Préparée à l'Institut Curie

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

From mechanistic modeling of cancer to clinical interpretation and impact evaluation

Soutenue par
Jonas BEAL

Le 15/09/2020

École doctorale n°515
Complexité du Vivant

Spécialité
Génomique

Composition du jury :

Test NOM Titre, Établissement	*Président*
Test NOM Titre, Établissement	*Rapporteur*
Test NOM Titre, Établissement	*Examinateur*
Test NOM Titre, Établissement	*Examinateur*
Emmanuel BARILLOT Institut Curie, INSERM, Mines ParisTech, PSL	*Directeur de thèse*
Aurélien LATOUCHE Institut Curie, Cnam	*Directeur de thèse*
Laurence CALZONE Institut Curie, INSERM, Mines ParisTech, PSL	*Co-encadrante de thèse*

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

Key-words:

Résumé

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

Mots-clés:

Acknowledgements

Many persons to thanks. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

Table of contents

List of Tables

Table **Page**

List of Figures

Part I

Cells and their models

CHAPTER 1

Scientific modeling: abstract the complexity

"Ce qui est simple est toujours faux. Ce qui ne l'est pas est inutilisable."
Paul Valéry (Mauvaises pensées et autres, 1942)

The notion of modeling is embedded in science, to the point that it has sometimes been used to define the very nature of scientific research. What is called a model can, however, correspond to very different realities which need to be defined before addressing the object of this thesis which will consist, if one wants to be mischievous, in analyzing models with other models. This semantic elucidation is all the more necessary as this thesis is interdisciplinary, suspended between systems biology and biostatistics. In order to convince the reader of the need for such a preamble, he is invited to ask a statistician and a systems biologist the question how they would define what a model is.

1.1 What is a model?

1.1.1 In your own words

A model is first of all an ambiguous object and a polysemous word. It therefore seems necessary to start with a semantic study. Among the many meanings and synonymous proposed by the dictionary (Figure 1.2), while

Figure 1.1: **A scientist and his model.** Joseph Wright of Derby, *A Philosopher Giving a Lecture at the Orrery (in which a lamp is put in place of the sun)*, c. 1763-65, oil on canvas, Derby Museums and Art Gallery

some definitions are more related to art, several find echoes in scientific practice. It is sometimes a question of the physical representation of an object, often on a reduced scale as in Figure 1.1, and sometimes of a theoretical description intended to facilitate the understanding of the way in which a system works [Collins, 2020]. It is even sometimes an ideal to be reached and therefore an ambitious prospect for an introduction.

The narrower perspective of the scientist does not reduce the completeness of the dictionary's description to an unambiguous object [Bailer-Jones, 2002]. In an attempt to approach these mult-faceted objects that are the models, Daniela Bailer-Jones interviewed different scientists and asked them the same question: what is a model? Across the different profiles and fields of study, the answers vary but some patterns begin to emerge (Figure 1.3). A model must capture the essence of the phenomenon being studied. Because it eludes, voluntarily or not, many details or complexity, it is by nature a simplification of the phenomenon. These limitations may restrict its validity to certain cases or suspend it to the fulfilment of some hypotheses. They are not necessarily predictive, but they must be able to generate new hypotheses, be tested and possibly questioned. Finally, and fundamen-

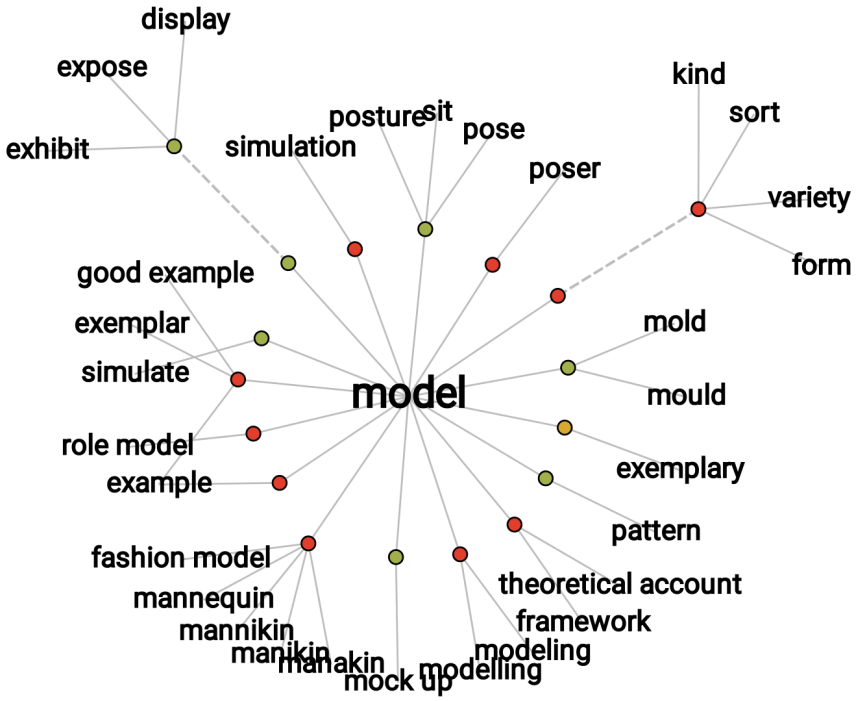

Figure 1.2: **Network visualization of *model* thesaurus entries.** Generated with the 'Visual Thesaurus' ressource

tally, they must provide insights about the object of study and contribute to its understanding.

These definitions circumscribe the *model* object, its use and its objectives, but they do not in any way describe its nature. And for good reason, because even if we agree on the described contours, the biodiversity of the models remains overwhelming for taxonomists:

> *Probing models, phenomenological models, computational models, developmental models, explanatory models, impoverished models, testing models, idealized models, theoretical models, scale models, heuristic models, caricature models, exploratory models, didactic models, fantasy models, minimal models, toy models, imaginary models, mathematical models, mechanistic models, substitute models, iconic models, formal models, analogue models, and instrumental models are but some of the notions that are used to categorize models.*
> [Frigg and Hartmann, 2020]

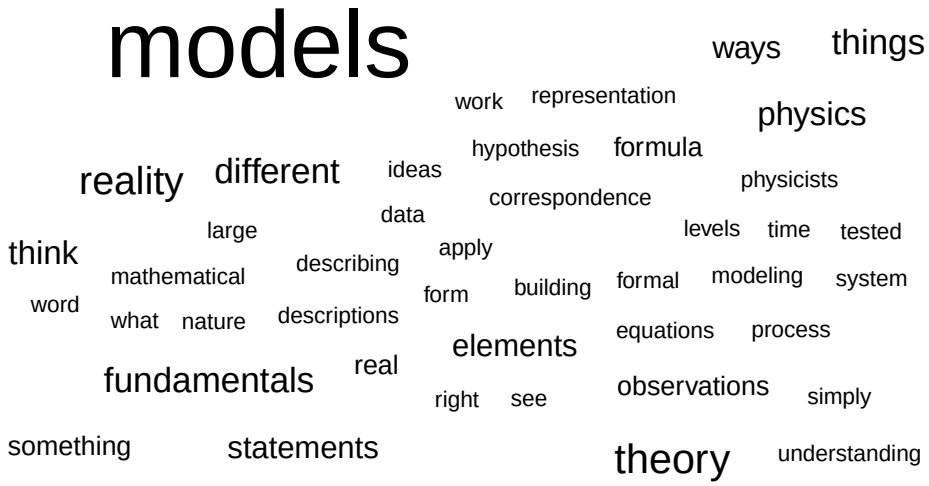

Figure 1.3: **Scientists talk about their models: words cloud.** Cloud of words summarizing the lexical fields used by scientists to talk about their models in dedicated interviews [Bailer-Jones, 2002].

1.1.2 Physical world and world of ideas

Without claiming to be exhaustive, we can make a first simple dichotomy between physical/material and formal/intellectual models [Rosenblueth and Wiener, 1945]. The former consist in replacing the object of study by another object, just as physical but nevertheless simpler or better known. These may be models involving a change of scale such as the simple miniature replica placed in a wind tunnel, or the metal double helix model used by Watson and Crick to visualize DNA. In all these cases the model allows to visualize the object of study (Figure 1.4 A and B) to manipulate it and play with it to better understand or explain, just like the scientist with his orrery (Figure 1.1). In the case of biology, we will think mainly of model organisms such as drosophila, zebrafish or mice, for example. We then benefit from the relative simplicity of their genomes, a shorter time scale or ethical differences, usually to elucidate mechanisms of interest in humans. Correspondence between the target system and its model can sometimes be more conceptual, such as that ones relying on mechanical–electrical analogies: a mechanical system (e.g. a spring-mass system) can sometimes be represented by an electric network (e.g. a RLC circuit).

The model is then no longer simply a mimetic replica but is based on an intellectual equivalence: we are gradually moving into the realm of formal models [Rosenblueth and Wiener, 1945]. These are of a more symbolic nature and they represent the original system with a set of logical or mathe-

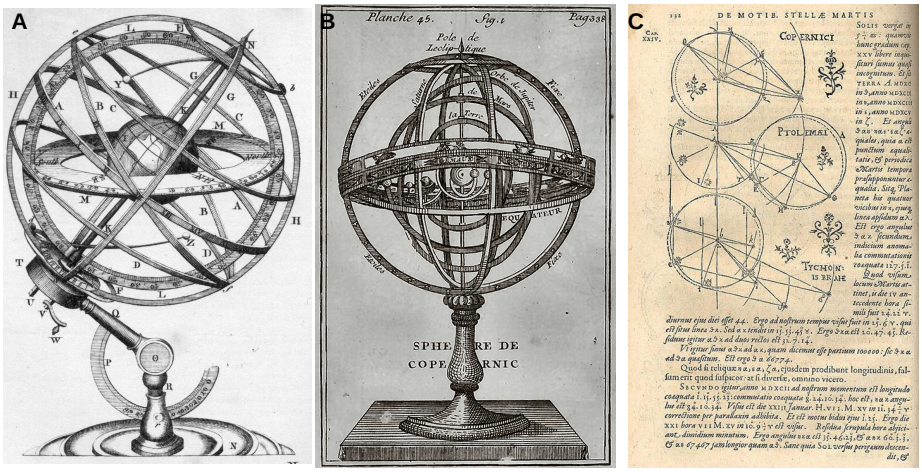

Figure 1.4: **Orrery, planets and models**. Physical models of planetary motion, either geocentric (Armillary sphere from *Plate LXXVII* in *Encyclopedia Britannica*, 1771) or heliocentric in panel B (Bion, 1751, catalogue Bnf) and some geometric representations by Johannes Kepler in panel C (in *Astronomia Nova*, 1609)

matical terms, describing the main driving forces or similar structural properties as geometrical models of planetary motions summarized by Kepler in Figure 1.4C. Historically these models have often been expressed by sets of mathematical equations or relationships. Increasingly, these have been implemented by computer. Despite their sometimes less analytical and more numerical nature, many so-called computational models could also belong to this category of formal models. There are then many formalisms, discrete or continuous, deterministic or stochastic, based on differential equations or Boolean algebra [Fowler et al., 1997]. Despite their more abstract nature, they offer similar scientific services: it is possible to play with their parameters, specifications or boundary conditions in order to better understand the phenomenon. One can also imagine these formal models from a different perspective, which starts from the data in a bottom-up approach instead of starting from the phenomenon in a top-down analysis. These models will then often be called statistical models or models of data[Frigg and Hartmann, 2020]. This distinction will be further clarified in section 1.2.

To summarize and continue a little longer with the astronomical metaphor, the study of a particularly complex system (the solar system) can be broken down into a variety of different models. Physical and mechanical models such as armillary spheres (1.4A and B), which make it

possible to touch the object of study. Moreover, we can observe the evolution of models which, when confronted with data, have progressed from a geocentric to a heliocentric representation to get closer to the current state of knowledge. Sometimes, models with more formal representations are used to give substance to ideas and hypotheses (1.4C). One of the most conceptual forms is then the mathematical language and one can thus consider that the different models find their culmination in Kepler's equations about orbits, areas and periods that describe the elliptical motion of the planets. We refer to them today as Kepler's laws. The model has become a law and therefore a paragon of mathematical modeling [Wan, 2018].

1.1.3 Preview about cancer models

As we get closer to the subject of our study, and in order to illustrate these definitions more concretely, we can take an interest in the meaning of the word *model* in the context of cancer research. For this, we restrict our corpus to articles responding to the "cancer model" search in the Pubmed article database. Among these, we look at the occurrences of the word *model* and the sentences in which it is included. This cancer-related context of model is represented as a tree in Figure 1.5. Some of the distinctions already mentioned can be found here. The *mouse* and *xenograft* models, which will be discussed later in this thesis, represent some of the most common physical models in cancer studies. These are animal models in which the occurrence and mechanisms of cancer, usually induced by the experimenter, are studied. On the other hand, *prediction*, *prognostic* or *risk score* models refer to formal models and borrow from statistical language.

Another way to classify cancer models may be to group them into the following categories: *in vivo*, *in vitro* and *in silico*. The first two clearly belong to the physical models but one uses whole living organisms (a human tumour implanted in an immunodeficient mouse) and the other separates the living from its organism in order to place it in a controlled environment (tumour cells in growth medium in a Petri dish). **In the thesis, data from both *in vivo* and *in vitro* models will be used. However, unless otherwise stated, a model will always refer to a representation *in silico*.** This third category, however, contains a very wide variety of models [Deisboeck et al., 2009], to which we will come back in chapter @ref(computational_cancer). A final ambiguity about the nature of the formal models used in this thesis needs to be clarified beforehand.

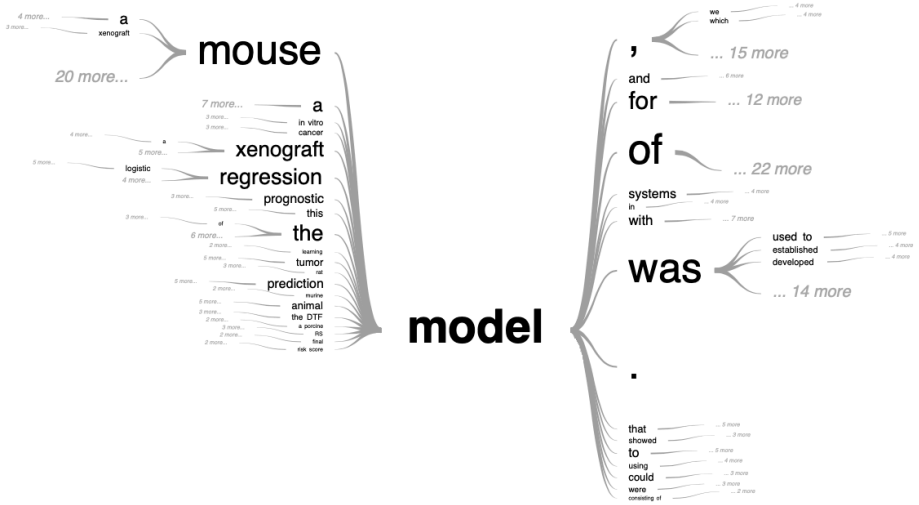

Figure 1.5: **Tree visualization of *model* semantic context in cancer-related literature** Generated with the 'PubTrees' tool by Ed Sperr, and based on most relevant PubMed entries for "cancer model" search.

1.2 Statistics or mechanistic

A rather frequent metaphor is to compare formal models to black boxes that take in input X predictors, or independent variables, and output response variable(s) Y, also named dependent variables. The models then split into two categories (Figure 1.6) depending on the answer to the question: are you modeling the inside of the box or not?

1.2.1 The inside of the box

The purpose of this section is to present in a schematic, and therefore somewhat caricatural, manner the two competing formal modeling approaches that will be used in this thesis and that we will call mechanistic modeling and statistical modeling. Assuming the unambiguous nature of the predictors and outputs we can imagine that the natural process consists in defining the result Y from the inputs X according to a function of a completely unknown form (Figure 1.6A).

The first modeling approach, that we will call mechanistic, consists in building the box by imitating what we think is the process of data generation (Figure 1.6B). This integration of a priori knowledge can take different forms. In this thesis it will often come back to presupposing certain relations between entities according to what is known about their behaviour.

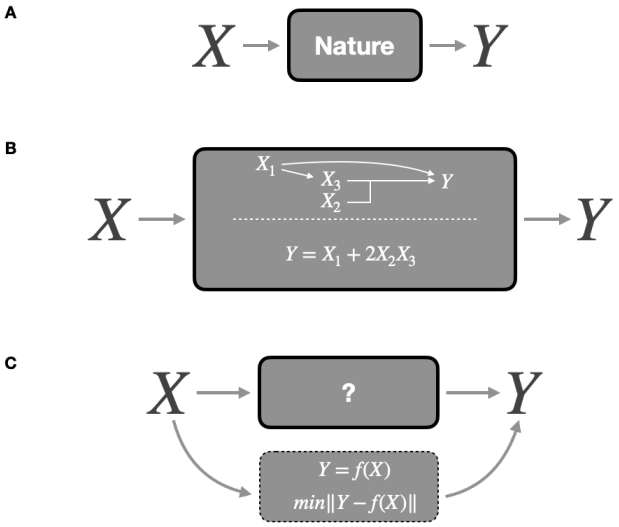

Figure 1.6: **Different modeling strategies.** (A) Data generation from predictors X to response Y in the natural phenomenon. (B) Mechanistic modeling defining mechanisms of data generation inside the box. (C) Statistical modeling finding the function f that gives the best predictions (adapted from Breiman [2001b]).

X_1 which acts on X_3 may correspond to the action of one biological entity on another, supposedly unidirectional; just as the joint action of X_2 and X_3 may reflect a known synergy in the expression of genes or the action of proteins. Mathematically this is expressed here with a perfectly deterministic model defined a priori. All in all, in a purely mechanistic approach, the nature of the relations between entities should be linked to biological processes and the parameters in the model all have biological definitions in such a way that it could even be considered to measure them.

The second approach, often called statistical modeling or machine learning, does not necessarily seek to reproduce the natural process of data generation but to find the function allowing the best prediction of Y from X (Figure 1.6C). Pushed to the limit, they are "idealized version of the data we gain from immediate observation" [Frigg and Hartmann, 2020], thus providing a phenomenological description. The methods and algorithms used are then intended to be sufficiently flexible and to make the fewest possible assumptions about the relationships between variables or the distribution of data. Without listing them exhaustively, the approaches such as boosting [Bühlmann and Hothorn, 2007], support vector machines [Cortes and Vapnik, 1995] or random forests [Breiman, 2001a], which will sometimes be

mentioned in this thesis, fall into this category which contains many others [Hastie et al., 2009].

Several discrepancies result from this difference in nature, some of which are summarized in the Table 1.1. In a somewhat schematic way, we can say that the mechanistic model first asks the question of *how* and then looks at the result for the output. The notion of causality is intrinsic to the definition of the model. Conversely, the statistical model first tries to approach the Y and then possibly analyses what can be deduced from it, regarding the importance of the variables or their relationships in a *post hoc* approach [Ishwaran, 2007, Manica et al. [2019]]. The causality is then not a by-product of the algorithm and must be evaluated according to dedicated frameworks [Hernán and Robins, 2020]. The greater flexibility of statistical methods makes it possible to better accept the heterogeneity of the variables, but this is generally done at the cost of a larger number of parameters and therefore requires more data. Moreover, we can contrast the inductive capability of statistical models able to use already generated data to identify patterns in it. Conversely, mechanistic models are more deductive in the sense that they can theoretically allow to extrapolate beyond the original data or knowledge used to build the model [Baker et al., 2018]. Finally, the most relevant way of assessing the value or adequacy of these models may be quite different. A statistical model is measured by its ability to predict output in a validation dataset different from the one used to train its parameters. The mechanistic model will also be evaluated on its capacity to approach the data but also to order, to give a meaning. If its pure predictive performance is generally inferior, how can the value of understanding be assessed? This question will be one of the threads of the dissertation.

Mechanistic and statistical models are not perfectly exclusive and rather form the two ends of a spectrum. The definitions and classification of some examples is therefore still partly personal and arbitrary. For instance, the example in 1.6B can be transformed into a model with a more ambiguous status:

$$logit(P[Y = 1]) = \beta_1 X_1 + \beta_{23} X_2 X_3$$

The logistic model shown in Figure 1.6B is deliberately ambiguous. It is a logistic model which is therefore natural defined as a statistical model. The definition of the interaction between X_2 and X_3 denotes a mechanistic presupposition. The very choice of a logistic and therefore parametric model could result from a knowledge of the phenomenon even if in practice it is often a default choice for a binary output. Finally, the nature of the parameters β_1 and β_{23} is likely to change the interpretation of the

Table 1.1: **Some pros and cons for mechanistic and statistical modeling** (adapted from Baker et al. [2018])

Mechanistic modeling	Statistical modeling
Definition	
seeks to establish a mechanistic relationship between inputs and outputs	seeks to establish statistical relationships between inputs and outputs
Pros and cons	
presupposes and investigates causal links between the variables	looks for patterns and establishes correlations between variables
capable of handling small datasets	requires large datasets
once validated, can be used as a predictive tool in new situations possibly difficult to access through experimentation	can only make predictions that relate to patterns within the data supplied
difficult to accurately incorporate information from multiple space and time scales due to constrained specifications	can tackle problems with multiple space and time scales thanks to flexible specifications
evaluated on closeness to data and ability to make sense of it	evaluated based on predictive performance

model. If they are deduced from the data and therefore optimized to fit Y as well as possible, one will think of a statistical model whose specification is nevertheless based on knowledge of the phenomenon. On the other hand, one could imagine that these parameters are taken from the literature, biochemical or other data. The model will then be more mechanistic. The boundary between these models is further blurred by the different possibilities of combining these approaches and making them complementary [Baker et al., 2018, Salvucci et al., 2019], we will come back to this later.

1.2.2 A tale of prey and predators

The following is a final general illustration of the concepts and procedures introduced with respect to statistical and mechanistic models through a famous and characteristic example: the Lotka-Volterra model of interactions between prey and predators. This model was, like many students, my first encounter with what could be called mathematical biology. The Italian

mathematician Vito Volterra states this system for the first time studying the unexpected characteristics of fish populations in the Adriatic Sea after the First World War. Interestingly, Alfred Lotka, an American physicist deduced the exact same system independantly, starting from very generic process of redistribution of matter among the several components derived from law of mass action [Knuuttila and Loettgers, 2017]. A detailed description of their works and historical formulation can be found in original articles [Lotka, 1925, Volterra, 1926] or dedicated reviews [Knuuttila and Loettgers, 2017].

The general objective is to understand the evolution of the populations of a species of prey and its predator, reasonably isolated from outside intervention. Here we will use Canada lynx (*Lynx canadensis*) and snowshow hare (*Lepus americanus*) populations for which an illustrative data set exists [Hewitt, 1917]. In fact, commercial records listing the quantities of furs sold by trappers to the Canadian Hudson Bay Company may represent a proxy for the populations of these two species as represented in Figure 1.7A. Denoting the population of lynx $L(t)$ and the population of hare $H(t)$ it can be hypothesized that prey, in the absence of predators, would increase in population, while predators on their own would decline in the absence of preys. A prey/predator interaction term can then be added, which will positively impact predators and negatively impact prey. The system can then be formalized with the following differential questions with all coefficients $a_1, a_2, b_1, b_2 > 0$:

$$\frac{dH}{dt} = a_1 H - a_2 HT$$

$$\frac{dL}{dt} = -b_1 L + b_2 HL$$

$a_1 H$ represents the growth rate of the hare population (prey), i.e. the population grows in proportion to the population itself according to usual birth modeling. The main losses of hares are due to predation by lynx, as represented with a negative coefficient in the $-a_2 HT$ term. It is therefore assumed that a fixed percentage of prey-predator encounters will result in the death of the prey. Conversely, it is assumed that the growth of the lynx population depends primarily on the availability of food for all lynxes, summarized in the $b_2 HL$ term. In the absence of hares, the lynx population decreases, as denoted by the coefficient $-b_1 L$. Here we find some of the important characteristics of a mechanistic model. The equations are based on a priori knowledge or assumptions about the structure of the problem and the parameters of the model can be interpreted. a_1, for

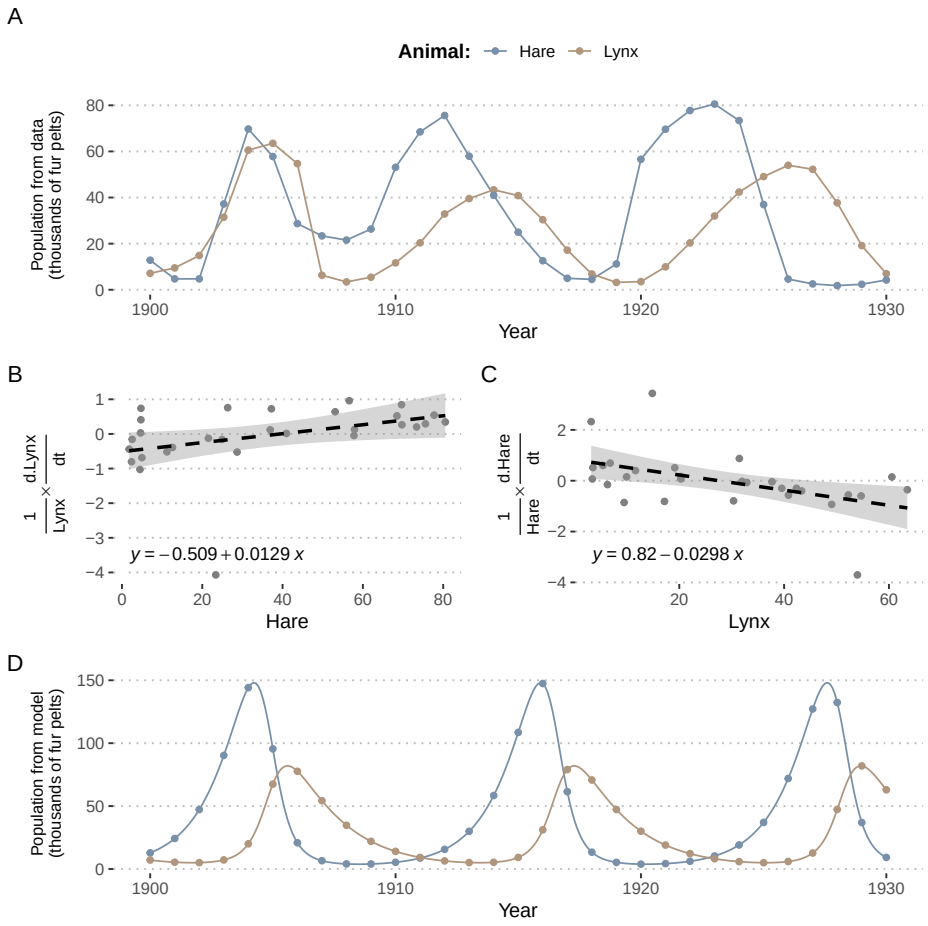

Figure 1.7: **Some analyses around Lotka-Volterra model of a prey-predator system**. (A) Evolution of lynx and hares populations based on Hudson Bay Company data about fur pelts. (B) and (C) Linear regression for estimation of parameters. (D) Evolution of lynx and hare populations as predicted by the model based on inferred parameters and initial conditions.

example, could correspond to the frequency of litters among hares and the number of offspring per litter.

This being said, the structure of the model having been defined a priori, it remains to determine its parameters. Two options would theoretically be possible: to propose values based on the interpretation of the parameters and ecological knowledge, or to fit the model to the data in order to find the best parameters. For the sake of simplicity, and because this example has only a pedagogical value in this presentation, we propose to determine them roughly using the following Taylor-based approximation:

$$\frac{1}{y(t)}\frac{dy}{dt} \simeq \frac{1}{y(t)}\frac{y(t+1)-y(t-1)}{2}$$

By applying this approximation to the two equations of the differential system and plotting the corresponding linear regressions (Figures 1.7B and C), we can obtain an evaluation of the parameters such as $a_1 = 0.82$, $a_2 = 0.0298$, $b_1 = 0.509$, $b_2 = 0.0129$. By matching the initial conditions to the data, the differential system can then be fully determined and solved numerically (Figures 1.7D). Comparison of data and modeling provides a good illustration of the virtues and weaknesses of a mechanistic model. Firstly, based on explicit and interpretable hypotheses, the model was able to recover the cyclical behaviour and dependencies between the two species: the increase in the lynx population always seems to be preceded by the increase in the hare population. However, the amplitude of the oscillations and their periods are not exactly those observed in the data. This may be related to approximations in the evaluation of parameters, random variation in the data or, of course, simplifications or errors in the structure of the model itself.

Besides, if one tries to carry out a statistical modeling of these data, it is very likely that it is possible to approach the curve of populations evolution much closer, especially for the hares. But should it be expressed simply as a function of time or should a joint modeling be proposed? The nature of the causal link between prey and predators will be extremely difficult to establish without strong hypotheses such as those of the mechanistic model. On the other hand, if populations in later years had to be predicted as accurately as possible, it is likely that a sufficiently well-trained statistical model would perform better. Finally, and this is a fundamental difference, the mechanical model makes it possible to test cases or hypotheses that go beyond the scope of the data. Quite simply, by playing with the variables or parameters of the model, we can predict the exponential decrease of predators in the absence of prey and the exponential growth of prey in the absence of prey. More generally, it is also possible to study analytically or numerically the bifurcation points of the system in order to determine the families of behaviours according to the relative values of the parameters [Flake, 1998]. It is not possible to infer these new or hypothetical behaviours directly from the data o of the statistical model. This is theoretically possible on the basis of the mechanistic model, provided that it is sufficiently relevant and that its operating hypotheses cover the cases under investigation. Now that the value of mechanistic models has been illustrated in a fairly theoretical example, all that remains is to explore in the next chapters how they can be built and used in the context of cancer.

1.3 Simplicity is the ultimate sophistication

Before concluding this modeling introduction, it is important to highlight one of the most important points already introduced in a concise manner by Valéry at the beginning of this chapter. Whatever its nature, a model is always a simplified representation of reality and by extension is always wrong to a certain extent. This is a generally well-accepted fact, but it is crucial to understand the implications for the modeller. This simplification is not a collateral effect but an intrinsic feature of any model:

> *No substantial part of the universe is so simple that it can be grasped and controlled without abstraction. Abstraction consists in replacing the part of the universe under consideration by a model of similar but simpler structure. Models, formal and intellectual on the one hand, or material on the other, are thus a central necessity of scientific procedure.*
> [Rosenblueth and Wiener, 1945]

Therefore, a model exists only because we are not able to deal directly with the phenomenon and simplification is a necessity to make it more tractable [Potochnik, 2017]. This simplification appeared many times in the studies of frictionless planes or theoretically isolated systems, in a totally deliberate strategy. However, this idealization can be viewed in several ways [weisberg2007three]. One of them, called Aristotelian or minimal idealization, is to eliminate all the properties of an object that we think are not relevant to the problem in question. This amounts to lying by omission or making assumptions of insignificance by focusing on key causal factors only [Frigg and Hartmann, 2020]. We therefore refer to the *a priori* idea that we have of the phenomenon. The other idealization, called Galilean, is to deliberately distort the theory to make it tractable as explicited by Galileo himself:

> *We are trying to investigate what would happen to moveables very diverse in weight, in a medium quite devoid of resistance, so that the whole difference of speed existing between these moveables would have to be referred to inequality of weight alone. Since we lack such a space, let us (instead) observe what happens in the thinnest and least resistant media, comparing this with what happens in others less thin and more resistant.*

This fairly pragmatic approach should make it possible to evolve iteratively, reducing distortions as and when possible. This could involve the

addition of other species or human intervention into the Lotka-Volterra system described above. A three-species Lotka-Volterra model can however become chaotic [Flake, 1998], and therefore extremely difficult to use and interpret, thus underlining the importance of simplifying the model.

We will have the opportunity to come back to the idealizations made in the course of the cancer models but it is already possible to give some orientations. The experimenter who seeks to study cancer using cell lines or animal models is clearly part of Galileo's lineage. The mathematical or *in silico* modeler has a more balanced profile. The design of qualitative mechanistic models based on prior knowledge, which is the core of the second part of the thesis, is more akin to minimal idealization, which seeks to highlight the salient features of a system. This pragmatism consisting in creating computationnaly-tractable models is also quite widespread, particularly in highly dimensional statistical approaches.

Because of the complexity of the phenomena, simplification is therefore a necessity. The objective then should not necessarily be to make the model more complex, but to match its level of simplification with its assumptions and objectives. Faced with the temptation of the author of the model, or his reviewer, to always extend and complicate the model, it could be replied with Lewis Carrol words[1]:

> *"That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?"*
> *"About six inches to the mile."*
> *"Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"*
> *"Have you used it much?" I enquired.*
> *"It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."*
> Lewis Carroll, *Sylvie and Bruno* (1893)

[1]More concisely, in [Rosenblueth and Wiener, 1945], "best material model for a cat is another cat, or preferably the same cat."

CHAPTER 2

Cancer as deregulation of complex machinery

"All happy families are alike; each unhappy family is unhappy in its own way."

Leo Tolstoy (Anna Karenina, 1877)

Armed with all these models, whether statistical or mechanistic, we are going to look at a particularly complex system that fully justifies their use: cancer. Since the first chapter recalled how important prior knowledge of the phenomenon under study is for designing models, whatever their nature, this chapter will briefly summarize some of the most important characteristics of this disease before returning to the models themselves in the next chapter. Without aiming for exhaustiveness, and after an epidemiological and statistical description, we will focus on the most useful information for the modeller, i.e. the underlying biological mechanisms and available data.

2.1 What is cancer?

Cancer can be described as a group of diseases characterized by uncontrolled cell divisions and growth which can spread to surrounding tissues. Descrip-

Figure 2.1: **Cancer is an old disease** Rembrandt, *Bathsheba at Her Bath*, c. 1654, oil on canvas, Louvre Museum, Paris

tions of this disease, and essentially associated solid tumours, have been found as far back as ancient Egyptian documents, at least 1600 BC and we know from the first century A.D. with Aulus Celsus that it is better to remove the tumors and this as soon as possible [Hajdu, 2011a]. Progress will accelerate during the Renaissance with the renewed interest in medicine, and anatomy in particular, which will advance the knowledge of tumour pathology and surgery [Hajdu, 2011b]. The progress of anatomical knowledge has also left brilliant testimonies in the field of painting, which make the renown of the Renaissance today. The precision of these artists' traits has also allowed some retrospective medical analyses, some of them going so far as to identify the signs of a tumour in certain subjects [Bianucci et al., 2018]. Such is the bluish stain on the left breast of the Bathsheba painted

by Rembrandt (Figure 2.1) which has been subject to controversial interpretations, sometimes described as an example of "skin discolouration, distortion of symmetry with axillary fullness and peau d'orange" [Braithwaite and Shugg, 1983] and sometimes spared by photonic and computationnal analyses [Heijblom et al., 2014]. The mechanisms of the disease only began to be elucidated with the appearance of the microscope in the 19th century, which revealed its cellular origin [Hajdu, 2012a]. The classification and description of cancers is then gradually refined and the first non-surgical treatments appear with the discovery of ionising radiation by the Curies [Hajdu, 2012b]. The 20th century is then the century of understanding the causes of cancer [Hajdu and Darvishian, 2013, Hajdu and Vadmal, 2013]. Some environmental exposures are characterized as asbestos or tobacco. Finally, the biological mechanisms become clearer with the identification of tmour-causing viruses and especially with the discovery of DNA [Watson and Crick, 1953]. The foundations of our current understanding of cancer date back to this period, which marks the beginning of the molecular biology of cancer. It is this branch of biology that contains the bulk of the knowledge that will be used to build our mechanistic models, and it will be later detailed in Section 2.3.

One of the ways to read this brief history of cancer is to see that theoretical and clinical progress has not followed the same timeframes.The medical and clinical management of cancers initially progressed slowly but surely, and this in the absence of an understanding of the mechanisms of cancer. Conversely, the theoretical progress of the last century has not always led to parallel medical progress, except on certain specific points. The interaction between the two is therefore not always obvious. The transformation of fundamental knowledge into medical and clinical impact is therefore of particular importance. This is what is called *translational medicine*, the aim of which is to go from laboratory bench to bedside [Cohrs et al., 2015]. It is, modestly, in this perspective that the mechanistic models of cancer that will receive particular attention in this thesis are placed. Their objective is to integrate biological knowledge, or at least a synthesis this knowledge, in order to transform it into with a relevant clinical information.

2.2 Cancer from a distance: epidemiology and main figures

Before going down to the molecular level, it is important to detail some figures and trends in the epidemiology of cancer today. Following the de-

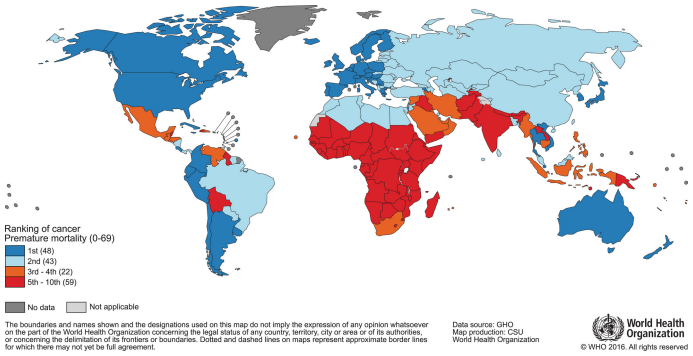

Figure 2.2: **World map and national rankings of cancer as a cause of premature death.** Classification of cancer as a cause of death before the age of 70, based on data for the year 2015. Original Figure, data and methods from Bray et al. [2018].

scription in the previous section, cancer is first and foremost defined as a disease. Considered to be a unique disease, it caused 18.1 million new cancer cases and 9.6 million cancer deaths in 2018 according to the Global Cancer Observatory affiliated to World Health Organization [Bray et al., 2018]. However, these aggregated data conceal disparities of various kinds. The first one is geographical. Indeed, mortality figures make cancer one of the leading causes of premature death in most countries of the world but its importance relative to other causes of death is even greater in the more developed countries (Figure 2.2). All in all, cancer is the first or second cause of premature death in almost 100 countries worldwide [Bray et al., 2018]. These differences call for careful consideration of the impact of population age structures and health-related covariates.

A second disparity lies in the different types of cancer. If we classify tumours solely according to their location, i.e. the organ affected first, we already obtain very wide differences. First of all, the incidence varies considerably (Figure 2.3A)). Cancers do not occur randomly anywhere in the body and certain environments or cell types appear to be more favourable [Tomasetti and Vogelstein, 2015]. Mortality is also highly variable but is not directly inferred from incidence. Not all types of cancer have the same prognosis (Figure 2.3A and B) and survival rates [Liu et al., 2018]. Although breast cancer is much more common than lung cancer, it causes fewer deaths because its prognosis is, on average, much better. The mechanisms at work in the emergence of cancer are therefore not necessarily the same as those that will govern its evolution or its response to treatment. And still on the response to treatment, Figure 2.3B highlights another disparity: not only

are the survival prognosis associated with each cancer very different, but the evolution (and generally the improvement) of these prognoses has been very uneven over the last few decades. This means that theoretical and therapeutic advances have not been applied to all types of cancer with the same success. It is one more indication of the diversity of biological mechanisms at work, which make it impossible to find a panacea, and which, on the contrary, encourage us to carefully consider the particularities of each tumour, both to understand them and to treat them. Under a generic name and in spite of common characteristics, the cancers thus appear as extremely heterogeneous. And to understand the sources of this heterogeneity, it will be necessary to place ourselves on a much smaller scale.

2.3 Basic molecular biology of cancer

[Weinberg and Weinberg, 2013]
[Tomasetti et al., 2015]

2.4 High-throughput data and multi-omics

2.5 From genetic to network disease

[Sanchez-Vega et al., 2018]

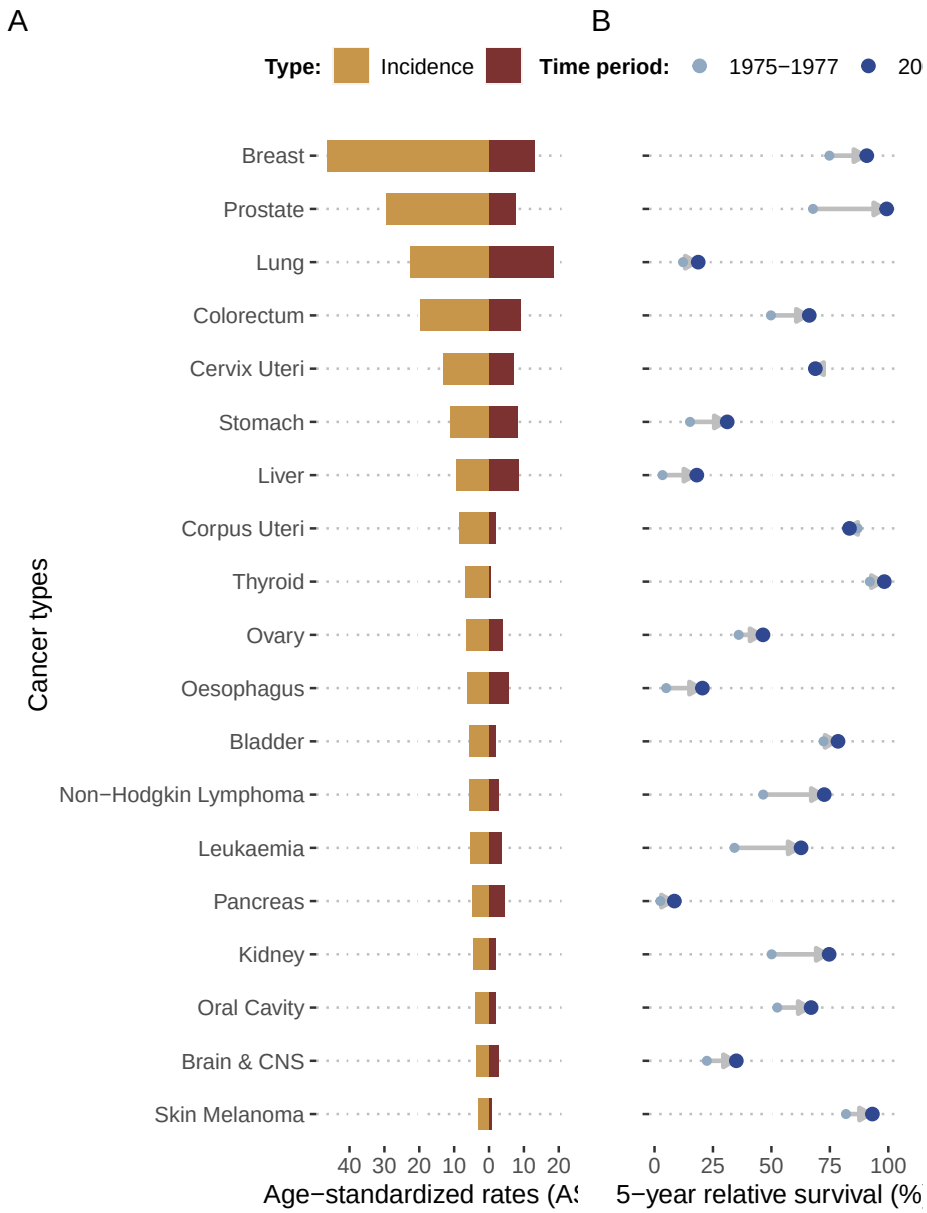

Figure 2.3: **Incidence, mortility and survival per cancer types**. (A) World incidence and mortality for the 19 most frequent cancer types in 2018, expressed age-standardized rates (adjusted age structure based on world population); data retrieved from Global Cancer Observatory. (B) Evolution of 5-years relative survival for the same cancer types based on US data from SEER registries in 1975-1977 and 2006-2012; data retrieved from Jemal et al. [2017].

Chapter 3

Test part

This is a test

3.1 Abc

Bla bla ref Miskovic et al. [2019] and [Miskovic et al., 2019].

But in Béal et al. [2019] we have the Figure 3.1 as referenced in Chapter 3

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

And an external figure 3.2

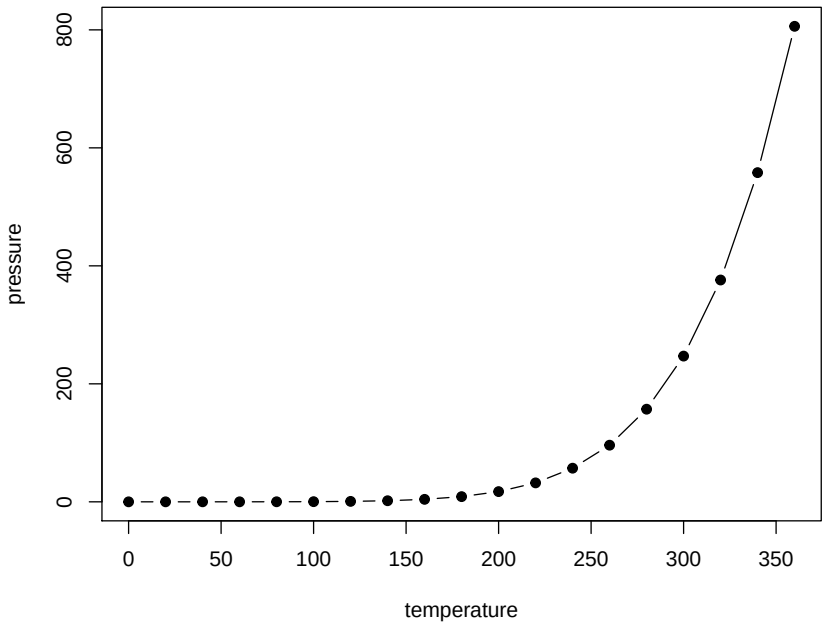

Figure 3.1: Here is a nice figure!

Figure 3.2: Example pic

Bibliography

Daniela M Bailer-Jones. Scientists' thoughts on scientific models. *Perspectives on Science*, 10(3):275–301, 2002.

Ruth E Baker, Jose-Maria Pena, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology letters*, 14(5):20170660, 2018.

Jonas Béal, Arnau Montagud, Pauline Traynard, Emmanuel Barillot, and Laurence Calzone. Personalization of Logical Models With Multi-Omics Data Allows Clinical Stratification of Patients. *Frontiers in Physiology*, 2019. ISSN 1664-042X. doi: 10.3389/fphys.2018.01965. URL https://www.frontiersin.org/article/10.3389/fphys.2018.01965/full.

Raffaella Bianucci, Antonio Perciaccante, Philippe Charlier, Otto Appenzeller, and Donatella Lippi. Earliest evidence of malignant breast cancer in renaissance paintings. *The Lancet Oncology*, 19(2):166–167, 2018.

Peter Allen Braithwaite and Dace Shugg. Rembrandt's bathsheba: the dark shadow of the left breast. *Annals of the Royal College of Surgeons of England*, 65(5):337, 1983.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001a.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001b.

Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

Randall J Cohrs, Tyler Martin, Parviz Ghahramani, Luc Bidaut, Paul J Higgins, and Aamir Shahzad. Translational medicine definition by the european society for translational medicine, 2015.

Collins. *The Collins English Dictionary*. HarperCollins, 2020. URL https://www.collinsdictionary.com/dictionary/english/model. Model.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Thomas S Deisboeck, Le Zhang, Jeongah Yoon, and Jose Costa. In silico cancer modeling: is it ready for prime time? *Nature Clinical Practice Oncology*, 6(1):34–42, 2009.

Gary William Flake. *The computational beauty of nature: Computer explorations of fractals, chaos, complex systems, and adaptation*. MIT press, 1998.

Andrew Cadle Fowler, Anna C Fowler, and AC Fowler. *Mathematical models in the applied sciences*, volume 17. Cambridge University Press, 1997.

Roman Frigg and Stephan Hartmann. Models in science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition, 2020.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 1. *Cancer*, 117(5):1097–1102, 2011a.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 2. *Cancer*, 117(12):2811–2820, 2011b.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 3. *Cancer*, 118(4):1155–1168, 2012a.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 4. *Cancer*, 118(20):4914–4928, 2012b.

Steven I Hajdu and Farbod Darvishian. A note from history: landmarks in history of cancer, part 5. *Cancer*, 119(8):1450–1466, 2013.

Steven I Hajdu and Manjunath Vadmal. A note from history: Landmarks in history of cancer, part 6. *Cancer*, 119(23):4058–4082, 2013.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Michelle Heijblom, Linda M Meijer, Ton G van Leeuwen, Wiendelt Steenbergen, and Srirang Manohar. Monte carlo simulations shed light on bathsheba's suspect breast. *Journal of biophotonics*, 7(5):323–331, 2014.

MA Hernán and JM Robins. Causal inference: What if. *Boca Raton: Chapman & Hill/CRC*, 2020.

C Gordon Hewitt. Conservation of wild life in canada, 1917.

Hemant Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

Ahmedin Jemal, Elizabeth M Ward, Christopher J Johnson, Kathleen A Cronin, Jiemin Ma, A Blythe Ryerson, Angela Mariotto, Andrew J Lake, Reda Wilson, Recinda L Sherman, et al. Annual report to the nation on the status of cancer, 1975–2014, featuring survival. *JNCI: Journal of the National Cancer Institute*, 109(9):djx030, 2017.

Tarja Knuuttila and Andrea Loettgers. Modelling as indirect representation? the lotka–volterra model revisited. *The British Journal for the Philosophy of Science*, 68(4):1007–1036, 2017.

Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.

AJ Lotka. Principles of physical biology. *Baltimore: Waverly*, 1925.

Matteo Manica, Ali Oskooei, Jannis Born, Vigneshwari Subramanian, Julio Sáez-Rodríguez, and María Rodríguez Martínez. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics*, 2019.

Ljubisa Miskovic, Jonas Béal, Michael Moret, and Vassily Hatzimanikatis. Uncertainty reduction in biochemical kinetic models: Enforcing desired model properties. *PLoS Computational Biology*, 2019. ISSN 15537358. doi: 10.1371/journal.pcbi.1007242. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007242.

Angela Potochnik. *Idealization and the Aims of Science*. University of Chicago Press, 2017.

Arturo Rosenblueth and Norbert Wiener. The role of models in science. *Philosophy of science*, 12(4):316–321, 1945.

Manuela Salvucci, Arman Rahman, Alexa J Resler, Girish M Udupi, Deborah A McNamara, Elaine W Kay, Pierre Laurent-Puig, Daniel B Longley, Patrick G Johnston, Mark Lawler, et al. A machine learning platform to optimize the translation of personalized network models to the clinic. *JCO clinical cancer informatics*, 3:1–17, 2019.

Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K Chatila, Augustin Luna, Konnor C La, Sofia Dimitriadoy, David L Liu, Havish S Kantheti, Sadegh Saghafinia, et al. Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2):321–337, 2018.

Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.

Cristian Tomasetti, Luigi Marchionni, Martin A Nowak, Giovanni Parmigiani, and Bert Vogelstein. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences*, 112(1):118–123, 2015.

Vito Volterra. Fluctuations in the abundance of a species considered mathematically, 1926.

Frederick YM Wan. *Mathematical models and their analysis*, volume 79. SIAM, 2018.

James D Watson and Francis HC Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

Robert A Weinberg and Robert A Weinberg. *The biology of cancer*. Garland science, 2013.

RÉSUMÉ

Cuius acerbitati uxor grave accesserat incentivum, germanitate Augusti turgida supra modum, quam Hannibaliano regi fratris filio antehac Constantinus iunxerat pater, Megaera quaedam mortalis, inflammatrix saevientis adsidua, humani cruoris avida nihil mitius quam maritus; qui paulatim eruditiores facti processu temporis ad nocendum per clandestinos versutosque rumigerulos conpertis leviter addere quaedam male suetos falsa et placentia sibi discentes, adfectati regni vel artium nefandarum calumnias insontibus adfligebant.

MOTS CLÉS

Caesar licentia post honoratis haec adhibens urbium honoratis nullum Caesar.

ABSTRACT

Verum ad istam omnem orationem brevis est defensio. Nam quoad aetas M. Caeli dare potuit isti suspicioni locum, fuit primum ipsius pudore, deinde etiam patris diligentia disciplinaque munita. Qui ut huic virilem togam dedit nihil dicam hoc loco de me; tantum sit, quantum vos existimatis; hoc dicam, hunc a patre continuo ad me esse deductum; nemo hunc M. Caelium in illo aetatis flore vidit nisi aut cum patre aut mecum aut in M. Crassi castissima domo, cum artibus honestissimis erudiretur.

KEYWORDS

Delatus delatus nominatus onere aut trahebatur quod tenus et bonorum.