Data Analytics and Models for Insurance

Presentation of the research chair

Christian ROBERT

ISFA-COLUMBIA Workshop
Monday June 27, 2016 - Lyon

2015 - 2020

CHAIR OF EXCELLENCE

Data Analytics & Models for Insurance

BNP PARIBAS CARDIF | LABORATOIRE SAF

Université Claude Bernard Lyon 1

Institut Louis Bachelier

2010 - 2015

Management of modelling in Life-insurance

BNP PARIBAS CARDIF

I.S.F.A.

Université Claude Bernard Lyon 1

LYON 1 FONDATION

Frédéric PLANCHET
Co-responsable scientifique
Membre du Laboratoire SAF

Professeur des Universités
ISFA, Université Lyon 1

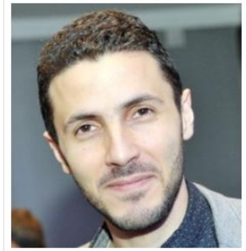
Christian ROBERT
Co-responsable scientifique
Directeur du Laboratoire SAF

Professeur des Universités
ISFA, Université Lyon 1

Alexis BIENVENUE
Membre du Laboratoire SAF

Maître de conférences
ISFA, Université Lyon 1

Nabil KAZI-TANI
Membre du Laboratoire SAF



Maître de conférences
ISFA, Université Lyon 1

Stéphane LOISEL
Membre du Laboratoire SAF

Professeur des Universités
ISFA, Université Lyon 1

Xavier MILHAUD
Membre du Laboratoire SAF

Maître de conférences associé
ISFA, Université Lyon 1

Didier RULLIERE
Membre du Laboratoire SAF

Maître de conférences
ISFA, Université Lyon 1

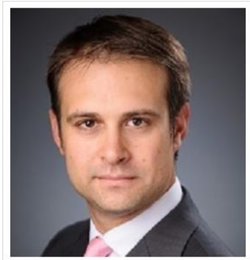
Jean-Louis RULLIERE
Membre du Laboratoire SAF

Professeur des Universités
ISFA, Université Lyon 1

Yahia SALHI
Membre du Laboratoire SAF

Maître de conférences associé
ISFA, Université Lyon 1

Pierre THEROND
Membre du Laboratoire SAF

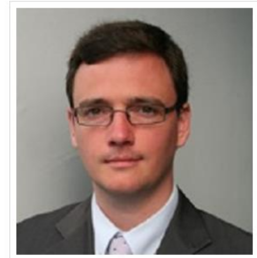


Enseignant-chercheur associé
ISFA, Université Lyon 1

Christophe GEISSLER
Membre associé

CEO,
Advestis

Donatien HAINAUT
Membre associé



Professeur de Finance
ESC Rennes

Bernard BOLLE-REDDAT
Membre du comité de pilotage

Chief Risks Officer,
BNP Paribas Cardif

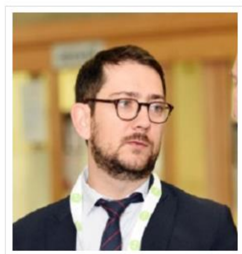
Michaël DE TOLDI
Membre du comité de pilotage

Chief Data Officer,
BNP Paribas Cardif

Jean-Paul FELIX
Membre du comité de pilotage

Head of Risk Tools & Processes ,
BNP Paribas Cardif

Sébastien CONORT
Membre associé

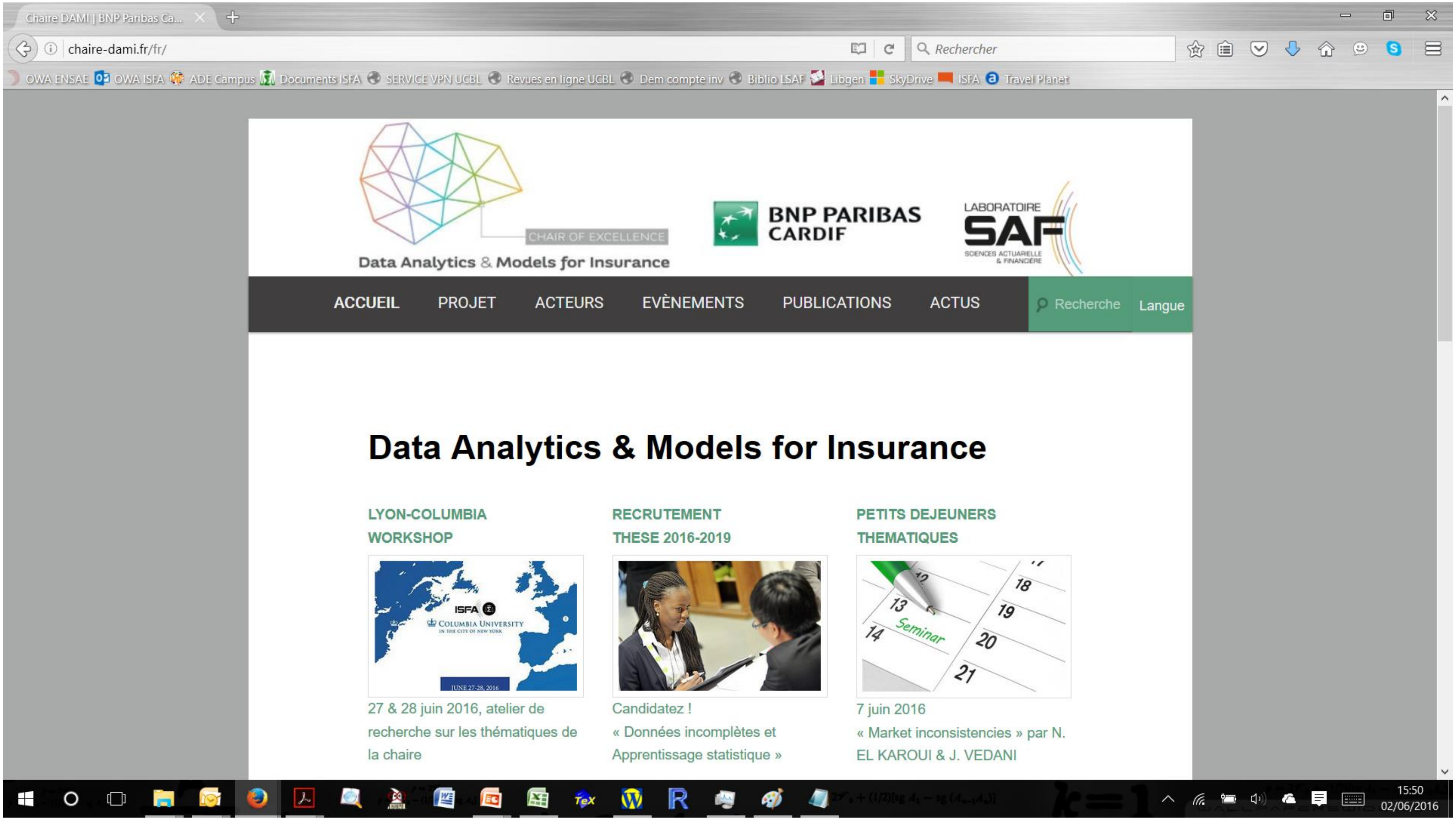


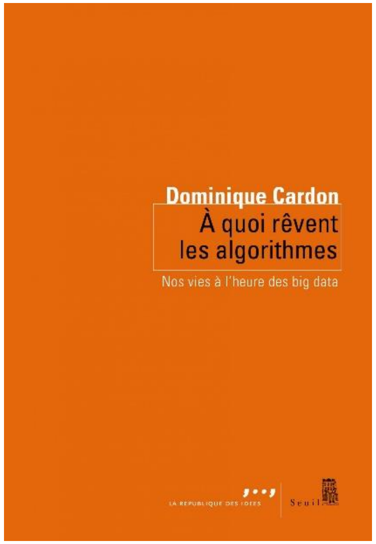
Chief Data scientist,
BNP Paribas Cardif

Lam DANG
BNP Paribas Cardif

Data scientist,
BNP Paribas Cardif

chaire-dami.fr





March 15, 2016

Seminar – Breakfast – « Politics of algorithms » by Dominique Cardon

HAL
archives-ouvertes.fr

Market inconsistencies of the market-consistent European life insurance economic valuations: pitfalls and practical solutions

Nicole El Karoui, Stéphane Loisel, Jean-Luc Prigent, Julien Vedani

▶ To cite this version:

Nicole El Karoui, Stéphane Loisel, Jean-Luc Prigent, Julien Vedani. Market inconsistencies of the market-consistent European life insurance economic valuations: pitfalls and practical solutions. 2015. <hal-01242023>

HAL Id: hal-01242023
https://hal.archives-ouvertes.fr/hal-01242023

Submitted on 11 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

June 7, 2016

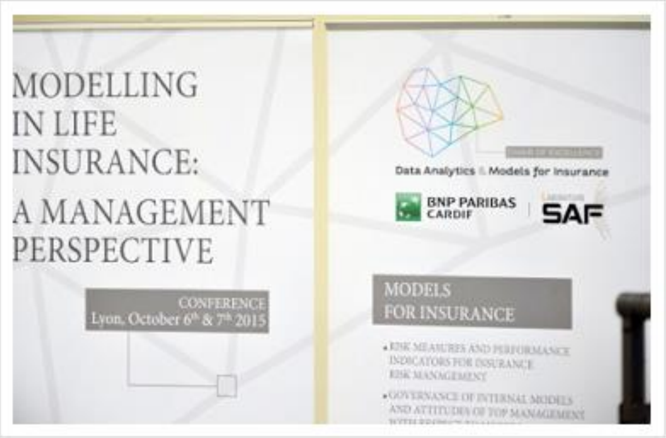
Seminar – Breakfast– « Market inconsistencies » by Nicole El Karoui & Julien Védani

March 23, 2016 – Topics

- Market inconsistencies of the market-consistent European life insurance economic valuations
- Proxys for SII
- Impact of volatility clustering on equity indexed annuities
- Assessment of beneficiary clauses in free text via Text Mining
- Optimization of treatment of web leads queue with scoring and simulation
- The experiments for observation of human behaviors

March 25 2015 – Topics

- Credit Losses Impairment
- Agents attitudes towards risk and models: Study of a new analysis and comparison
- Asymmetry & Big Data : which impact for insurance ?
- Working group on the risk-neutral approach
- Longevity risk
- Financial information and Risk in insurance : Change for the better and for worse
- Kaggle AXA competition : methodology of the research lab



October 6 & 7, 2015

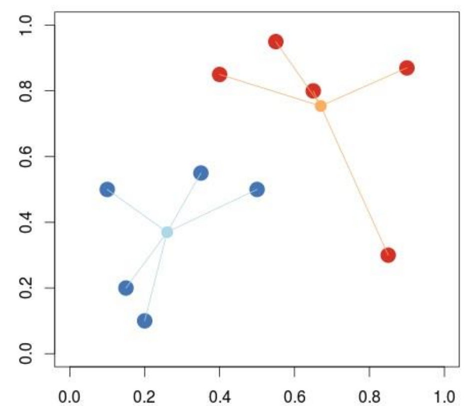
David INGRAM (Willis Re) « Bridging the gap between managers and models »
Bernard BOLLE-REDDAT (BNP Paribas Cardif) « Management and models »
Clément PETIT – Guillaume ALABERGERE (ACPR) « Validation in life modelling, a supervisory point of view »
Antoon PELSSER (Maastricht University) « The difference between LSMC and replicating portfolio in insurance liability modelling »
Michaël SCHMUTZ (FINMA) « Group solvency tests, intragroup transfers and intragroup diversification: A set-valued perspective »
Georges DIONNE (HEC Montréal) « Governance of risk management »
Thomas BREUER (FHV) « Systemic stress testing and model risk »
Andreas TSANAKAS (Cass Business School) « Model risk & culture »
Michaël de TOLDI (BNP Paribas Cardif) « Governance for data & analytics in insurance »

Risk measures and performance indicators for insurance risk management

Governance of internal models and attitudes of top management with respect to models

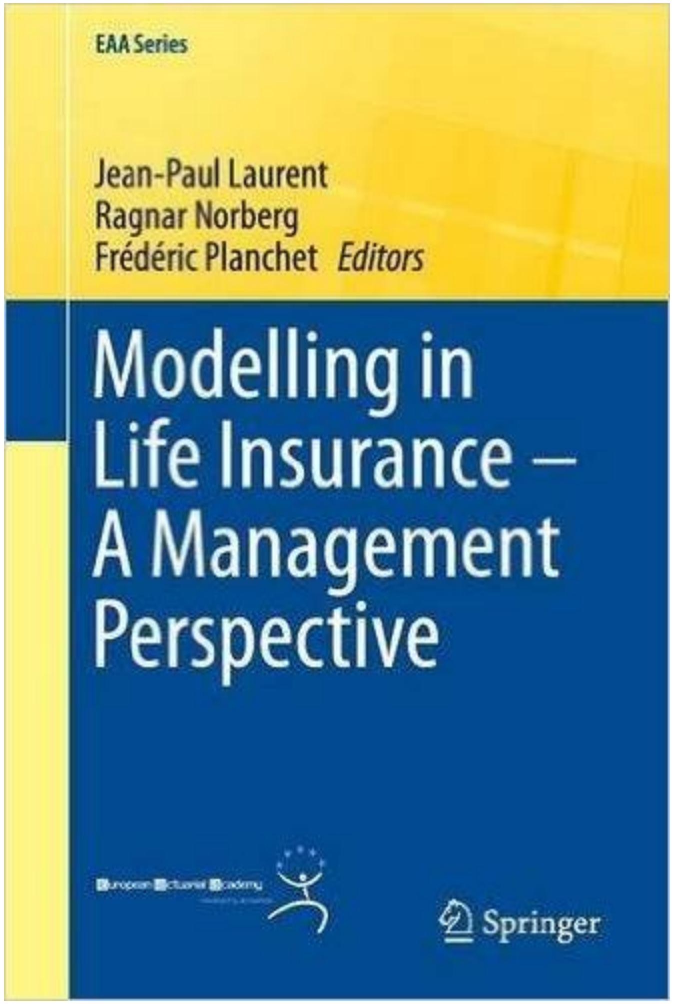
Insurance models

the impact of the regulatory and accounting environment on their development and management

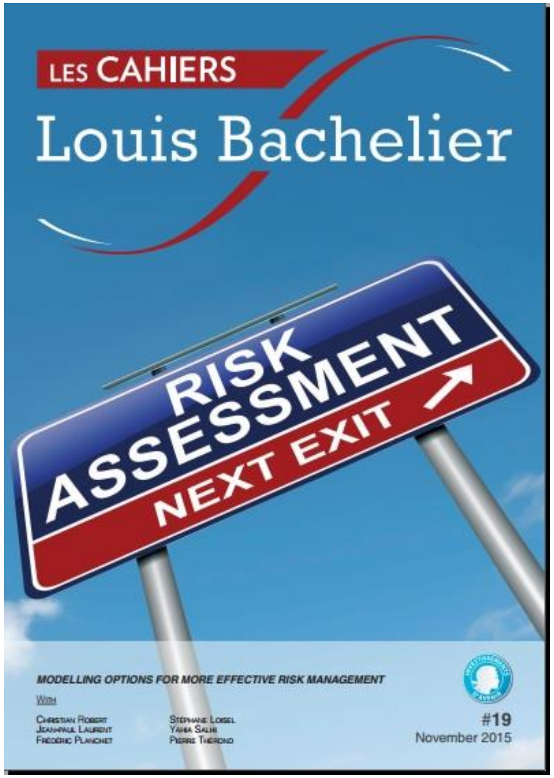


Customer behaviour and risk attitudes

Proxies, model points and advanced simulation techniques for risk management


EAA Series
Jean-Paul Laurent
Ragnar Norberg
Frédéric Planchet Editors
Modelling in Life Insurance – A Management Perspective
Springer

Contents



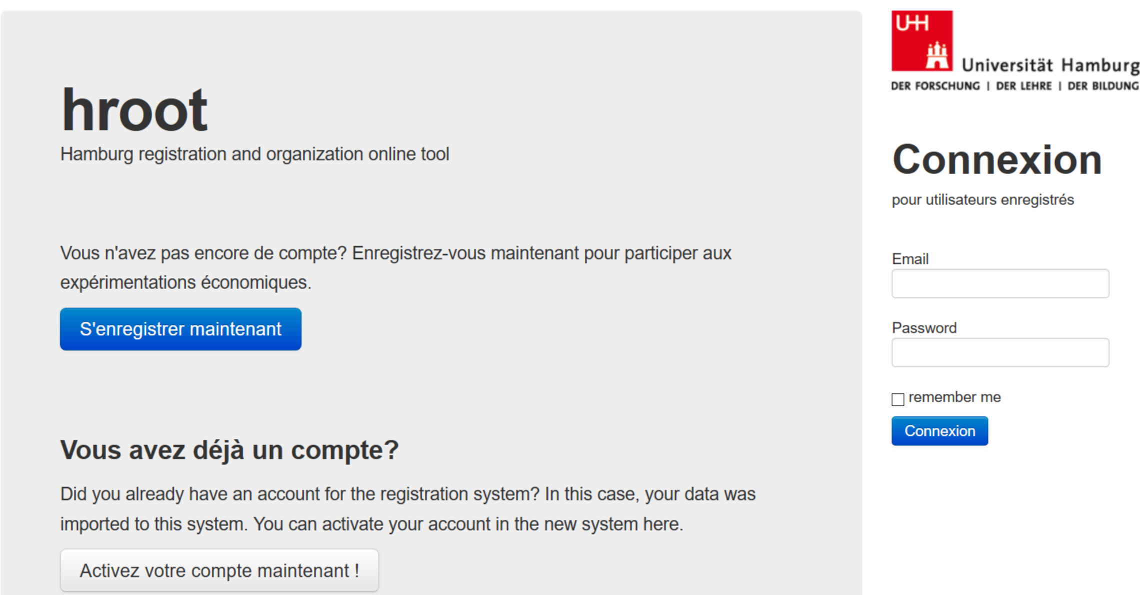
Les cahiers de l’ILB – #19 – November 2015

INDEX

Experiments in the lab

Experimental Economics is a branch of economics that focuses on individual behavior in a controlled laboratory setting or out in the field.

Experimental economics helps to prove or disprove economic theories and create predictions and insights about real-world behavior.



WHAT DO WE STUDY ?

- Individual choices

(choosing under risk, arbitrage, intertemporal choice ...)

- Strategic interactions

(Negotiation, conflict, contract, incentives, ...)

- Market designs

(trade efficiency, public good provision, market design ...)

Privacy concerns, data anonymization, open data

Data analytics in insurance

Governance for data analytics, new business models with big data and analytics

Risk-based pricing, predictive analytics, machine learning

Traineeship: Textual analysis of published and working paper in Machine Learning research

1. Identification of the leading Machine Learning research journals
2. Recovery of titles, abstracts, names of authors and their affiliations
3. Creation of a text-mining tool identifying the key issues and key research center
4. Creation of a visualization tool and mapping of research in Machine Learning in the world
5. Identification of subjects with potential applications for insurance


```
@Article{Tian2016,
author="Tian, Zhang-peng
and Wang, Jing
and Zhang, Hong-yu
and Wang, Jian-qiang",
title="Multi-criteria decision-making based on generalized
prioritized aggregation operators under simplified neutrosophic
uncertain linguistic environment",
journal="International Journal of Machine Learning and
Cybernetics",
year="2016",
pages="1--17",
abstract="A simplified neutrosophic uncertain linguistic set that
integrates quantitative and qualitative evaluation can serve as
an extension of both an uncertain linguistic variable and a
simplified neutrosophic set. It can describe the real preferences
of decision-makers and reflect their uncertainty, incompleteness
and inconsistency. This paper focuses on multi-criteria decision-
making (MCDM) problems in which the criteria occupy different
priority levels and the criteria values take the form of
simplified neutrosophic uncertain linguistic elements. Having
reviewed the relevant literatures, this paper develops some
generalized simplified neutrosophic uncertain linguistic
prioritized weighted aggregation operators and applies them to
solve MCDM problems. Finally, an illustrative example is given,
and two cases of comparison analysis are conducted with other
representative methods to demonstrate the effectiveness and
feasibility of the developed approach.",
issn="1868-808X",
doi="10.1007/s13042-016-0552-9",
url="http://dx.doi.org/10.1007/s13042-016-0552-9"
}
```

Incomplete data, Machine Learning and Insurance

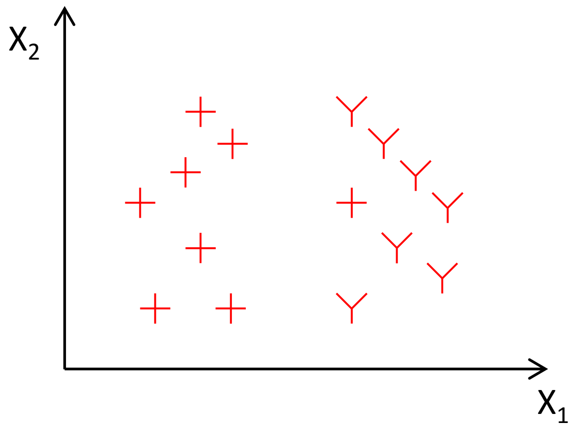
A research project on data science

Christian ROBERT

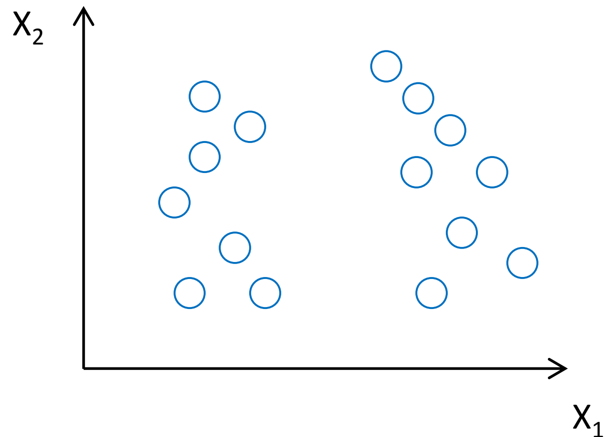
ISFA-COLUMBIA Workshop
Monday June 27, 2016 - Lyon

Data types

Labeled data



Unlabeled data

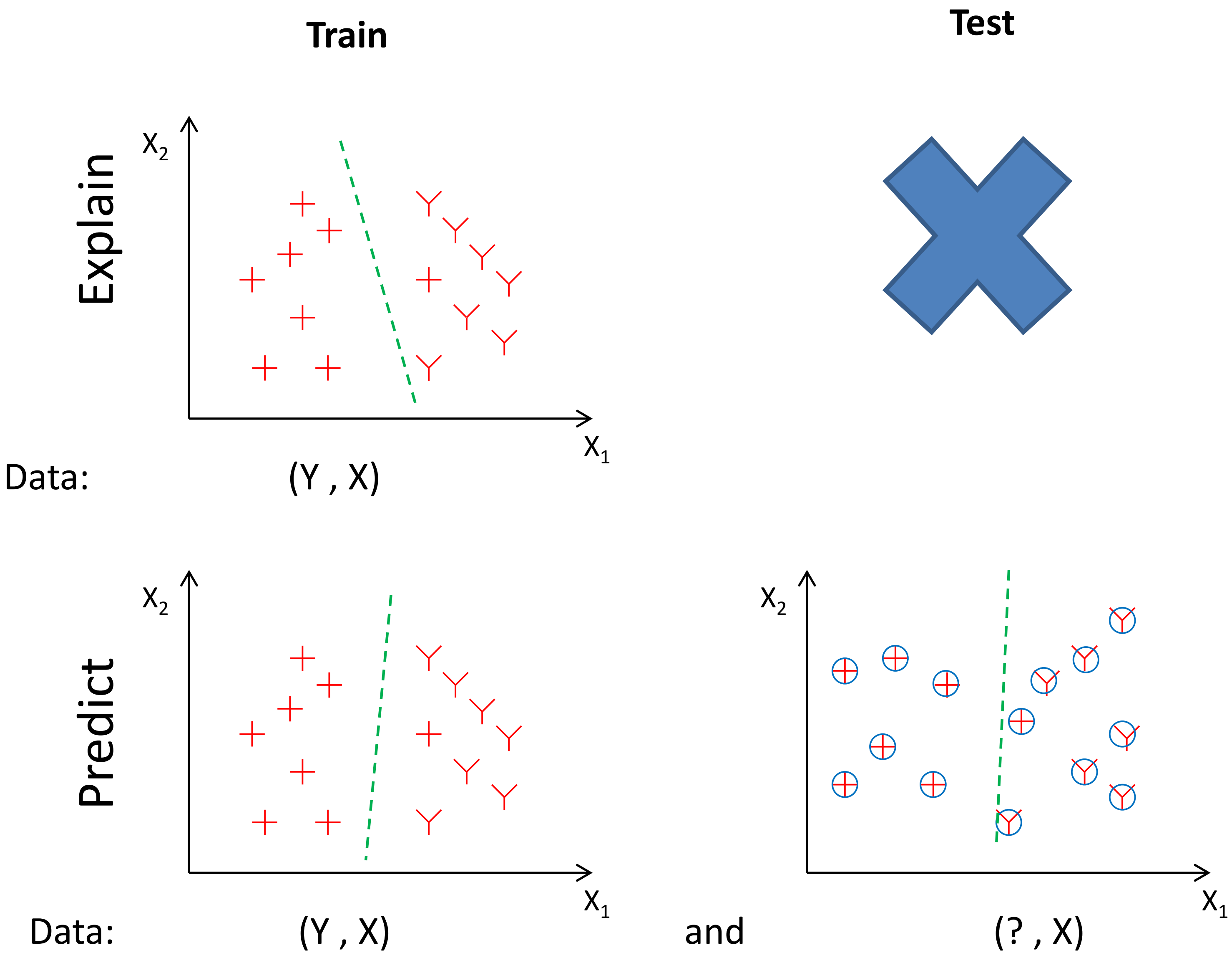


Data: (Y , X) (X)

Y : labels = + or Y , response variable, output variable

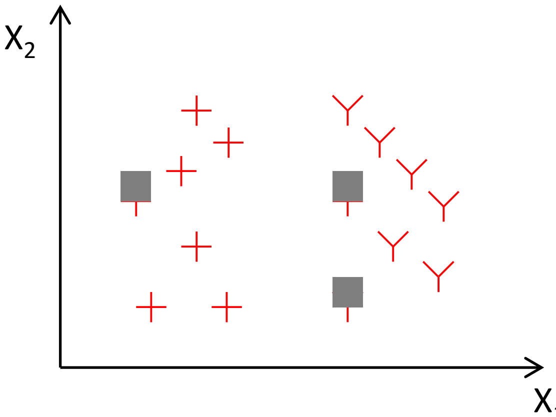
X : explanatory variables, input variables, covariates, independent variables, control variables, features,…

Data to be explained and/or to be predicted

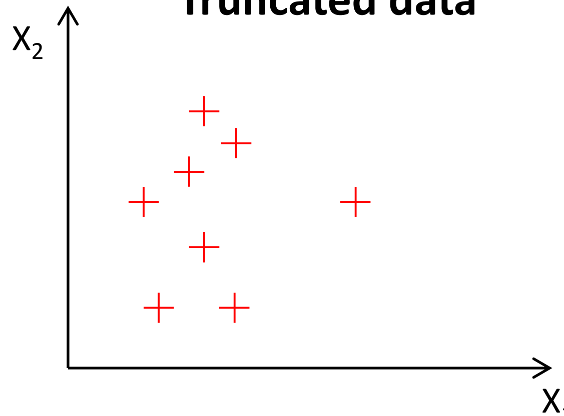


Imperfect labeled data

Censored data



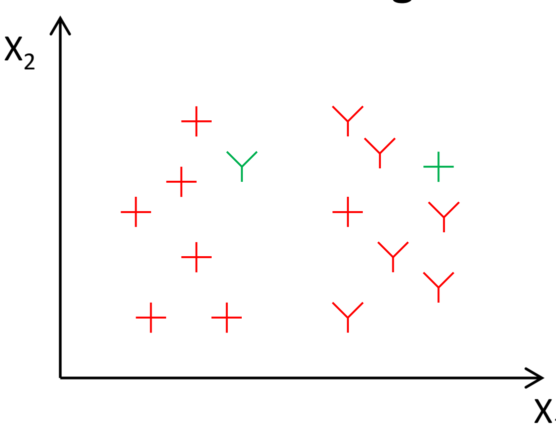
Truncated data



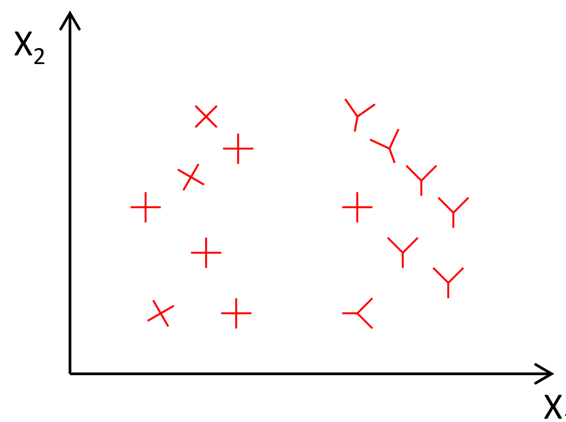
Data: $((\min(Y, C), 1_{Y > C}), X)$ with $Y \perp C$

$((Y, C, X) \mid Y > C)$ with $Y \perp C$

Random wrong label



Noisy labeled data with endogenous errors



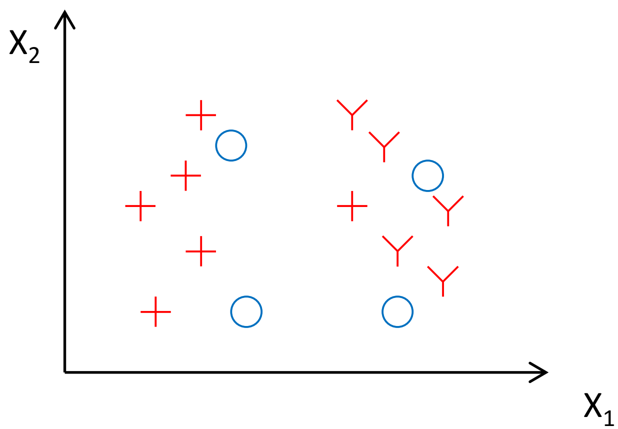
Data: $(Y^{*} = Y\, 1_{\varepsilon = 1} + Y^{\wedge}\, 1_{\varepsilon = -1}, X)$ with $Y \perp Y^{\wedge}$

$(Y^{*} = Y + \varepsilon, X)$ with $\varepsilon \not\perp X$

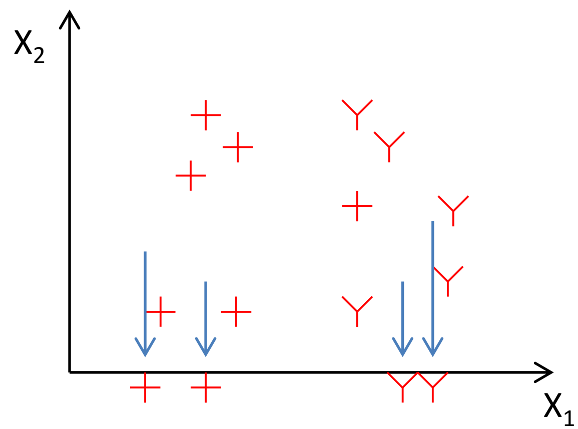
➡ Only probabilistic schemes?

Labeled with unlabeled data / Missing values

Some labels Y are not observed



Some components of X are not observed



Missing completly at random

$(Y^* = Y\, 1_{\varepsilon = 1} + \emptyset\, 1_{\varepsilon = -1}\, ,\, X)\ \ \varepsilon \perp X$

$(Y\, ,\, X^* = X\, 1_{\varepsilon = 1} + \emptyset\, 1_{\varepsilon = -1}\,)\ \varepsilon \perp X$

Missing at random

$(Y^* = Y\, 1_{\varepsilon = 1} + \emptyset\, 1_{\varepsilon = -1}\, ,\, X)\ \ \varepsilon \not\perp X$

$(Y\, ,\, X^* = X\, 1_{\varepsilon = 1} + \emptyset\, 1_{\varepsilon = -1}\,)\ \varepsilon \not\perp X$

Missing not a random

$(Y^* = Y\, 1_{Y < c} + \emptyset\, 1_{Y > c}\, ,\, X)$

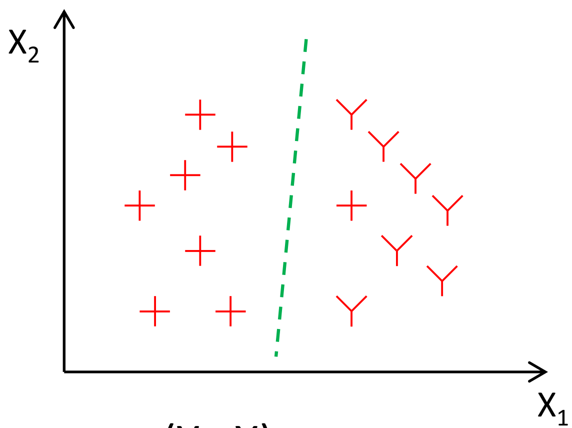
$(Y\, ,\, X^* = X\, 1_{Y < c} + \emptyset\, 1_{Y > c}\,)$

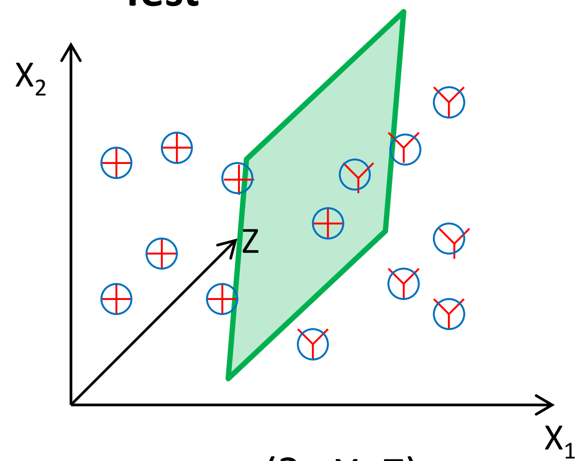
When train and test data bases differ

Train

Test

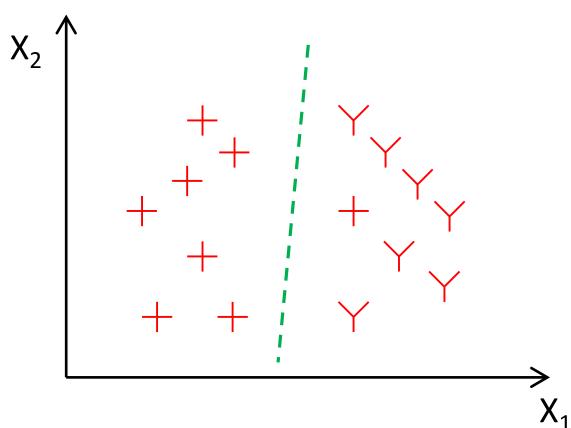
Predict controlled data

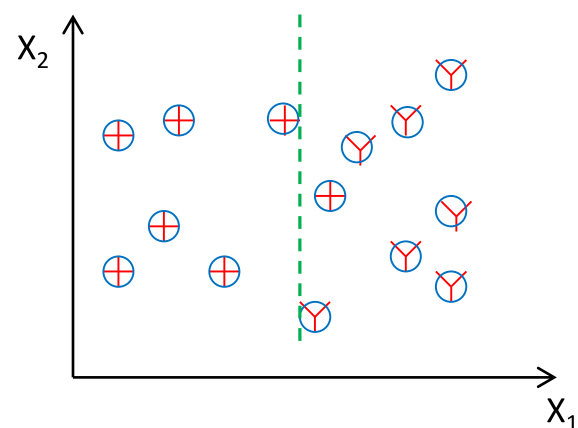




Data: (Y , X) and (? , X, Z)

Predict test data with a different generating process





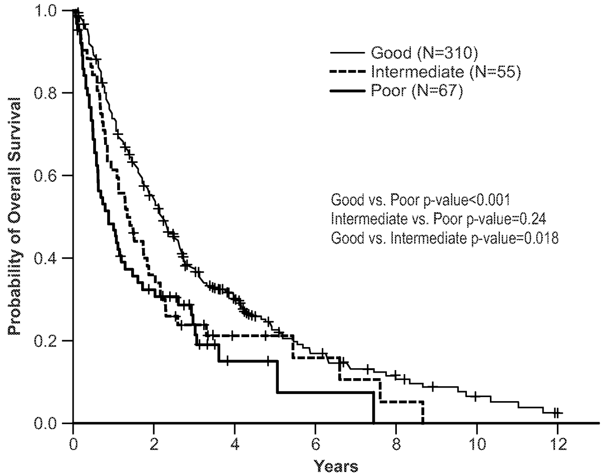
Data: (Y , X) and (? , X)

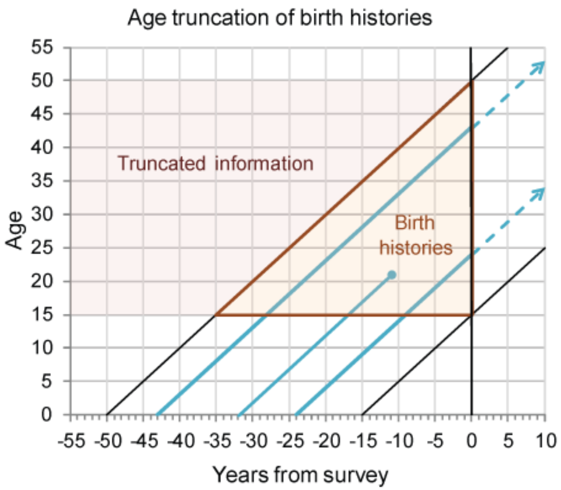
$Y = f_1(X)$

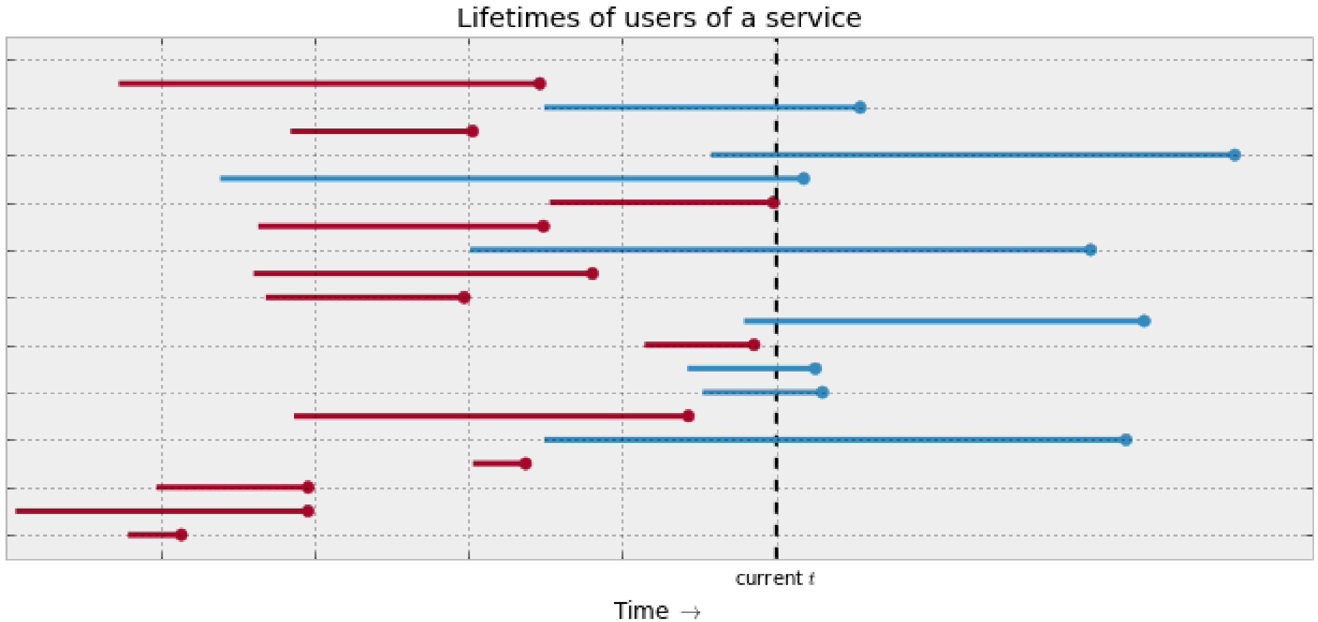
$Y = f_2(X)$ $D(f_1, f_2) < A$

Mining imperfect data in insurance

Truncated / censored data

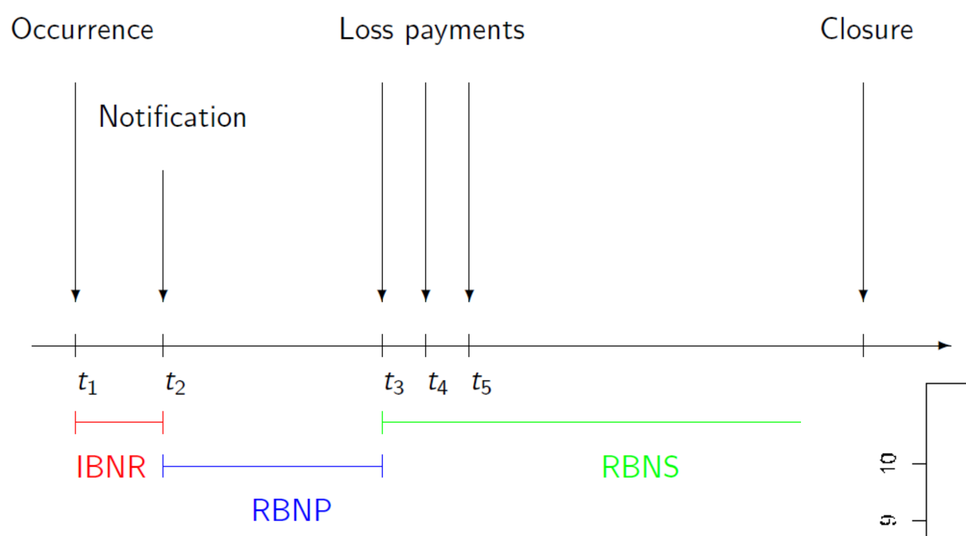




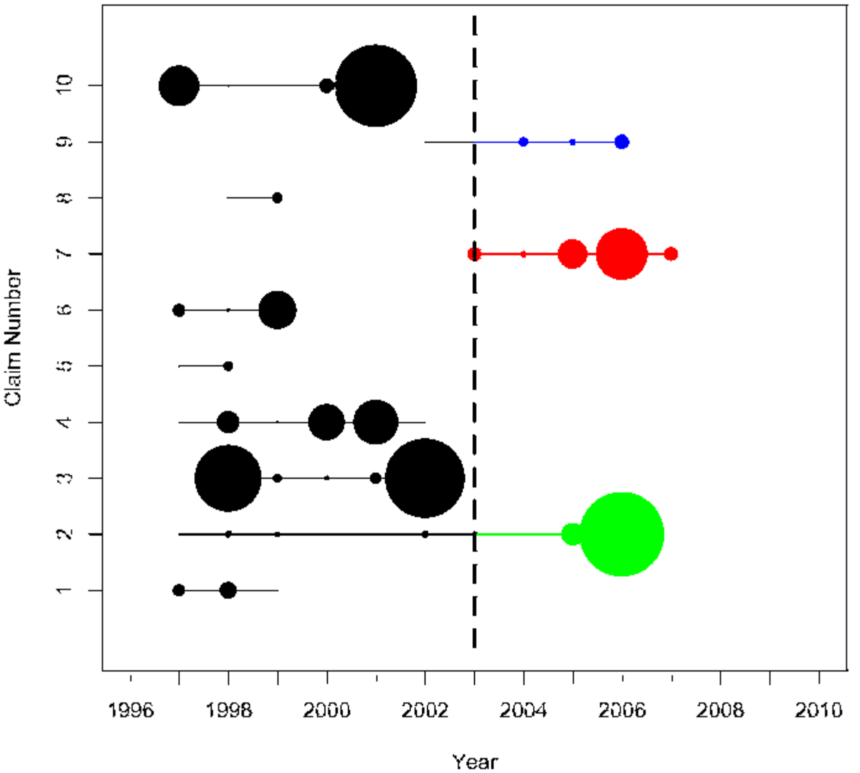


Mining imperfect data in insurance

Individual claim process



Incurred But Not Reported (IBNR) claims
Reported But Not Paid (RBNP) claims
Reported But Not Settled (RBNS) claims

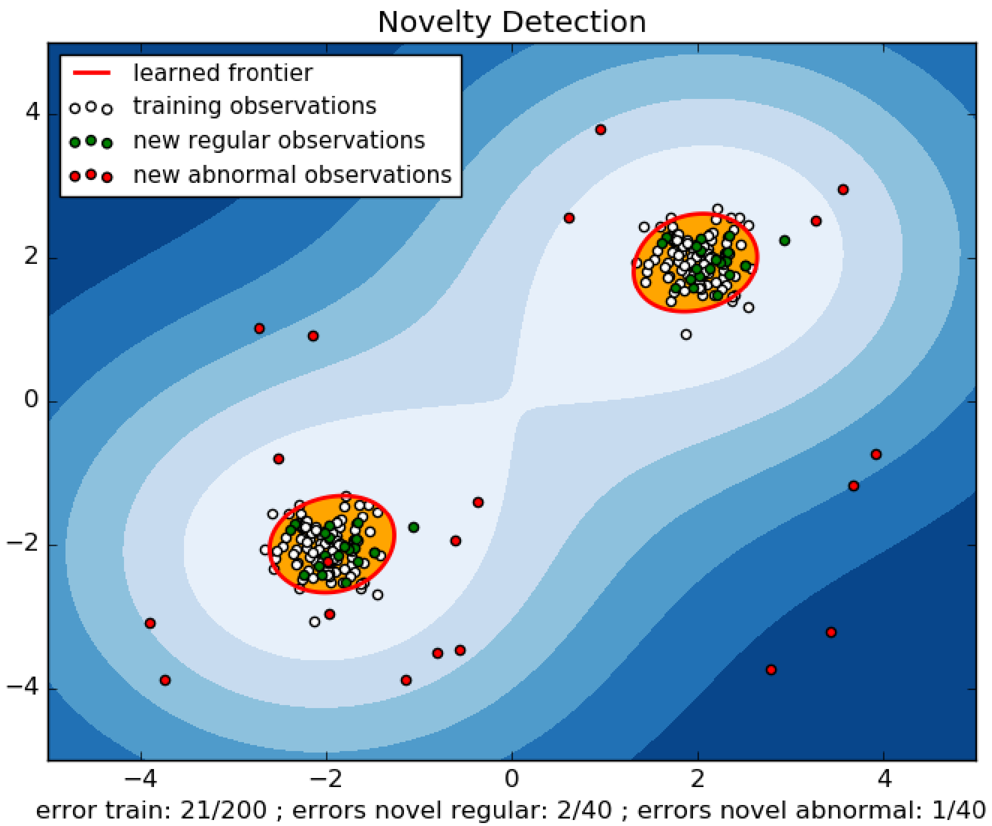


Mining imperfect data in insurance

Insurance products with several generations of policies / customers

Mining imperfect data in insurance

Novelty / Fraud detection



Machine Learning vs Statistics/Econometrics

Subfields

Machine Learning *is a subfield of computer science and artificial intelligence which deals with building systems that can learn from data, instead of explicitly programmed instructions.*

Statistical Modelling *is a subfield of mathematics which deals with finding relationship between variables to predict an outcome*

Data mechanism/data generating process

Machine Learning *uses algorithmic models and treats the data mechanism as unknown.*

Statistical Modelling *assumes that the data are generated by a given stochastic data model.*

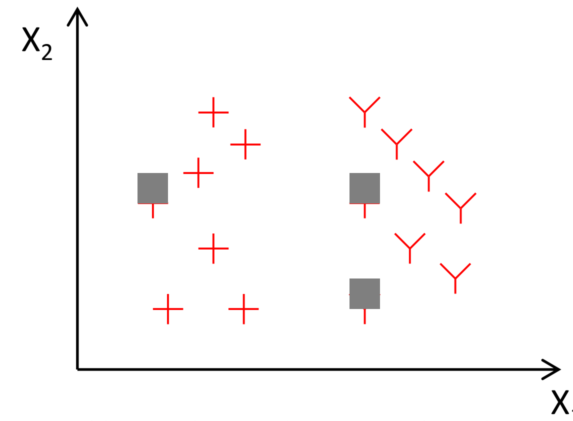
Model choice

Machine Learning *focuses on Predictive Accuracy even in the face of lack of interpretability of models. Model Choice is based on Cross Validation of Predictive Accuracy using Partitioned Data Sets.*

Statistical Modelling *focuses on hypothesis testing of causes and effects and interpretability of models. Model Choice is based on parameter significance and/or confidence intervals, and In-sample Goodness-of-fit.*

Tree-based censored regression/Survival random forest

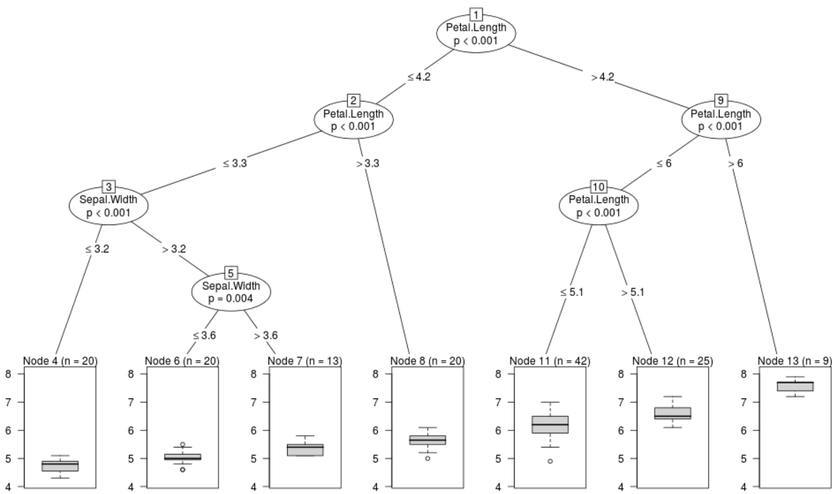
• Random forests have been extended to the survival context by Ishwaran et al. (2008), who prove consistency of **Random Survival Forests** (RSF) algorithm assuming that all variables are categorical.



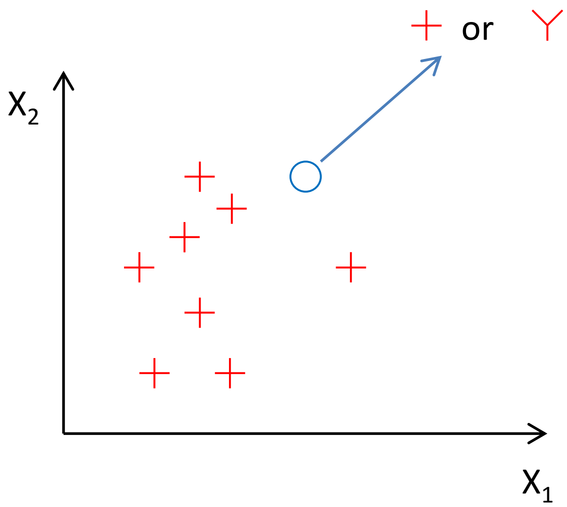
$((\min(Y , C), 1_{Y > C}), X)$ with $Y \perp C$

• Yang et al. (2010) showed that by incorporating **kernel functions into RSF**, their algorithm KIRSF achieves better results in many situations.

• Lopez et al. (2015) used an approach that is based on the **IPCW strategy** (Inverse Probability of Censoring Weighting") and that consists in determining **a weighting scheme** that compensates the lack of complete observations in the sample.



One-class classification



One-class classification tries to identify objects of a specific class amongst all objects, by learning from a **training set containing only the objects of that class**.

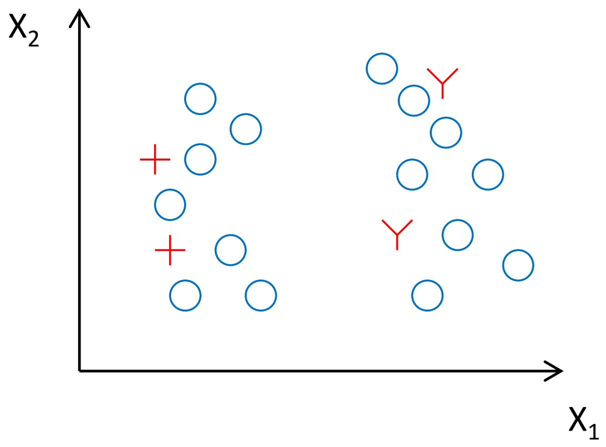
It is also known as Outlier detection, Novelty detection, Concept learning, Single class classification, or Unary classification.

An example is the automatic diagnosis of a disease. It is relatively easy to compile positive data (all patients who are known to have a 'common' disease) but negative data may be difficult to obtain since other patients in the database cannot be assumed to be negative cases if they have never been tested, and such tests can be expensive.

<u>Algorithms that can be used</u>

- One-class Support Vector Machines (OSVMs)
- Neural networks
- Decision trees
- Nearest neighbors

Semi-supervised learning



It is a class of supervised learning tasks and techniques that also make use of unlabeled data for training – **typically a small amount of labeled data with a large amount of unlabeled data**.

Goal: Using both labeled and unlabeled data to build better learners, than using each one alone.

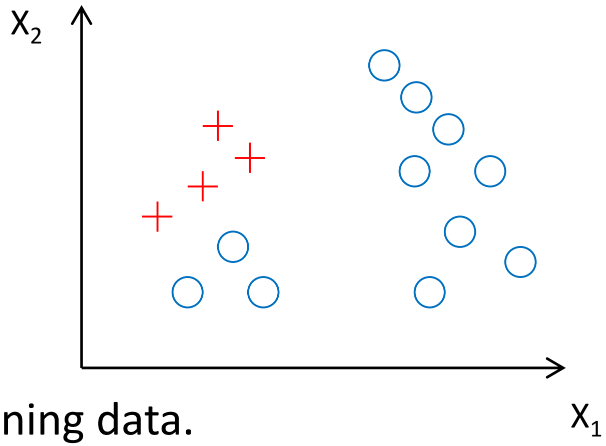
In order to make any use of unlabeled data, it is implicitly assumed some structure to the underlying distribution of data: Smoothness assumption, Cluster assumption, Manifold assumption.

<u>Algorithms that can be used</u>

- self-training models,
- EM with generative mixture,
- co-training,
- transductive support vector machines,
- graph-based methods.

Learning from Positive and Unlabeled data

One has a set of examples of a class +, and a set of unlabeled examples with instances of a class + and also not from + (negative examples).



Goal: Build a classifier to classify the unlabeled examples and/or future (test) data.

Key feature of the problem: no labeled negative training data.

This problem is known as PU-learning.

An example is when a company has a database with details on its customer – positive examples, and a database with details on individuals who are not customers, but could become or not customers if they were proposed some products.

2-step strategy for text classification
Step 1: Identifying a set of reliable negative documents from the unlabeled set.
Step 2: Building a sequence of classifiers by iteratively applying a classification algorithm and then selecting a good classifier.