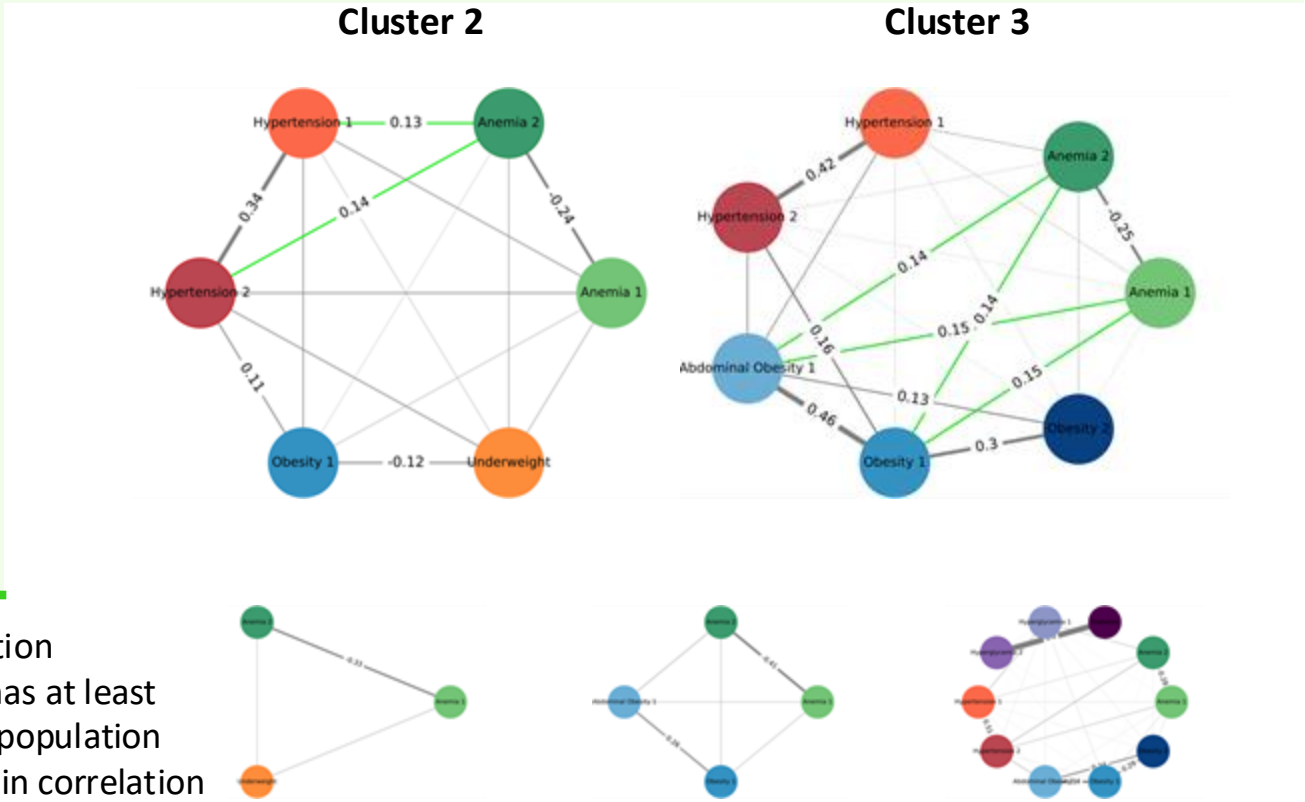
Visualisation individual disease patterns



Data Product Communication & Evaluation

C

Graph based visualisation of correlations between disease



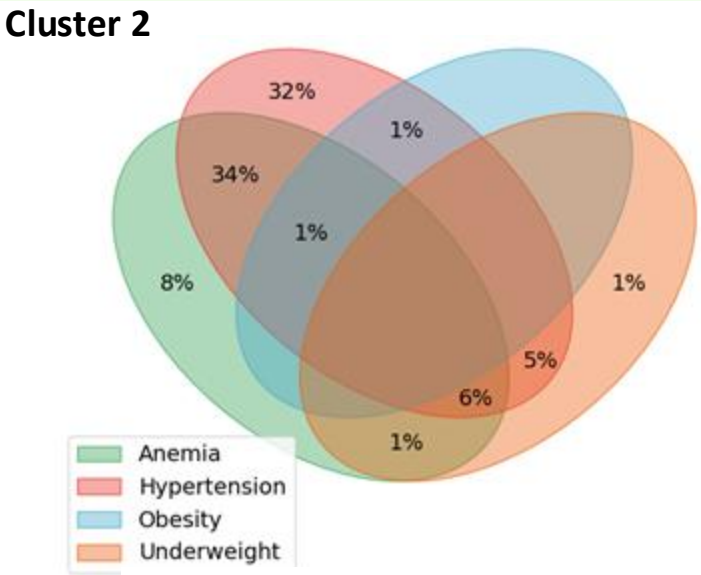
- Use graph-based visualisation
- Show only disease which has at least ten percent of the cluster population
- Green: represent changes in correlation strength compared to overall correlation

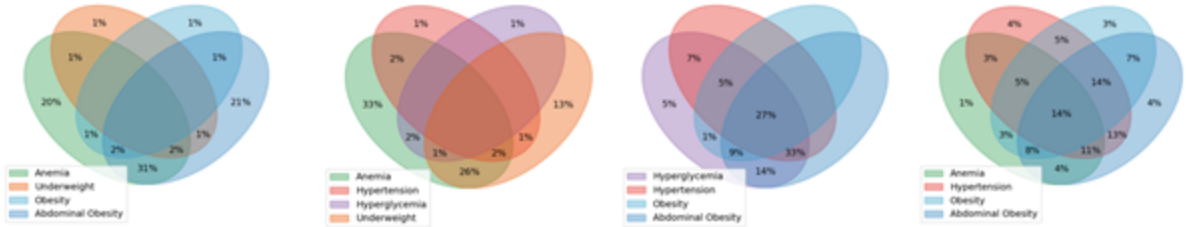
Data Product Communication & Evaluation

D

Visualisation
Multi-Morbidity

- Use Venn Diagram
- Show only four most prevalent disease within each cluster





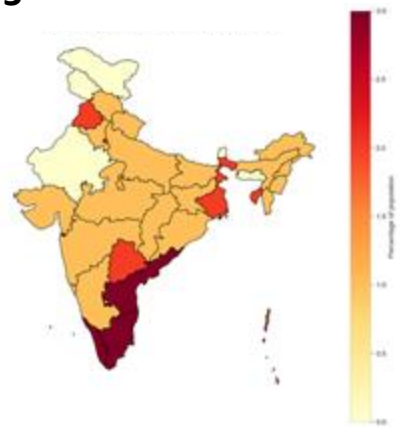
Data Product Communication & Evaluation

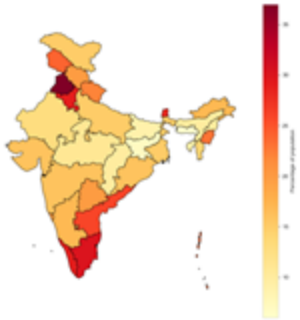
E

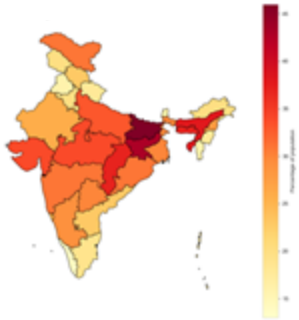
Visualisation cluster distribution in country

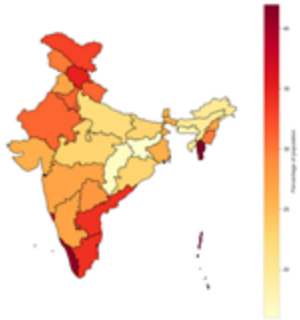
Compute the percentage that the people within each cluster contribute to population within each state

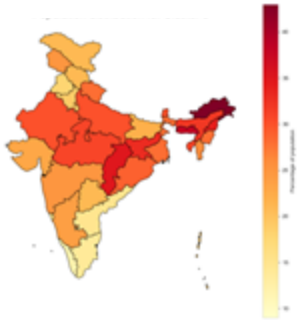
Cluster 5







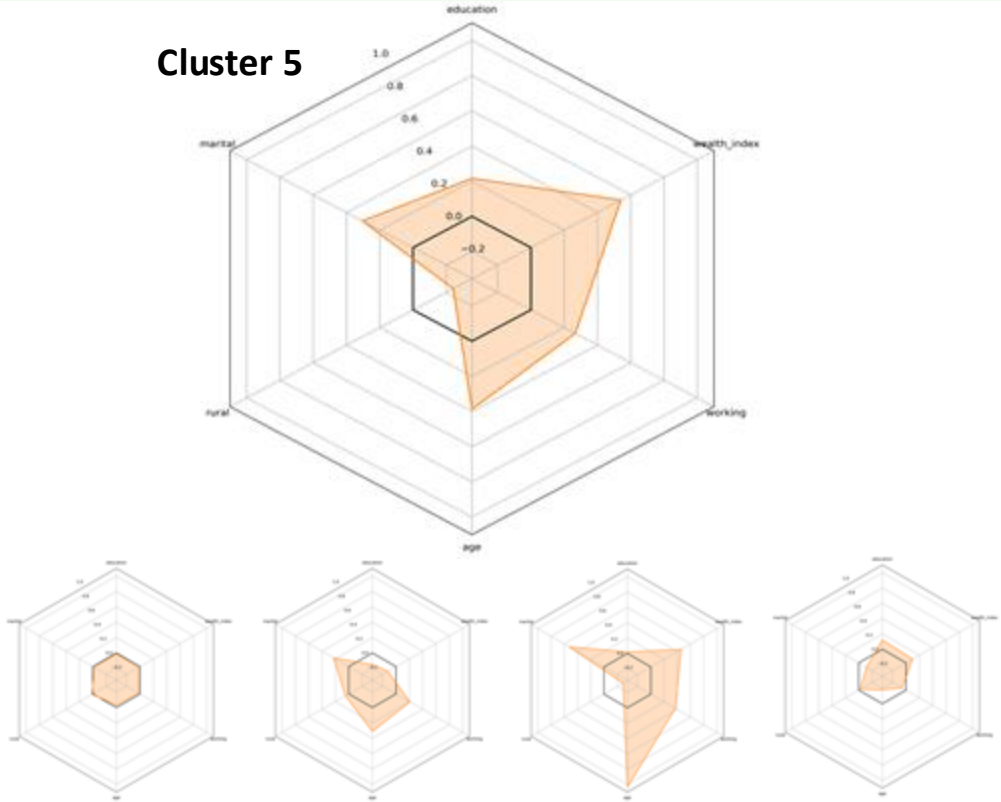




Data Product Communication & Evaluation

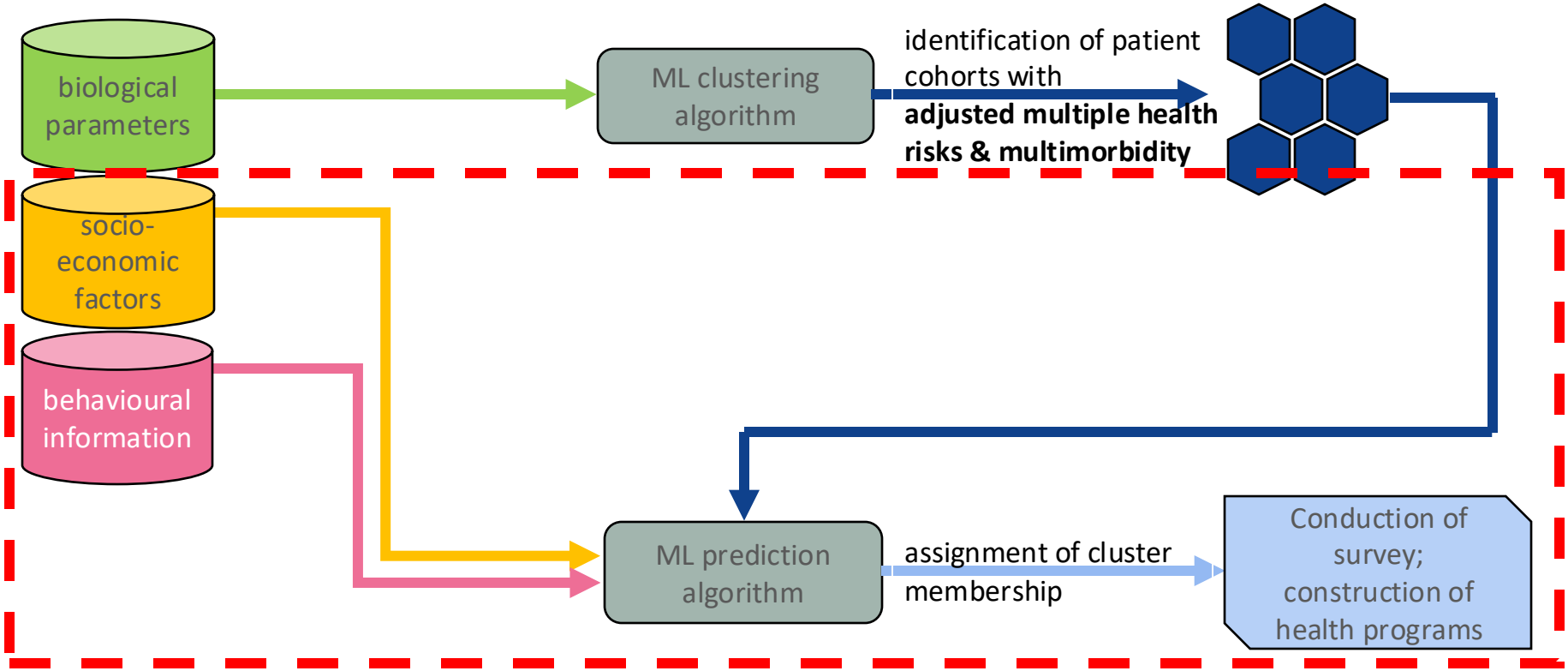
F

Visualisation of other parameters within clusters



- Standardise ordinal socio economic factors to same scale
- Compare patterns across clusters

Full Data Analysis



Prediction of Clusters

NFHS-5 India: 635.000 women; 85.000 men

socio-economic factors

behavioural information

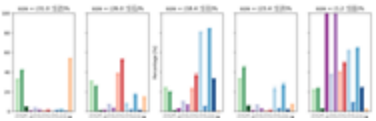
wealth index [7]	urban vs. rural [8]	gender [9]	age [9]
caste [7]	sanitation & hygiene [8]	educational level [9]	marital
smoking	alcohol	physical inactivity	Unbalanced nutrition
religion			

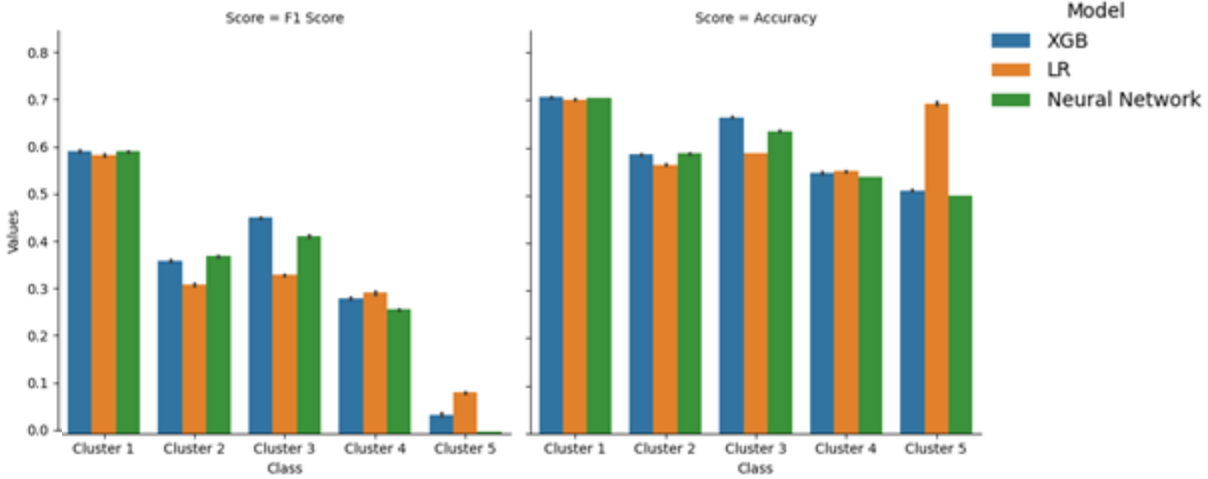
[7] BARU et al. (2010).; [8] Cowling et al. (8 October 2014); [9] Raza et al. (December 2015).

Prediction of Clusters

socio-economic factors

behavioural information





- Be aware of the performance scores that you are using
- If there are small clusters: F1 Score is more reliable
- **socio-economic features can identify cluster membership**

Odds Ratio

a/c = exposure ratio diseased
(cases)

b/d = exposure ration not diseased
(control)

Odds Ratio = ad/bc

In our case:

cases = belongs to cluster;
control = does not belong to cluster

exposure = behavioural or socio-economic factors

Example B: Odds Ratio

	Diseased	Not Diseased	Total
Exposed	300	503	803
Unexposed	200	497	697
Total	500	1000	1500

figure extracted from presentation by Vidaisha Naidoo; Tuesday

Odds Ratio

a/c = exposure ratio diseased
(cases)

b/d = exposure ration not diseased
(control)

Odds Ratio = ad/bc

In our case:

cases = belongs to cluster;
control = does not belong to cluster

exposure = behavioural or socio-economic factors

Supplemental Figure 16: Odds Ratios for predicting clusters in women

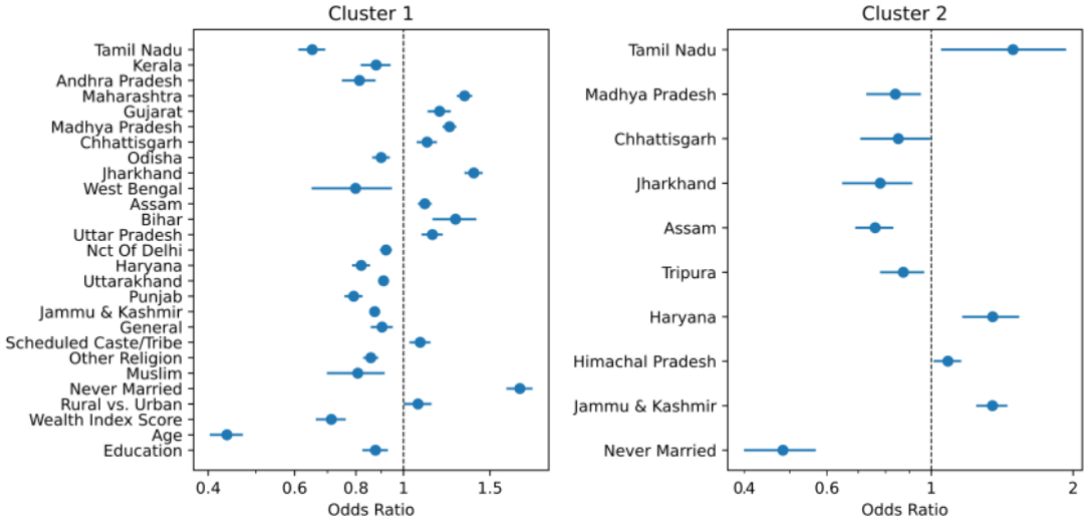
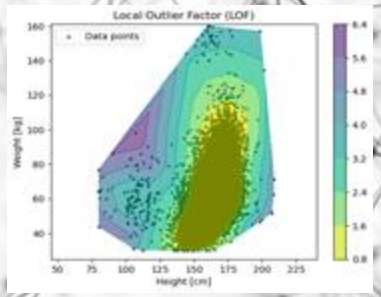
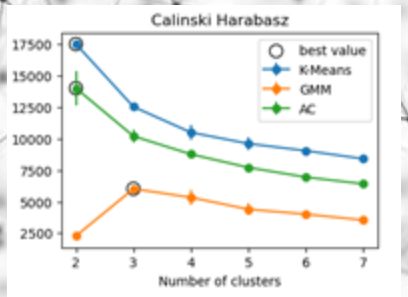


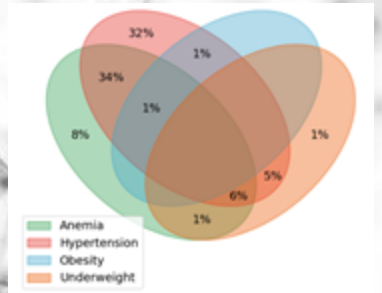
figure extracted from presentation by Vidaisha Naidoo; Tuesday

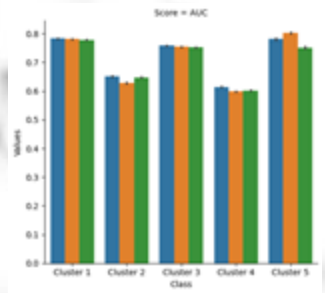
TAKE HOME MESSAGE

- Steps in the data preprocessing have a strong impact on the clustering result
- For the validation of the clustering results, check stability and similarity of results across different methods
- By using appropriate visualisations, distinct information and conclusions can be drawn from clusters
- Socio-economic features can be used to identify cluster membership









TEAM AND SUPPORT

Anna-Katharina Nitschke, Carlos Brandl, Jonathan Elias Berthold, Jannis Demel, Carola Behr and Matthias Weidemüller

Physikalisches Institut, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 226, 69120 Heidelberg, Germany

in collaboration with Till Bärnighausen, Kavita Singh, Ullrich Köthe, Michaela Theilmann, Jen Manne-Goehler, Martin Siegel, Sujata

THANK YOU!

Stay updated - publication will follow soon

STRUCTURES CLUSTER OF EXCELLENCE

DFG Deutsche Forschungsgemeinschaft

bwForCluster Helix

The DHS Program Demographic and Health Surveys

SPONSORED BY THE Federal Ministry of Education and Research

Baden-Württemberg MINISTRY OF SCIENCE, RESEARCH AND ARTS