Chapter 8
In Silico Hypothesis Discovery

Philip R.O. Payne

By the End of This Chapter, Readers Should Be Able to:

Understand the role of conceptual knowledge collections in terms of informing the design and use of reasoning systems for the purpose of in silico hypothesis discovery

- Select appropriate evaluation methodologies that can be used the assess the performance of in silico hypothesis discovery tools and platforms
- Identify open research questions related to the future of high-throughput hypothesis generation and the impact of such innovations on current and future scientific and healthcare delivery paradigms.

8.1 Introduction

As noted in the preceding chapters, the fundamental methods needed to conduct basic science, and clinical and translational research are very complex, involving a multitude of actors, workflows and data types. For example, the translational research paradigm focuses on cyclical flow of data, information and knowledge between laboratory researchers, clinical investigators and clinical or public health practitioners, and is predicated on systems-level approaches that involve diverse information needs, sources and management requirements [1]. A variety of reports and scholarly works have enumerated challenges that may prevent the effective conduct of translational research. As introduced in Chap. 1, one such challenge is commonly known as the "T1 block" and is concerned with issues that impact the ability to move data, information and knowledge between basic science and clinical research settings. Similarly, a second challenge, often known as the "T2 block", focuses upon impediments affecting the movement of data, information and knowledge between

P.R.O. Payne, PhD, FACMI
Department of Biomedical Informatics,
The Ohio State University Wexner Medical Center, Columbus, OH, USA
e-mail: philip.payne@osumc.edu

P.R.O. Payne, P.J. Embi (eds.), *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*, Health Informatics,
DOI 10.1007/978-1-4471-4646-9_8, © Springer-Verlag London 2015

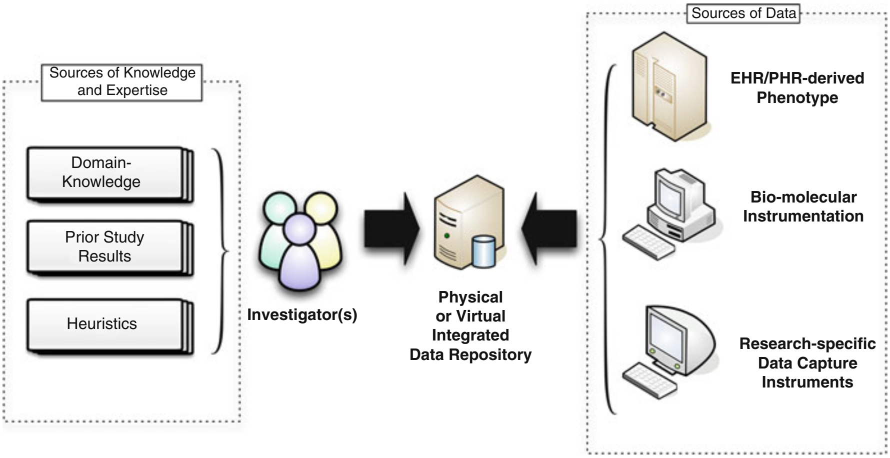

Sources of Knowledge and Expertise
Domain-Knowledge
Prior Study Results
Heuristics
Investigator(s)
Physical or Virtual Integrated Data Repository
Sources of Data
EHR/PHR-derived Phenotype
Bio-molecular Instrumentation
Research-specific Data Capture Instruments

Fig. 8.1 Overview of traditional, investigator-driven approach to asking and answering questions regarding complex and large scale data sets. In this model, the investigator (or research team) serves as the primary integration of various sources of knowledge and expertise, formulating and asking questions concerning available data using a combination of their domain knowledge, experiential knowledge from prior studies, and heuristics that they may have formulated relative to an application domain

the clinical research environment and clinical or public health practice [2]. For both of these categories of challenges, the methods required to address them are extremely reliant on the provision of tools and methods that can facilitate the collection, formalization, analysis and dissemination of large-scale and integrative data sets [3]. The potential impact of informatics-based approaches in terms of addressing such information needs has been well established; yet those same tools and methods remain largely under-utilized by the research and practice communities [4–12].

Within this broad context, one major area of concern is the way in which we formulate and test hypotheses relative to "big" biomedical data. This concern is amplified by the fact that the volume, velocity and variability of biomedical data continue to expand at a rapid rate. This growth is in large part a function of the proliferation of computerized sources of biomedical data, such as Electronic Health Records (EHRs), Personal Health Records (PHRs), Clinical Trial or Research Management Systems (CTMS/CRMS), high-throughput bio-molecular instrumentation, and ubiquitous sensor technologies. While computational methods continue to be devised and applied to support or enable the capture, storage and transaction of these data sets, there has not been a corresponding focus on improvements in the ways in which we ask and answer important questions utilizing this data. In fact, the traditional, reductionist approach to intuitive hypothesis generation based on the expertise or insights of an individual or small number of investigators remains the norm (Fig. 8.1). However, this approach is highly linear, and limited by the cognitive capacities of such investigators or teams, leading to an underutilization of available and costly to assemble data sets.

In effect, we continue to create and maintain bigger and more complex data sets at great expense, while we ask and answer small numbers of questions regarding the

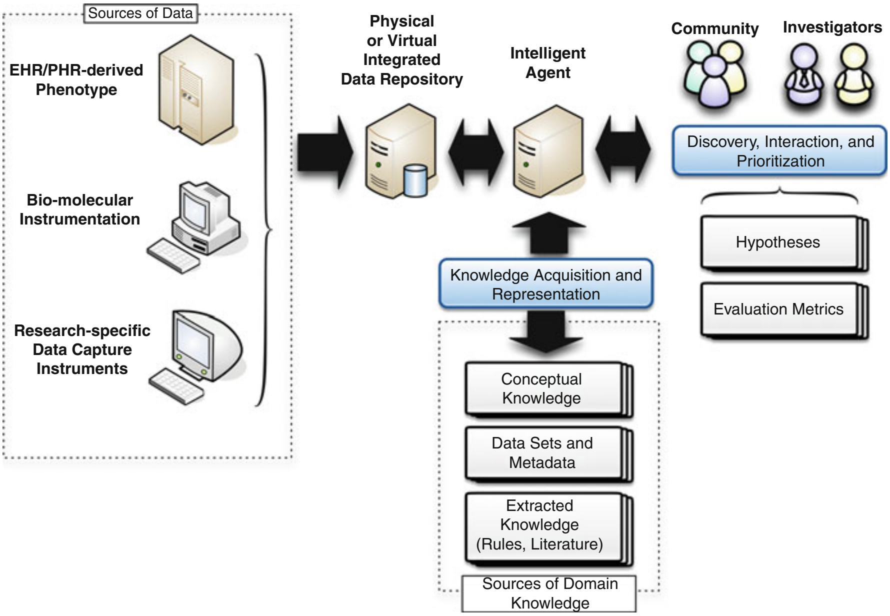

Sources of Data
EHR/PHR-derived Phenotype
Bio-molecular Instrumentation
Research-specific Data Capture Instruments
Physical or Virtual Integrated Data Repository
Intelligent Agent
Community
Investigators
Discovery, Interaction, and Prioritization
Hypotheses
Evaluation Metrics
Knowledge Acquisition and Representation
Conceptual Knowledge
Data Sets and Metadata
Extracted Knowledge (Rules, Literature)
Sources of Domain Knowledge

Fig. 8.2 Alternative, high-throughput approach to asking and answering questions regarding "big data" resources, using in silico hypothesis discovery methods. In this model, intelligent computational agents draw upon a variety of domain knowledge collections, using formally represented variants of those collections, in order to identify potential relationships of interest between elements or collections of elements in an integrated repository. These relationships are then presented, along with corresponding evaluation metrics that serve to characterize their potential accuracy and novelty, to both investigators and their teams as well as broader groups of interested community members, who can then discover, interact with, and prioritize such hypotheses concerning data-level interactions for subsequent investigation

contents of those data sets using methods that are not far removed from those used around the time of the dawn of modern science [13]. This concerning juxtaposition is the driver for an emerging body of research that seeks to couple high-throughput data generation with new and similarly high-throughput hypothesis generation techniques, which can at a high level be referred to as ***in silico hypothesis discovery methods*** (Fig. 8.2).

Such high-throughput approaches to asking and answering questions corresponding to "big data" resources are essential to the synthesis of novel biomedical knowledge, such as that required to support personalized medicine paradigms. Such precision approaches to wellness promotion and care delivery aim to improve quality, outcomes and cost of care [2, 3, 14–16]. Acting upon this vision of high-throughput in silico hypothesis discovery requires:

1. An understanding of the design and appropriate use of domain-specific conceptual knowledge collections;
2. The application of intelligent agents that are informed by such knowledge collections and based upon formal computational methods; and
3. The evaluation of ensuing hypothesis using appropriate metrics and measures.

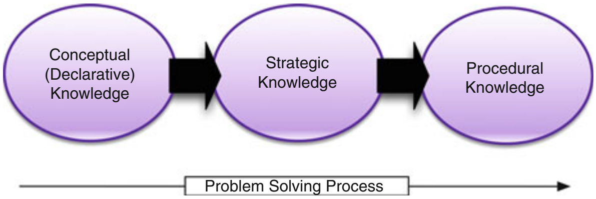


Fig. 8.3 Spectrum of knowledge types, spanning from conceptual to strategic to procedural knowledge, where conceptual knowledge is the most abstract form of understanding a domain, and procedural knowledge is the most application- or problem-oriented understanding of a given need or task

In the following sections, we will explore critical aspects of all three of the aforementioned foundational dimensions that underpin the design and use of in silico hypothesis discovery tools and platforms.

8.2 Conceptual Knowledge in Biomedicine

Conceptual knowledge has been defined in the computational, psychology, and education literature as being comprised of a combination of atomic units of information *and* the meaningful relationships between those units. The same literature goes on to define two additional types of complementary knowledge, known as procedural and strategic knowledge respectively. *Procedural knowledge* is a process-oriented understanding of a given problem domain [17–20], effectively concerned with the methods and approaches used to solve a given problem or address a task. *Strategic knowledge* is that which is used by individuals in order to translate conceptual knowledge into procedural knowledge [19] (Fig. 8.3).

Of note, these definitions are based upon a wide-ranging collection of empirical research on learning and problem-solving in complex scientific and quantitative domains such as mathematics and engineering [18, 20]. The cognitive science literature provides a very similar and confirmatory differentiation of knowledge types, making the distinction between procedural and declarative knowledge. Declarative knowledge in this context is synonymous with conceptual knowledge as defined previously [21].

Conceptual knowledge collections in the biomedical domain include a variety of constructs such as ontologies, controlled terminologies, semantic networks and database schemas. A common theme when considering the existing state-of-the-art relative to the design and use of conceptual knowledge collections in the biomedical domain is the need for systematic and rigorous processes for representing conceptual knowledge in a computable form. It is also important to note when considering the need for such knowledge representation best practices that conceptual knowledge collections rarely exist in isolation. Instead, they usually occur within structures that contain multiple types of knowledge. For example, a modern clinical decision support system (CDSS) might include: (1) a database of potential find-

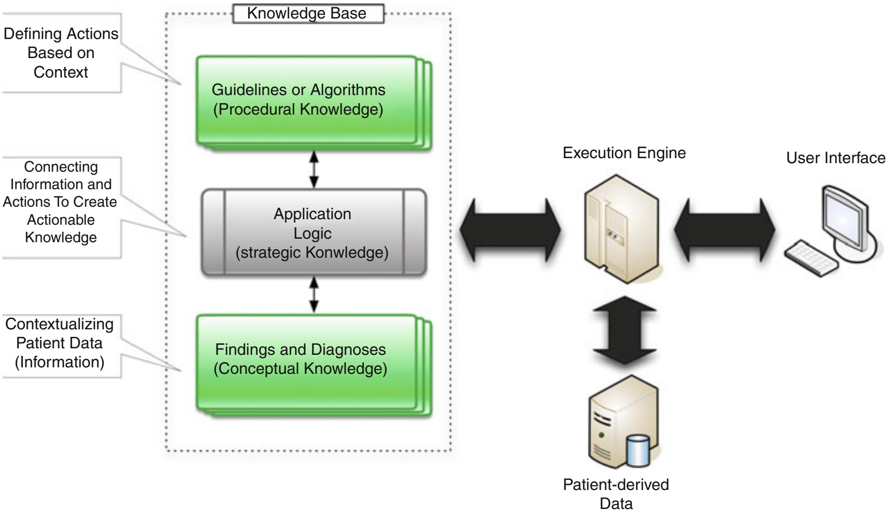

Knowledge Base
Defining Actions Based on Context
Guidelines or Algorithms (Procedural Knowledge)
Connecting Information and Actions To Create Actionable Knowledge
Application Logic (strategic Konwledge)
Contextualizing Patient Data (Information)
Findings and Diagnoses (Conceptual Knowledge)
Execution Engine
User Interface
Patient-derived Data

Fig. 8.4 Overview of a prototypical CDSS platform, incorporating conceptual, procedural, and strategic knowledge types in order to generate actionable knowledge from patient-derived data

ings, diagnoses and the relationships between them (*conceptual knowledge*); (2) a set of guidelines or algorithms used to reason upon the preceding database (*procedural knowledge*); and (3) a formal definition of the logic used to operationalize the preceding two knowledge collections (*strategic knowledge*) (Fig. 8.4).

It is only when these three types of knowledge are combined that it is possible to realize a functional decision support system [22]. Given the close similarities between such CDSS and the previously introduced framework for in silico hypothesis discovery methods or tools (as is illustrated in Fig. 8.2), this phenomenon is important to keep in mind for the remainder of this chapter.

8.2.1 Knowledge Engineering

The core theories and methods that underlie the ability to systematically and rigorously represent conceptual knowledge inform a set of application-level techniques known as knowledge engineering (KE). The KE process (Fig. 8.5) incorporates four major steps:

1. Acquisition of knowledge (KA)
2. Representation of that knowledge (KR) in a computable form
3. Implementation or refinement of intelligent agents (e.g., applications that use formally represented knowledge to reason upon data sets and generate results of interest to end-users) or applications
4. Verification and validation of the output of those knowledge-based agents or applications against one or more reference standards.

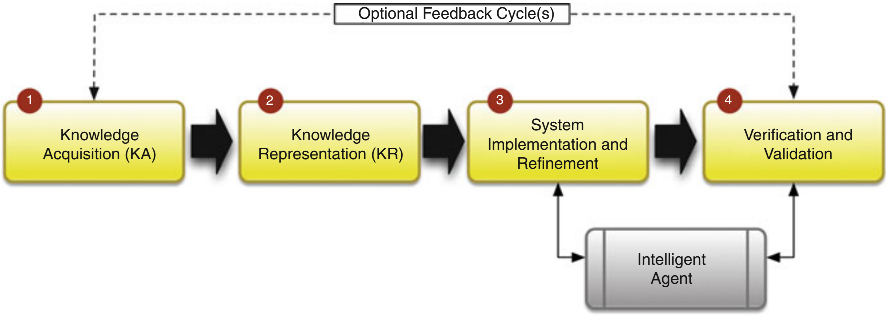

Optional Feedback Cycle(s)
1
Knowledge Acquisition (KA)
2
Knowledge Representation (KR)
3
System Implementation and Refinement
4
Verification and Validation
Intelligent Agent

Fig. 8.5 Overview of the Knowledge Engineering (*KE*) process, consisting of knowledge acquisition (*KA*), knowledge representation (*KR*), system implementation and refinement, and the verification and validation of those systems (numbered per the steps enumerated in Sect. 8.2.1). Of note, there is an optional feedback mechanism from the verification and validation results back to the initial KA component, which helps to inform subsequent KA activities and the refinement of existing knowledge bases

With regards to the final step of the KE process (verification and validation), the reference standards used to evaluate the performance of an intelligent agent can include expert performance measures, requirements acquired before designing the knowledge-based system and/or requirements that were realized upon implementation of the knowledge-based system. In this context, verification is the process of ensuring that the knowledge-based system meets the initial requirements of the potential end-user community. In comparison, validation is the process of ensuring that the knowledge-based system meets the realized requirements of the end-user community once a knowledge-based system has been implemented [23].

8.2.2 Theoretical Frameworks for KE

Underlying the KE process is a set of theories concerning the ability to acquire and represent knowledge in a computable format, which is known as the physical symbol hypothesis. First proposed by Newell and Simon [24], and expanded upon by Compton and Jansen [25], the physical symbol hypothesis argues that knowledge consists of both symbols of reality, and relationships between those symbols. This definition of knowledge thus allows for the creation of “physical symbol systems” (e.g., conceptual knowledge collections), which are defined as:

> …a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure). Thus, a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another). At any instant of time the system will contain a collection of these symbol structures. [26]

In a similar manner, it has been argued within the KE literature that the psychological constructs used by experts can be used as the basis for informing the design and composition of conceptual knowledge collections [27]. This argument is based on a framework for expertise transfer known as Kelly's Personal Construct Theory (PCT). PCT defines humans as "anticipatory systems", where individuals create templates, or constructs that allow them to recognize situations or patterns in the "information world" surrounding them. These templates are then used to anticipate the outcome of a potential action given knowledge of similar previous experiences [28]. Kelly views all people as "personal scientists" who make sense of the world around them through the use of a hypothetico-deductive reasoning system. The details of PCT help to explain how experts create and use such constructs. Specifically, Kelly's fundamental postulate is that "*a person's processes are psychologically channelized by the way in which he anticipated events*" [28]. This is complemented by the theory's first corollary, which is summarized by his statement that [28]:

> Man looks at his world through transparent templates which he creates and then attempts to fit over the realities of which the world is composed… Constructs are used for predictions of things to come… The construct is a basis for making a distinction… not a class of objects, or an abstraction of a class, but a dichotomous reference axis.

Building upon these basic concepts, Kelly goes on to state in his Dichotomy Corollary that "*a person's construction system is composed of a finite number of dichotomous constructs*" [28]. Finally, the parallel nature of personal constructs and conceptual knowledge is illustrated in Kelly's Organization Corollary, which states, "*each person characteristically evolves, for his convenience of anticipating events, a construction system embracing ordinal relationships between constructs*" [27, 28].

When taken as a whole, the two preceding theoretical frameworks provide the basic premises for arguing that:

1. Domain experts (e.g., humans) use personal constructs that roughly approximate those constructs that define formal knowledge (e.g., conceptual, strategic, and procedural knowledge), so as to make sense of the "information world" surrounding them;
2. Formal knowledge can be represented in a computationally tractable format, based upon the physical symbol hypothesis, and again, such symbolic systems closely approximate the definitions of conceptual knowledge; and
3. Knowledge engineering methods, and in particular, knowledge acquisition techniques, provide a set of tools for the elicitation and representation (in computable formats) of domain expert knowledge, helping to bridge the two preceding and complementary postulates.

Thus, it is possible to systematically and rigorously collect, formalize, and represent domain knowledge in a manner such that computers can reason upon those knowledge collections in a high throughput manner, thus replicating expert hypothesis generation processes in a way that is not constrained by innate human cognitive limitations and/or potential biases. Such a conclusion "opens the door" for an exploration of ensuing in silico hypothesis discovery methods, as will be introduced in Sect. 8.3. Additionally, Payne et al. [29] provide a more comprehensive review of the theories, frameworks, and methods that make up the biomedical KE domain.

8.3 Design and Use of Intelligent Agents for In Silico Hypothesis Generation

While there exist a broad variety of methods that can be used for the purposes of in silico hypothesis discovery, spanning a spectrum from machine learning and data mining to iterative human-computer interaction in order to discovery high level patterns within complex data sets, for the purposes of this chapter, we will focus on a specific and exemplar type of methodology known as *knowledge discovery in databases* (KDD). This specific method has been selected in order to highlight the generalizable features of a much broad class of knowledge-based software and intelligent agents that can be used for in silico hypothesis generation. At a high level KDD is concerned with the utilization of intelligent agents, which are software applications that are designed to replicate human problem solving through the leverage of conceptual knowledge collections as an integral part of their architecture and function. In KDD, intelligent agents are used specifically to derive knowledge from the contents of databases, including database metadata. The use of domain-specific conceptual knowledge collections, such as ontologies, is central to the KDD induction process since commonly used database modeling approaches do not incorporate semantic knowledge corresponding to the database contents. This overall approach is the basis for a specific KDD methodology known as *constructive induction* (CI). In CI, data elements defined by a database schema are mapped to concepts defined by one or more ontologies or equivalent conceptual knowledge collections. Subsequently, the relationships included in the mapped ontologies are used to induce semantically meaningful relationships between the mapped data elements. The induction process generates what are known as “facts” concerning the contents of the database, which are defined in terms of data elements and semantic relationships that significantly link those elements together (Fig. 8.6).

These “facts” (which are a type of conceptual knowledge) can then be used to support higher level reasoning about the data defined by the targeted database schema. It is important to note that such “facts” can exploit the transitive closure principles associated with the graph-like representation of most ontologies, and therefore may include intermediate concepts that do not map to a database element but serve to create a semantically related concept triplet or high-order relationship that begins and terminates with concepts that do map to database elements.

The implementation of an intelligent agent that utilizes the preceding CI methodology often follows the multi-step process illustrated in Fig. 8.7 (which each phase numbered to reflect the following description) and outlined below:

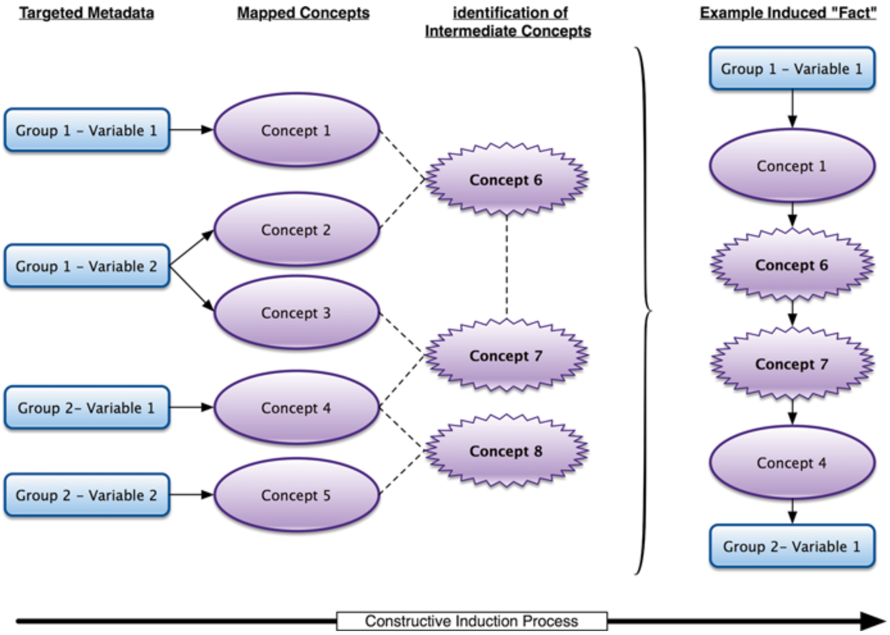


Fig. 8.6 Overview of constructive induction process whereby mapping between database elements as described via their metadata and corresponding ontology concepts are used to induce new "facts" concerning the contents of the database. In this general case, concepts 6–8, which is included in the ontology but does not map to the database construct, is used as an intermediate concept to define a concept triplet or higher order construct involving multiple intermediate entities that begins and terminates with data elements that map to concepts in the ontology construct

- **Phase 1 – Metadata to Conceptual Knowledge Entity Mapping**: In the first phase of implementing a CI-based agent, the metadata that serves to define a knowledge source of interest (e.g., a data dictionary or equivalent description of the contents of a data set or sets) must be mapped using either manual or automated processes to the entities that comprise one or more conceptual knowledge collections (e.g., syntactic or semantic matching of metadata definitions to entities in a terminology, ontology, or equivalent construct). This process usually results in one-to-many mappings, in which each metadata items corresponds to more than one conceptual knowledge entity. For example, if mapping a clinical data set with the specific variable corresponding to a "White Blood Cell Count", depending on the mapping approach being used and the intent of the KE initiative, that variable could be linked to multiple ontology-anchored concepts, such as the molecular entity "White Blood Cell", the laboratory procedure "White Blood Cell Count", as well as the clinical findings of "White Blood Cell Count Normal", "White Blood Cell Count High", and "White Blood Cell Low." This process generates a "knowledge map" that resolves individual variables of interest in the metadata being utilized to a corresponding set of atomic conceptual knowledge entities.

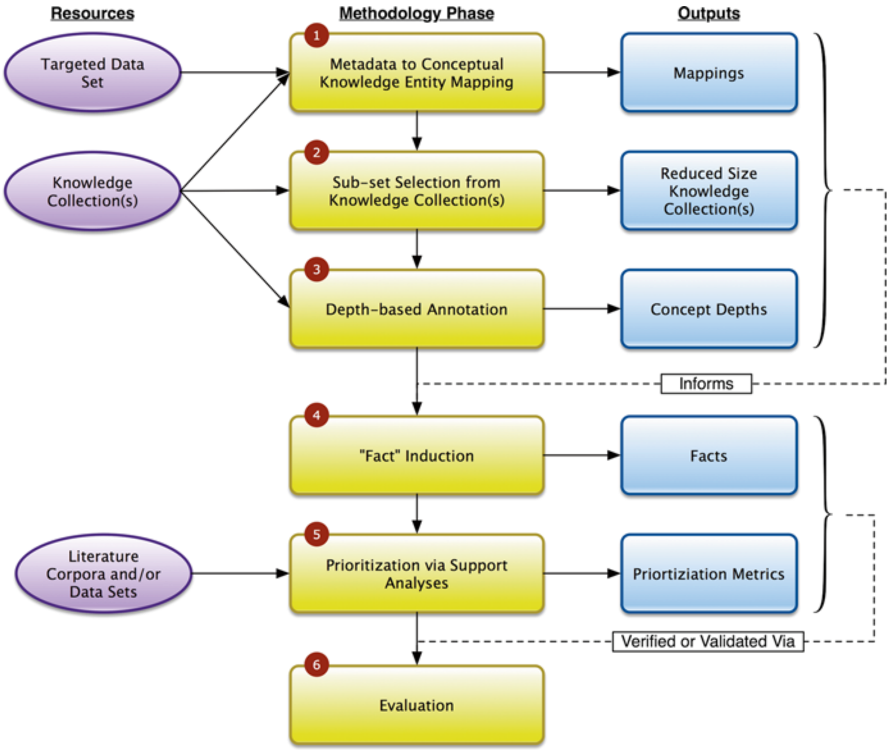

Resources
Methodology Phase
Outputs
Targeted Data Set
Knowledge Collection(s)
Literature Corpora and/or Data Sets
1
Metadata to Conceptual Knowledge Entity Mapping
2
Sub-set Selection from Knowledge Collection(s)
3
Depth-based Annotation
4
"Fact" Induction
5
Prioritization via Support Analyses
6
Evaluation
Mappings
Reduced Size Knowledge Collection(s)
Concept Depths
Facts
Priortiziation Metrics
Informs
Verified or Validated Via

Fig. 8.7 Overview of major steps, resources, and outputs associated with the design and use of a CI-based agent

- **Phase 2 – Subset Selection from Knowledge Collection(s)**: Given that many conceptual knowledge collections contain thousands, if not hundreds of thousands, of distinct atomic entities and corresponding hierarchical or semantic relationships, such constructs can present computational challenges, such as the tractability, computational cost, or timeliness of computational tasks applied to such knowledge collections, which can be addressed through a process known as *search space reduction*. Effectively, once Phase 1 (Metadata to Conceptual Knowledge Entity Mapping) is complete, we can select a subset of those conceptual knowledge collections that directly correspond to: (1) the atomic elements mapped to the targeted metadata; (2) the hierarchical and/or semantic relationships that serve to link those atoms together; and (3) any additional atoms necessary to complete the linking paths identified via [2]. This allows refinement of the initial knowledge collections to one that is constrained to the problem-solving task at hand.
- **Phase 3 – Depth-based Annotation**: Once we have reduced the overall search space (Phase 2), an additional computational challenge must be addressed, concerned with the granularity of concepts being used for reasoning purposes. If we extend our prior example of “White Blood Cell Count” and its mapping

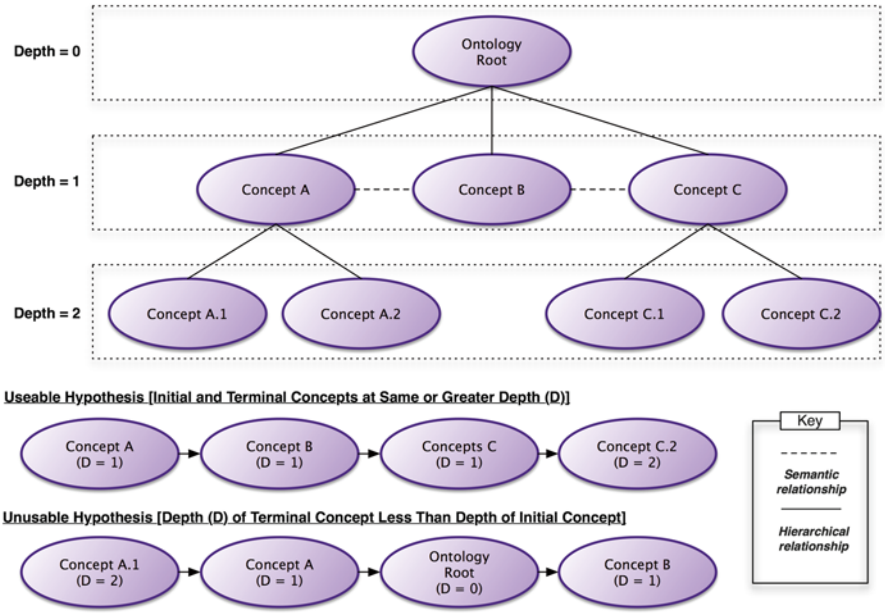


Fig. 8.8 Illustration of depth-based annotation and its implications for the induction of useable vs. unusable (e.g., overly general) “facts” or hypotheses

to an ontology-anchored concept of the laboratory procedure that has that same name, such a mapping could be used, when traversing the atomic units of information and relationships that comprise an ontology, to assert a relationship between “White Blood Cell Count” and the broad category of “Laboratory Procedures”, which then in turn allows for the resolution of relationships with every other known laboratory procedures subsumed by that concept. This would be a factually accurate relationship to assert, but one that is functionally useless for hypothesis discovery, as it is overly broad and general. Why is this the case? Simply put, the concepts of “White Blood Cell Count” and “Laboratory Procedure” are not of an equivalent level of granularity (e.g., the former is much more specific than the latter). One approach that can serve as a surrogate for concept granularity in the source ontologies employed by a CI-based agent is the relative depth from the ontology root of those concepts (Fig. 8.8). Using such measurements, we can then constrain “fact induction” (Phase 4) to include only relationships between conceptual entities that exist at a similar or deeper depth from the ontology root and therefore can be expected to express useful and not overly generic hypothetical relationships. Doing so, however, requires us to first calculate the depth to the ontology root (or roots) for every conceptual entity selected in Phase 2 of this process, usually using the shortest such path as the preferred measurement when there exist more than one path from the concept to the root of the source ontology or equivalent conceptual knowledge construct.

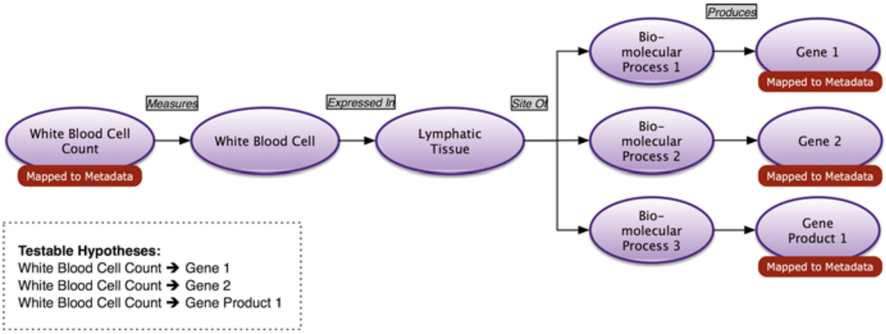


Fig. 8.9 Example of "fact" induction for prototypical example, in this case, creating testable hypotheses linking the initial and terminal concepts via multiple intermediate concepts and relationships (please note, this example assumes satisfaction of the depth based granularity controls associated with Phase 3 of the overall CI process)

- **Phase 4 – "Fact" Induction**: Once we have completed Phase 1–3, we can begin the "fact" induction process. In this phase we begin with a collection of variables contained within the target metadata of interest. For example, we could select all of the clinical measurements available that might serve to characterize how a patient would response to a therapy (such as laboratory findings or disease-specific performance or functional status indicators). Now, beginning with those variables, we can select a second set of variables that might serve as biomarkers of interest for predicting such treatment outcome, for example, indicators of genomic expression. Then, using the conceptual knowledge collection(s) that are mapped and sub-selected due to their connections to such variables, we can begin to explore the graph like representation of that knowledge to identify pathways that may link together variables in those two respective target "sets", being mindful of the granularity controls introduced in Phase 3. It is important to note in this process that such pathways are often "higher order" and can include multiple "intermediate" concepts and relationships that serve to link together an initial and terminal concept. For example, using our favorite case of "White Blood Cell Count", we might find that it is linked to the molecular entity "White Blood Cell" via a relationship labeled as "measures", and that "White Blood Cell" in turn has a relationship labeled as "expressed in" that connects it to the entity "Lymphatic Tissue." Subsequently "Lymphatic Tissue" could be linked via multiple "site of" relationships to a variety of bio-molecular processes that in turn may have relationships to certain genes or gene products that serve to measure the function or outcomes of those processes. Thus, we can then assert a "fact" that may infer a testable hypothesis linking our initial and terminal concepts and that could be tested using information contained in the source data sets(s) characterized by the metadata first identified in Phase 1 of this process (as illustrated in Fig. 8.9).
- **Phase 5 – Prioritization via Support Analyses**: In this nearly final step, it is often necessary to prioritize the hypotheses (or "facts") generated in Phase 4,

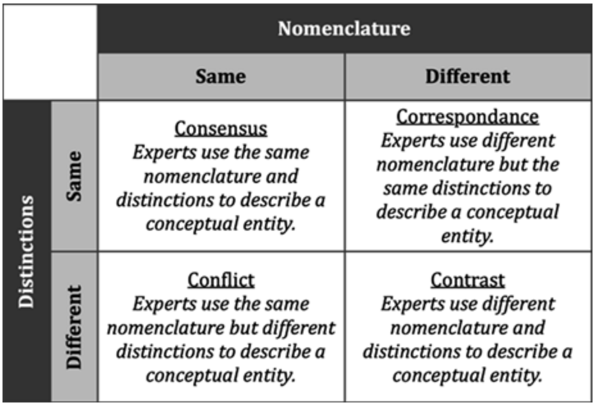
using some sort of quantifiable metric. This is necessary as CI-based agents can often generate thousands of hypotheses when reasoning over even a hundred or more initial and terminal variables. It is unlikely that human beings will take the time and expense (or have the energy and focus) to review and test all possible hypotheses. In response to this need, we often go back to the source data or alternatively, look at published literature and the knowledge that can be extracted from that literature (for example, the statistical distributions or co-occurrence of two variables of interest in the data or literature respectively) to calculate a support metric. Such support metrics tell us how common or uncommon those data or concepts are, and can be used to judge either the likelihood of the hypothesis being testable and/or novel. Then, depending on our use case, we can apply such metrics to prioritize or rank hypotheses for exploration and testing.

- **Phase 6 – Evaluation**: Finally (and perhaps most importantly), we must evaluate the output of CI-based agents using a variety of verification and validation methodologies. Such evaluations must incorporate multiple dimensions, include the factual accuracy or validity of system output, its likelihood in terms of informing novel hypotheses, and its overall utility as judged by the targeted end users. Further details on specific approaches to addressing this particular need are provided in Sect. 8.4.

8.4 Evaluating the Output of In Silico Hypothesis Generation Tools and Methods

The verification and validation of conceptual knowledge collections and the results of intelligent agents that leverage such knowledge to reason over data sets is ideally approached as an iterative and multi-method evaluation approach. First and foremost, when designing and applying such evaluation plans, it is very important to recognize and understand what types of process or outcomes measures are being targeted. Attaining such an understanding, in the context of intelligent agent design, requires us to differentiate between verification and validation. To summarize the definitions provided earlier, *verification* is the evaluation of whether an intelligent agent meets the perceived requirements of end-users, and *validation* is the evaluation of whether that same agent meets the realized (i.e., “real-world”) requirements of the end-users. The only difference between these techniques is that during verification, results are compared to initial design requirements, whereas during validation the results are compared to the requirements for the system that are realized after its implementation.

Examples of verification and validation criteria include the degree of interrelatedness of the relationships discovered by the intelligent agent, the logical consistency of those relationships, and multiple-source or expert agreement with the results generated therein. Often, the degree of interrelatedness between relationships generated by an intelligent agent for hypothesis discovery purposes is used as a measure of its “quality”, with such “quality” being defined by the degree to



		Nomenclature	
		Same	**Different**
Distinctions	**Same**	Consensus *Experts use the same nomenclature and distinctions to describe a conceptual entity.*	Correspondance *Experts use different nomenclature but the same distinctions to describe a conceptual entity.*
	Different	Conflict *Experts use the same nomenclature but different distinctions to describe a conceptual entity.*	Contrast *Experts use different nomenclature and distinctions to describe a conceptual entity.*

Table 8.1 Differentiation of types of agreement in multi-expert KA studies. In this model, the use of the "same" nomenclature or distinctions refers to the sources or experts using semantically similar or compatible means of describing or classifying concepts in a domain. Similarly, the use of "different" nomenclature or distinctions refers to the sources or experts using semantically dissimilar or incompatible means of describing or classifying concepts in a domain

which possible relationships between entities are enumerated or otherwise defined within the underlying knowledge collections. The logical, or axiomatic consistency of the relationships that comprise a hypothesis is often used as a measure of the accuracy of the output of the agent, again as defined by the correspondence of axioms that may be derived from the source knowledge collection(s) with the hierarchical and semantic assertions that make up such conceptual knowledge. Finally, multiple-source or expert agreement is most commonly used to validate the utility or impact of the output of the intelligent agent in "real world" application-oriented scenarios. This later set of measures is a critical criterion when attempting to measure the likely utility or impact of results generated by an intelligent agent. Unfortunately, there is not a single approach for measuring multiple-source, or expert agreement – since most evaluation methods corresponding to this type of metric involve the engagement of multiple (human) subject matter experts (SMEs). Instead, metrics must be chosen based upon variables such as data type as well as the number and types of knowledge sources being used. Most importantly, such analyses must be formulated in a manner consistent with the relative importance of four different types of agreement: (1) consensus; (2) correspondence; (3) conflict; and (4) contrast. Definitions of each of these types of agreement are provided in Table 8.1. A detailed discussion of the techniques that may be applied to measure agreement can be found in the reviews provided by Hripcsak et al. [30, 31].

At the highest level, the specific methods that can be used to satisfy the types of evaluation measures introduced above can be organized into a taxonomy consisting of the following major categories: heuristic, quantitative, information theoretic, graph theoretic and logical (Fig. 8.10). Brief descriptions of the techniques included in each category are provided below:

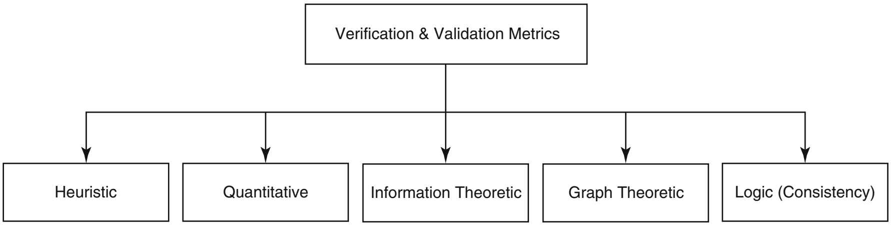


Fig. 8.10 Taxonomy of verification and validation metrics for the results generated by intelligent agents that leverage conceptual knowledge collections

8.4.1 Heuristic Methods

Heuristic metrics are probably the most common approach to verifying or validating the output of intelligent agents such as in silico hypothesis discovery tools. In this case, we use the term heuristic to refer to "rules of thumb" or more formally, rules that are informed by the expertise or commonly held knowledge of human SMEs. The advantages of using heuristics are the ability to incorporate domain-specific knowledge or conventions, and their simplicity (i.e., knowledge engineers or experts manually review the knowledge collection to determine if the contents are consistent with the heuristics). However, since such measures are difficult to automate or scale to larger data sets, such heuristic techniques are limited in their tractability when applied to "big data" contexts. Furthermore, heuristically comparing "quality" across multiple hypotheses or underlying knowledge collections is difficult, as a result of the relative and qualitative nature of the evaluation. Specific heuristic criteria for verifying or validating the output of intelligent agents have previously been proposed by Gruber [32] and include the following factors:

- Clarity
- Coherence
- Extendibility
- Minimal encoding bias
- Minimal deviation from ontological commitment, where ontological commitment refers to the situation were all observable actions of a knowledge-based system utilizing the given ontology are consistent with the relationships and definitions contained within that ontology.

8.4.2 Quantitative Methods

Quantitative methods of evaluating the results generated by intelligent agents are best suited for measuring both multi-source agreement and the degree of interrelatedness of ensuing hypotheses. Such measures can include simple statistics such as the precision, accuracy and chance-corrected agreement of the multiple sources used during reasoning processes [31–36]. Using frequency-based measures (e.g., measuring the frequency with which a given entity is related to other entities within the knowledge collection) in addition to simple statistics can allow for the assessment of the degree of interrelatedness of a set of multiple hypotheses [37].

8.4.3 Information Theoretic Methods

Information theoretic methods are most commonly applied to measure multi-source agreement in an aggregate collection of multiple hypotheses. The use of information theory to evaluate the agreement between multiple sources is based on the argument that if such agreement exists, it will be manifested as repetitive patterns within the resulting information constructs. To utilize this verification and validation approach, the relationships between units of knowledge that make up each constituent hypothesis must be represented as a numerical matrix, where each cell contains a numerical indication of the strength of the relationship between the two units of knowledge identified by the corresponding row and column indices. Given such a matrix, repeating patterns can be quantified based on their effect on information content or complexity. Matrix complexity is determined by calculating the number of repeating patterns within the matrix less the contribution of the overall environment within which the matrix is constructed. The probability of each repeating pattern detected in the actual matrix occurring randomly or as a result of the environmental contribution can be computed by generating multiple random matrices. As matrix complexity decreases, the degree of multi-source agreement increases [35]. This type of evaluation method is summarized in Fig. 8.11, and further detail can be found in the work reported on by Kudikyala et al. [35].

8.4.4 Graph Theoretic Methods

Graph theoretic methods are based on the ability to represent knowledge-based formulations, such as the output of intelligent agents, as graph constructs, where individual units of information or knowledge are represented as nodes, and the relationships between these units as arcs. Such graph representation of knowledge collections has been described in a number of areas, including ontologies [32, 38], taxonomies [39, 40], controlled terminologies [41] and semantic networks [40, 42]. Given a particular graph representation of a hypothesis or set of hypotheses, the degree of interrelatedness of those knowledge-based products can be assessed using a group of graph-theoretic techniques known as class cohesion measures. Such metrics are used to assess the degree of cohesion, a property representative of connectivity within a graph. Specific class cohesion measurement algorithms include the Lack of Cohesion of Methods (LCOM), Configurational-Bias Monte Carlo (CBMC), Improved Configurational-Bias Monte Carlo (ICBMC) and Geometrical

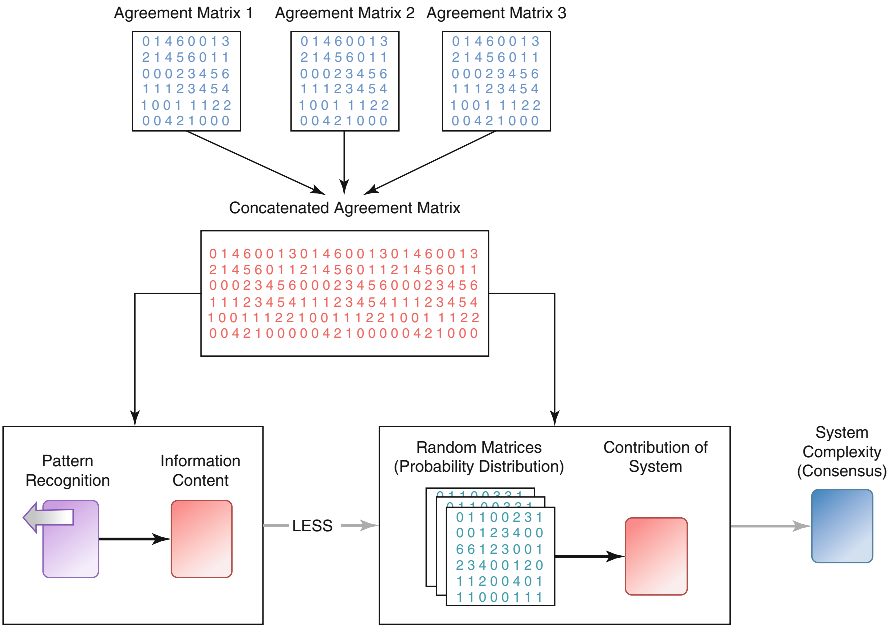


Fig. 8.11 Overview of information theoretic evaluation method for determining the degree of multi-source or expert agreement within a knowledge collection or system

Design Rule Checking (DRC) algorithms [43]. All of these algorithms use some combination of the number of and distance between interrelated vertices within the graph as the basis for determining cohesion. Most cohesive graphs generally possess more interrelated vertices with relatively short edges between them. However, it is important to note that a precise definition of what constitutes “cohesion” in a graph is not necessarily universally agreed upon. Due to this lack of agreement, class cohesion algorithms tend to utilize different measures for cohesion. The applicability of these metrics varies depending on the specific evaluation context. As a result, the selection of an appropriate cohesion measure is highly dependent on the specific nature of the data set and application scenario being evaluated. Further details concerning the theoretical basis and application of graph theory-based cohesion measures can be found in the review provided by Zhou et al. [43].

8.4.5 Logical Methods

The application of logic-based verification and validation techniques for the output of intelligent agents focuses on the detection of axiomatic consistency. These techniques require the extraction of logical axioms from the knowledge collection that has informed such in silico hypothesis discovery operations. Once axioms have been extracted, they are then applied within the targeted domain in order to evaluate their consistency and performance. In addition, logical methods can be utilized to examine axioms and assess the existence of unnecessary or redundant relationships within the knowledge collection. One of the most common approaches to implementing this type of evaluation is the representation of the individual hypothesis generated by the agent as formal ontological constructs within the Protegé knowledge editor [44]. Once such hypotheses have been represented in Protégé, logical axioms can be extracted and evaluated using the Protegé Axiom Language (PAL) extension [45]. An example of this method can be found in the formal evaluation of the logical consistency of the Gene Ontology (GO) [46] reported by Yeh et al. [45].

8.4.6 Hybrid Methods

As described earlier, hybrid methods for verifying or validating knowledge collections involve the use of techniques belonging to two or more of the classes of measures as described above. An example of such a hybrid method is the novel computational simulation approach to validating the results of multi-expert categorical sorting studies as proposed by Payne and Starren [47]. This approach measures multi-source agreement using a combination of quantitative and graph theoretic methods. Another example of a hybrid technique is the use of hypothesis discovery methods, such as hierarchical clustering [48] to determine the degree of interrelatedness of a knowledge collection. Such evaluative methods combine statistical, heuristic and graph theoretic techniques.

8.5 Implications for Stakeholders

It can be seen that each of the different stakeholders described in Chap. 1 benefits realizing the vision of a Translational Informatics model that enables and facilitates knowledge-driven healthcare. With specific regard to the concepts associated with in silico hypothesis discovery, these benefits are multi-fold, and largely focus upon the accelerated pace and ease with which new diagnostic and therapeutic discoveries can be generated from existing or new data sets. Specific benefits at all of the levels introduced in Chap. 1 include:

8.5.1 Evidence and Policy Generators

- **Investments in the creation of large-scale and multi-dimensional data sets can exhibit much higher returns on investment** owing to the ability to generate a larger number of testable and potentially clinically actionable hypotheses from those resources;

- **Novel evidence and/or policy frameworks can be inferred based upon previously undiscovered patterns or motifs in historical data sets**, thus allowing such knowledge or decision making to be informed by the best possible information.

8.5.2 Providers and Healthcare Organizations

- Providers are able to **engage in the delivery of evidence-based and precision medicine informed by a full spectrum of scientific knowledge** that has been formulated by identifying and testing large numbers of hypotheses against all available data types and resources
- **Healthcare organizations can leverage their investments in EHR technologies and bio-molecular instrumentation** so as to rapidly learn from all patient-centered data being created during the course of normal clinical operations; that is, achieving the vision of a "learning healthcare system" wherein every patient encounter is an opportunity to both create new knowledge and improve care for that patient, their family, and their community.

8.5.3 Patients and Their Communities

- **Patients are able to be part of the "learning healthcare system"** such that they both become an integral component of research processes and benefit from the knowledge generated therein
- **Interested community members can begin to identify novel or interesting associations between disparate data that spans healthcare providers and the world-at-large**, thus becoming part of the research enterprise. For example, community members could use in silico hypothesis discovery tools to identify relationships between healthcare outcomes and socio-demographic factors that could inform advocacy and/or community development activities intended to promote wellness.

8.6 Conclusions

As has been discussed in a variety of ways throughout this book, the ongoing growth and increasing complexity of biomedical data presents a wealth of challenges and opportunities relative to informing a Translational Informatics vision for knowledge-drive healthcare. In this chapter, we have discussed a specific aspect of those challenges and opportunities, concerned with the disconnect between the volume of data being generated in numerous settings and the current state-of-the-art in terms of hypothesis formulation and testing relative to such resources, which remains extremely basic. As has been illustrated, most if not all hypotheses that are evaluated in the modern scientific setting are generated in a low-throughput manner based upon the intuition or belief systems of an individual or team of investigators. Despite historical precedence for such approaches, they are discordant with the modern, high-throughput data types we regularly encounter, and that are being generated by EHRs, PHRs, sensor technologies and bio-molecular instrumentation (to name a few of innumerable examples). In response to this challenge, we can look to a set of core concepts that underlie alternative and high-throughput approaches that can lead to in silico hypothesis discovery paradigms. These types of methods employ domain-specific conceptual knowledge collections, such as ontologies or knowledge that can be extracted from the domain literature using machine learning or natural language processing methods, in order to reason upon and generate hypothesis corresponding to a data set or data sets in an extremely high throughput manner, usually realized via the implementation of knowledge-based and intelligent software agents. While these types of in silico hypothesis discovery methods remain very early in their development, they also hold great promise in terms of accelerating the pace, breadth and depth of scientific discovery in the "big data" era, and thus represent a critical dimension of the vision for Translational Informatics.

Discussion Points

- What are the major barriers to the generation and testing of hypotheses in large-scale and/or heterogeneous data sets?
- What differentiates procedural, strategic, and conceptual knowledge? How are these knowledge types related across a continuum of operationalization?
- What role can conceptual knowledge collections play in overcoming the preceding barriers?
- As an example of an in silico hypothesis discovery method, what considerations must be addressed when employing Constructive Induction (CI) relative to concept granularity and/or the evaluation of ensuing hypotheses?
- When evaluating the output of knowledge-based intelligent agents used for in silico hypothesis generation, what is the fundamental difference between the verification versus validation of such constructs?

References

1. Zerhouni EA. US biomedical research: basic, translational, and clinical sciences. JAMA. 2005;294(11):1352–8. PubMed PMID: 16174693.
2. Sung NS, Crowley Jr WF, Genel M, Salber P, Sandy L, Sherwood LM, et al. Central challenges facing the national clinical research enterprise. JAMA. 2003;289(10):1278–87. PubMed PMID: 12633190.
3. Payne PR, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. J Investig Med. 2005;53(4):192–200. PubMed PMID: 15974245.

4. Liu H, Motoda H. Feature extraction, construction and selection: a data mining perspective. Norwell: Kluwer Academic Publishers; 1998.
5. Payne PR, Borlawsky TB, Kwok A, Dhaval R, Greaves AW, editors. Ontology-anchored approaches to conceptual knowledge discovery in a multi-dimensional research data repository. In: 2008 AMIA Translational Bioinformatics Summit. San Francisco: American Medical Informatics Association; 2008.
6. Payne PR, Borlawsky TB, Kwok A, Greaves AW. Supporting the design of translational clinical studies through the generation and verification of conceptual knowledge-anchored hypotheses. AMIA Annu Symp Proc. 2008:566–70. PubMed PMID: 18998958. Pubmed Central PMCID: 2656058. Epub 2008/11/13. eng.
7. Payne PR, Borlawsky TB, Rice R, Embi PJ. Evaluating the impact of conceptual knowledge engineering on the design and usability of a clinical and translational science collaboration portal. AMIA Clinical Research Informatics Summit. San Francisco: American Medical Informatics Association; 2010.
8. Payne PR, Embi PJ, Johnson SB, Mendonca EA, Starren JB. Improving the usability of clinical trial participant tracking tools using knowledge-anchored design methodologies. Appl Clin Inform. 2010;1(2).
9. Payne PR, Huang K, Keen-Circle K, Kundu A, Zhang K, Borlawsky TB. Multi-dimensional discovery of biomarker and phenotype complexes. AMIA Translational Bioinformatics Summit. San Francisco: American Medical Informatics Association; 2010.
10. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article 17. PubMed PMID: 16646834. Epub 2006/05/02. eng.
11. Zhang J, Ding L, Keen-Circle K, Borlawsky TB, Xiang Y, Ozer G, et al. Predicting biomarkers for chronic lymphocytic leukemia using gene co-expression network analyses for ZAP70. AMIA Translational Bioinformatics Summit. San Francisco: American Medical Informatics Association; 2010.
12. Zhang J, Xiang Y, Jin R, Huang K. Using frequent co-expression network to identify gene clusters for breast cancer prognosis. Proc Int Joint Conf Bioinforma Syst Biol Intell Comput. 2009;428–34.
13. Goldstein T. Dawn of modern science: from the ancient Greeks to the Renaissance. New York: Da Capo Press; 1995.
14. Chung TK, Kukafka R, Johnson SB. Reengineering clinical research with informatics. J Investig Med. 2006;54(6):327–33. PubMed PMID: 17134616.
15. Embi PJ, Payne PR. Clinical research informatics: challenges, opportunities and definition for an emerging domain. J Am Med Inform Assoc. 2009;16(3):316–27. PubMed PMID: 19261934. Epub 2009/03/06.
16. Zerhouni EA. Translational and clinical science – time for a new vision. N Engl J Med. 2005;353(15):1621–3. PubMed PMID: 16221788.
17. Glaser R. Education and thinking: the role of knowledge. Am Psychol. 1984;39(2):93–104.
18. Hiebert J. Procedural and conceptual knowledge: the case of mathematics. London: Lawrence Erlbaum Associates; 1986.
19. McCormick R. Conceptual and procedural knowledge. Int J Technol Des Educ. 1997;7:141–59.
20. Scribner S. Knowledge at work. Anthropol Educ Q. 1985;16(3):199–206.
21. Barsalow LW, Simmons WK, Barbey AK, Wilson CD. Grounding conceptual knowledge in modality-specific systems. Trends Cogn Sci. 2003;7(2):84–91.
22. Borlawsky T, Li J, Jalan S, Stern E, Williams R, Lussier YA. Partitioning knowledge bases between advanced notification and clinical decision support systems. AMIA Annu Symp Proc. 2005:901. PubMed PMID: 16779188.
23. Preece A. Evaluating verification and validation methods in knowledge engineering. In industrial knowledge management. 2001:91–104. Springer London
24. Newell A, Simon HA. Computer science as empirical inquiry: symbols and search. In: Haugeland J, editor. Mind Design. Cambridge: MIT Press/Bradfor Books; 1981. p. 35–66.
25. Compton P, Jansen R. A philosophical basis for knowledge acquisition. Knowl Acquis. 1990;2(3):241–57.

26. Newell A, Simon HA, editors. Computer science as empirical inquiry: symbols and search. ACM annual conference. Minneapolis; 1975.
27. Gaines BR, Shaw MLG. Knowledge acquisition tools based on personal construct psychology 1993 [Cited 2005 8/23/2005]. Available from: http://www.repgrid.com/reports/KBS/KER/.
28. Kelly GA. The psychology of personal constructs. 1st ed. New York: Norton; 1955. 2 v. (1218) p.
29. Payne PR, Mendonca EA, Johnson SB, Starren JB. Conceptual knowledge acquisition in biomedicine: a methodological review. J Biomed Inform. 2007;40(5):582–602. PubMed PMID: 17482521.
30. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002;35(2):99–110. PubMed PMID: 12474424.
31. Hripcsak G, Wilcox A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc. 2002;9(1):1–15. PubMed PMID: 11751799.
32. Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. Int J human-computer studies. 1995;43(5):907–28.
33. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005;12(3):296–8. PubMed PMID: 15684123.
34. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR large scale vocabulary test. J Am Med Inform Assoc. 1997;4:484–500. PubMed PMID: 9391936.
35. Kudikyala UK, Allen EB, Vaughn RB, editors. Measuring consensus during verification and validation of requirements. Proceedings of the tenth IEEE International Software Metrics symposium (METRICS 2004). Chicago; 2004.
36. Morgan MS, Wm. Benjamin Martz, Jr. Group Consensus: do we know it when we see it? In: Proceedings of the Proceedings of the 37th annual Hawaii international conference on System Sciences (HICSS'04) – Track 1 – vol. 1: IEEE Computer Society. Waikoloa: Hawaii; 2004.
37. Brachman RJ, McGuinness DL. Knowledge representation, connectionism and conceptual retrieval. In: Proceedings of the 11th annual international ACM SIGIR conference on Research and Development in Information Retrieval. Grenoble: ACM Press; 1988.
38. Ian N, Adam P. Towards a standard upper ontology. In: Proceedings of the international conference on Formal Ontology in Information Systems – vol. 2001. Ogunquit: ACM Press; 2001.
39. Alan LR, Chris W, Jeremy R, Angus R. Untangling taxonomies and relationships: personal and practical problems in loosely coupled development of large ontologies. In: Proceedings of the international conference on Knowledge Capture. Victoria: ACM Press; 2001.
40. Burgun A, Bodenreider O. Aspects of the taxonomic relation in the biomedical domain. Ogunquit: ACM Press; 2001. p. 222–33.
41. Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. J Am Med Inform Assoc. 2000;7(3):288–97. PubMed PMID: 10833166.
42. Griffith RL. Three principles of representation for semantic networks. ACM Trans Database Syst. 1982;7(3):417–42.
43. Zhou Y, Lu J, Xu HB. A comparative study of graph theory-based class cohesion measures. SIGSOFT Softw Eng Notes. 2004;29(2):13.
44. Noy NF, Crubezy M, Fergerson RW, Knublauch H, Tu SW, Vendetti J, et al. Protege-2000: an open-source ontology-development and knowledge-acquisition environment. AMIA Annu Symp Proc. 2003:953. PubMed PMID: 14728458.
45. Yeh I, Karp PD, Noy NF, Altman RB. Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). Bioinformatics. 2003;19(2):241–8. PubMed PMID: 12538245.
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Gene Ontol Cosortium Nat Genet. 2000;25(1):25–9. PubMed PMID: 10802651.

47. Payne PR, Starren JB. Quantifying visual similarity in clinical iconic graphics. J Am Med Inform Assoc. 2005;12(3):338–45. PubMed PMID: 15684136.
48. Everitt B, Landau S, Leese M. Cluster analysis. 4th ed. New York: Oxford University Press; 2001. p. 237.

Additional Reading

Everitt B, Landau S, Leese M. Cluster analysis. 1st ed. New York: Oxford University Press; 2001. 2 v. (1218) p.

Glaser R. Education and thinking: the role of knowledge. Am Psychol. 1984;39(2):93–104.

Goldstein T. Dawn of modern science: from the ancient Greeks to the Renaissance. New York: Da Capo Press; 1995.

Hiebert J. Procedural and conceptual knowledge: the case of mathematics. London: Lawrence Erlbaum Associates; 1986.

Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002;35(2):99–110.

Hripcsak G, Wilcox A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc. 2002;9(1):1–15.

Liu H, Motoda H. Feature extraction, construction and selection: a data mining perspective. Norwell, MA: Kluwer Academic Publishers; 1998.

Newell A, Simon HA. Computer science as empirical inquiry: symbols and search. In: Haugeland J, editor. Mind design, vol. 1. Cambridge: MIT Press/Bradfor Books; 1981.

Payne PR, Mendonca EA, Johnson SB, Starren JB. Conceptual knowledge acquisition in biomedicine: a methodological review. J Biomed Inform. 2007;40(5):582–602.