METHODS AND TOOLS FOR HYPOTHESIS-DRIVEN RESEARCH SUPPORT: A SURVEY

L. Kalinichenko¹, D. Kovalev¹, D. Kovaleva², and O. Malkov²

Abstract: Data intensive research (DIR) is being developed in frame of the new paradigm of research study known as the Fourth paradigm, emphasizing an increasing role of observational, experimental, and computer simulated data practically in all research domains. The principal goal of DIR is an extraction (inference) of knowledge from data. The intention of this work is to make an overview of the existing approaches, methods, and infrastructures of the data analysis in DIR accentuating the role of hypotheses in such process and efficient support of hypothesis formation, evaluation, and selection in course of the natural phenomena modeling and experiments carrying out. An introduction into various concepts, methods, and tools intended for effective organization of hypothesis-driven experiments in DIR is presented.

Keywords: data intensive research; Fourth paradigm; hypotheses; models; theories; hypothetico-deductive method; hypothesis testing; hypothesis lattice; Galaxy model; connectome analysis; automated hypothesis generation

DOI: 10.14357/19922264150104

1 Hypotheses, Theories, Models and Laws in Data Intensive Science

Data intensive research is being developed in accordance with the Fourth Paradigm [1] of research study (following three previous historical paradigms of the science development (empirical science, theoretical science, and computational science)) emphasizing that science as a whole is becoming increasingly dependent on data as the core source for discovery. Emerging of the Fourth Paradigm is motivated by the huge amount of data coming from scientific instruments, sensors, simulations, as well as from people accumulating data in Web or social nets. The basic objective of DIR is to infer knowledge from the integrated data organized in networked infrastructures (such as warehouses, grids, clouds). At the same time, “Big Data” movement has emerged as a recognition of the increased significance of massive data in various domains. Open access to large volumes of data, therefore, becomes a key prerequisite for discoveries in the XXI century. Data intensive research denotes a crosscut of DIR/IT areas aimed at the creation of effective data analysis technologies for DIR covering scientific and other data intensive domains (including finance, economy, social environment, business, etc.).

Science endeavors to give a meaningful description of the world of natural phenomena using that are known as laws, hypotheses, and theories. Hypotheses, theories, and laws in their essence have the same fundamental character (Fig. 1) [2].

A scientific hypothesis is a proposed explanation of a phenomenon which still has to be rigorously tested. In contrast, a scientific theory has undergone extensive testing and is generally accepted to be the accurate explanation behind an observation. A scientific law is a proposition, which points out any such orderliness or regularity in nature, the prevalence of an invariable association between a particular set of conditions and particular phenomena. In the exact sciences, laws can often be expressed in the form of mathematical relationships. Hypotheses explain laws, and well-tested, corroborated hypotheses become theories (see Fig. 1). At the same time, the laws do not cease to be laws, just because they did not appear first as hypotheses and pass through the stage of theories.

---

¹This work has been partially supported by the RFBR grants 13-07-00579 and 14-07-00548.

¹Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

²Institute of Astronomy, Russian Academy of Sciences, 48 Pyatnitskaya Str., Moscow 119017, Russian Federation
Though theories and laws are different kinds of knowledge, actually, they represent different forms of the same knowledge construct. Laws are generalizations, principles, or patterns in nature, and theories are the explanations of those generalizations. However, classification expressed in Fig. 1 is subjective. Article [3] provides examples showing that the differences between laws, hypotheses, and theories consist only in that they stand at different levels in their claim for acceptance depending on how much empirical evidence is amassed. Therefore, there is no essential difference between constructs used for expressing hypotheses, theories, and laws. Important role of hypotheses in scientific research can scarcely be overestimated. In the edition of M. Poincaré’s book [4], it is stressed that without hypotheses, there is no science. Thus, it is not surprising that so much attention in the scientific research and the respective publications is devoted to the methods for hypothesis manipulation in experimenting and modeling of various phenomena applying the means of informatics. The idea that the new approaches are needed that can address both data- and hypothesis-driven sciences runs all through this paper. Such symbiosis alongside with the hypothesis-driven tradition of science (“first hypothesize-then-experiment”) might cause wide application of another one that is typified by “first experiment-then-hypothesize” mode of research. Often, the “first experiment” ordering in DIR is motivated by the necessity of analysis of the existing massive data to generate a hypothesis.

In the course of the present study, paying attention to the issue of inductive and deductive reasoning in hypothesis-driven sciences will be emphasized. In Fig. 2, such ways of knowledge production are shown [2]. Here, “generalization” means any subset of hypotheses, theories, and laws and “Evidence” is any subset of all facts accumulated in a specific DIR.

All researchers collect and interpret empirical evidence through the process called induction. This is a technique by which individual pieces of evidence are collected and examined until a law is discovered or a theory is invented. Frances Bacon first formalized induction [5]. The method of (naïve) induction (see Fig. 2), he suggested, is, in part, the principal way by which humans traditionally have produced generalizations that permit predictions. The problem with induction is that it is impossible to collect all observations pertaining to a given situation in all time — past, present, and future.

The formulation of a new law begins through induction as facts are heaped upon other relevant facts. Deduction is useful in checking the validity of a law. Figure 2 shows that a valid law would permit the accurate prediction of facts not yet known. Also an abdication [6] is the process of validating a given hypothesis through reasoning by successive approximation. Under this principle, an explanation is valid if it is the best possible explanation of a set of known data. Abductive validation is common practice in hypothesis formation in science. Hypothesis related logic reasoning issues are considered in more details in section 3.

In [4], the useful hypotheses of science are considered to be of two kinds:

1. the hypotheses which are valuable precisely because they are either verifiable or, else, refutable through a definite appeal to the tests furnished by experience; and
2. the hypotheses which, despite the fact that experience suggests them, are valuable despite, or even because, of the fact that experience can neither confirm nor refute them.

Aspects of science which are determined by the use of the hypotheses of the second kind are considered in [4] as “constituting an essential human way of viewing nature, an interpretation rather than a portrayal or a prediction of the objective facts of nature, an adjustment of our conceptions of things to the internal needs of our intelligence.” According to Poincaré’s discussion, the central problem of the logic of science becomes the problem of the relation between the two fundamentally distinct kinds of hypotheses, i.e., between those which cannot be verified or refuted through experience and those which can be empirically tested.

The analysis in this paper will be focused mostly on the modeling of hypotheses of the first kind, leaving issues of analysis of the relations between such two kinds of hypotheses to further study. The rest of the paper is organized as follows. Section 2 discusses the basic concepts defining the role of hypotheses in the formation of scientific knowledge and the respective organization of the scientific experiments. Approaches for hypothesis formulation, logical reasoning, hypothesis modeling, and testing are briefly introduced in section 3. In section 4, a general overview of the basic facilities provided by informatics for the hypothesis-driven experimentation scenarios, including conceptual modeling, simulations, statistics and machine learning methods is given. In section 5, several examples of organization of hypothesis-driven scientific experiments are included. Concluding remarks summarize the discussion.
2 Role of Hypotheses in Scientific Experiments: Basic Principles

Normally, scientific hypotheses have the form of a mathematical model. Sometimes, one can also formulate them as existential statements, stating that some particular instance of the phenomenon under examination has some characteristic or causal explanations, which have the general form of universal statements, stating that every instance of the phenomenon has a particular characteristic (e.g., for all x, if x is a swan, then x is white). Scientific hypothesis considered as a declarative statement identifies the predicted relationship (associative or causal) between two or more variables (independent and dependent). In causal relationship, a change caused by the independent variable is predicted in the dependent variable. Variables are more commonly related in noncausal (associative) way [7].

In experimental studies, the researcher manipulates the independent variable. The dependent variable is often referred to as consequence or the presumed effect that varies with a change of the independent variable. The dependent variable is not manipulated. It is observed and assumed to vary with changes in the independent variable. Predictions are made from the independent variable to the dependent variable. It is the dependent variable that the researcher is interested in understanding, explaining, or predicting [7].

In case when a possible correlation or similar relation between variables is investigated (such as, for example, whether a proposed medication is effective in treating a disease, that is, at least to some extent and for some patients), a few cases in which the tested remedy shows no effect do not falsify the hypothesis. Instead, statistical tests are used to determine how likely it is that the overall effect would be observed if no real relation as hypothesized exists. If that likelihood is sufficiently small, the existence of a relation may be assumed. In statistical hypothesis testing, two hypotheses are compared, which are called the null hypothesis and the alternative hypothesis. The null hypothesis states that there is no relationship between the phenomena (variables) whose relation is under investigation or, at least, not of the form given by the alternative hypothesis. The alternative hypothesis, as the name suggests, is the alternative to the null hypothesis: it states that there is some kind of relation.

Alternative hypotheses are generally used more often than null hypotheses because they are more desirable to state the researcher’s expectations. But in any study that involves statistical analysis, the underlying null hypothesis is usually assumed [7]. It is important that the conclusion “do not reject the null hypothesis” does not necessarily mean that the null hypothesis is true. It suggests that there is not sufficient evidence against the null hypothesis in favor of the alternative hypothesis. Rejecting the null hypothesis suggests that the alternative hypothesis may be true.

Any useful hypothesis will enable predictions by reasoning (including deductive reasoning). It might predict the outcome of an experiment in a laboratory setting or the observation of a phenomenon in nature. The prediction may also invoke statistics assuming that a hypothesis must be falsifiable [8] and that one cannot regard a proposition or theory as scientific if it does not admit the possibility of being shown false. The way to demarcate between hypotheses is to call scientific those for which we can specify (beforehand) one or more potential falsifiers as the respective experiments. Falsification was supposed to proceed deductively instead of inductively.

Other philosophers of science have rejected the criterion of falsifiability or supplemented it with other criteria, such as verifiability (only statements about the world that are empirically confirmable or logically necessary are cognitively meaningful). They claim that science proceeds by “induction” — that is, by finding confirming instances of a conjecture. Popper treated confirmation as never certain [8]. However, a falsification can be sudden and definitive. Einstein said: “No amount of experimentation can ever prove me right; a single experiment can prove me wrong.” To scientists and philosophers outside the Popperian belief [8], science operates mainly by induction (confirmation), and also and less often by disconfirmation (falsification). Its language is almost always one of induction. For this survey both philosophical treatment of hypotheses are acceptable. Sometimes such way of reasoning is called the hypothetico-deductive method. According to it, scientific inquiry proceeds by formulating a hypothesis in a form that could conceivably be falsified by a test on observable data. A test that could and does run contrary to predictions of the hypothesis is taken as a falsification of the hypothesis. A test that could but does not run contrary to the hypothesis corroborates the theory.

A scientific method involves experiment to test the ability of some hypothesis to adequately answer the question under investigation. A prediction enabled by hypothesis suggests a test (observation or experiment) for the hypothesis thus becoming testable. If a hypothesis does not generate any observational tests, there is nothing that a scientist can do with it.

For example, not testable hypothesis: “Our universe is surrounded by another, larger universe, with which we can have absolutely no contact;” not verifiable (though testable) hypothesis: “There are other inhabited planets in the universe;” scientific hypothesis (both testable and verifiable): “Any two objects dropped from the same
height above the surface of the earth will hit the ground at the same time as long as air resistance is not a factor” (http://www.batesville.k12.in.us/physics/phynet/aboutscience/hypotheses.html).

A problem (research question) should be formulated as an issue of what relation exists between two or more variables. The problem statement should be such as to imply possibilities of empirical testing; otherwise, this will not be a scientific problem. Problems and hypotheses being generalized relational statements enable to deduce specific empirical manifestations implied by the problem and hypotheses. In this process, hypotheses can be deduced from theory and from other hypotheses. A problem cannot be scientifically solved unless it is reduced to hypothesis form, because a problem is not directly testable [9].

Most formal hypotheses connect concepts by specifying the expected relationships between propositions. When a set of hypotheses are grouped together, they become a type of conceptual framework. When a conceptual framework is complex and incorporates causality or explanation, it is generally referred to as a theory [10]. In general, hypotheses have to reflect the multivariate complexity of the reality. A scientific theory summarizes a hypothesis or a group of hypotheses that have been supported with repeated testing. A theory is valid as long as there is no evidence to dispute it. Scientific paradigm explains the working set of theories under which science operates.

Elements of hypothesis-driven research and their relationships are shown in Fig. 3 [11, 12]. The hypothesis triangle relations, explains, formulates, and represents, are functional in the scientist’s final decision in adopting a particular model $m_1$ to formulate a hypothesis $h_1$, which is meant to explain phenomenon $p_1$.

In [12], the lattice structure for hypothesis interconnection is proposed as shown in Fig. 4. A hypothesis lattice is formed by considering a set of hypotheses equipped with wasDerivedFrom as a strict order (from the bottom to the top). Hypotheses directly derived from exactly one hypothesis are atomic, while those directly derived from at least two hypotheses are complex.

The hypothesis lattice is unfolded into model and phenomena isomorphic lattices according to the hypothesis triangle (see Fig. 3) [12]. The lattices are isomorphic if one takes subsets of $M$ (Model), $H$ (Hypotheses), and $P$ (Phenomenon) such that formulates, explains, and represents are both one-to-one and onto mappings (i.e., bijections), seen as structure-preserving mappings (morphisms). Example of the isomorphic lattice is shown in Fig. 5 [12]. This particular lattice corresponds to the case in Computational Hemodynamics considered in [12]. Here, model $m_1$ formulates hypothesis $h_1$, which explains phenomenon $p_1$. Similarly, $m_2$ formulates $h_2$, which explains $p_2$, and so on. Properties of the hypothesis lattices and operations over them are considered in [13].

Models are one of the principal instruments of modern science. Models can perform two fundamentally different representational functions: a model can be a representation of a selected part of the world, or a model can represent a theory in the sense that it interprets the laws and hypotheses of that theory.

Here, let consider scientific models to be representations in both senses at the same time. One of the most perplexing questions in connection with models is how they relate to theories. In this respect, models can be considered as a complement to theories, as preliminary theories, can be used as substitutions of theories when the latter are too complicated to handle. Learning about the model is done through experiments, thought experiments, and simulation. Given a set of parameters, a model can generate expectations about how the system will behave in a particular situation. A model and the hypotheses it is based upon are supported when the model generates expectations that match the behavior of its real-world counterpart.
A law generalizes a body of observations. Generally, a law represents a group of related undisputable hypotheses using a handful of fundamental concepts and equations to define the rules governing a set of phenomena. A law does not attempt to explain why something happens—it simply states that it does.

Facilities for support of the hypothesis-driven experimentation will be discussed in the remaining sections.

3 Hypothesis Manipulation in Scientific Experiments

3.1 Hypothesis generation

Researchers that support rationality of scientific discovery presented several methods for hypothesis generation, including discovery as abduction, induction, anomaly detection, heuristics programming, and use of analogies [14].

Discovery as abduction characterizes reasoning processes that take place before a new hypothesis is justified. The abductive model of reasoning that leads to plausible hypotheses formulation is conceptualized as an inference beginning with data. According to [15], an abduction happens as follows:

1. some phenomena $p_1, p_2, p_3, \ldots$ are encountered for which there is no or little explanation;
2. however, $p_1, p_2, p_3, \ldots$ would not be surprising if a hypothesis $H$ were added. They would certainly follow from something like $H$ and would be explained by it; and
3. therefore, there is a good reason for elaborating a hypothesis $H$—for proposing it as a possible hypothesis from which the assumption $p_1, p_2, p_3, \ldots$ might follow.

The abductive model of reasoning is primarily a process of explaining anomalies or surprising phenomena [16]. The scientists’ reasoning proceeds abductively from an
anomaly to an explanatory hypothesis in light of which the phenomena would no longer be surprising. There can be several different hypotheses that can serve as the explanations for phenomena; so, additionally some criteria for choosing among different hypotheses are required.

One way to implement abductive model of reasoning is the abductive logic programming [17]. Hypothesis generation in abduction logical framework is organized as follows. During the experiment, some new observations are encountered. Let \(B\) represents the background knowledge and \(O\) is the set of facts that represents observations. Both \(B\) and \(O\) are the logic programs (set of rules in some rule language). In addition, \(\Gamma\) stands for a set of literals representing the set of abducibles, which are candidate assumptions to be added to \(B\) for explaining \(O\). Given \(B, O, \) and \(\Gamma\), the hypothesis-generation problem is to find a set \(H\) of literals (called a hypothesis) such that:

1. \(B\) and \(H\) entail \(O\);
2. \(B\) and \(H\) are consistent; and
3. \(H\) is some subset of \(\Gamma\).

If all conditions are met, then \(H\) is an explanation of \(O\) (with respect to \(B\) and \(\Gamma\)). Examples of abductive logic programming systems include ACLP [18], A-system [19], ABDUAL [20], and ProLogICA [21]. Abductive logic programming can also be implemented by means of Answer Set Programming systems, e.g., by the DLV system [22].

The example abductive logic program in ProLogICA describes a simple model of the lactose metabolism of the bacterium E.Coli [21]. The background knowledge \(B\) describes that E.coli can feed on the sugar lactose if it makes two enzymes permease and galactosidase. Like all enzymes (E), these are made if they are coded by a gene (G) that is expressed. These enzymes are coded by two genes (lac(y) and lac(z)) in cluster of genes (lac(X)) called an operon that is expressed when the amounts (amt) of glucose are low and lactose are high or when they are both at medium level. The abducibles, \(\Gamma\), declare all ground instances of the predicates “amount” as assumable. This reflects the fact that in the model, it is not known what are the amounts at any time of the various substances. This is incomplete information that should be found out in each problem case that is examined. The integrity constraints state that the amount of a substance (S) can only take one value.

```
## Background Knowledge (B)
feed(lactose):- make(permease),
make(galactosidase).
make(Enzyme):- code(Gene,Enzyme),express(Gene).
express(lac(X)):-amount(glucose,low),
amount(lactose,hi).
express(lac(X)):-amount(glucose,medium),
amount(lactose,medium).
```
— probabilities refer to relative frequencies of events. They are objective properties of the real world;
— parameters of hypotheses (models) are fixed, unknown constants. Because they are not fluctuating, probability statements about parameters are meaningless; and
— statistical procedures should have well-defined long-run frequency properties.

In contrast, Bayesian approach takes the following assumptions:
— probability describes the degree of subjective belief, not the limiting frequency. Probability statements can be made about things other than data, including hypotheses (models) themselves as well as their parameters; and
— inferences about a parameter are made by producing its probability distribution — this distribution quantifies the uncertainty of our knowledge about that parameter. Various point estimates, such as expectation value, may then be readily extracted from this distribution.

The Bayesian interpretation of probability can be seen as an extension of propositional logic that enables reasoning with hypotheses, i.e., the propositions whose truth or falsity is uncertain.

Bayesian probability belongs to the category of evidential probabilities; to evaluate the probability of a hypothesis, the Bayesian probabilist specifies some prior probability, which is then updated in the light of new, relevant data (evidence) [28]. The Bayesian interpretation provides a standard set of procedures and formulae to perform this calculation.

Hypothesis testing in classical statistic style. After null and alternative hypotheses are stated, some statistical assumptions about data samples should be done, e.g., assumptions about statistical independence or distributions of observations. Failure in providing correct assumptions leads to the invalid test results.

A common problem in classical statistics is to ask whether a given sample is consistent with some hypothesis. For example, one might be interested in whether a measured value \( x_i \), or the whole set \( \{x_i\} \), is consistent with being drawn from a Gaussian distribution \( N(\mu, \sigma) \).

Here, \( N(\mu, \sigma) \) is the null hypothesis.

It is always assumed that we know how to compute the probability of a given outcome from the null hypothesis: for example, given the cumulative distribution function, \( 0 \leq H_0(x) \leq 1 \), the probability that we would get a value at least as large as \( x_i \) is \( p(x > x_i) = 1 - H_0(x_i) \) and is called the \( p \)-value. Typically, a threshold \( p \) value is adopted, called the significance level \( \alpha \), and the null hypothesis is rejected when \( p \leq \alpha \) (e.g., if \( \alpha = 0.05 \) and \( p < 0.05 \), the null hypothesis is rejected at a 0.05 significance level). If one fails to reject a hypothesis, it does not mean that its correctness is proved because it may be that the sample is simply not large enough to detect an effect.

When performing these tests, one can meet with two types of errors, which statisticians call Type I and Type II errors. Type I errors are the cases when the null hypothesis is true but incorrectly rejected. In the context of source detection, these errors represent spurious sources or, more generally, false positives (with respect to the alternative hypothesis). The false-positive probability when testing a single datum is limited by the adopted significance level \( \alpha \). Cases when the null hypothesis is false but it is not rejected are called Type II errors (missed sources, or false negatives (again, with respect to the alternative hypothesis)). The false-negative probability when testing a single datum is usually called \( \beta \) and is related to the power of \( \alpha \) test as \((1 - \beta)\).

Hypothesis testing is intimately related to comparisons of distributions.

As the significance level \( \alpha \) is decreased (the criterion for rejecting the null hypothesis becomes more conservative), the number of false positives decreases and the number of false negatives increases. Therefore, there is a trade-off to be made to find an optimal value of \( \alpha \), which depends on the relative importance of false negatives and positives in a particular problem. Both the acceptance of false hypotheses and the rejection of true ones are errors that scientists should try to avoid. There is discussion as to what states of affairs is less desirable; many people think that the acceptance of a false hypothesis is always worse than failure to accept a true one and that science should in the first place try to avoid the former kind of error.

When many instances of hypothesis testing are performed, a process called multiple hypothesis testing, the fraction of false positives can significantly exceed the value of \( \alpha \). The fraction of false positives depends not only on \( \alpha \) and the number of data points, but also on the number of true positives (the latter is proportional to the number of instances when an alternative hypothesis is true).

Depending on data type (discrete vs. continuous random variables) and what one can assume (or not) about the underlying distributions, and the specific question one asks, different statistical tests can be used. The underlying idea of statistical tests is to use data to compute an appropriate statistic and then compare the resulting data-based value to its expected distribution. The expected distribution is evaluated by assuming that the null hypothesis is true. When this expected distribution implies that the data-based value is unlikely to have arisen from it by chance (i.e., the corresponding \( p \) value is small), the null hypothesis is rejected with some thresh-
The number of various statistical tests in the literature is overwhelming and their applicability is often hard to decide (see [29, 30] for variety of statistical methods in SPSS (Statistical Package for the Social Sciences)). When the distributions are not known, tests are called nonparametric, or distribution-free tests. The most popular nonparametric test is the Kolmogorov–Smirnov (K-S) test, which compares the cumulative distribution function, $F(x)$, for two samples, $\{x_{i1}\}, i = 1, \ldots, N_1$, and $\{x_{i2}\}, i = 1, \ldots, N_2$. The K-S test is not the only option for nonparametric comparison of distributions. The Cramér–von Mises criterion, the Watson test, and the Anderson–Darling test are similar in spirit to the K-S test, but consider somewhat different statistics. The Mann–Whitney–Wilcoxon test (or the Wilcoxon rank-sum test) is a nonparametric test for testing whether two data sets are drawn from distributions with different location parameters (if these distributions are known to be Gaussian, the standard classical test is called the $t$ test). A few standard statistical tests can be used when it is known, or can be assumed, that both $h(x)$ and $f(x)$ are the Gaussian distributions (e.g., the Anderson–Darling test, the Shapiro–Wilk test) [27]. More on statistical tests can be found in [27, 29, 30, 31].

**Hypothesis (model) selection and testing in Bayesian style.** The Bayesian approach can be thought of as formalizing the process of continually refining our state of knowledge about the world, beginning with no data (as encoded by the prior), then updating that by multiplying in the likelihood once the data are observed to obtain the posterior. When more data are taken, then the posterior based on the first data set can be used as the prior for the second analysis. Indeed, the data sets can be different.

The question often arises as to which is the ‘best’ model (hypothesis) to use; ‘model selection’ is a technique that can be used when we wish to discriminate between competing models (hypotheses) and identify the best model (hypothesis) in a set, $\{M_1, \ldots, M_n\}$, given the data.

Let remind the basic notation. The Bayes theorem can be applied to calculate the posterior probability $p(M_j|d)$ for each model (or hypothesis) $M_j$ representing our state of knowledge about the truth of the model (hypothesis) in the light of the data $d$ as follows:

$$p(M_j|d) = \frac{p(d|M_j)p(M_j)}{\sum_i p(d|M_i)p(M_i)}$$

where $p(M_j)$ is the prior belief in the model (hypothesis) that represents our state of knowledge (or ignorance) about the truth of the model (hypothesis) before the current data have been analyzed; $p(d|M_j)$ is the model (hypothesis) likelihood (represents the probability that some data are produced under the assumption of this model); and $p(d)$ is the normalization constant given by

$$p(d) = \sum_i p(d|M_i)p(M_i).$$

The relative ‘goodness’ of models is given by a comparison of their posterior probabilities; so, to compare two models $M_a$ and $M_b$, let look at the ratio of the model posterior probabilities:

$$\frac{p(M_a|d)}{p(M_b|d)} = \frac{p(d|M_a)p(M_a)}{p(d|M_b)p(M_b)}.$$

The Bayes factor, $B_{ab}$, can be computed as the ratio of the model likelihoods:

$$B_{ab} = \frac{p(d|M_a)}{p(d|M_b)}.$$

Empirical scale for evaluating the strength of evidence from the Bayes factor $B_{ij}$ between two models is shown in the table [32].

<table>
<thead>
<tr>
<th>$\ln B_{ij}$</th>
<th>Odds</th>
<th>Strength of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt; 1.0$</td>
<td>$&lt; 3 : 1$</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>$1.0$</td>
<td>$\sim 3 : 1$</td>
<td>Weak evidence</td>
</tr>
<tr>
<td>$2.5$</td>
<td>$\sim 12 : 1$</td>
<td>Moderate evidence</td>
</tr>
<tr>
<td>$5.0$</td>
<td>$\sim 150 : 1$</td>
<td>Strong evidence</td>
</tr>
</tbody>
</table>

The Bayes factor gives a measure of the ‘goodness’ of a model regardless of the prior belief about the model; the higher the Bayes factor, the better the model is. In many cases, the prior belief in each model in the set of proposed models will be equal; so, the Bayes factor will be equivalent to the ratio of the posterior probabilities of the models. The ‘best’ model in the Bayesian sense is the one which gives the best fit to the data with the smallest parameter space.

A special case of model (hypothesis) selection is **Bayesian hypothesis testing** [27, 33]. Taking $M_1$ to be the “null” hypothesis, one can ask whether the data supports the alternative hypothesis $M_2$, i.e., whether one can reject the null hypothesis. Taking equal priors $p(M_1) = p(M_2)$, the odds ratio is

$$B_{21} = \frac{p(d|M_1)}{p(d|M_2)}.$$

The inability to reject $M_1$ in the absence of an alternative hypothesis is very different from the hypothesis testing procedure in classical statistics. The latter procedure rejects the null hypothesis if it does not provide a good description of the data, that is, when it is very unlikely that the given data could have been generated as prescribed by the null hypothesis. In contrast, the Bayesian approach is based on the posterior rather than
on the data likelihood and cannot reject a hypothesis if there are no alternative explanations for observed data [27].

Comparing classical and Bayesian approaches [27], it is rare for a mission-critical analysis be done in the “fully Bayesian” manner, i.e., without the use of the frequentist tools at the various stages. Philosophy and beauty aside, the reliability and efficiency of the underlying computations required by the Bayesian framework are the main practical issues. A central technical issue at the heart of this is that it is much easier to do optimization (reliably and efficiently) in high dimensions than it is to do integration in high dimensions. Thus, the usable machine learning methods, while there are ongoing efforts to adapt them to Bayesian framework, are almost all rooted in frequentist methods.

Most users of Bayesian estimation methods, in practice, are likely to use a mix of Bayesian and frequentist tools. The reverse is also true — frequentist data analysts, even if they stay formally within the frequentist framework, are often influenced by “Bayesian thinking,” referring to “priors” and “posteriors.” The most advisable position is probably to know both paradigms well, in order to make informed judgments about which tools to apply in which situations [27]. More details on Bayesian style of hypothesis testing can be found in [27, 28, 33].

3.2.2 Logic-based hypothesis testing

According to the hypothetico-deductive approach, the hypotheses are tested by deducing predictions or other empirical consequences from general theories. If these predictions are verified by experiments, this supports the hypothesis. It should be noted that not everything that is logically entailed by a hypothesis can be confirmed by a proper test for it. The relation between hypothesis and evidence is often empirical rather than logical. A clean deduction of empirical consequences from a hypothesis, as it may sometimes exist in physics, is practically inapplicable in biology. Thus, entailment of the evidence by hypotheses under test is neither sufficient nor necessary for a good test. Inference to the best explanation is usually construed as a form of inductive inference (see abduction in subsection 3.1) where hypothesis’ explanatory credentials are taken to indicate its truth [34].

An inductive logic is a system of evidential support that extends deductive logic to less-than-certain inferences. For valid deductive arguments, the premises logically entail the conclusion where the entailment means that the truth of the premises provides a guarantee of the truth of the conclusion. Similarly, in a good inductive argument, the premises should provide some degree of support for the conclusion, where such support means that the truth of the premises indicates with some degree of strength that the conclusion is true. If the logic of good inductive arguments is to be of any real value, the measure of support it articulates should meet the Criterion of Adequacy (CoA): as evidence accumulates, the degree to which the collection of true evidence statements comes to support a hypothesis, as measured by the logic, should tend to indicate that the hypotheses are probably false or probably true. In [35], the extent to which a kind of logic based on the Bayes theorem can estimate how the implications of hypotheses about evidence claims influences the degree to which hypotheses are supported is discussed in detail. In particular, it is shown how such a logic may be applied to satisfy the CoA: as evidence accumulates, false hypotheses will very probably come to have evidential support values (as measured by their posterior probabilities) that approach 0; and as this happens, a true hypothesis will very probably acquire evidential support values (measured by their posterior probabilities) that approach 1.

3.2.3 Parameter estimation

Models (hypotheses) are typically described by parameters $\theta$ whose values are to be estimated from data. The authors describe this process according to [27]. For a particular model $M$ and prior information $I$, one gets:

$$p(M, \theta | I) = \frac{p(d|M, \theta, I)p(M, \theta | I)}{p(d | I)}.$$  

The result $p(M, \theta | d, I)$ is called the posterior probability density function (pdf) for model $M$ and parameters $\theta$, given data $d$ and other prior information $I$. This term is a $(k+1)$-dimensional pdf in the space spanned by $k$ model parameters and the model $M$. The term $p(d|M, \theta, I)$ is the likelihood of data given some model $M$ and some fixed values of parameters $\theta$ describing it and all other prior information $I$. The term $p(M, \theta | I)$ is the a priori joint probability for model $M$ and its parameters $\theta$ in the absence of any of the data used to compute likelihood and is often simply called the prior.

In the Bayesian formalism, $p(M, \theta | d, I)$ corresponds to the state of our knowledge (i.e., belief) about a model and its parameters, given data $d$. To simplify the notation, $M(\theta)$ will be substituted by $M$ whenever the absence of explicit dependence on $\theta$ is not confusing. A completely Bayesian data analysis has the following conceptual steps.

1. Formulation of the data likelihood $p(d|M, I)$.
2. Choice of the prior $p(\theta | M, I)$, which incorporates all other knowledge that might exist, but is not used when computing the likelihood (e.g., prior measurements of the same type, different measurements, or simply an uninformative prior). Several methods for constructing “objective” priors have been proposed. One of them is the principle of maximum entropy for assigning uninformative priors by maximizing the entropy over a suitable set of pdfs,
finding the distribution that is least informative (given the constraints). Entropy maximization with no testable information takes place under a single constraint: the sum of the probabilities must be one. Under this constraint, the maximum entropy for a discrete probability distribution is given by the uniform distribution.

3. Determination of the posterior $p(M|d, I)$, using Bayes theorem above. In practice, this step can be computationally intensive for complex multidimensional problems.

4. The search for the best model $M$ parameters, which maximizes $p(M|d, I)$, yielding the maximum a posteriori (MAP) estimate. This point estimate is the natural analog to the maximum likelihood estimate (MLE) from classical statistics.

5. Quantification of uncertainty in parameter estimates, via credible regions. As in MLE, such an estimate can be obtained analytically by doing mathematical derivations specific to the chosen model. The same as in MLE, various numerical techniques can be used to simulate samples from the posterior. This can be viewed as an analogy to the frequentist approach, which can simulate draws of samples from the true underlying distribution of the data. In both cases, various descriptive statistics can then be computed on such samples to examine the uncertainties surrounding the data and estimators of model parameters based on that data.

6. Hypothesis testing as needed to make other conclusions about the model (hypothesis) or parameter estimates.

3.3 Algorithmic generation and evaluation of hypotheses

Two cultures of data analysis (formulaic modeling\(^1\) and algorithmic modeling) distinguished here in accordance with [36] can be applied to the hypothesis extraction and generation based on data.

Formulaic modeling is a process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the formulae $y = f(x)$ that give a relation specifying a vector of dependent variables $y$ in terms of a vector of independent variables $x$. In a statistics experiment (based on various regression techniques), the dependent variable defines the event studied and is expected to change whenever the independent variable (predictor variables, extraneous variables) is altered. Such methods as linear regression, logistic regression, and multiple regression are the well-known examples of the representatives of this modeling approach.

In the algorithmic modeling culture, the approach is to find an algorithm that operates on $x$ to predict the responses $y$. What is observed is a set of $x$’s that go in and a subsequent set of $y$’s that come out. Predictive accuracy and properties of the algorithms (such as, for example, their convergence if they are iterative) are the issues to be investigated. Machine learning algorithms focus on prediction, based on known properties learned from the training data. Such machine learning algorithms as decision tree, association rule, neural networks, support vector machines as well as other techniques of learning in Bayesian and probabilistic models [37, 38] are examples of the methods that belong to this second culture.

The models that best emulate the nature in terms of predictive accuracy are also the most complex and unscrutable. Nature forms the outputs $y$ from the inputs $x$ by means of a black box with complex and unknown interior. Current accurate prediction methods are also complex black boxes (such as neural nets, forests, support vectors). So, we are facing two black boxes, where ours seem only slightly less inscrutable than nature’s [36]. In a choice between accuracy and interpretability, in applications, people sometimes prefer interpretability.

However, the goal of a model is not interpretability (a way of getting information), but getting useful, accurate information about the relation between the response and predictor variables. It is stated in [36] that algorithmic models can give better predictive accuracy than formulaic models, providing also better information about the underlying mechanism. And actually, this is what the goal of statistical analysis is. The researchers should be focused on solving the problems instead of asking what regression model they can create.

An objection to this idea (expressed by Cox) is that prediction without some understanding of underlying process and linking with other sources of information becomes more and more tentative. Due to that, it is suggested to construct the stochastic calculation models that summarize the understanding of the phenomena under study. One of the objectives of such approach might be an understanding and test of hypotheses about underlying process. Given the relatively small sample size, following such direction could be productive. But data characteristics are rapidly changing. In many of the most interesting current problems, the idea of starting with a formal model is not tenable. The methods used in statistics for small sample sizes and a small number of variables are not applicable. Data analytics need to be more pragmatic. Given a statistical problem, find a good solution, whether it is a formulaic model, an algorithmic model, or a Bayesian model or a completely different approach.

In the context of the hypothesis-driven analysis, one should pay attention to the question how far can we

\(^1\)In [36], instead of “formulaic modeling,” the term “data modeling” is used that looks misleading in the computer science context.
go applying the algorithmic modeling for hypothesis generation and testing. Various approaches to machine learning use related to hypothesis formation and selection can be found in [27, 36, 38].

Besides machine learning, an interesting example of algorithmic generation of hypotheses can be found in the IBM Watson project [39] where the symbiosis of the general-purpose reusable natural language processing (NLP) and knowledge representation and reasoning (KRR) technologies (under the name DeepQA) is exploited for answering arbitrary questions over the existing natural language documents as well as structured data resources. Hypothesis generation takes the results of question analysis and produces candidate answers by searching the available data sources and extracting answer-sized snippets from the search results. Each candidate answer plugged back into the question is considered a hypothesis, which the system has to prove correct with some degree of confidence. After merging, the system must rank the hypotheses and estimate confidence based on their merged scores. A machine-learning approach adopted is based on running the system over a set of training questions with known answers and training a model based on the scores. An important consideration in dealing with NLP-based scorers is that the features they produce may be quite sparse, and so, accurate confidence estimation requires the application of confidence-weighted learning techniques [39] — a new class of online learning methods that maintain a probabilistic measure of confidence in each parameter. It is important to note that instead of statistics based hypothesis testing, contextual evaluation of a wide range of loosely coupled probabilistic question and semantic based content analytics is applied for scoring different questions (hypotheses) and content interpretations. Training different models on different portions of the data in parallel and combining the learned classifiers into a single classifier allow to make the process applicable to the large collections of data. More details on that can be found in [39, 40] as well as in other Watson project related publications.

3.4 Bayesian motivation for discovery

One way for discriminating between competing models of some phenomenon is to use Bayesian model selection approach (see paragraph 3.2.1), the Bayesian evidences for each of the proposed models (hypotheses) can be computed and the models can then be ranked by their Bayesian evidence. This is a good method for identifying which is the best model in a given set of models, but it gives no indication of the overall goodness of the model. Bayesian model selection says nothing about the best model in the set may merely be the best of in a set of poor models. Knowing that the best model in the current set of models is not particularly good model would provide motivation to search for a better model and, hence, may lead to model discovery.

One way of assigning some measure of the absolute goodness of a model is to use the concept of Bayesian doubt first introduced in [41]. Bayesian doubt works by comparing all the known models in a set with an idealized model, which acts as a benchmark model.

An application of the Bayesian doubt method for the cosmological model building is given in [32, 42]. One of the most important questions in cosmology is to identify the fundamental model underpinning the vast amount of observations nowadays available. The so-called ‘cosmological concordance model’ is based on the cosmological principle (i.e., the Universe is isotropic and homogeneous, at least on large enough scales) and on the hot big bang scenario, complemented by an inflationary epoch. This remarkably simple model is able to explain with only half a dozen free parameter observations spanning a huge range of time and length-scales. Since both a cold dark matter (CDM) and a cosmological constant (Λ) component are required to fit the data, the concordance model is often referred to as ‘the ΛCDM model.’

Several different types of explanation are possible for the apparent late time acceleration of the Universe, including different classes of dark energy model such as ΛCDM, wCDM; theories of modified gravity; void models or the back reaction [32]. The methodology of Bayesian doubt which gives an absolute measure of the degree of goodness of a model has been applied to the issue of whether the ΛCDM model should be doubted.

The methodology of Bayesian doubt dictates that an unknown idealized model X should be introduced against which the other models may be compared. Following [41], ‘doubt’ may be defined as the posterior probability of the unknown model:

\[ D \equiv p(X|d) = \frac{p(d|X)p(X)}{p(d)}. \]

Here, \( p(X) \) is the prior doubt, i.e., the prior on the unknown model, which represents the degree of belief that the list of known models does not contain the true model. The sum of all the model priors must be unity.

The methodology of Bayesian doubt requires a baseline model (the best model in the set of known models), for which, in this application, the ΛCDM has been chosen. The average Bayes factor between ΛCDM and each of the known models is given by:

\[ \langle B_{i\Lambda} \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} B_{i\Lambda} . \]
The ratio $R$ between the posterior doubt and prior doubt, which is called the relative change in doubt, is:

$$R \equiv \frac{D}{p(X)}.$$  

For doubt to grow, i.e., the posterior doubt to be greater than the prior doubt ($R \ll 1$), the Bayes factor between the unknown model $X$ and the baseline model must be much greater than the average Bayes factor:

$$\frac{\langle B_{\Lambda \Lambda} \rangle}{B_{XX}} \ll 1.$$  

To genuinely doubt the baseline model, $\Lambda CDM$, it is not sufficient that $R > 1$, but additionally, the probability of $\Lambda CDM$ must also decrease such that its posterior probability is greater than its prior probability, i.e., $p(\Lambda|d) < p(\Lambda)$. One can define:

$$R_\Lambda \equiv \frac{p(\Lambda|d)}{p(\Lambda)}.$$  

For $\Lambda CDM$ to be doubted, the following two conditions must be fulfilled:

$$R > 1; \quad R_\Lambda < 1.$$  

If these two conditions are fulfilled, then it suggests that the set of known models is incomplete, and gives motivation to search for a better model not yet included, which may lead to model discovery.

In [41], a way of computing an absolute upper bound for $p(d|X)$ achievable among the class of known models has been proposed. Finally, it was found that current cosmic microwave background (CMB), matter power spectrum (mpk), and Type Ia supernovae (SNIa) observations do not require the introduction of an alternative model to the baseline $\Lambda CDM$ model. The upper bound of the Bayesian evidence for a presently unknown dark energy model against $\Lambda CDM$ gives only weak evidence in favor of the unknown model. Since this is an absolute upper bound, it was concluded that $\Lambda CDM$ remains a sufficient phenomenological description of currently available observations.

It becomes paramount to offer scientists mechanisms to manage the variety of knowledge produced during such investigations. Specific conceptual modeling facilities [43] are investigated to allow scientists to represent scientific hypotheses, models, and associated computational or simulation interpretations which can be compared against phenomena observations (see Fig. 3). The model allows scientists to record the existing knowledge about an observable investigated phenomenon, including a formal mathematical interpretation of it, if any. Model evolution and model sharing need also to be supported taking either a mathematical or computational view (e.g., expressed by scientific workflows). Declarative representation of scientific model allows scientists to concentrate on the scientific issues to be investigated. Hypotheses can be used also to bridge the gap between an ontological description of studied phenomena and the simulations. Conceptual views on scientific domain entities allow for searching for definitions supporting scientific models sharing among different scientific groups.

In [12], the engineering of hypothesis as linked data is addressed. A semantic view on scientific hypotheses shows their existence apart from a particular statement formulation in some mathematical framework. The mathematical expression is considered as not enough to identify the hypothesis: first, because it must be physically interpreted, and second, because there can be many ways to formulate the same hypothesis. The link to a mathematical expression, however, brings to the hypothesis concept higher semantic precision. Another link, in addition, to an explicit description of the explained phenomenon (emphasizing its “physical interpretation”) can bring forth the intended meaning. By dealing with that hypothesis as a conceptual entity, the scientists make it possible to change its statement or even to assert a semantic mapping to another incarnation of the hypothesis in case someone else reformulates it.

In [43], the following elements related to hypothesis-driven science are conceptualized: a phenomenon observed, a model interpreting this phenomenon, the metadata defining the related computation together with the simulation definition (for simulation, a declarative logical language is proposed). In this work, specific attention is devoted to hypothesis definition. The explanation, a scientific hypothesis conveys, is a relationship between the causal phenomena and the simulated one, namely, that the simulated phenomenon is caused by or produced under the conditions set by the causal phenomena. By running the simulations defined by the antecedents in the causal relationship, the scientist aims at providing hypothetical analysis of the studied phenomenon.

Thus, the scientific hypothesis becomes an element of the scientific model that may replace a phenomenon.

4 Facilities for the Scientific Hypothesis-Driven Experiment Support

4.1 Conceptualization of scientific experiments

Data intensive research increasingly becomes dependent on computational resources to aid complex researches.
When computing a simulation based on a scientific hypothesis, i.e., according to the causal relationship it establishes, the output results may be compared against phenomenon observations to assess the quality of the hypothesis. Such interpretation provides for bridging the gap between qualitative description of the phenomenon domain (scientific hypotheses may be used in qualitative (i.e., ontological) assertions) and the corresponding quantitative valuation obtained through simulations. According to the approach [43], complex scientific models can be expressed as the composition of computation models similarly to database views.

4.2 Hypothesis space browsers

In the HyBrow (Hypothesis Space Browser) project [44], the hypotheses for the biology domain are represented as a set of first-order predicate calculus sentences. In conjunction with an axiom set specified as rules that model known biological facts over the same universe and experimental data, the knowledge base may contradict or validate some of the sentences in hypotheses, leaving the remaining ones as candidates for new discovery. As more experimental data are obtained and rules are identified, discoveries become positive facts or are contradicted. In the case of contradictions, the rules that caused the problems must be identified and eliminated from the theory formed by the hypotheses. In such model-theoretical approach, the validation of hypotheses considers the satisfiability of the logical implications defined in the model with respect to an interpretation. This might be applicable also for simulation-based research, in which validation is solved based on the quantitative analysis between the simulation results and the observations [43]. HyBrow is based on an OWL ontology and application-level rules to contradict or validate hypothetical statements. HyBrow provides for designing hypotheses and evaluating them for consistency with existing knowledge and uses an ontology of hypotheses to represent hypotheses in machine understandable form as relations between objects (agents) and processes [45].

As an upgrade of HyBrow, the HyQue [46] framework adopts linked data technologies and employs Bio2RDF linked data to add to HyBrow semantic interoperability capabilities. HyBrow/HyQue’s hypotheses are domain-specific statements that correlate biological processes (seen as events) in the First-Order Logic (FOL). Hypotheses are formulated as instances of the HyQue Hypothesis Ontology and are evaluated through a set of SPARQL queries against biologically-typed OWL and HyBrow data. The query results are scored in terms of how the set of events correspond to background expectations. A score indicates the level of support the data lend the hypothesis. Each event is evaluated independently in order to quantify the degree of support it provides for the hypothesis posed. Hypothesis scores are linked as properties to the respective hypothesis.

OBI (the Ontology for Biomedical Investigations) project (http://obi-ontology.org) aims to model the design of an investigation: the protocols, the instrumentation, and the materials used in experiments and the data generated [47]. Ontologies such as EXPO and OBI enable the recording of the whole structure of scientific investigations: how and why an investigation was executed, what conclusions were made, the basis for these conclusions, etc. As a result of these generic ontology development efforts, the Minimum Information about a Genotyping Experiment (MIEx) recommends the use of terms defined in OBI. The use of a generic or a compliant ontology to supply terms will stimulate cross-disciplinary data-sharing and reuse. As much detail about an investigation as possible in order to make the investigation more reproducible and reusable can be collected [48].

Hypothesis modeling is embedded into the knowledge infrastructures being developed in various branches of science. One example of such infrastructure is considered under the name SWAN — a Semantic Web Application in Neuromedicine [47]. SWAN is a project for developing an integrated knowledge infrastructure for the Alzheimer disease (AD) research community. SWAN incorporates the full biomedical research knowledge lifecycle in its ontological model, including support for personal data organization, hypothesis generation, experimentation, laboratory data organization, and digital prepublication collaboration. The common ontology is specified in an RDF Schema. SWAN’s content is intended to cover all stages of the “truth discovery” process in biomedical research, from formulation of questions and hypotheses to capture of experimental data, sharing data with colleagues, and ultimately, the full discovery and publication process.

Several information categories created and managed in SWAN are defined as subclasses of Assertion. They include Publication, Hypothesis, Claim, Concept, Manuscript, Dataset, and Annotation. An Assertion may be made upon any other Assertion, or upon any object specifiable by URL. For example, a scientist can make a Comment upon, or classify, the Hypothesis of another scientist. Linking to objects “outside” SWAN by URL allows one to use SWAN as metadata to organize, for example, all one’s PDFs of publications, or the Excel files in which one’s laboratory data are stored, or all the websites of tools relevant to Neuroscience. Annotation may be structured or unstructured. Structured annotation means attaching a Concept (tag or term) to an Assertion. Unstructured annotation means attaching free text. Concepts are nodes in controlled vocabularies, which may also be hierarchical (taxonomies).
4.3 Scientific hypothesis formalization

An example showing in Fig. 6 the diversity of the components of a scientific hypothesis model has been borrowed from the applications in Neuroscience [43, 49] and in a human cardiovascular system in Computational Hemodynamics [12, 50]. The formalization of a scientific hypothesis was provided by a mathematical model, by a set of differential equations for continuous processes, quantifying the variations of physical quantities in continuous space–time, and by the mathematical solver (HEMOLAB) for discrete processes. The mathematical equations were represented in MathML, enabling models interchange and reuse.

In [51], the formalism of quantitative process models is presented that provides for encoding of scientific models formally as a set of equations and informally in terms of processes expressing those equations. The model revision works as follows. For input, it is required an initial model; a set of constraints representing acceptable changes to the initial model in terms of processes; a set of generic processes that may be added to the initial model; and observations to which the revised model should fit. These data provide the approach with a heuristic that guides search toward parts of the model space that are consistent with the observations. The algorithm generates a set of revised models that are sorted by their distance from the initial model and presented with their mean squared error on the training data. The distance between a revised model and the initial model is defined as the number of processes that are present in one but not in the other. The abilities of the approach have been successfully checked in several environmental domains.

Formalisms for hypothesis formation are mostly monotonic and are considered to be not quite suitable for knowledge representation, especially in dealing with incomplete knowledge, which is often the case with respect to biochemical networks. In [52], knowledge-based framework for the general problem of hypothesis formation is presented. The framework has been implemented by extending BioSigNet-RR — a knowledge-based system that supports elaboration tolerant representation and nonmonotonic reasoning. The main features of the extended system provide:

1. seamless integration of hypothesis formation with knowledge representation and reasoning;

---

**Figure 6** Elements of the scientific hypothesis model
use of various resources of biological data as well as human expertise to intelligently generate hypotheses; and

(3) support for ranking hypotheses and for designing experiments to verify hypotheses.

The extended system is positioned as a prototype of an intelligent research assistant of molecular biologists.

4.4 Hypothesis-driven robots

The Robot Scientist [53] oriented on genomic applications is a physically implemented system which is capable of running cycles of scientific experimentation and discovery in a fully automatic manner: hypothesis formation, experiment selection to test these hypotheses, experiment execution using robotic system, results analysis and interpretation, repeating the cycle (closed-loop in which the results obtained are used for learning from them and feeding the resulting knowledge back into the experimental models). Deduction, induction, and abduction are the types of logical reasoning used in scientific discovery (see section 3). The full automation of science requires ‘closed-loop learning,’’ where the computer not only analyses the results, but learns from them and feeds the resulting knowledge back into the next cycle of the process (Fig. 7).

In the Robot Scientist, the automated formation of hypotheses is based on the following key components:

(1) machine-computable representation of the domain knowledge;

(2) abductive or inductive inference of novel hypotheses;

(3) an algorithm for the selection of hypotheses; and

(4) deduction of the experimental consequences of hypotheses.

Adam, the first Robot Scientist prototype, was designed to carry out microbial growth experiments to study functional genomics in the yeast *Saccharomyces cerevisiae*, specifically to identify the genes encoding ‘locally orphan enzymes.’ Adam uses a comprehensive logical model of yeast metabolism, coupled with a bioinformatic database (Kyoto Encyclopaedia of Genes and Genomes — KEGG) and standard bioinformatics homology search techniques (PSI-BLAST and FASTA) to hypothesize likely candidate genes that may encode the locally orphan enzymes. This hypothesis generation process is abductive.

To formalize Adam’s functional genomics experiments, the LABORS ontology (LABoratory Ontology for Robot Scientists) has been developed. LABORS is a version of the ontology EXPO (as an upper layer ontology) customized for Robot scientists to describe biological knowledge. LABORS is expressed in OWL-DL. LABORS defines various structural research units, e.g., trial, study, cycle of study and replicate as well as design strategy, plate layout, expected actual results. The respective concepts and relations in the functional genomics data and metadata are also defined. Both LABORS and the corresponding database (used for storing the instances of the classes) are translated into Datalog in order to use the SWI-Prolog reasoner for required applications [48].

There were two types of hypotheses generated. The first level links an orphan enzyme, represented by its enzyme class (E.C.) number, to a gene (ORF) that potentially encodes it. This relation is expressed as a two-place predicate where the first argument is the ORF and the second is the E.C. number. An example of

![Figure 7 Hypothesis-driven closed-loop learning](image-url)

In the diagram:
- System model and knowledge base
- Initial START point: hypotheses generation
- Experiment generation and design
- Cycles of automated hypotheses generation and experimentation
- Execution of experiments on an automated robotic system
- Collection of experimental observations and other metadata
- New knowledge
- Analysis of results by statistics and machine learning
hypothesis at this level is: \textit{encodesORFtoEC('YBR166C', '1.1.1.25')}.

The second level of hypothesis involves the association between a specific strain, referenced via the name of its missing ORF, and a chemical compound which should affect the growth of the strain, if added as a nutrient to its environment. This level of hypothesis is derived from the first by logical inference using a specific model of yeast metabolism. An example of such a hypothesis is: \textit{affects growth('C00108', 'YBR166C')}, where the first argument is the compound (names according to KEGG) and the second argument is the strain considered.

Adam then designs the experimental assays required to test these hypotheses for execution on the laboratory robotic system. These experiments are based on a two-factor design that compares multiple replicates of the strains with and without metabolites compared against wild type strain controls with and without metabolites.

Adam follows a hypothetico-deductive methodology (see section 2). Adam abductively hypothesizes new facts about yeast functional biology, then it deduces the experimental consequences of these facts using its model of metabolism, which it then experimentally tests. To select experiments, Adam takes into account the variable cost of experiments, and the different probabilities of hypotheses. Adam chooses its experiments to minimize the expected cost of eliminating all but one hypothesis. This is, in general, an NP complete problem and Adam uses heuristics to find a solution [45].

It is now likely that the majority of hypotheses in biology are computer-generated. Computers are increasingly automating the process of hypothesis formation, for example: machine learning programs (based on induction) are used in chemistry to help design drugs; and in biology, genome annotation is essentially a vast process of (abductive) hypothesis formation. Such computer-generated hypotheses have been necessarily expressed in a computationally amenable way, but it is still not common practice to deposit them into a public database and make them available for processing by other applications [45].

The details describing the software and informatics decisions in the Robot Scientist project can be found in [45, 53] and online at the website http://www.aber.ac.uk/compsci/Research/bio/robotsci/data/informatics/. The details for developing the formalization used for Adam’s functional genomics investigations can be found in [48, 54]. An ontology-based formalization based on graph theory and logical modeling makes it possible to keep an accurate track of all the result units used for different goals, while preserving the semantics of all the experimental entities involved in all the investigations. It is shown how experimentation and machine learning are used to identify additional knowledge to improve the metabolic model [54].

4.5 Hypotheses as data in probabilistic databases

Another view of hypotheses encoding and management is presented in [55]. Authors use probabilistic database techniques for hypotheses systematic construction and management. MayBMS [56], a probabilistic database management system, is used as a core for hypothesis management. This methodology (called \(\gamma\)-DB) enables researchers to maintain several hypotheses explaining some phenomena and provides evaluation mechanism based on Bayesian approach to rank them.

The construction of \(\gamma\)-DB database comprises several steps. In the first step, phenomenon and hypothesis entities are provided as input to the system. Hypothesis is a set of mathematical equations expressed as functions in W3C MathML-based format and is associated with one or more simulation trial dataset, consisting of tuples with input variables of equation and its corresponding output as functionally dependent variables (the predictions). Phenomenon is represented by at least one empirical dataset similar to simulation trials. In the next step, the system deals with hypotheses and phenomena in the following way:

1. researcher has to provide some metadata about hypotheses and phenomena; e.g., hypotheses need to be associated with the respective phenomena and assigned a prior confidence distribution (uniform by default according to the principle of maximum entropy (see paragraph 3.2.3));
2. functional dependencies (FD) are extracted from equations in order to obtain database schema to store simulations and experimental data; it should be mentioned that to precisely identify hypothesis formulation, the special attributes for phenomena and hypothesis references are introduced into FD;
3. tuples are synthesized from simulation trials and observational data by uncertain pseudotransitive closure and reasoning; and finally,
4. the probabilistic \(\gamma\)-DB database is formed.

Once phenomenon and hypothesis (with empirical datasets and simulation trials) are produced, it becomes possible to manipulate them with database tools.

MayBMS provides tools to evaluate competing hypotheses for the explanation of a single phenomenon. With prior probabilities already provided, the system allows to make one or more (if new observational data appears) Bayesian inference steps. In each step, the prior probability is updated to posterior according to Bayes’ theorem. As a result, hypotheses which better explain phenomenon get higher probabilities enabling researchers to make more confident decisions (see also paragraph 3.2.1). The \(\gamma\)-DB approach provides a promising way to analyze hypotheses in large-scale...
5 Examples of Hypothesis-Driven Scientific Research

5.1 Hypotheses in Besançon Galaxy model

Various models in astronomy heavily rely on hypotheses. One of the most impressive is the Besançon galaxy model (BGM) [57–59] evolving for many years and representing the population and structure synthesis model for the Milky Way. It allows astronomers to test hypotheses on the star formation history, star evolution, and chemical and dynamical evolution of the Galaxy. As the result of simulation process, one can get the following: multidimensional histograms of intrinsic star properties or observable properties, a catalog of pseudoobservations, or the integrated luminosity in a specified photometric band [60]. From the beginning, the aim of the BGM was not only to be able to simulate reasonable star counts but further to test scenarios of Galactic evolution from assumptions on the rate of star formation (SFR), initial mass function (IMF), and stellar evolution.

The model has explicit and implicit hypotheses associated with it. Explicit hypotheses are usually some sets of equations, taken from the literature studies and put as the ingredient of the model. Some of explicit hypotheses are passed as the input of the model, e.g., star formation rate, initial mass function, evolutionary tracks, chemical evolution, atmosphere models, density laws, interstellar extinction model.

The model has some implicit hypotheses as well. For example, it is assumed that no star population comes from the outside of the Galaxy. There are several more implicit hypotheses about disk formation and dark matter assumptions encoded inside the model. It is usually much harder to get all the implicit hypotheses, since many of them are not described in the articles and are difficult to pin from the code.

BGM has not only the large number of explicit and implicit hypotheses, but also a complex interrelations between them. So, some of the hypotheses are being independent, e.g., IMF and SFR; so, it is possible to change them independently. On the other side, some of the hypotheses are connected, e.g., the age distribution, the density laws, and the potential are linked with the age–velocity dispersion via the Boltzmann equation and need to be consistent. Such kind of dependencies make the model hard to be tested and to keep it consistent while varying different parameters during model fitting. Another example of interrelations of hypotheses is competing hypotheses.

BGM has changed drastically over the last 30 years. This has happened because of the appearance of new data surveys, technologies, and methods of observation development. As an example of such evolution, the model developed in 2014 compared to previous versions handles variations of the SFR, IMF, evolutionary tracks, and atmosphere models. These hypotheses are passed as input parameters to the model; so, the user can vary them.

The second improvement of the model is the implementation of the stellar binarity, being an important change since binaries can account for about 50% of the total stellar content of the Milky Way. The authors of the new version underline the importance of understanding interrelationships between different hypotheses and need for model evolution tools [60]: “In practice, to build a Galaxy from the fundamental building-blocks, we had to reconstruct the previous model and apply important changes in the code arrangement. That required to understand well the underlying relations between all mentioned components.”

It is planned further to focus on the renewed BGM [59], in which authors draw their attention to the Galaxy thin disk treatment and use of Tycho-2 as a testing dataset. The parameters of BGM (such as IMF, SFR and evolutionary track sets) explicitly and model ingredients implicitly can be treated as hypotheses. Model ingredients include the treatment of binarity, the local stellar mass densities of thin disk, extinction model, age-metallicity and age-velocity relations, radial scale length, the age of the Galaxy thin disc, different sets of the star atmosphere models, etc.

Tycho-2 dataset and \( \chi^2 \)-type statistics test is used to test various versions of these hypotheses in order to choose the most appropriate ones and update model to better fit the provided data. The tests were made by comparing star counts and \( (B−V)T \) color distributions between data and simulations. Two different tests were used to evaluate the adequacy of the stellar densities globally and to test the shape of the color distribution. Other parameters to be tested are: star counts, radio velocity, magnitudes, colors, proper motions, parallax, effective temperatures, gravity, and metallicity. Authors use histograms, 2 goodness of fit (maximum likelihood and \( \chi^2 \)-test) and for velocity parameter, Kolmogorov–Smirnov and Henderson–Darling tests.

Due to the fact that some ingredients of the model are highly correlated (such as the IMF, SFR, and the local mass density), the authors defined default models as a combination of a new set of ingredients that significantly improve the fit to Tycho data. So, 11 IMF functions, 2 SFR functions, 2 evolutionary track sets, 3 sets of atmosphere models, 3 values for the age of the formation of the thin disk, and 3 sets of values of the thin disk local stellar volume mass density were tested. As a result of testing, the two most appropriate IMS and SFR hypotheses were chosen.
BGM authors have plans to incorporate other star surveys and test the model against them. To do simulations directly comparable with data, the selected magnitudes from the surveys need to be complete in terms of magnitude. Among these surveys, there are the Geneva–Copenhagen survey, SDSS–II/III, SEGUE/SEGUE2, APOGEE, RAVE, LAMOST, Gaia, Gaia-ESO, GALAH LSST, WEAVE, 4MOST, and MOONS surveys [61].

5.2 Connectome analysis based on network data

In the neuroscience community, the development of common paradigms for interrogating the myriad functional systems in the brain remains to be the core challenge. Building on the term “connectome,” coined to describe the comprehensive map of neural connections in the human brain, the “functional connectome” denotes the collective set of functional connections in the human brain (its “wiring diagram”) [62]. More broadly, a connectome would include the mapping of all neural connections within an organism’s nervous system. The production and study of connectomes, known as connectomics, may range in scale from a detailed map of the full set of neurons and synapses within part or all of the nervous system of an organism to a macroscale description [63] of the functional and structural connectivity between all cortical areas and subcortical structures. The ultimate goal of connectomics is to map the human brain. In functional magnetic resonance imaging (fMRI), associations are thought to represent functional connectivity in the sense that the two regions of the brain participate together in the achievement of some higher-order function, often in the context of performing some task. fMRI has emerged as a powerful tool used to interrogate a multitude of functional circuits simultaneously. This has elicited the interest of statisticians working in that area. At the level of basic measurements, neuroimaging data can be considered to consist typically of a set of signals (usually, time series) at each of a collection of pixels (in two dimensions) or voxels (in three dimensions). Building from such data, various forms of higher-level data representations are employed in neuroimaging. In recent years, a substantial interest in network-based representations has emerged in neuroimaging to use networks to summarize relational information in a set of measurements, typically assumed to be reflective of either functional or structural relationships between regions of interest in the brain. With neuroimaging, now, a standard tool in clinical neuroscience, quickly moving towards a time in which we will have available databases composed of large collections of secondary data in the form of network-based data objects, is predictable.

One of the most basic tasks of interest in the analysis of such data is the testing of hypotheses in answer to questions such as “Is there a difference between the networks of these two groups of subjects?” Networks are not Euclidean objects and, hence, classical methods of statistics do not directly apply. Network-based analogues of classical tools for statistical estimation and hypothesis testing are investigated in [64, 65]. Such research is motivated by the 1000 Functional Connectomes Project (FCP) launched in 2010 [62]. The 1000 FCP [66] constitutes the largest data set of its kind similarly to large data sets in genetics. Other projects (such as the Human Connectome Project (HCP)) are aimed to build a network map of the human brain in healthy, living adults. The total volume of data produced by the HCP will likely be multiple petabytes [67]. HCP informatics platform includes data management system ConnectomeDB that is based on the XNAT (eXtensive Neuroimaging Archive Toolkit) imaging informatics platform [68], a widely used open source system for managing and sharing imaging and related data.

Now, HCP has information about more than 500 subjects including structural scans (T1w and T2w), resting-state fMRI (rfMRI), task fMRI (tfMRI), and high angular resolution diffusion imaging (dMRI). In addition, some resting-state MEG (rMEG) and/or task MEG (tMEG) data are available.

Data come in several formats: “unprocessed” raw data, “minimally preprocessed,” and “analysis” datasets. Preprocessed datasets have spatial distortions minimized and data have been aligned across modalities and across subjects using appropriate volume-based and surface-based registration methods. HCP consortium recommends to use the preprocessing dataset.

Visualization, processing, and analysis of high-dimensional data such as images often require some kind of preprocessing to reduce the dimensionality of the data and find a mapping from the original representation to a low-dimensional vector space. The assumption is that the original data resides in a low-dimensional subspace or manifold [69], embedded in the original space. This topic of research is called dimensionality reduction, nonlinear dimensionality reduction, including methods for parameterization of data using low-dimensional manifolds as models. Within the neural information processing community, this has become known as manifold learning. Methods for manifold learning are able to find nonlinear manifold parameterizations of datapoints residing in high-dimensional spaces, very much like Principal Component Analysis (PCA) is able to learn or identify the most important linear subspace of a set of data points (projecting data on a n-dimensional linear subspace which maximizes the variance of the data in the new space).
In [64], necessary mathematical properties associated with a certain notion of a ‘space’ of networks used to interpret functional neuroimaging connectome-oriented data are established. Extension of the classical statistics tools to network-based datasets, however, appeared to be highly nontrivial. The main challenge in such an extension is due to the fact that networks are not Euclidean objects (for which classical methods were developed) — rather, they are combinatorial objects, defined through their sets of vertices and edges. In [64], it was shown that networks can be associated with certain natural subsets of Euclidean space and demonstrated that through a combination of tools from geometry, probability on manifolds, and high-dimensional statistical analysis, it is possible to develop a principled and practical framework in analogy to classical tools. In particular, an asymptotic framework for one- and two-sample hypothesis testing has been developed. Key to this approach is the correspondence between an undirected graph and its Laplacian, where the latter is defined as a matrix (associating with a network). Graph Laplacian appeared to be particularly appropriate to be used for such matrices. The space of graph Laplacians is used working in certain subsets of Euclidian space which are some submanifolds of the standard Euclidian space.

The 1000 FCP describes functional neuroimaging data from 1093 subjects, located in 24 community-based centers. The mean age of the participants was 29 years, and all subjects were 18 years old or older. It is of interest to compare the subject-specific networks of males and females in the 1000 FCP data set. In [64], for the 1000 FCP, database comparing networks with respect to the sex of the subjects, over different age group, and over various collection sites is considered. It is shown that it is necessary to compute the means in each subgroup of networks. This was done by constructing the Euclidean mean of the Laplacians for each group of subjects in different age groups. Such group-specific mean Laplacians can then be interpreted as the mean functional connectivity in each group. Such approach provides for building the hypothesis tests about the average of networks or groups of networks to investigate the effect of sex differences on entire networks.

For the 1000 FCP data set, it was tested using the two-sample test for Laplacians whether sex differences were significant to influence patterns of brain connectivity. The null hypothesis of no group differences was rejected with high probability. Similarly for the three different age cohorts, the null hypothesis of no cohort differences was also rejected with high probability.

On such examples, it was shown [64] that the proposed global test has sufficient power to reject the null hypothesis in cases when mass-univariate approach (considered to be the gold standard in fMRI research [70]) fails to detect the differences at the local level. According to the mass-univariate approach, statistical analysis is performed iteratively on all voxels to identify brain regions whose fMRI detected responses display significant statistical effects. Thus, it was shown that a framework for network-based statistical testing is more statistically powerful than a mass-univariate approach.

It is expected that in the near future, there will be a plethora of databases of network-based objects in neuroscience motivating the development and extension of various tools from classical statistics to global network data.

In paper [71] discussing the relationship between neuroimaging and Big Data areas, it is analyzed how modern neuroimaging research represents a multifactorial and broad ranging data challenge, involving the growing size of the data being acquired; sociological and logistical sharing issues; infrastructural challenges for multisite, multitype archiving; and the means by which to explore and mine these data. As neuroimaging advances further, e.g., aging, genetics, and age-related disease, new vision is needed to manage and process this information while marshalling of these resources into novel results. It is predicted that on this way, “big data” can become “big” brain science.

In [72], authors formulate a hypothesis about the brain connectivity and evaluate it against HCP data. They use the task fMRI data, there is specific data about a well-validated task used to probe animate motion detection. The audience was shown short videoclips (20 s) of objects (squares, circles, and triangles) either interacting in some way (animate motion) or moving mechanically (inanimate motion). Participants rated the video by selecting if there was any social interaction, no interaction, or not sure for interaction. There were 2 sessions comprised of 5 videoblocks.

Hypothesis states that some regions of the brain (V5 and pSTS) are effectively connected and impacted by animate motion.

To test it, general linear models were used. The time series were modeled with regressors All motion — No motion, Animate—Inanimate motion. Together with regressors about head, tongue, and finger movement, these regressors were used to build general linear model. A group level ANOVA was performed to identify significant regional effects for the All Motion contrast and a contrast for Animate—Inanimate motion. For effective connectivity discovery, Dynamic Causal Modeling (DCM) technique was used. DCM tells about self-, forwards, and backward connections between active brain regions during an experiment, enabling to infer the way of brain regions impact each other mostly. As the result of DCM modeling, 16 models were created and passed as the input to Bayesian Model Selection procedure, which chose the winning model among them.
The results show that there is a connectivity between V5 and the pSTS brain regions in both hemispheres, which was independent of the type of motion. Animate motion stimulates the forward and backward connection between V5 and the pSTS in both hemispheres.

5.3 Climate in Australia

Another view on hypothesis representation and evaluation is presented in [73]. Authors argue that as long as in DIR data relevant to some hypotheses get continuously aggregated as time passes, hypotheses should be represented as programs that are executed repeatedly, as new relevant amounts of data get aggregated. Their method and techniques are illustrated by examining hypotheses about temperature trends in Australia during the 20th century. The hypothesis being tested comes from [74], stated that the temperature series is not stationary and is integrated of order 1 (I(1)). Nonstationarity means that the level of the time series is not stable in time and can show increasing and decreasing trends; I(1) means that by differentiating the stochastic process, a stationary process (main statistical properties of the series remain unchanged) is obtained. Phillips–Perron test and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test are used and both of them are executed in R. Several data sources are crawled: (i) The National Oceanographic and Atmospheric Administration marine and weather information; and (ii) Australian Bureau of Meteorology dataset. The framework consists of R interpreter and R SPARQL, tseries packages. Authors also used agINFRA for computation and rich semantics to support traditional scientific workflows for natural sciences. Authors received further evidence on different independent dataset that time series is integrated of order 1.

5.4 Financial market

Efficient-market hypothesis (EMH) is one of the most prominent in finance and “asserts that financial markets are informationally efficient.” In [75], authors test the weak form of EMH, stating that prices on traded assets (e.g., stocks, bonds, or property) already reflect all past publicly available information. The null hypothesis states that successive prices changes are independent (random walk). The alternative hypothesis states that they are dependent. To check if the successive closing prices are dependent of each other, the following statistical tests were used: a serial correlation test, a runs test, an augmented Dickey–Fuller test, and the multiple variance ratio test. Tests were performed on daily closing prices from the six European stock markets (France, Germany, U.K., Greece, Portugal, and Spain) during the period between 1993 and 2007. The result of each test states whether successive closing prices are dependent of each other.

Test provides evidence that for monthly prices and returns, the null hypothesis should not be rejected for all six markets. If daily prices are concerned, the null hypothesis is not rejected for France, Germany, U.K., and Spain, but this hypothesis is rejected for Greece and Portugal. However, on the 2003–2007 dataset, the null hypothesis for these two countries is not rejected as well.

In [76], Bollen et al. use different approach to test EMH. Authors investigate whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict the stock market. They build public mood time series by sentiment analysis of tweets from February 28 to December 19, 2008 and try to show that it can predict Dow Jones Index corresponding values. The null hypothesis states that the mood time series do not predict DJIA (Dow Jones Industrial Average) values. Granger causality analysis in which Dow Jones values and mood time series are correlated is used to test the null hypothesis. Granger causality analysis is used to determine if one time series can predict another time series. Its results reject the null hypothesis and claim that public opinion is predictive of changes in DJIA closing values.

5.5 Publication-based automated hypothesis generation in life sciences

Researchers and scientists from leading academic, pharmaceutical, and other research centers have begun deploying IBM’s Watson Discovery Advisor to rapidly analyze and test hypotheses using data in millions of scientific papers available in public databases. A new scientific research paper is published nearly every 30 s, which equals more than a million annually. According to the National Institutes of Health, a typical researcher reads about 23 scientific papers per month, which translates to nearly 300 per year, making it humanly impossible to keep up with the ever-growing body of scientific material available. Building on Watson’s ability to understand nuances in natural language, Watson Discovery Advisor can understand the language of science, such as how chemical compounds interact, making it a uniquely powerful tool for researchers in life sciences and other research and industrial domains. Specifically, the Watson Discovery Advisor for Life Sciences is armed with expertise and understands field-specific lexicon in areas such as clinical trial data, genomics, drugs, and human anatomy.

Recently, scientists of Baylor College of Medicine and IBM using the Baylor Knowledge Integration Toolkit (KnIT), based on Watson technology, identified new enzymes (called kinases) that can modify p53, an important protein related to many cancers [77]. There are
over 240,000 papers that mention one or more of 500+ known human kinases in their Medline abstract. There are over 70,000 papers published on p53 to make their analysis manually is completely unrealistic task. Watson analyzed the scientific articles related to p53 to predict proteins that turn on or off p53’s activity. This automated analysis led the Baylor cancer researchers to identify six potential p53 kinases to target for new research. These results are notable, considering that over the last 30 years, scientists averaged one p53 kinase discovery per year. Knowing which proteins are modified by each kinase and, therefore, which kinases would make good drug targets is a difficult and unsolved problem. There are over 500 known human kinases and tens of thousands of possible proteins they can target.

KnIT collects the abstracts to be mined applying queries. A specific kinase name and its synonyms are used in this process. Entity resolution process looks as follows. The words and phrases that make up the document feature space are determined by counting the number of documents in which each word appears and identifying the words with the highest counts. A phrase is considered to be a sequence of two words. Only the $N$ most frequent words and phrases are selected. This becomes the feature space.

Once a feature space is received, a representation of each kinase by averaging the feature vectors of all documents that contain the kinase is created. This is the kinase centroid. Next, a distance matrix is calculated that measures the distance between each kinase and every other kinase in the space.

Finally, a meaningful picture of kinase–kinase relationships is obtained. Thus, it is possible to identify a set of kinases that may modify p53. However, some sort of principled ranking scheme is needed in order to prioritize the kinases for further experimentation. To provide such a scheme, the graph diffusion method [78] was used. Graph diffusion is a semisupervised learning approach for classification based on labeled and unlabeled data. It takes known information (initial labels) and then constrains the new labels to be smooth in respect to a defined structure (e.g., a network). In the case considered, it is known which kinases can modify p53 (initial labels); one would like to know which other proteins can modify p53 (final labels). The distance matrix based on the literature gives the structure of the kinase network. The initial labels are extracted from current knowledge found in review articles.

To test the algorithm, it was first applied in a retrospective analysis to show whether recent annotations of new p53 kinases occurring after a certain date (2003) could be predicted from a model that only took into account papers written before that date, at a time when these discoveries of p53 kinases were still unknown. Next, it was asked whether some variations in the algorithm could improve p53 kinase prediction as its performance was compared to the common approach used most typically to identify functionally similar proteins in biology. Finally, the analysis was expanded to a larger set of proteins to test scalability.

This research represents the first stage in the IBM–Baylor collaborative effort and as such, it proves the principle that mining past literature is a viable strategy for predicting previously unknown biological events. It was shown that p53 kinases predicted with the text mining methods are supported by laboratory findings. In the future, it should be possible to make many other kinds of predictions on a much larger scale as the infrastructure and capabilities will be increased. In the future, it is planned to focus on a wider area of proteins and functions, building up comprehensive networks of interactions and predicting where new connections ought to exist based on everything else that is known. It is expected that this will ultimately accelerate the pace of cancer discoveries by an order of magnitude and allow scientists to come to a much more complete understanding of the mechanisms behind this disease.

Expanding KnIT to other areas of biology or the physical sciences is not straightforward. For example, to generalize to more proteins and genes is a big problem. In subjects like physics, results tend to be presented using equations and graphs rather than words. However, data-mining groups are working to retrieve information from these, too.

6 Concluding Remarks

The objective of this study is to analyze, collect, and systematize information on the role of hypotheses in the DIR process as well as on support of hypothesis formation, evaluation, selection, and refinement in course of the natural phenomena modeling and scientific experiments. The discussion is started with the basic concepts defining the role of hypotheses in the formation of scientific knowledge and organization of the scientific experiments. Based on such concepts, the basic approaches for hypothesis formulation applying logical reasoning, various methods for hypothesis modeling and testing (including classical statistics, Bayesian hypothesis, and parameter estimation methods, hypothetico-deductive approaches) are briefly introduced. Special attention is given to discussion of the data mining and machine learning methods role in process of generation, selection, and evaluation of hypotheses as well as the methods for motivation of new hypothesis formulation. Facilities of informatics for support of hypothesis-driven experiments, considered in the paper, are aimed at the conceptualization of scientific experiments, hypothesis formulation, and browsing in various domains (includ-
ing biology, biomedical investigations, neuromedicine, and astronomy), automatic organization of hypothesis-driven experiments. Examples of scientific research applying hypotheses considered in the paper include modeling of population and structure synthesis of the Galaxy, connectome-related hypothesis testing, studying of temperature trends in Australia, analysis of stock markets applying the EMH, as well as algorithmic generation of hypotheses in the collaborative project based on IBM Watson–Baylor Knowledge Integration Tool-kit applying the NLP and knowledge representation and reasoning technologies. An introduction into the state of the art of the hypothesis-driven research presented in the paper opens a way for investigation of the generalized approaches for efficient organization of hypothesis-driven experiments applicable for various branches of DIR.

References


Contributors

Kalinichenko Leonid A. (b. 1937) — Doctor of Science in physics and mathematics, professor; Head of Laboratory, Institute of Informatics Problems, Russian Academy of Sciences; 44-2 Vavilov Str., Moscow 119333, Russian Federation; professor, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; leonidandk@gmail.com

Kovalev Dmitry Yu. (b. 1988) — junior scientist, Institute of Informatics Problems, Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; dkovalev@ipiran.ru

Kovaleva Dana A. (b. 1973) — Candidate of Science (PhD) in physics and mathematics, scientist, Institute of Astronomy, Russian Academy of Sciences, 48 Pyatnitskaya Str., Moscow 119017, Russian Federation; dana@inasan.ru

Malkov Oleg Yu. (b. 1961) — Doctor of Science in physics and mathematics, associate professor; Head of Department, Institute of Astronomy, Russian Academy of Sciences; 48 Pyatnitskaya Str., Moscow 119017, Russian Federation; professor, Faculty of Physics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; malkov@inasan.ru

Received February 10, 2015
МЕТОДЫ И СРЕДСТВА ПОДДЕРЖКИ ИССЛЕДОВАНИЙ, ДВИЖИМЫХ ГИПОТЕЗАМИ: ОБЗОР

Л. А. Калиниченко1, Д. Ю. Ковалев2, Д. А. Ковалева3, О. Ю. Малков4

1 Институт проблем информатики Российской академии наук; leonidandk@gmail.com
2 Институт проблем информатики Российской академии наук; dkovalev@ipiran.ru
3 Институт астрономии Российской академии наук; dana@inasan.ru
4 Институт астрономии Российской академии наук; malkov@inasan.ru

Аннотация: Исследования с интенсивным использованием данных (ИИИД), развиваемые в рамках новой парадигмы изучения естественных явлений, именуемой Четвертой парадигмой, придают особое значение все возрастающей роли, которую играют данные, полученные в результате наблюдений, экспериментов или компьютерного моделирования, практически во всех областях анализа и накопления информации. Главной целью ИИИД является извлечение (вывод) знаний из данных. Целью настоящей работы является обзор существующих подходов, методов и инфраструктур анализа данных в ИИИД с акцентом на роли гипотез в процессе анализа информации и эффективной поддержки формирования, оценки и выбора гипотез при моделировании естественных явлений и проведении экспериментов. Статья включает введение в разнообразные понятия, методы и средства эффективной организации движимых гипотезами экспериментов в ИИИД.

Ключевые слова: исследования с интенсивным использованием данных; Четвертая парадигма; гипотезы; модели; теории; гипотетико-дедуктивный метод; проверка гипотез; решетка гипотез; модель Галактики, анализ коннектома; автоматизированная генерация гипотез

DOI: 10.14357/19922264150104

Литература


Поступила в редакцию 10.02.2015