Introduction into Analysis of Methods and Tools for Hypothesis-Driven Scientific Experiment Support

© Kalinichenko L.A. © Kovalev D.Y. © Kovaleva D.A. © Malkov O.Y.
Institute of Informatics Problems RAS Institute of Astronomy RAS
Moscow
leonidk@synth.ipi.ac.ru dm.kovalev@gmail.com dana@inasan.ru malkov@inasan.ru

Abstract

Data intensive sciences (DIS) are being developed in frame of the new paradigm of scientific study known as the Fourth paradigm, emphasizing an increasing role of observational, experimental and computer simulated data practically in all fields of scientific study. The principal goal of data intensive research (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 research 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 in the paper.

1 Hypotheses, theories, models and laws in data intensive science

Data intensive science (DIS) is being developed in accordance with the 4th Paradigm [29] of scientific study (following three previous historical paradigms of the science development (empirical science, theoretical science, computational science)) emphasizing that science as a whole is becoming increasingly dependent on data as the core source for discovery. Emerging of the 4th Paradigm is motivated by the huge amounts of data coming from scientific instruments, sensors, simulations, as well as from people accumulating data in Web or social nets. The basic objective of DIS 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 21st century. Data Intensive Research (DIR) denotes a crosscut of DIS/IT areas aimed at the creation of effective data analysis technologies for DIS and other data intensive domains.

Science endeavors to give a meaningful description of the world of natural phenomena using what are known as laws, hypotheses and theories. Hypotheses, theories and laws in their essence have the same fundamental character (Fig. 1) [48].

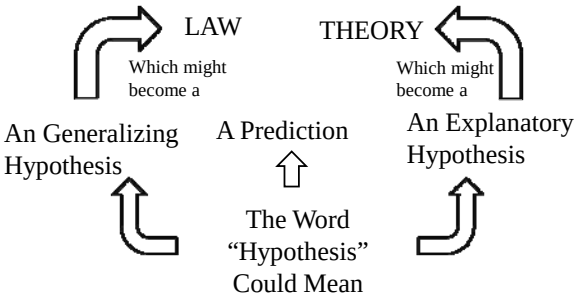

Fig. 1. Multiple incarnations of hypotheses

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, t*he 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 (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.

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 at the Fig. 1 is subjective. [40] 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

Proceedings of the 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2014, Dubna, Russia, October 13–16, 2014.

hypotheses, theories and laws. Important role of hypotheses in scientific research can scarcely be overestimated. In the edition of M. Poincaré's book [52] 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 driven 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 DIS is motivated by the necessity of analysis of the existing massive data to generate a hypothesis.

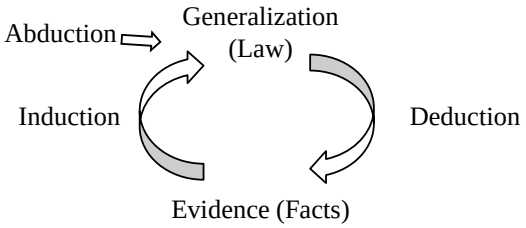

Fig. 2 Enhanced knowledge production diagram

In the course of our study paying attention to the issue of inductive and deductive reasoning in hypothesis driven sciences will be emphasized. On Fig. 2 such ways of knowledge production are shown [48]. “Generalization” here means any subset of hypotheses, theories and laws and “Evidence” is any subset of all facts accumulated in a specific DIS.

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 [4]. The method of (naïve) induction (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 both 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. The Fig. 2 shows that a valid law would permit the accurate prediction of facts not yet known. Also an *abduction* [49] 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 [52] 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;

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 the M. Poincaré's book [52] 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 M. 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 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. Into Section 5 several examples of organization of hypothesis driven scientific experiments are included. Conclusion summarizes 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 and 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 non-causal (associative) way [25].

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 [25].

In case when a possible correlation or similar relation between variables is investigated (such as, e.g., 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 [25]. 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* [53], 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 [53]. 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 [53], 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 [37].

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* [28]. 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 on Fig. 3 [23, 57]. The hypothesis triangle relations, *explains, formulates, represents* are functional in the scientist's final decision in adopting a particular model *m1* to formulate a hypothesis *h1*, which is meant to explain phenomenon *p1*.

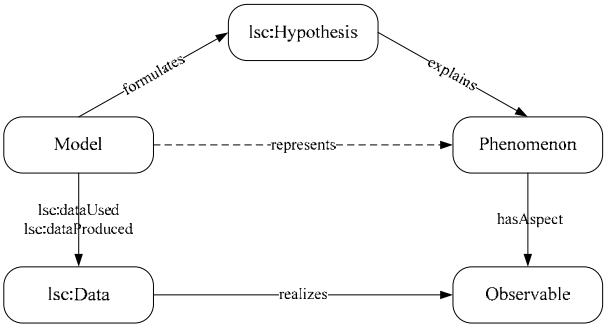

Fig. 3. Elements of hypothesis-driven research

In [23] the lattice structure for hypothesis interconnection is proposed as shown on 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*.

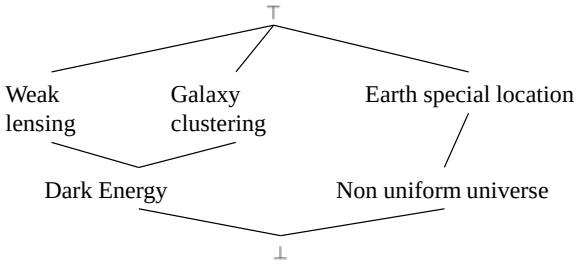

Fig. 4. A lattice theoretic representation for hypothesis relationship

The hypothesis lattice is unfolded into model and phenomena isomorphic lattices according to the hypothesis triangle (Fig. 3) [23]. 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 on the Fig. 5 [23]. This particular lattice corresponds to the case in Computational Hemodynamics considered in [23]. Here model *m1* formulates hypothesis *h1*, which explains phenomenon *p1*. Similarly, *m2* formulates *h2*, which explains *p2*, and so on. Properties of the hypothesis lattices and operations over them are considered in [24].

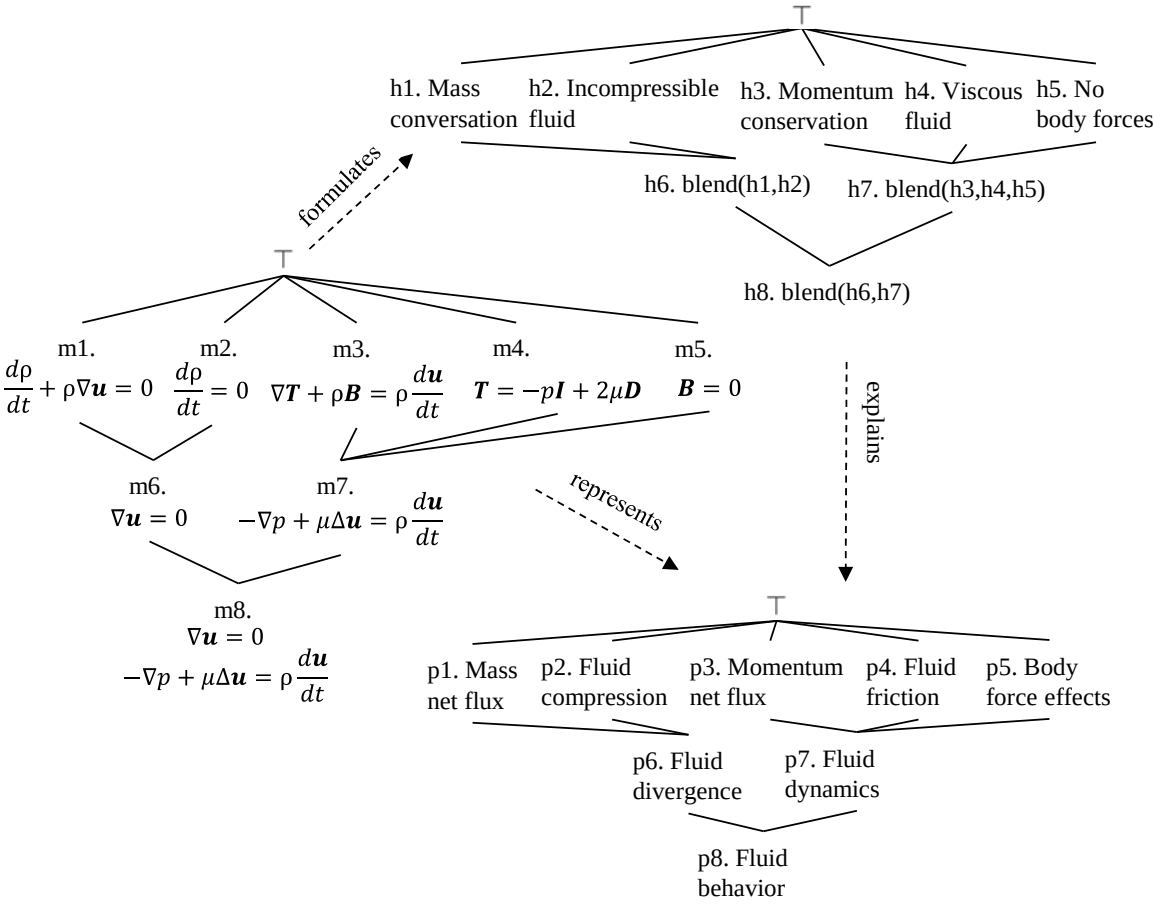

Fig. 5. Hypothesis lattice unfolded into model and phenomenon isomorphic lattice

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 we 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 [73].

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 [50] an abduction happens as follows: 1) Some phenomena *p1, p2, p3,* … are encountered for which there is no or little explanation; 2) However, *p1, p2, p3,* … would not be surprising if a hypothesis *H* were added. They would certainly follow from something like *H* and would be explained by it; 3) Therefore there is good reason for elaborating an hypothesis *H* – for proposing it as a possible hypothesis from which the assumption *p1, p2, p3,* … might follow. The abductive model of reasoning is primarily a process of explaining anomalies or surprising phenomena [63]. 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 [36]. Hypothesis generation in abduction logical framework is organized as follows. During the experiment, some new observations are encountered. Let *B* represents the background knowledge; *O* is the set of facts that represents observations. Both *B* and *O* are logic programs (set of rules in some rule language). In addition, *Γ* 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 *Γ*, 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* is consistent, and 3) *H* is some subset of *Γ*. If all conditions are met then *H* is an explanation of *O* (with respect to *B* and *Γ*). Examples of abductive logic programming systems include ACLP [35], A-system [71], ABDUAL [2] and ProLogICA [59]. Abductive logic programming can also be implemented by means of Answer Set Programming systems, e.g. by the DLV system [14].

The example abductive logic program in ProLogICA describes a simple model of the lactose metabolism of the bacterium E.Coli [59]. 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, *Γ*, 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 we want to find out in each problem case that we are examining. 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).
code(lac(y),permease).
code(lac(z),galactosidase).
temperature(low):-amount(glucose,low).
false :- amount(S,V1), amount(S,V2), V1 != V2.

##  Abducibles (Γ)
abducible_predicate(amount).

## Observation (O)
feed(lactose).

This goal generates two possible hypotheses:
{amount(lactose,hi), amount(glucose,low)}
{amount(lactose,medium),amount(glucose,medium)}
```

Just a couple of another examples of real rule-based systems, where abductive logic programming is used.

Robot Scientist (see 4.4) abductively hypothesizes new facts about the yeast functional biology by inferring what is missing from a model [38]. In [68], both abduction and induction are used to formulate hypotheses about inhibition in metabolic pathways. Augmenting background knowledge is done with abduction, after that induction is used for learning general rules. In [33] authors use SOLAR reasoning system to abductively generate hypotheses about the inhibitory effects of toxins on the rat metabolisms.

The process of discovery is deeply connected also with the search of *anomalies*. There are a lot of methods and algorithms to discover anomalies. Anomaly detection is an important research problem in data mining that aims to find objects that are considerably dissimilar, exceptional and inconsistent with respect to the majority data in an input database [6].

Analogies play several roles in science. Not only do they contribute to discovery but they also play a role in the development and evaluation of scientific theories (new hypotheses) by analogical reasoning.

3.2 Hypothesis evaluation

Being testable and falsifiable, a scientific hypothesis provides a solid basis to its further modeling and testing. There are several ways to do it, including the use of statistics, machine learning and logic reasoning techniques.

3.2.1 Statistical testing of hypotheses

The classical (*frequentist*) and *Bayesian* statistic approaches are applicable for hypothesis testing and selection. Brief summary of the basic differences between these approaches are as follows [34].

Classical (frequentist) statistics is based on the following beliefs:

− 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;

− 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;

− 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) [64]. 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. Failing to provide 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, we 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 our *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* α, 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 we fail to reject a hypothesis, it does not mean that we proved its correctness because it may be that our sample is simply not large enough to detect an effect.

When performing these tests, we can meet with two types of errors, which statisticians call *Type I and Type II errors*. Type I errors are 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 α. 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 β, and is related to *the power of* α *test* as $(1 - \beta)$. Hypothesis testing is intimately related to comparisons of distributions.

As the significance level α 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 α, 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 α. The fraction of false positives depends not only on α 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 we can assume (or not) about the underlying distributions, and the specific question we ask, we can use different statistical tests. 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 threshold probability α, typically 0.05 or 0.01 ($p < \alpha$). Note again that $p > \alpha$ does *not* mean that the hypothesis is *proven* to be correct.

The number of various statistical tests in the literature is overwhelming and their applicability is often hard to decide (see [19, 31] for variety of statistical methods in SPSS). 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, $\{x_{1i}\}$, $i = 1, ..., N_1$ and $\{x_{2i}\}$, $i = 1, ..., N_2$. 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 we know, or can assume, that both $h(x)$ and $f(x)$ are Gaussian distributions (e.g., the Anderson–Darling test, the Shapiro–Wilk test) [34]. More on statistical tests can be found in [19, 31, 32, 34].

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, $\{M_1, ..., M_n\}$, given the data.

We need to 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:

$$p(M_j|d) = p(d|M_j)\, p(M_j) / p(d),$$

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 we have analyzed the current data, $p(d|M_j)$ is the model (hypothesis) *likelihood* (represents the probability that some data are produced under the assumption of this model) and $p(d)$ is a normalization constant given by:

$$p(d) = \sum\nolimits_i p(d|M_i)\, p(M_i)\,.$$

The relative 'goodness' of models is given by a comparison of their posterior probabilities, so to compare two models M_a and M_b, we look at the ratio of the model posterior probabilities:

$$p(M_a|d) / p(M_b|d) = p(d|M_a)\, p(M_a) / p(d|M_b)\, p(M_b).$$

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

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

Empirical scale for evaluating the strength of evidence from the Bayes factor B_{ij} between two models is shown in Tabl. 1 [45].

Tabl. 1. Strength of evidence for Bayes factor B_{ij} for two models

$\lvert\ln B_{ij}\rvert$	Odds	Strength of evidence
< 1.0	< 3 : 1	Inconclusive
1.0	~ 3 : 1	Weak evidence
2.5	~ 12 : 1	Moderate evidence
5.0	~ 150 : 1	Strong evidence

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* [34, 62]. Taking $M1$ to be the "null" hypothesis, we can ask whether the data supports the alternative hypothesis $M2$, i.e., whether we can reject the null hypothesis. Taking equal priors $p(M_1) = p(M_2)$, the odds ratio is

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

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 [34].

Comparing classical and Bayesian approaches [34], 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 [34]. More details on Bayesian style of hypothesis testing can be found in [34, 62, 64].

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 anything that is logically entailed by a hypothesis can be confirmed by a proper test for it. The relation between hypothesis and evidence is often *empirica l* 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 3.1) where a hypothesis' explanatory credentials are taken to indicate its truth [72].

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 [27] 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 $\boldsymbol{\theta}$ whose values are to be estimated from data. We describe this process according to [34]. For a particular model M and prior information I we get:

$$p(M, \boldsymbol{\theta}|d, I) = p(d|M, \boldsymbol{\theta}, I)\, p(M, \boldsymbol{\theta}|I) / p(d|I)$$

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

In the Bayesian formalism, $p(M, \boldsymbol{\theta}|d, I)$ corresponds to the state of our *knowledge* (i.e., belief) about a model and its parameters, given data d. To simplify the notation, $M(\boldsymbol{\theta})$ will be substituted by M whenever the absence of explicit dependence on $\boldsymbol{\theta}$ 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 $p(\boldsymbol{\theta}|M,I)$, 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. Also 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 [10] 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 $\boldsymbol{y} = f(\boldsymbol{x})$ that give a relation specifying a vector of dependent variables $\boldsymbol{y}$ in terms of a vector of independent variables $\boldsymbol{x}$. 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, multiple regression are 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 $\boldsymbol{x}$ to predict the responses $\boldsymbol{y}$. What is observed is a set of $\boldsymbol{x}$'s that go in and a subsequent set of $\boldsymbol{y}$'s that come out. Predictive accuracy and properties of the algorithms (such as, e.g., 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 [5, 26] 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 inscrutable. Nature forms the outputs $\boldsymbol{y}$ from the inputs $\boldsymbol{x}$ 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 seems only slightly less inscrutable than nature's [10]. 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 [10] 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 we should pay attention to the question how far can we 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 [5, 10, 34].

Besides machine learning, an interesting example of algorithmic generation of hypotheses can be found in the IBM Watson project [18] where the symbiosis of the general-purpose reusable natural language processing

[1] In [10] instead of "formulaic modeling" the term "data modeling" is used that looks misleading in the computer science context.

(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 [18] – 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 allows to make the process applicable to the large collections of data. More details on that can be found in [17, 18] 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 (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 *absolute goodness* of the model. Bayesian model selection says nothing about the *overall quality* of the set of models (hypotheses) as a whole —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 by [67]. 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 [44, 45]. 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, *w*CDM; theories of modified gravity; void models or the back reaction [45]. 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 [67], 'doubt' may be defined as the posterior probability of the unknown model:

$$D \equiv p(X|d) = p(d|X)\, p(X) / p(d).$$

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:

$$\langle B_{i\Lambda} \rangle \equiv 1/\mathrm{N} \sum_{i=1}^{N} B_{i\Lambda} \ .$$

The ratio R between the posterior doubt and prior doubt, which is called the relative change in doubt, is:

$$R \equiv D/p(X).$$

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

$$\langle B_{i\Lambda} \rangle / B_{X\Lambda} << 1.$$

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

$$R_\Lambda \equiv p(\Lambda|d) / p(\Lambda).$$

For ΛCDM to be doubted, the following two conditions must be fulfilled:

$$R > 1, \ R_\Lambda < 1.$$

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 [67] 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 ΛCDM model. The upper bound of the Bayesian evidence for a presently unknown dark energy model against ΛCDM gives only weak evidence in favor of the unknown model. Since this is an absolute upper bound, it was concluded that ΛCDM remains a sufficient phenomenological description of currently available observations.

4 Facilities for the scientific hypothesis-driven experiment support

4.1 Conceptualization of scientific experiments

DIS increasingly becomes dependent on computational resources to aid complex researches. It becomes paramount to offer scientists mechanisms to manage the variety of knowledge produced during such investigations. Specific conceptual modeling facilities [54] are investigated to allow scientists to represent scientific hypotheses, models and associated computational or simulation interpretations which can be compared against phenomena observations (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 [23] 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 equation is considered as not enough to identify the hypothesis, first because it must be physically interpreted, 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 formulation or even to assert a semantic mapping to another incarnation of the hypothesis in case someone else reformulates it.

In [54] 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 logic-based language is proposed). Specific attention in this work 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. 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 [54], 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 [58] 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 is obtained and rules 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 decided based on the quantitative analysis between the simulation results and the observations [54]. 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, uses an ontology of hypotheses to represent hypotheses in machine understandable form as relations between objects (agents) and processes [65].

As an upgrade of HyBrow, the HyQue [12] 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 lends 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 materials used in experiments and the data generated [20]. 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 (MIGen) 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 [39].

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 SemanticWeb Application in Neuromedicine [20]. 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 pre-publication 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 is 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 on Fig. 6 the diversity of the components of a scientific hypothesis model has been borrowed from the applications in Neuroscience [54, 55] and in a human cardiovascular system in Computational Hemodynamics [23, 56]. 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 [3] 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; 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 [69] 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 non-monotonic reasoning. The main features of the extended system provide: (1) seamless integration of hypothesis formation with knowledge representation and reasoning; (2) use of various resources of biological data as well as human expertise to intelligently generate hypotheses; (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.

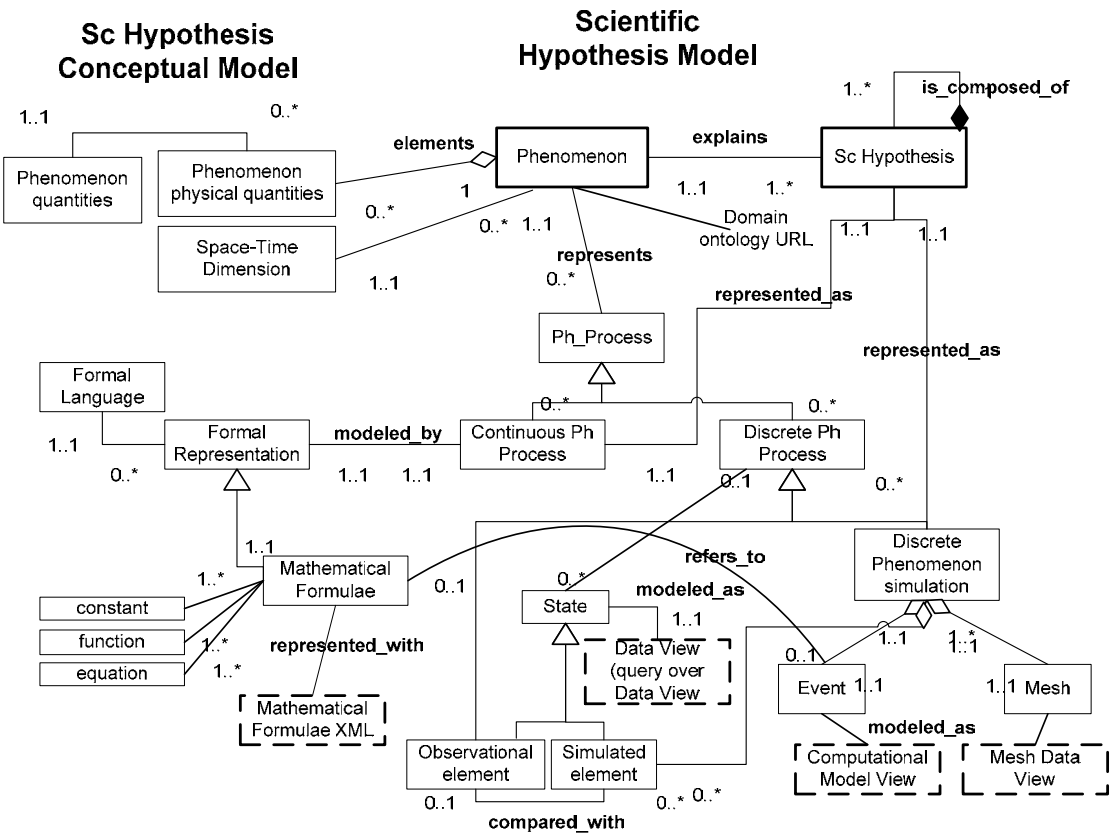

Fig. 6. Elements of the scientific hypothesis model

4.4 Hypothesis-driven robots

The Robot Scientist [66] 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 types of logical reasoning used in scientific discovery (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. 6).

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.

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 [39].

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 the E.C. number. An example of hypothesis at this level is: *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: *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.

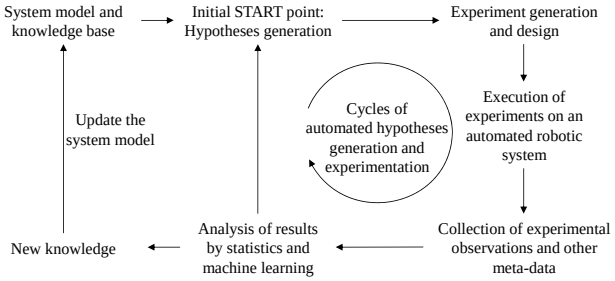

Fig. 6. Hypothesis driven closed-loop learning

Adam follows a hypothetico-deductive methodology (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 a NP complete problem and Adam uses heuristics to find a solution [65].

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 [65].

The details describing the software and informatics decisions in the Robot Scientist project can be found in [65, 66] 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 [13, 39]. 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 [13].

4.5 Hypotheses as data in probabilistic databases

Another view of hypotheses encoding and management is presented in [51]. Authors use probabilistic database techniques for hypotheses systematic construction and management. MayBMS [30], a probabilistic database management system, is used as a core for hypothesis management. This methodology (called γ-DB) enables researchers to maintain several hypotheses explaining some phenomena and provides evaluation mechanism based on Bayesian approach to rank them.

The construction of γ-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 (FD) 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 meta data 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 (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 pseudo-transitive closure and reasoning; 4) finally, the probabilistic γ-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 3.2.1). The γ-DB approach provides a promising way to analyse hypotheses in large scale DIR as uncertain predictive database in face of empirical data.

5 Examples of hypothesis-driven scientific research

5.1 Besançon Galaxy model

Various models in astronomy heavily rely on hypotheses. One of the most impressive is the Besançon galaxy model (BGM) [16, 60, 61] 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. 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.

We will further focus on the renewed BGM [16], 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 χ^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 $(B-V)_T$ colour 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 colour distribution.

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, 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. Based on this experience, an investigation of the thick disc is underway using SDSS and 2MASS surveys.

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") [7]. 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 macro scale description [15] 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 [21, 22]. Such research is motivated by the 1000 Functional Connectomes Project (FCP) launched in 2010 [7].The 1000 FCP [74] 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 [46]. HCP informatics platform includes data management system ConnectomeDB that is based on the XNAT imaging informatics platform [47], a widely used open source system for managing and sharing imaging and related data.

Visualization, processing and analysis of high-dimensional data such as images often requires 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 [11], embedded in the original space. This topic of research is called dimensionality reduction, non-linear 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 non-linear 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 [21] 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 non-trivial. 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 [21] 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 is 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 [21] for the 1000 FCP database comparing of 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 also was rejected with high probability.

On such examples it was shown [21] 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 [43]) 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 the [70] paper discussion 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 multi-site, multi-datatype 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.

5.3 Climate in Australia

Another view on hypothesis representation and evaluation is presented in [41]. Authors argue, that as long as in DIS data relevant to some hypotheses gets continuously aggregated as time passes, hypotheses should be represented as programs that are executed repeatedly, as new relevant amounts of data gets 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 [42], stated that the temperature series is not stationary and is integrated of order 1 (I(1)). Non-stationarity 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: 1) The National Oceanographic and Atmospheric Administration marine and weather information, 2) 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 [8] 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 and UK, 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, UK 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 [8] 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, 2008 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 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.

6 Conclusion

The objective of this study is to analyze, collect and systematize information on the role of hypotheses in the data intensive research 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. Facilties 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 (including biology, biomedical investigations, neuromedicine, astronomy), automatic organization of hypothesis-driven experiments. Examples of scientific researches 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 EMN (Efficient market hypothesis), as well as algorithmic generation of hypotheses in the IBM Watson project 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 DIS.

References

[1] Agresti, A., Finlay, B. Statistical Methods for the Social Sciences (4th Edition), 2008. – P. 624.

[2] Alferes, J. J., Pereira, L. M., Swift, T. Abduction in well-founded semantics and generalized stable models via tabled dual programs. In: TPLP, 2004. – Vol. 4, No. 4. – P. 383–428.

[3] Asgharbeygi, N., Langley, P., Bay, S., Arrigo, K. Inductive revision of quantitative process models. In: Ecological modelling – Vol. 194, No. 1. – P. 70–79.

[4] Bacon, F. The new organon. In: R. M. Hutchins, (ed.), Great books of the western world. The works of Francis Bacon. Chicago, Encyclopedia Britannica, Inc., 1952 – Vol. 30. – P. 107–195.

[5] Barber, D. Bayesian Reasoning and Machine Learning. Cambridge University Press, 2010. – P. 720.

[6] Bartha, P. Analogy and Analogical Reasoning. In: The Stanford Encyclopedia of Philosophy, 2013. – http://plato.stanford.edu/archives/ fall2013/entries/reasoning-analogy/

[7] Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Windischberger, C. Toward discovery science of human brain function. In: Proceedings of the

National Academy of Sciences, 2010. – V. 107, No. 10. – P. 4734–4739.

[8] Bollen, J., Mao, H., Zeng, X. Twitter mood predicts the stock market. In: Journal of Computational Science, 2011. – V. 2, No. 1. – P. 1–8.

[9] Borges, M. R. Efficient market hypothesis in European stock markets. In: The European Journal of Finance, 2010. – V. 16, No. 7. – P. 711–726.

[10] Breiman, L. Statistical Modeling: The Two Cultures. In: Statistical Science, 2001. – V. 16, No. 3. – P. 199–231.

[11] Brun, A. Manifold learning and representations for image analysis and visualization. Department of Biomedical Engineering, Linköpings universitet, 2006.

[12] Callahan, A., Duumontier, M., Shah, N. HyQue: Evaluating hypotheses using Semantic Web technologies. In: J. Biomedical Semantics, 2011. – V. 2, No. S-2. – P. S3.

[13] Castrillo, J.I. , S.G. Oliver (eds.). Yeast Systems Biology: Methods and Protocols. In: Methods in Molecular Biology, Springer, 2011. – V. 759. – P. 535.

[14] Citrigno, S., Eiter, T., Faber, W., Gottlob, G., Koch, C., Leone, N., Scarcello, F. The dlv system: Model generator and application frontends. In: Proceedings of the 12th Workshop on Logic Programming, 1997. – P. 128–137.

[15] Craddock, R.C., Jbabdi, S., Yan, C.G., Vogelstein, J.T., Castellanos, F.X., Di Martino, A., Milham, M.P. Imaging human connectomes at the macroscale. In: Nature methods, 2013. – V. 10, No. 6. – P. 524–539.

[16] Czekaj, M.A., Robin, A.C., Figueras, F., Luri, X., Haywood, M. The Besançon Galaxy model renewed I. Constraints on the local star formation history from Tycho data. In: arXiv preprint arXiv:1402.3257, 2014.

[17] Dredze, M, Crammer, K., Pereira, F. Confidence-Weighted Linear Classification. In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 2008. – P. 264-271.

[18] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Welty, C. Building Watson: An overview of the DeepQA project. In: AI magazine, 2010. –V. 31, No. 3. – P. 59–79.

[19] Field, A. Discovering statistics using IBM SPSS statistics. In: Sage, 2013. – P. 915.

[20] Gao, Y., Kinoshita, J., Wu, E., Miller, E., Lee, R., Seaborne, A., Clark, T. SWAN: A distributed knowledge infrastructure for Alzheimer disease research. In: Web Semantics: Science, Services and Agents on the World Wide Web, 2006. - V. 4, No. 3. – P. 222–228.

[21] Ginestet, C.E., Balanchandran, P., Rosenberg, S., Kolaczyk, E.D. Hypothesis Testing For Network Data in Functional Neuroimaging. In: arXiv preprint arXiv:1407.5525, 2014.

[22] Ginestet, C. E., Fournel, A. P., Simmons, A. Statistical network analysis for functional MRI: summary networks and group comparisons. In: Frontiers in computational neuroscience, 2014. – Vol. 8.

[23] Gonçalves, B., Porto, F. A Lattice-Theoretic Approach for Representing and Managing Hypothesis-driven Research. In: AMW, 2013.

[24] Gonçalves, B., Porto, F., Moura, A. M. C. On the semantic engineering of scientific hypotheses as linked data. In: Proceedings of the 2nd International Workshop on Linked Science, 2012.

[25] Haber, J. Research Questions, Hypotheses, and Clinical Questions. In: Evolve Resources for Nursing Research, 2010. – P. 27–55.

[26] Hastie, T., Tibshirani, R., Friedman, J., Franklin, J. The elements of statistical learning: data mining, inference and prediction. In: The Mathematical Intelligencer, 2005. – Vol. 27, No. 2. – P. 83–85.

[27] Hawthorne, J. Inductive Logic. In: The Stanford Encyclopedia of Philosophy, 2014 – http://plato.stanford.edu/archives/sum2014/entries/logic-inductive/

[28] Hempel, C. G. Fundamentals of concept formation in empirical science. In: Int. Encyclopedia Unified Science, 1952. – V. 2, No. 7.

[29] Hey, T., Tansley, S., Tolle, K. (eds.). The fourth paradigm: Data-intensive scientific discovery. Redmond, Microsoft Research, 2009. – P. 252.

[30] Huang, J., Antova, L., Koch, C., Olteanu, D. MayBMS: a probabilistic database management system. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, 2009. – P. 1071–1074.

[31] IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp. IBM SPSS Statistics base. IBM Corp., 2013.

[32] Ihaka, R., Gentleman, R. R: a language for data analysis and graphics. In: Journal of computational and graphical statistics, 1996. – Vol. 5, No. 3. – P. 299–314.

[33] Inoue K., Sato T., Ishihata M., Kameya Y., Nabeshima H. Evaluating abductive hypotheses using and EM algorithm on BDDs. In: Proceedings of IJCAI-09, 2009. – P. 810–815.

[34] Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., Gray, A. Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. Princeton University Press, 2014. – P. 552.

[35] Kakas, A.C., Michael, A., Mourlas, C. ACLP: Abductive constraint logic programming. In: The Journal of Logic Programming, 2000. – Vol. 44, No. 1. – P. 129–177.

[36] Kakas, A.C., Kowalski, R.A., Toni, F. Abductive Logic Programming. In: Journal of Logic and Computation, 1993. – Vol. 2, No. 6. – P. 719–770.

[37] Kerlinger, F.N., Lee, H.B. Foundations of behavioral research: Educational and psychological inquiry. New York: Holt, Rinehart and Winston, 1964. – P. 739.

[38] King, R.D., Liakata, M., Lu, C., Oliver, S.G., Soldatova, L.N. On the formalization and reuse of scientific research. In: Journal of The Royal Society Interface, 2011. – Vol. 8, No. 63. – P. 1440–1448.

[39] King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G., Bryant, C.H., Muggleton, S.H., Oliver, S.G. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature, 2004. – Vol. 427, No. 6971. – P. 247–252.

[40] Lakshmana Rao, J R. Scientific 'Laws', 'Hypotheses' and 'Theories'. In: Meanings and Distinctions. Resonance, 1998. – Vol. 3. – P. 69–74.

[41] Lappalainen, J., Sicilia, M.Á., Hernández, B. Automatic Hypothesis Checking Using eScience Research Infrastructures, Ontologies, and Linked Data: A Case Study in Climate Change Research. In: Procedia Computer Science, 2013. – Vol. 18. – P. 1172–1178.

[42] Lenten, L.J., Moosa, I.A. An empirical investigation into long-term climate change in Australia. In: Environmental Modelling & Software, 2003. – Vol. 18, No. 1. – P. 59–70.

[43] Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., Brovelli, A. Multivoxel Pattern Analysis for fMRI Data: A Review. In: Computational and mathematical methods in medicine, 2012.

[44] March, M.C. Advanced Statistical Methods for Astrophysical Probes of Cosmology. In: Springer Theses, 2013. – Vol. 20. – P. 177.

[45] March, M.C., Starkman, G.D., Trotta, R., Vaudrevange, P. M. Should we doubt the cosmological constant?. In: Monthly Notices of the Royal Astronomical Society, 2011. – Vol. 410, No. 4. – P. 2488–2496.

[46] Marcus, D.S., Harwell, J., Olsen, T., Hodge, M., Glasser, M.F., Prior, F., Van Essen, D.C. Informatics and data mining tools and strategies for the human connectome project. In: Frontiers in neuroinformatics, 2011. – Vol. 5.

[47] Marcus, D.S., Olsen, T.R., Ramaratnam, M., Buckner, R.L. The extensible neuroimaging archive toolkit. In: Neuroinformatics, 2007. – Vol. 5, No. 1. – P. 11–33.

[48] McComas, W.F. The principal elements of the nature of science: dispelling the myths. In: The Nature of Science in Science Education, 1998. – P. 53–70.

[49] Menzies, T. Applications of Abduction: Knowledge-Level Modeling, In: International Journal of Human-Computer Studies, 1996. – V. 45, No. 3. – P. 305–335.

[50] Nickles, T. (ed.). Scientific discovery: Case studies. Taylor & Francis, 1980. – Vol. 2. – P. 501.

[51] Plotkin, G.D. A note on inductive generalization. In: Machine Intelligence. Edinburgh University Press, 1970. – Vol. 5. – P. 153–163.

[52] Poincaré, Henri. The Foundations of Science: Science and Hypothesis, The Value of Science, Science and Method. The Project Gutenberg EBook, 2012. – Vol. 39713. – P 554.

[53] Popper, K.. The Logic of Scientific Discovery (Taylor & Francis e-Library ed.). London and New York: Routledge / Taylor & Francis e-Library, 2005.

[54] Porto, F. Big Data in Astronomy. The LIneA-DEXL case. Presentation at the EMC Summer School on BIG DATA – NCE/UFRJ, 2013.

[55] Porto, F. Moura, A. M. C., Gonçalves, B., Costa, R., Spaccapietra, S. A Scientific Hypothesis Conceptual Model. In: ER Workshops, 2012. – Vol. 7518. – P. 101–110.

[56] Porto, F., Moura, A. M. C. Scientific Hypothesis Database. Report, 2011.

[57] Porto, F., Spaccapietra, S. Data model for scientific models and hypotheses. In: The evolution of conceptual modeling, 2011. – Vol. 6520. – P. 285–305.

[58] Racunas, S.A., Shah, N.H., Albert, I., Fedoroff, N.V. Hybrow: a prototype system for computer-aided hypothesis evaluation. In: Bioinformatics, 2004. – Vol. 20, No. 1. – P. 257–264.

[59] Ray, O., Kakas, A. ProLogICA: a practical system for Abductive Logic Programming. In: Proceedings of the 11th International Workshop on Non-monotonic Reasoning, 2006. – P. 304–312.

[60] Robin, A.C., Reylé, C., Derrière, S., Picaud, S. A synthetic view on structure and evolution of the Milky Way. arXiv preprint astro-ph/0401052, 2004.

[61] Robin, A., Crézé, M. Stellar populations in the Milky Way-A synthetic model. In: Astronomy and Astrophysics, 1986. – Vol. 157. – P. 71–90.

[62] Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. In: Psychonomic bulletin & review, 2009. – Vol. 16, No. 2. – P. 225–237.

[63] Schickore, J. Scientific Discovery. The Stanford Encyclopedia of Philosophy, 2014 – http://plato.stanford.edu/archives/spr2014/entries/scientific-discovery/

[64] Sivia, D.S., Skilling, J. Data Analysis. A Bayesian Tutorial. Oxford University Press Inc., New York, 2006. – P. 264.

[65] Soldatova, L.N., Rzhetsky, A., King, R. D. Representation of research hypotheses. In: J. Biomedical Semantics, 2011. – Vol. 2, No. S-2. – P. S9.

[66] Sparkes, A., Aubrey, W., Byrne, E., Clare, A., Khan, M. N., Liakata, M., King, R. D. Towards Robot Scientists for autonomous scientific discovery. In: Autom Exp, 2010. – Vol. 2, No 1.

[67] Starkman, G.D., Trotta, R., Vaudrevange, P.M. Introducing doubt in Bayesian model comparison. arXiv preprint arXiv:0811.2415, 2008.

[68] Tamaddoni-Nezhad, A., Chaleil, R., Kakas, A., Muggleton, S.H. Application of abductive ILP to learning metabolic network inhibition from temporal data. In: Machine Learning, 2006. – Vol. 64. – P. 209–230.

[69] Tran, N., Baral, C., Nagaraj, V.J., Joshi, L. Knowledge-based integrative framework for hypothesis formation in biochemical networks. In: Data Integration in the Life Sciences, 2005. – P. 121–136.

[70] Van Horn, J.D., Toga, A.W. Human neuroimaging as a “Big Data” science. In: Brain imaging and behavior, 2014. – Vol. 8, No. 2. – P. 323–331.

[71] Van Nuffelen, B., Kakas, A. A-system: Declarative programming with abduction. In: Logic Programming and Nonmotonic Reasoning, 2001. – P. 393–397.

[72] Weber, M. Experiment in Biology. The Stanford Encyclopedia of Philosophy, 2014. – http://plato.stanford.edu/archives/fall2014/entries/biology-experiment/

[73] Woodward, J. Scientific Explanation. The Stanford Encyclopedia of Philosophy, 2011. – http://plato.stanford.edu/archives/win2011/entries/scientific-explanation/

[74] Yan, C.G., Craddock, R.C., Zuo, X.N., Zang, Y.F., Milham, M.P. Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. In: Neuroimage, 2013. – Vol. 80. – P. 246–262.