# Machine Learning for Insurers and Actuaries

Arthur Charpentier

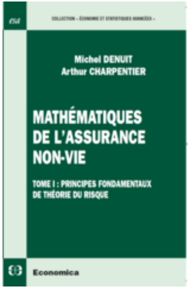
2025

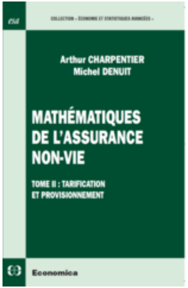
SCOR FOUNDATION FOR SCIENCE

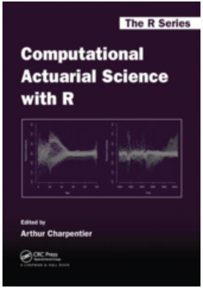
# Short Bio

**Arthur Charpentier** Professor at Université du Québec à Montréal

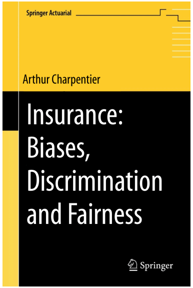
- Denuit and Charpentier (2004, 2005) Mathématiques de l'Assurance Non-Vie,
- Charpentier (2014) Computational Actuarial Science with R,
- Bénéplanc et al. (2022) Manuel d'Assurance,
- Charpentier (2024) Insurance: Biases, Discrimination and Fairness.







# Short Bio

- Insurance, Biases, Discrimination and Fairness (Szkoła Nauk Aktuarialnych, ENSAE Paris, UQAM Montréal)
- Econometrics and Machine Learning (Società Italiana di Econometria, Universitat de Barcelona)
- Insurance Data Science (Summer School of the Swiss Association of Actuaries)
- Data Science for Actuaries (AXA Group, Singapore, Istanbul, Paris)
- Former senior editor Journal of Risk and Insurance
- Member of the scientific committee Insurance Data Science Conference
- Recent fundings from SCOR Foundation for Science and AXA Research Fund

# Preambule

Runing some codes is (almost) easy. Having mathematical guarantees is challenging.

See, e.g., François-Lavet et al. (2018)

"$I(m) = \mathbb{E}_{\boldsymbol{X}}\big[\mathbb{E}_{\mathcal{D}}\big[\mathbb{E}_{Y|\boldsymbol{X}}\big[\ell(Y, m(\boldsymbol{X}) \mid \mathcal{D})\big]\big]\big]$"

corresponding to the bias-variance decomposition if $\ell$ is the square loss.

There may be a few mathematical formulas in the presentation

There could be some "⭐⭐⭐" to mention mathematical digressions

### 2.1 Supervised learning and the concepts of bias and overfitting

In its most abstract form, supervised learning consists in finding a function $f : \mathcal{X} \to \mathcal{Y}$ that takes as input $x \in \mathcal{X}$ and gives as output $y \in \mathcal{Y}$ ($\mathcal{X}$ and $\mathcal{Y}$ depend on the application):

$$y = f(x). \tag{2.1}$$

A supervised learning algorithm can be viewed as a function that maps a dataset $D_{LS}$ of learning samples $(x, y) \overset{\text{i.i.d.}}{\sim} (X, Y)$ into a model. The prediction of such a model at a point $x \in \mathcal{X}$ of the input space is denoted by $f(x \mid D_{LS})$. Assuming a random sampling scheme, s.t. $D_{LS} \sim \mathcal{D}_{LS}$, $f(x \mid D_{LS})$ is a random variable, and so is its average error over the input space. The expected value of this quantity is given by:

$$I[f] = \mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{D_{LS}} \mathop{\mathbb{E}}_{Y|X} L\left(Y, f(X \mid D_{LS})\right), \tag{2.2}$$

where $L(\cdot,\cdot)$ is the loss function. If $L(y, \hat{y}) = (y - \hat{y})^2$, the error decomposes naturally into a sum of a bias term and a variance term[1]. This bias-variance decomposition can be useful because it highlights a tradeoff between an error due to erroneous assumptions in the model selection/learning algorithm (the bias) and an error due to the fact that only a finite set of data is available to learn that model (the parametric variance). Note that the parametric variance is also called the overfitting error[2]. Even though there is no such direct decomposition for other loss functions (James, 2003), there is always a tradeoff between a sufficiently rich model (to reduce the model bias, which is present even when the amount of data would be unlimited) and a model not too complex (so as

# Overview and Roadmap

3 **Learning, with an actuarial perspective**

- what are the different concepts (buzz words) we hear in actuarial science, and insurance, related to machine learning?

2 **Learning, with an algorithmic perspective**

- what are the popular algorithms we can use, what do they do, can I use them?

1 **Learning, with a mathematical perspective**

- what are the appropriate concepts we need to discuss?

"When we speak of the 'probability of death', the exact meaning of this expression can be defined in the following way only. We must not think of an individual, but of a certain class as a whole [...] The phrase 'probability of death', when it refers to a single person, has no meaning for us at all," von Mises (1928, 1939)

"If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class," Reichenbach (1971)

# Overview and Roadmap

3 **Learning, with an actuarial perspective** $\leftarrow$ important questions

- Causality and Counterfactuals
- ○ "causal" question with "observational data"
- Interpretability and Explainability
- ○ how do we interpret the outcome of a "black box" model
- Transfert Learning
- ○ train a model with $(Y_i, \boldsymbol{X}_i) \sim \mathbb{P}$ but use it in a population where $(Y_i, \boldsymbol{X}_i) \sim \mathbb{Q}$
- Calibration and Conformal Prediction
- ○ probabilistic interpretation of non-probabilistic algorithms
- Fairness and Discrimination
- ○ is it possible to insure that a model does not discriminate with respect to a sensitive (protected attribute), with big datasets (proxy discrimination)

2 **Learning, with an algorithmic perspective**

1 **Learning, with a mathematical perspective**

# Overview and Roadmap

3 **Learning, with an actuarial perspective**

2 **Learning, with an algorithmic perspective** ← important techniques

- Supervised Learning
  - loss, accuracy, cross-validation
  - regression, classification, ensemble techniques (sequential or parallel), neural nets,
- Non Supervised Learning
  - cluster and dimension reduction
- More on Learning (Adversarial, Sequential)
  - adversarial, reinforcement learning

1 **Learning, with a mathematical perspective**

# Overview and Roadmap

3 **Learning, with an actuarial perspective**

2 **Learning, with an algorithmic perspective**

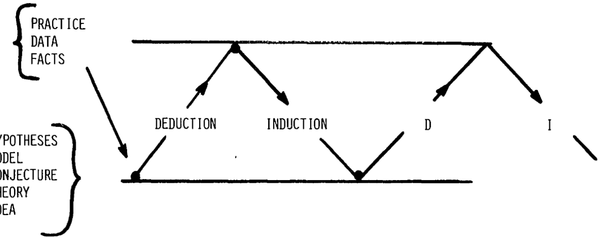
1 **Learning, with a mathematical perspective** $\leftarrow$ important concepts

- Probability and Statistics
  - epistemology of uncertainty, risk and insurance
- Simulations and Computational Probability
  - boostrap techniques, importance sampling
- Metrics, Similarities and Distances
  - distances between observations, variables, distributions
- Optimization
  - standard techniques, convex problems
- Matrices
  - properties on rectangular matrices (data) $n \times k$
  - and square matrices (operators) $k \times k$

# Data Science

"The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning...Data-driven predictions can succeed–and they can fail. It is when we deny our role in the process that the odds of failure rise. Before we demand more of our data, we need to demand more of ourselves," Silver (2012)

See Tukey (1962) on Explanatory Data Analysis




And inductive-deductive reasoning cycle from Box (1976)

THE FUTURE OF DATA ANALYSIS[1]

BY JOHN W. TUKEY

*Princeton University and Bell Telephone Laboratories*

I. General Considerations
1. Introduction
2. Special growth areas
3. How can new data analysis be initiated?
4. Sciences, mathematics, and the arts
5. Dangers of optimization
6. Why optimization?
7. The absence of judgment
8. The reflection of judgment upon theory
9. Teaching data analysis
10. Practicing data analysis
11. Facing uncertainty
II. Spotty Data
12. What is it?
13. An appropriate step forward
14. Trimming and Winsorizing samples
15. How soon should such techniques be put into service?
III. Spotty Data in More Complex Situations
16. Modified normal plotting
17. Automated examination
18. FUNOP
19. FUNOR-FUNOM in a two-way table
20. Example of use of FUNOR-FUNOM
IV. Multiple-Response Data
21. Where are we, and why?
22. The case of two samples
23. Factor analysis: the two parts
24. Factor analysis: regression
25. Factor analysis: the middle lines
26. Taxonomy; classification; incomplete data
V. Some Other Promising Areas
27. Stochastic-process data
28. Selection and screening problems
29. External, internal, and confounded estimates of error
30. The consequences of half-normal plotting
31. Heterogeneous data
32. Two samples with unequal variability
VI. Flexibility of Attack
33. Choice of modes of expression
34. Sizes, nomination, budgeting
35. A caveat about indications
36. FUNOP as an aid to group comparison
37. Continuation
VII. A Specific Sort of Flexibility
38. The vacuum cleaner
39. Vacuum cleaning: the subprocedure
40. The basic vacuum cleaner, and its attachments
41. The vacuum cleaner: an example
42. The example continued
VIII. How Shall We Proceed?
43. What are the necessary tools?
44. The role of empirical sampling
45. What are the necessary attitudes?
46. How might data analysis be taught?
47. The impact of the computer
48. What of the future?
References

# Regression or Machine Learning?

- Econometrics Working (1927), Tinbergen (1939) and, as in Morgan (1990):

"it has been considered legitimate to use some of the tools developed in statistical theory without accepting the very foundation upon which statistical theory is built [...] The reluctance among economists to accept probability models as a basis for economic research has, it seems, been founded upon a very narrow concept of probability and random variables," Haavelmo (1944)

- Machine Learning ("data mining" in Friedman (1998)), Charpentier et al. (2018):

"logistic regression can also be interpreted from a probabilistic perspective," Watt et al. (2016)

- $(y_i, \boldsymbol{x}_i)$, realizations of $(Y_i, \boldsymbol{X}_i)$ on $(\Omega, \mathcal{F}, \mathbb{P})$, Gourieroux and Monfort (1995),

$$\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}[Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}] = \overbrace{\frac{\exp[\boldsymbol{x}^\top \beta]}{1 + \exp[\boldsymbol{x}^\top \beta]} = \boxed{s(\boldsymbol{x})}}^{\text{"probabilistic scores"}}$$

# Buzz Words

Types of learning:

- **Classification**: Assigning a discrete label to items. Example: Exoplanet yes or no, topics in document classification, or content in image classification.
- **Regression**: Predicting a real value. Example: Prediction of the value of a stock value, temperature or other physical values.
- **Ranking**: Order items according to a criterion. Example: page rank to order webpages according to how well they fit a search query.
- **Clustering**: partitioning of items into subsets. See Figure 1. Example: social networks.
- **Dimensionality reduction/manifold learning**: transform high dimensional data set into a low dimensional representation.

# Buzz Words

Other important concepts

- **Features**: The set of attributes, or variables, or covariates.
- **Labels**: Values or categories assigned to the examples
- **Hyperparameters**: Parameters that define the learning algorithm. These are not learned (but, somehow, they can still be optimized). E.g., number of neurons of the neural networks, when to stop training, etc.
- **Training sample**: These are the examples that are used to train the learning algorithm.
- **Validation sample**: These examples are only indirectly used in the learning algorithm, to tune its hyperparameters.
- **Test sample**: These examples are not accessed during training. After training they are used to determine the accuracy of the algorithm.

# Buzz Words

- **Loss function**: This function is used to measure the distance between the predicted and true label. If $\mathcal{Y}$ is the set of labels, then $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ (or $\ell : \overline{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}_+$). Examples include the zero-one loss, $\ell_{0-1}(\widehat{y}, y) = 1 - xy$ when $\mathcal{Y} = \{-1, +1\}$ or the square loss $\ell_2(\widehat{y}, y) = \|x - y\|^2$, when $\mathcal{Y} = \mathbb{R}$. One could also consider the binary cross entropy (if $\mathcal{Y} = \{0, 1\}$) and $\overline{\mathcal{Y}} = (0, 1)$) where $\ell_e(\widehat{y}, y) = -[y \log(\widehat{y}) + (1 - y) \log(1 - \widehat{y})]$.
- **Hypothesis set**: A set of functions that map features to labels.
- **Supervised learning**: The learner has access to labels for every training and evaluation sample. This was the case in the exoplanet study.
- **Unsupervised learning**: Here we do not have labels. A typical example is clustering.
- **Semi-supervised learning**: Here some of the data have labels. Here the labels and the structure of the data need to be used.

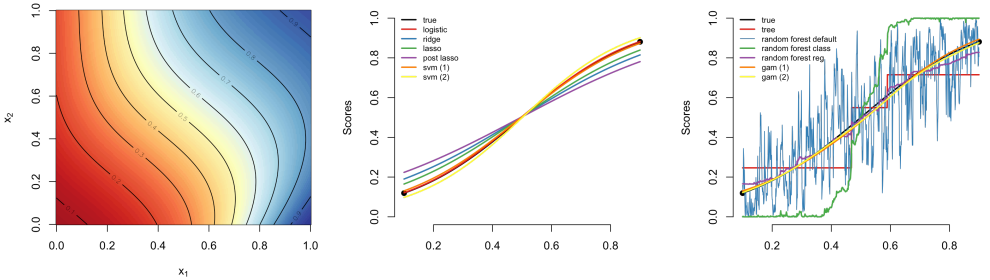
  

# Buzz Words

- **Online learning**: Here training and testing are performed iteratively in rounds. In each round we receive new data. We make a prediction receive an evaluation and update our model. The goal is to reduce the so-called regret. This describes how much worse one performed than an expert would in hindsight.
- **Reinforcement learning**: Similar to online learning in the sense that training and testing phases are mixed. The learner receives a reward for each action and seeks to maximise this reward. This if often used to train algorithms to play computer games.
- **Active learning**: An oracle exists that can be queried by the learner for labels to samples chosen by the learner.
- **Generalisation**: describes the performance of the learned algorithm outside of the training set.

# Example

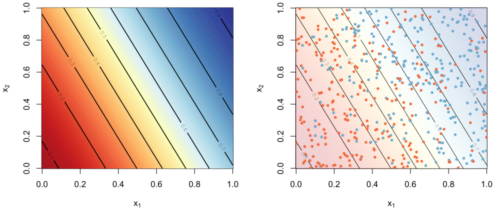
$$(y_i, x_{1,i}, x_{2,i}),\ \text{where } \mu(x_1,x_2) = \frac{\exp[\underbrace{x_1+x_2}_{\text{plain linear}} + \underbrace{\psi(x_1,x_2)}_{\text{nonlinear component}}]}{1+\exp[x_1+x_2+\psi(x_1,x_2)]},$$



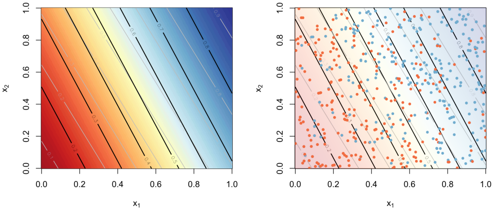

Evolution of $x \mapsto \widehat{m}(x,x)$, on the diagonal.

# Example

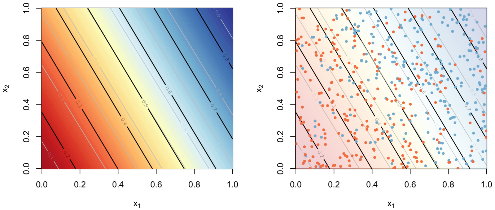
- (plain) Logistic




- Logistic with **lasso** ($\ell_2$ penalty)

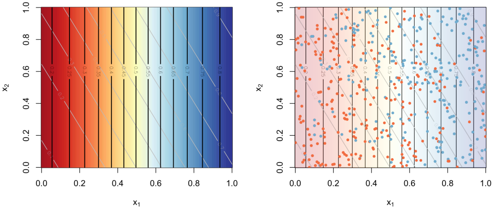



- Logistic with **Ridge** ($\ell_1$ penalty)



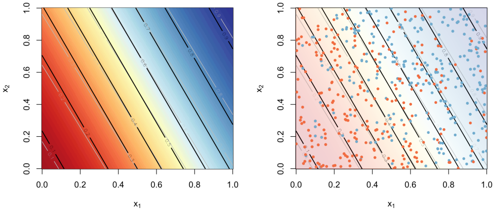

- Logistic with **post-lasso**

  (variable selection, here $x_1$)

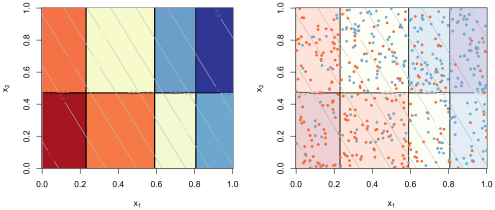


# Example

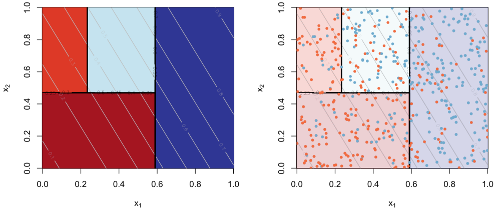
- Linear **discriminant analysis**



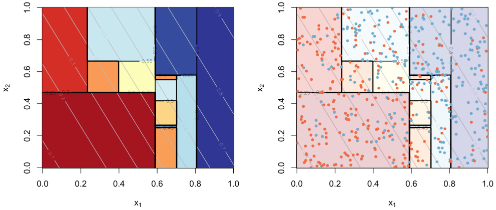

- Logistic with **categorical variables**
  (cut, $x_{j,k} = \mathbf{1}(x_j \in [a_k, a_{k+1}))$)




- **Classification Tree (1)**

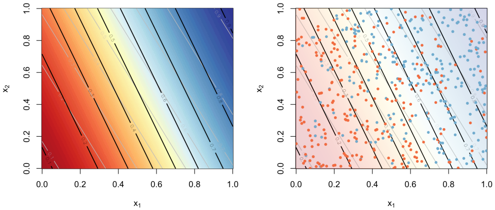



- **Classification Tree (2)**

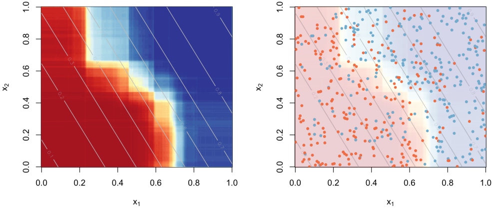


# Example

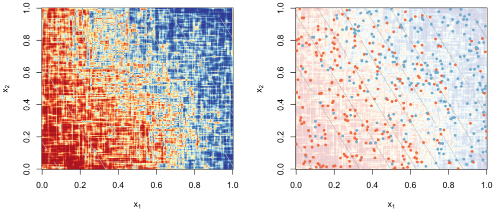
- **Support Vector Machine** (SVM) plain vanilla



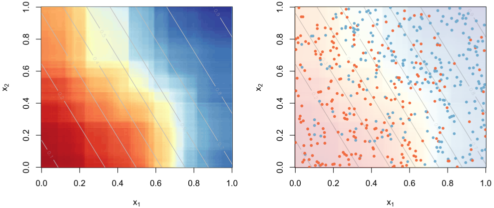

- **Classification Random Forest** with maximum nodes option




- **Classification Random Forest**

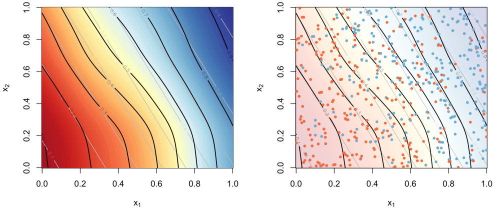



- **Regression Random Forest**

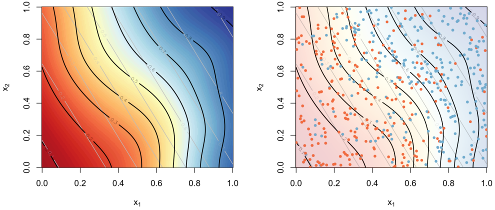


# Example

- Logistic **GAM** with additive splines




- Logistic **GAM** with bivariate splines



# Some datasets

```
> loc ="http://dutangc.perso.math.cnrs.fr/RRepository/pub/"
> install.packages("CASdatasets", repos = loc, type="source")
> library(CASdatasets)
> library(devtools)
> devtools::install_github("freakonometrics/InsurFair")
> library(InsurFair)
```

e.g.

```
> data(frenchmotor)
```

```
> loc = "http://freakonometrics.free.fr/titanic.RData"
> download.file(loc, "titanic.RData")
> load("titanic.RData")
> titanic = base[,1:7]
> loc = ""http://freakonometrics.free.fr/Davis.txt"
> davis = read.table(loc)
```

# Some datasets

```
> loc = "http://freakonometrics.free.fr/saporta.csv"
> myocarde = read.table(loc, head = TRUE, sep = ";")
> loc = "http://freakonometrics.free.fr/german_credit.csv"
> credit = read.csv(loc, header = TRUE, sep = ",")
```

# References


Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Béneplanc, G., Charpentier, A., and Thourot, P. (2022). *Manuel d'assurance*. Presses Universitaires de France.

Charpentier, A. (2014). *Computational actuarial science with R*. CRC press.

Charpentier, A. (2024). *Insurance: biases, discrimination and fairness*. Springer Verlag.

Charpentier, A., Flachaire, E., and Ly, A. (2018). Econometrics and machine learning. *Economie et Statistique*, 505(1):147–169.

Denuit, M. and Charpentier, A. (2004). *Mathématiques de l'assurance non-vie: Tome I Principes fondamentaux de théorie du risque*. Economica.

Denuit, M. and Charpentier, A. (2005). *Mathématiques de l'assurance non-vie: Tome II Tarification et provisionnement*. Economica.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., Pineau, J., et al. (2018). An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4):219–354.

# References

Friedman, J. H. (1998). Data mining and statistics: What's the connection? *Computing science and statistics*, 29(1):3–9.

Gourieroux, C. and Monfort, A. (1995). *Statistics and econometric models*, volume 1. Cambridge University Press.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115.

Morgan, M. S. (1990). *The history of econometric ideas*. Cambridge University Press.

Reichenbach, H. (1971). *The theory of probability*. University of California Press.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Tinbergen, J. (1939). *Statistical Testing of Business-Cycle Theories*. Oxford University Press.

Tukey, J. W. (1962). The future of data analysis. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 408–452. Springer.

von Mises, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*. Springer.

von Mises, R. (1939). *Probability, statistics and truth*. Macmillan.

Watt, J., Borhani, R., and Katsaggelos, A. K. (2016). *Machine learning refined: Foundations, algorithms, and applications*. Cambridge University Press.

Working, E. J. (1927). What do statistical "demand curves" show? *The Quarterly Journal of Economics*, 41(2):212–235.