Data mining and vector-borne diseases : what perspectives?

Frédéric Flouvat[1], Nazha Selmaoui-Folcher[1], Hugo Alatrista Salas[2], Sandra Bringay[3], Maguelonne Teisseire[4]

[1] Pôle Pluridisciplinaire de la Matière et de l'Environnement (PPME), Université de la Nouvelle-Calédonie

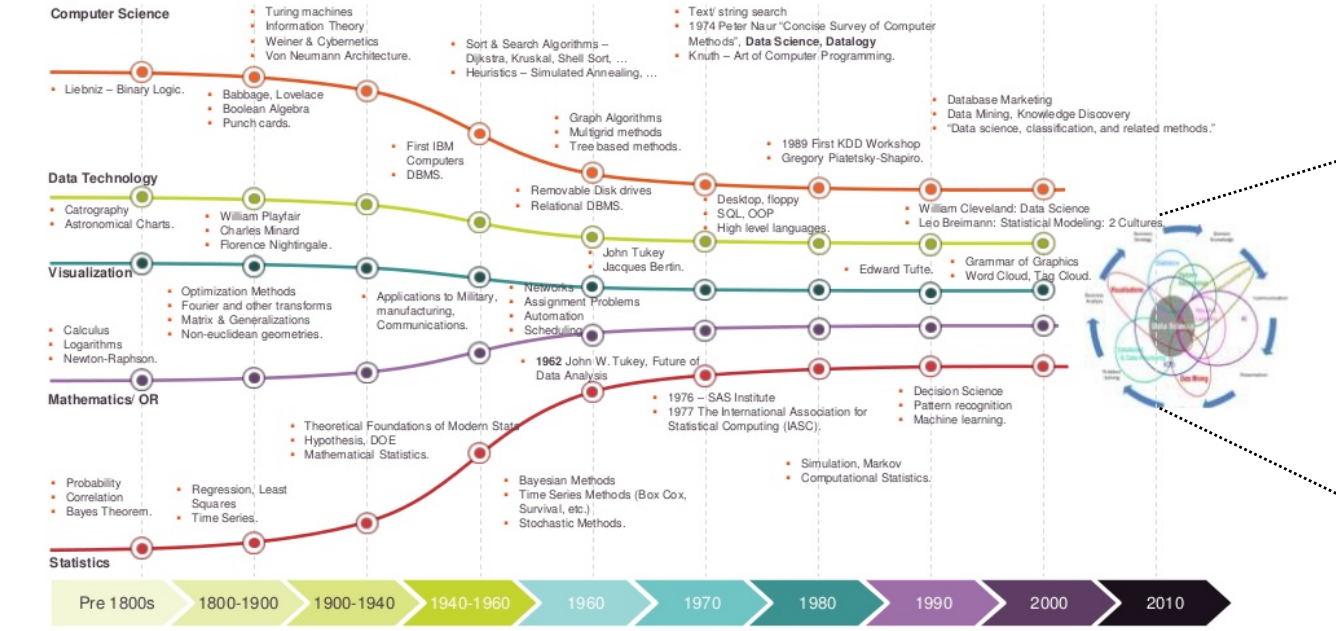
[2] Pontifical Catholic University of Peru (PUCP)

[3] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Université de Montpellier

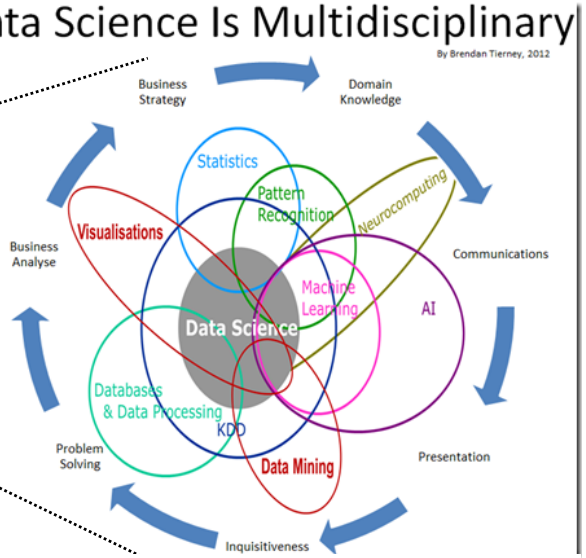
[4] UMR TETIS (Territoires, Environnement, Télédétection et Information Spatiale), IRSTEA

Data Science, Data Mining, Machine learning, KDD, etc

A brief history of Data Science

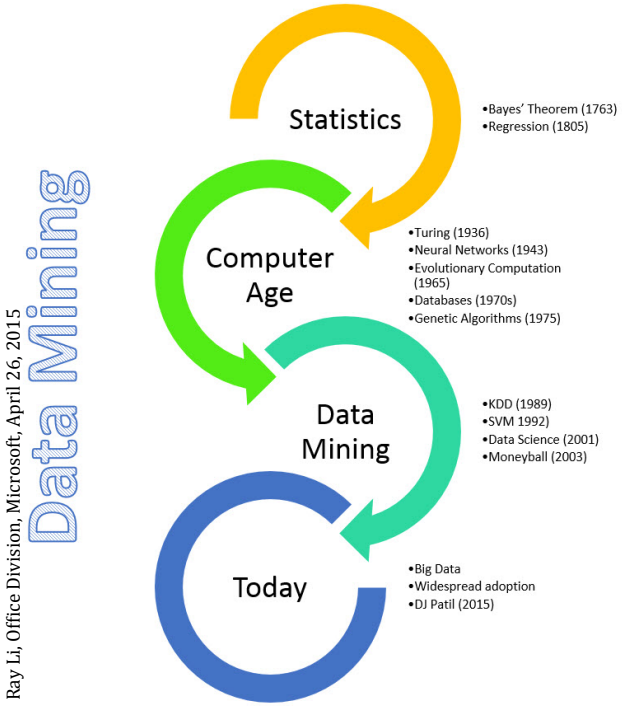


Data Science Is Multidisciplinary



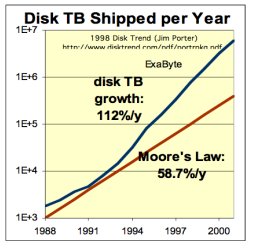
Copyright © 2014 Capgemini. All rights reserved.

Data Science, Data Mining, Machine learning, KDD, etc



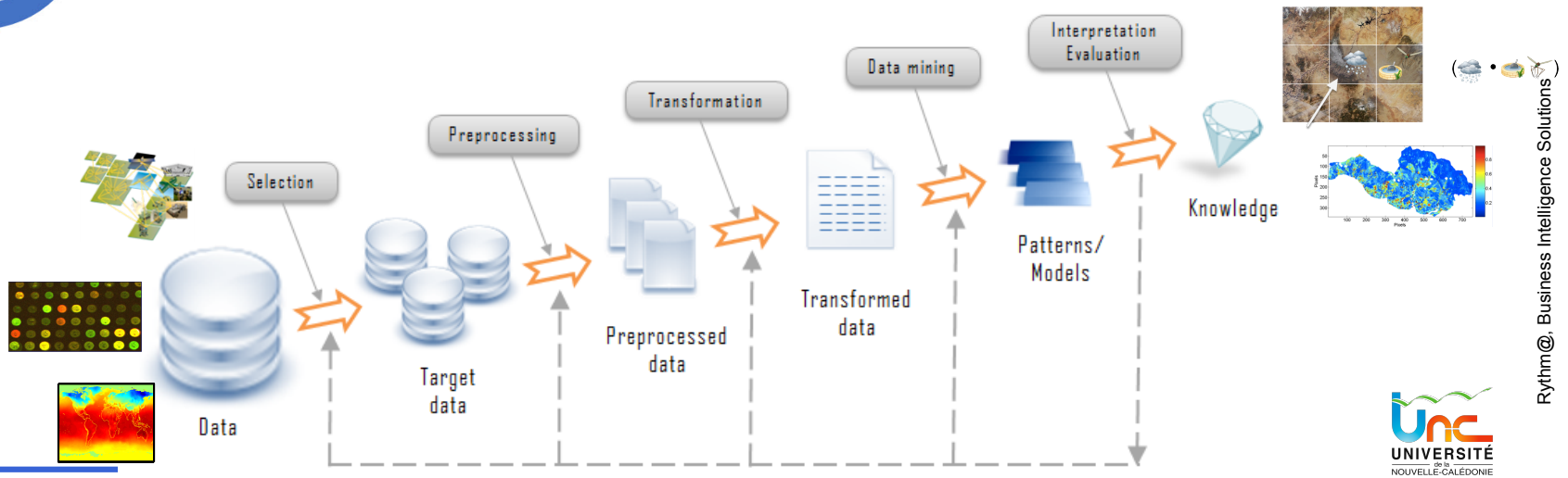
Data Mining

- Algorithms analyzing complex and/or voluminous data
- One of the steps of a knowledge discovery process
- Successfully applied to **public health**, environmental monitoring, customer relationship management, …



Knowledge Discovery in Databases (KDD)

- An iterative and interactive process



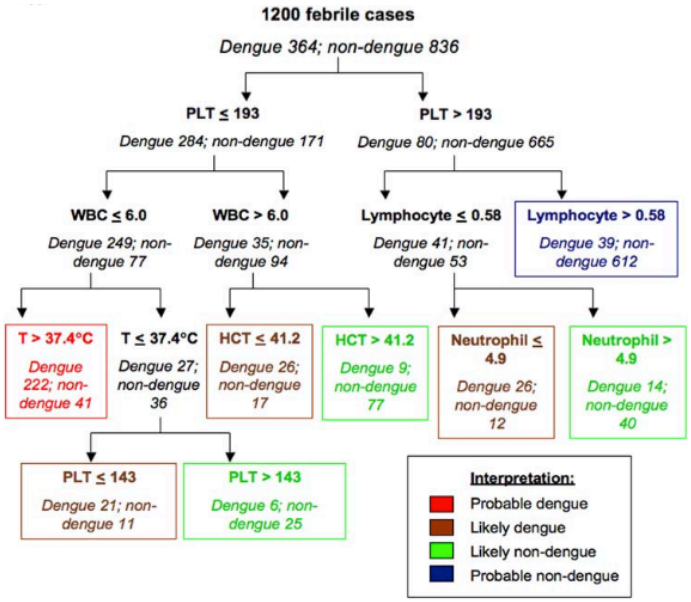
Rythm@, Business Intelligence Solutions

Using data mining to predict

- **Classification / Prediction :**
  - Extracting global models describing important data classes or to predict data future trends *(Data Mining: Concepts and techniques, J. Han et M. Kamber, 2006)*

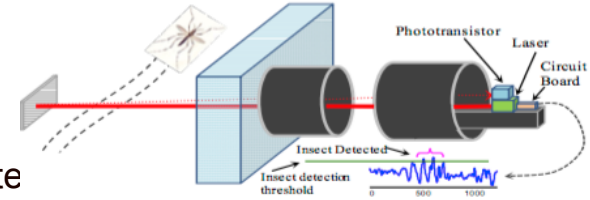
"*Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness*" (Tanner et al., PLoS Neglected Tropical Diseases, 2008)

- Objective: Differentiate dengue from other febrile illness in the primary care setting and predict severe disease
- Data:
  - 1200 patients, 4 weeks, Singapore and Vietnam
  - clinical, hematological and virological data
- Method: C4.5 decision tree (Quinlan et al., 1993)
  - 84.7% accuracy

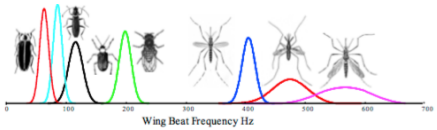


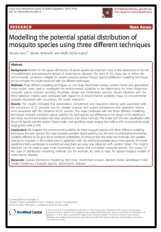
Using data mining to predict

"*Sensors and Software to Allow Computational Entomology, an Emerging Application of Data Mining*" (Batista et al., SIGKDD Demo, 2011)



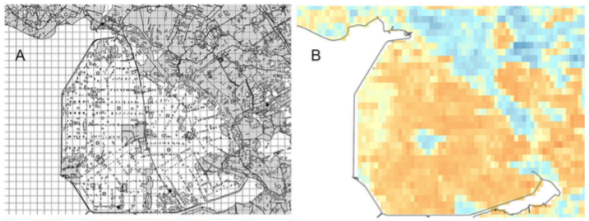
- Objective: Identify and count vectors of viruses (e.g. Aedes Aegypti)
- Data: 1220 insects, seven species of insect (e.g Drosophila melanogaste Culex quinquefascitus, Anopheles stephensi, Aedes aegypti)
- Method: Wing-beat sensor + Bayes classifier
  - 96.04% accuracy (89.92% for Aedes Aegypti)



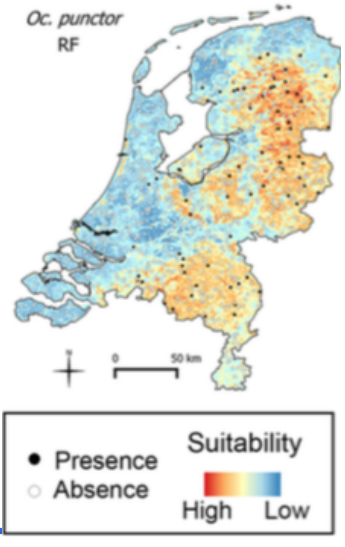


"*Modelling the potential spatial distribution of mosquito species using three different techniques*" (Cianci et al., Journal of Health Geographics, 2015)

- Objective: Model spatial distribution of vector species w.r.t. environmental conditions in Netherlands
- Data: 766 locations, 3 species
  - abundance data, population density, climate data, land cover, remote sensing data



- Methods: Generalized linear model, non-linear discriminant analysis (Rogers, 2000), random forest (Breiman, 2001)
  - best accuracy with random forest (89% - 94%)

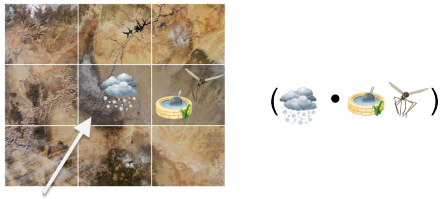


Using data mining to understand / explain

- **Mining frequent patterns, associations and correlations :**
  - Search for recurring relationships in a given data (i.e. "local models")

"*The Pattern Next Door: Towards Spatio-sequential Pattern Discovery*"
(Alatrista Salas et al., PAKDD, 2012)

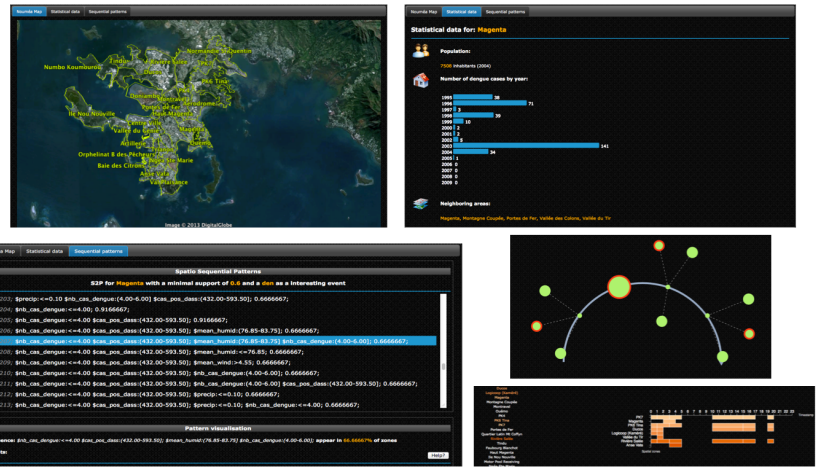
- Objective: Mine frequent evolutions over time w.r.t. environmental conditions of locations and their neighborhoods
  - application to a dengue epidemic in the districts of Nouméa
  - *(application to the monitoring of Saône river water quality)*



- Data: 32 districts of Nouméa, 2003
  - population data, entomological data, climate data, land cover, dengue cases

- Methods: Sequential pattern mining

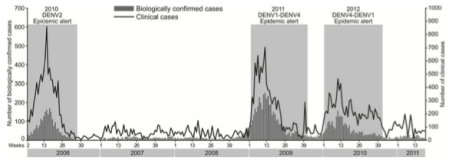
<(urban_water) (pt_trash) (Dengue)>
<(urban_water) (Θ.pt_trash) (Θ.Dengue)>



Using data mining to understand / explain

"*Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana*" (Flamand et al., Journal of the American Medical Informatic Association, 2014)

- Objective: Identify local meteorological drivers of Dengue Fever in French Guiana
- Data: 20 territories, 6 meteorological stations, 2006 – 2011 weekly
  - individual information, dengue serotype, clinical cases, rainfall, min-max temperatures, sunstroke average, relative humidity, etc



- Methods:
  - statistical analysis (bivarariate analyses, spearman rank correlation method)
  - extraction of frequent sequence of events for different context (pre-epidemic, epidemic, descending phase, …)

➢ showed temporal association between local weather conditions and evolution of dengue

relatively high (but not too extreme, 158 mm – 327 mm) level of cumulative rainfall frequently associated with beginning of outbreaks

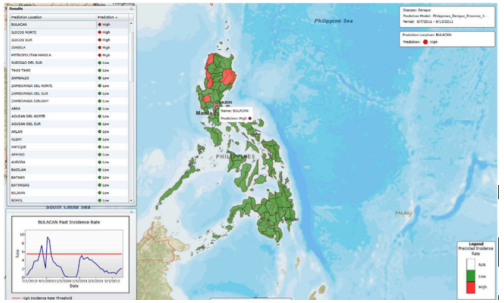
Minimal context	Epidemic associated sequential patterns	Support	C-specificity
Beginning of epidemic (4-week period)	(Var_BCC>40%) (BCC_i (0.3–1.9‰))	0.76	0.56
	(BCC_i (0.3–1.9‰)) (BCC_i (0.3–1.9‰))	0.86	0.51
	(BCC_i (0.3–1.9‰)) (Var_UX (0.1%; 0.4%))>	0.71	0.46
	(Var_BCC>40%) (RR (158–327 mm))	0.67	0.18
	(Var_BCC>40%) (Var_CC (0–33%)	0.81	0.16
	(Var_BCC>40%) (Var_UN (7–40%))	0.67	0.09
	(Var_BCC>40%) (UN (62–67%))	0.57	0.05
	(Var_GLOT>12%)	0.62	0.02
	(Var_BCC>40%) (Var_UX (0.1–0.4%))	0.62	0.01
	(UX<96%)	0.57	0.01
Epidemic peak (7-week period)	(BCC>8) (TX (30.3°; 31.2°), BCC>8)	0.6	0.25
	(BCC_i (1.8‰; 4.3‰)) (BCC_i (1.8‰; 4.3‰))	0.6	0.17
	(UN (62–67%)) (BCC>8)	0.65	0.10
	(Var_BCC (1–40%))	0.8	0.04
	(Var_GLOT>(3–12%)))	0.65	0.04
Descendant phase	(Var_BCC<−33%) (Var_UN (2–7%))	0.85	0.30
	(Var_BCC<−33%) (Var_TX (0%))	0.85	0.30
	(Var_BCC<−33%) (Var_BCC<−33%)	0.90	0.23
	(Var_CC (−4–0%))	0.70	0.21
	(Var_BCC<−33%) (Var_TX>2%)>	0.85	0.20

BCC, biologically confirmed case; CC, clinical case; GLOT, global brilliance; RR, cumulative rainfall; TX, maximum temperature; UN/UX, minimum and maximum relative humidity.

Hybrid approaches: pattern-based classification

- **Using local patterns as input data to a classifier**
  - ➢ improving classification results

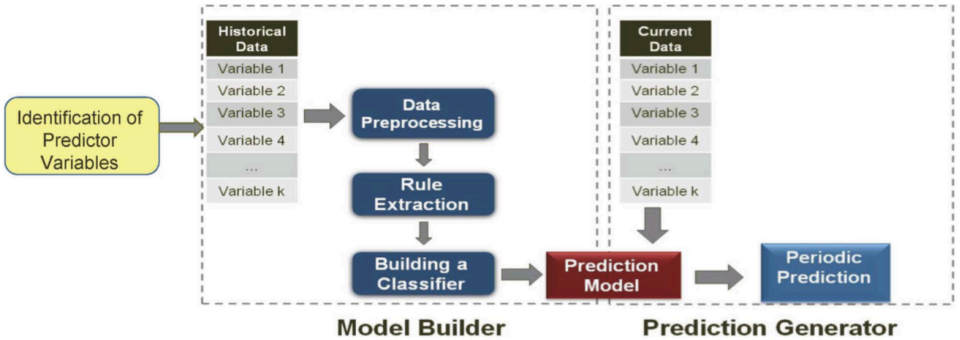
"*Prediction of High Incidence of Dengue in the Philippines*"
(Buczak et al., PLOS Neglected Tropical Diseases, 2014)

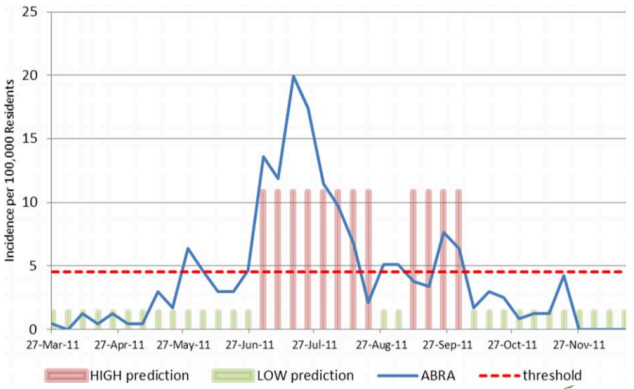


- Objective: Predict dengue incidence levels weeks in advance of an outbreak
- Data: 40 provinces, 2003 – 2011 weekly
  - rainfall, temperature, typhoon status and wind, NDVI, EVI, Southern Oscillation Index, sea surface temperature anomaly, altitude, socio-economic data and political stability data
- Methods: fuzzy association rule-based classifier **IF(X is *A)* ➔ (Y is *B*)**

IF (Past Dengue Incidence T1 **is** HIGH) **AND** (Adjacent Dengue Incidence T1 **is** HIGH) **AND** (Max Typhoon Status T3 **is** NO) **AND** (Week **is** WEEK 30-33) ➔ (Predicted Dengue Incidence T4 **is** HIGH),

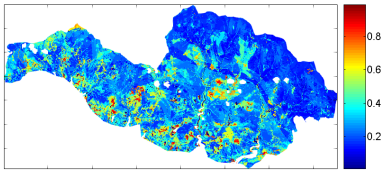
confidence = 0.9, support = 0.002, lift = 27.4

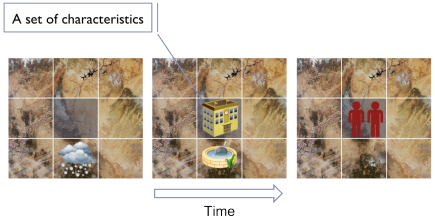




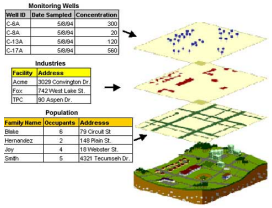
Conclusion & Perspectives

- Data Science and Data Mining develop methods/algorithms to analyze complex data
  - Predict / Classify
  - Understand / Explain

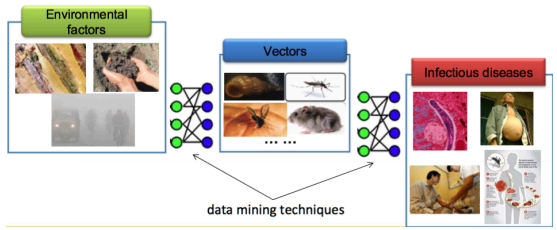




- Benefits
  - Give relevant insight for various temporal intervals and spatial units
  - Avoid analysts to multiply stratified analyses with traditional methods
  - Enable analysis of many outcomes and various explanatory variables simultaneously
  - Avoid a priori hypotheses on variables
  - Quantify known relationships and discover new ones



- Transfer of methods from data scientist communities to public health communities is still ongoing
  - Several ongoing research projects worldwide
    - e.g. "Monitoring and Early Warning of Vector-borne Diseases in China by Earth Observation Data Mining" 2012 – 2016
  - In current works, "classical" data science methods are mainly used



- Lot of perspectives w.r.t. Data Science and vector-borne diseases
  - Using recent data science methods for big data and more complex data (e.g. highly heterogeneous, unbalanced, noisy)
  - Coupling data science, epidemiological models and simulation