Chapter 9.
Scholarship, Science, Simulation, Scholarly Data Sets, and Domain Informatics

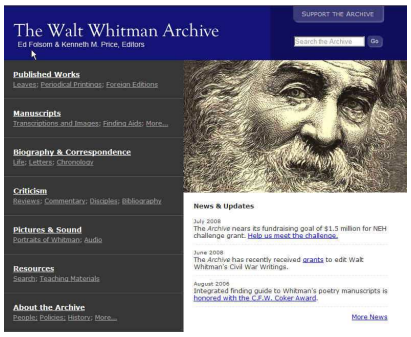

Figure 9.1: Walt Whitman Digital Archive (just upper portion) (check permission).

Scholarship is generally based on thoughtful, systematic investigations. Such reflection can appear overly dry but it has a role in ensuring carefully evaluation of complex issues. It uses a cumulative body of knowledge to develop new interpretations. Content-specific issues dominate. Science is one approach to scholarship which we examine in the next section. Formal information seeking everyday or citizen information seeking. Personal scientists. Data management as a type of library. This is how society captures knowledge. Institutions for collecting and organizing knowledge. Domain informatics.

9.1. Scholarship

Scholarship and science help get the right answer to difficult problems. Potentially, it can detect and understand broad trends. Both depend on systematic development and management of information. This can be thought of as critical thinking (5.12.0). Knowledge institutions especially universities (8.13.2). Although scholarship is often described as an ivory tower, it generally does aim to have an impact on societal issues. STS (9.4.0) and the social context of scholarship. Scholarship often comes with particularly high standards for identifying challenging problems and for evidence that addresses them.

9.1.1. Scholarly Communities and Communication

Scholarship can be solitary but the results of even individual scholarship eventually need to be disseminated to the broader community. Indeed, solutions to many complex issues are probably possible only by an ongoing scholarly discourse. Scholarly communities are communities of practice (5.8.2). Indeed, because scholarship emphasizes the value of ideas many scholarly communities have a norm of minimizing other marks of social status.

They are also particularly information intensive and focus on documenting their work through scholarly literature. The nature of scholarship, that of an empirically based presentation of intellectual work, requires a rigorous presentation and also a relatively free exchange of ideas. The publication of this work, by institutions both public and private, is at the same time the culmination of individual scholarship and the motor that drives it. Scholarship, be it in science or the humanities, is a highly collaborative endeavor. It is rare (and perhaps impossible) for scholarship to be completely new, and to not use assumptions or facts based on the work of past or present colleagues in the field. No document stands on its own — it is only by referencing the work of others that any validity is conferred. Published scholarly work contains footnotes and citations that link to previous relevant findings. Publish or perish.

Although the creation of scholarly material may be done alone, when it is completed, it is shared with

a community. In that community, scholarship is a cultural value. Associations of scholars are known as "learned societies". Scholarly communities can uphold ethical and quality standards. They may encourage data sharing but this must be balanced by constraints such as intellectual property rights and security. Scholarly organizations support the values of the profession (5.8.2).

Specialized descriptive systems and catalogs.

Academic honesty. One example is plagiarism (5.12.3). Scholars have ethics. These may range from citing previous research to not stealing other researcher's findings.

Universities' role as a knowledge institution (8.13.2). Scientific disciplines have different cultures.

Characteristics of Scholarly Publishing

Scholarly literature seeks to promote the scholarship of a given field. Thus, it needs to be clear and to reflect the standards of the scholarly community. Scholarship is cumulative; it builds on recording and sharing previous work. Scholarly publishing is the traditional approach way for sharing knowledge. Thus, there is a cycle of production and distribution of knowledge. Value proposition in terms of how literature supports future research. One of the main goals of scholarly activity is to disseminate advances in the field. Thus, scholarly conferences and publications are an integral aspect of scholarship. Reuse of scientific results.

Scholarly articles are summaries of the major points about research.

Range of types of communication: From classes to presentations to dissertations to books to research reports. Academic libraries (7.2.1).

Style: Maximize clarity. Avoid philosophy.

Even scientists are often less that forthright about admitting the limitations for their research.

These publications are often distributed as "proceedings" or "communications". to the work of "epistemic communities" [26]. Digital resources in scholarship. Academic libraries (7.2.1).

Contributions to scholarship should be original work which make a new contribution. The data speak for themselves, the paper should use a neutral tone. For instance, avoid honorary titles. That work would be published is published in the primary literature.

Who owns the copyright and publication in multiple venues such as arXiv.org

Scholarship and publication is confounded with implications for tenure and job security.

Standards of authorship. Inclusion as an author implies that the person has made a substantial contribution to the work.

Expectations for scholarly work to be published. Substantial contribution, original, novel. Nonetheless, there can be incorrect information but replication of the procedure can detect that.

Scientific commons. Scientific social media. Scholarly annotations. Systematic policies.

Controlling the community of editors to an extent. Published but not peer reviewed. Supporting scholarship with scholarly commons. Alternatives to peer review such as community reviews.

To a surprising extent, there are differences in the conventions in scholarly communities. Physicists use online publishing but chemists use journals. There are distinct norms within the different scholarly communities.

Beyond the traditional research report, there are many other types of scientific writing. Secondary Scholarly Literature magazines, textbooks, encyclopedias. As with journalism, for this material, the reputation of the publisher is in selecting quality material even its technical content is not reviewed

Figure 9.2: SSRN and arXiv are two well-known pre-print repositories.

There are also, unpublished works such as grant proposals. If publications are are written for a non-official purpose or destination and not peer reviewed, they are considered "gray literature," and are not accorded as much trust as peer-reviewed literature.

Material which has not been formally reviewed is known as gray literature although that distinction is becoming less distinct as more publication options are becoming available.

Scholarly Publishing and Scholarly Digital Libraries

Acceptance for publication confers status on a publication as being an important contribution. Content selection is based purely on merit. There should be no financial consideration. If advertising appears at all, it should be clearly separated from the text. These are derived from the core values of scholarly communities. Electronic resources may have more impact because they are more widely available.

Prior to publication most formal scholarly literature is reviewed Or refereed by independent readers. Reviews check that a scholarly contribution is made. Fig. 9.3. The most common approach is blind reviewing in which the author is known to the reviewers but the reviewers are not know to the author. In some cases, double-blind reviewing is used. In this technique, neither the authors nor the reviewers are supposed to know each other's identities. However, it is often easy to guess authorship of technical work even if the authors' names are not included. Systems to support reviews.

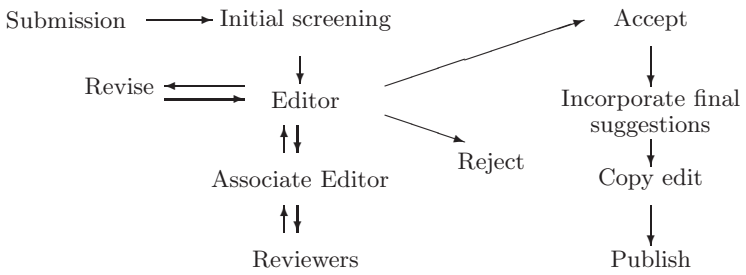

Figure 9.3: Schematic of a typical peer-review process for a scholarly publication.

Trust in a publisher's brand is often a factor in acquisition and use of an information resource. In a brader sense, the review process is a system for certification by the scientific community. It is often difficult to find expert reviewers for highly technical work because there are only a few experts. As a result, the only competent reviews are other researchers — the peers of the scholar — so we say that the work is be peer reviewed. While there is sometimes criticism of peer review because it appears that a small club is rewarding itself, there seems to be no better way to make these decisions. Group consensus and gaming detection mechanisms.

HIVE project[37].

Business Models for Scholarly Publishing and Digital Libraries The business models for scholarly publications are changing rapidly. Information systems have dramatically changed the production and

distribution of books. Because of the costs of production have been reduced so dramatically that prices are reduced. On one hand, scholarly research is generally created in response to the expectation that scholars will communicate their research. On the other hand, its distribution may be restricted by copyright. Open Access Publishing. Make scholarly, especially scientific, reports freely available to the public. One example is the Public Library of Science (PLOS) which employs an "author-pays" business model (Fig. 9.4). Earlier, we discussed business models for publishing (8.13.5). Feedback to publishers for the design of publication. Difficult to evaluate the value of scholarly research. Information economics (8.13.3). Alternative sources of value from one's knowledge work.

Figure 9.4: The Public Library of Science (PLOS) is an open access collection of medical and biological articles.

Evidence for long-term viability of open-access publications. Because scholarly literature supports scholarly communities, it can be thought of as a type of knowledge commons.

However, some publishers also require transfer of the copyright from the authors. And, increasingly, articles are only available online as part of publisher's digital library. The costs for access to those articles can be very expensive.

Scholarly communities and community-run journals. This collaboration serves to create an extensive community of practice, as scholars share ideas, give commentary, and suggest possibilities for more research. They are also useful for education. Publishing authors begin to develop a niche for themselves within their larger field. Sometimes that niche becomes polarized by two vying epistemic communities with competing theories. Other times there is a sort of community consensus about what lies within the norm of current thought, with each publication building from those that have gone before it. This ubiquitous and extensive interconnection of articles and ideas can provide intriguing insight into a given discipline or topic.

Scholars Workbench. Query management system.

Scholarly communities and publications exist in the context of institutions such as universities and libraries.

Is there a citation advantage for materials which are published in open access, online repositories?

9.1.2. Citations and Citation Analysis

Scholarship is cumulative. Scholarship requires providing evidence for claims. That evidence is provided by citations to prior work.

There are many reasons for citations but at some reflect the author's decisions about the orientation and emphasis of the research. Taken collectively, citations across the works in a field can characterize the values, influences and directions of that field. Analysis of the scholarly literature. It provides context and continuity for new contributions. It also provides us with a view of community.

There is a strong similarity between citations links and Web links (2.6.3).

A "citation index" can be particularly effective for researching a complex topic — the report could list all studies published in the past two years that contain a particular reference. The importance or impact of an individual work can also be determined by analyzing the number of times it has been referenced, similar to the ranking of Web pages on the Internet (2.6.3). [80] [18]

Citations support literature searching which we discussed earlier (3.3.1).

Citations are used by authors to indicate the intellectual foundations for scholarship. Citations link scholarly articles to previously published articles on related topics. However, the judgment about what citations to include is somewhat subjective; thus they may reveal something about the process and logic of science. They can show who is publishing in what fields and reveal communities of scholars who are doing related work. The existence of these communities of practice notable in peer-review.

Scholarship is cumulative and citations in articles provide links to documents published in the past. Citations can be an indication of similarity of documents. A better strategy for analyzing citations compares works which are cited together. This provides much richer data than simple citation analysis [?, ?]. Another strategy also includes links among authors. This leads to author co-citation analysis as compared to document co-citation.

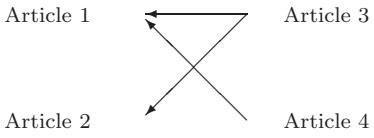

Figure 9.5: It is helpful to have a measure for how similar two articles are to each other. One way is by analyzing citation links. One measure is similarity is bibliographic coupling. Article 1 is cited by 3 and 4 so we say that 3 and 4 and similar. Articles 1 and 2 show a co-citation relationship with Article 3 It turns out that the co-citation measure is better than bibliographic coupling. (under construction)

Citation roles and types.

Citation analysis looks at the reference lists of published academic articles to determine if there are inherent patterns to what is cited, in what subject area, and with what other references. Analysis of this variety uncovers relationships between whole academic fields as well as specific subject areas. Citation analysis is related to bibliometrics in that it seeks to quantitatively study a vast bulk of information.

Citations and social networks[81]. Citations may also be used as an indicator of the rich social network in science. Co-authorship as an indication of collaboration. Small-world networks for co-authorship. Citation context can suggest relevant local information[66]. Author-topic models.

Citations across disciplines (8.13.2).

Finding Undiscovered Public Knowledge

Presumably, there are many important findings remaining to be found by combining knowledge which is recoded in books and journals. Typically, a researcher creates a hypothesis by being familiar with the current state of research in a given area of study. This requires familiarity with the literature. which in no way excludes the other, is by being familiar with that area's research literature. The development of automated data-mining techniques has revealed important, but previously unnoticed, relationships among concepts. An unsuspected relationship between magnesium and migraine headaches[76] was determined entirely from citation analysis in the literature. This connection was confirmed by subsequent medical research. These are bibliographically disconnected literatures.

The first step could be identifying important concepts (10.5.3). Without full knowledge extraction, we may still be able to make some useful inferences from factors such as patterns of citations. As an example: Fish oil. Undiscovered public knowledge (Fig. 9.6). If we find evidence in the literature which suggests that there is an association such that A causes B and B causes C then we may believe it is worth investigating whether A causes C.

9.1.3. Bibliometrics

Bibliometrics is, literally, counting things about books and other information resources. It often deals with characteristics of the authors or publishers. It can include counts such as circulation for printed materials or hits on web pages. Scimetircs ((sec:scimetrics)).

Figure 9.6: There may be "undiscovered public knowledge" can we actually merge facts from one domain into another. In a classic example, research about the effects of calcium on headaches can be applied to the hypotheses about the effects of magnesium (which is chemically related to calcium) on headaches. (Magnesium as a calcium blocker) (check)

h-measure.

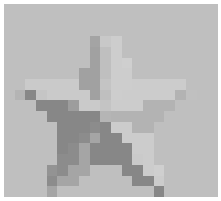

Figure 9.7: Google Scholar.

Measures derived from collections of citations may have several implications. For instance they are often used as indicators of scholarly productivity and impact. Zipf's Law.

Bibliometrics is often used to analyze trends within a discipline, or even the emergence of a new field. Mining citations.

Impact Factors

Impact factor. Who are the most commonly cited authors? Bibliometrics can also provide measures of the cohesion of scholarly communities (cf., -A.3.5).

Measuring impact of research with the number of papers which cite it. Counting citations. While citations are generally a positive indicator of quality, they are not satisfactory as a primary indicator. Citation counts are not very effective at predicting Nobel Prize winners among scientists[68]. Indeed, some papers which turn out to be highly influential are often not widely cited when they first appear. Citation counts have been proposed as a measure of quality of linking but they are not clear predictors. This raises the broader question of what are valid metrics of scholarly activity.

Earlier, we discussed the rate of change of individual documents. Bibliometrics can provide an indication of the evolution of a scholarly field. We may see changes in citation frequency. Note the pivot points (Fig. 9.8). Transition points between networks sub-graphs of the network. Rapid development in a field and the stability of co-citations[2]. Time-series visualization. Network Visualization. Pathfinder networks. Citation analysis and topic tracking (10.11.2). Burstiness as a sign of the rate of change in a field. Cumulatively, citations should be able to tell us about the intellectual underpinnings of a field. Interactive visualization tools. Citations can become symbols for concepts. Visualizations to support literature-related discovery.

Scholarly journals are the main path for distributing scholarly work. Some scholarly journals consistently have more impact than others. With a limited budget, we would prefer to access those publications with a high impact. We would like indicators of these publications' impact (Eq. 9.1). While we would hope that journal impact is related to broader societal impact, that is often not the case.

$$Journal\ Impact = \frac{number\ of\ citations\ to\ the\ journal\ from\ all\ sources}{number\ of\ articles\ published\ by\ the\ journal} \tag{9.1}$$

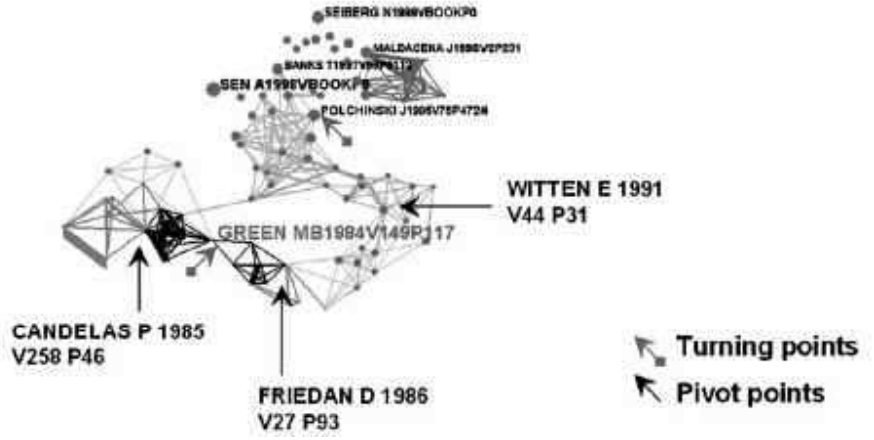

Figure 9.8: Citations networks can provide maps of a scholarly field. Networks of citations can suggest the evolution of a scholarly field[27]. Here we see clusters of research and transitions to new areas of research. (check permission)

Furthermore, there is a decreasing value of journals for a given researcher. Bradford's law. It is similar to the PageRank calculations for determining the importance of Web sites (10.10.2).

9.1.4. Research Space

Managing digital scholarship. Commons.

RDF triples for explantions.

9.2. Science

Science as a outcome or as a process. Science is a specific type of scholarship. Science develops statements about entities in the world and relationships among the entities. Furthermore, it applies specific criteria to those statements. Statements should be as general as possible but also as simple as possible. They should also be consistent with other scientific statements and falsifiable. Finally, they may be supported by interference but ultimately by experimental evidence. These properties give science distinctive characteristics. For instance, history also tries to make causal statements but there is no way to replicate the phenomena it studies. In addition, the activity of science should be distinguished from the informal use of the tern science as the theories that have been developed rather than about the process of discovering and modifying them.

Values of science activity[50]. Encouraging communication of results. Conflict of interest. Increasingly, politicians seem to be manipulating but the public's perception of science and even the reported science.

Science as practical art. Science allows us to manipulate the world.

9.2.1. The Scope of Science

Natural Science

In natural science, it's common to assume that the most valid observations are based on our senses. Models are built based on those sensory experiences and ideally, the models are tested by further observation. Thus science is distinguished from faith which is not confirmed by observation. Integral to the "Age of Reason".

Behavioral, and Social Science and Humanities

Natural science has been particularly effective in developing models and abstractions for the physical world. Behavioral science is based on consistencies in behavior. It has been less effective for describing human social behavior. As we have discussed, social phenomena are much more complex and adaptive

than natural systems. Indeed, social phenomena can be considered as an emergent and not readily derivable from the underlying natural processes.

The value of social science in relationship to natural science is has been the subject of so-called "science wars" [57]. One possibility for social science is to develop grounded theories [35] rather than trying to do hypothesis testing. Separation by levels of complexity. Social science has many epiphenomenon. Social science hypotheses sometimes do not allow falsification. Social science as understanding the nature and patterns of complex social systems. Nature of social science constructs. Increasingly, science includes understanding and modeling complex systems.

Thus, it is much more difficult to capture regularities for social systems. For instance, economics is sometimes called a "dismal science." Certainly, social activity appears to be very different from natural phenomena because human beings are highly adaptable and often unpredictable. As we have seen, many aspects of social systems are mutable. There is a debate about whether social science should attempt to identify isolated causal factors as compared to factors which are part of a broader milieu. Is the notion of causation in social science socially constructed?

Is an empirical social science possible? Causation in social science[64]. Moreover, its very difficult to do experiments in the same ways they are done in natural science. Social science addresses the variability in human behavior by trying to focus on those situations in which human behavior is relatively consistent and even then by statistical analyses[23]. Thus, explanations of human social organization need different sorts of models. Typically, social science is based on statistical testing.

Social science described with activity theory. Qualitative research. Subjective vs. intersubjective evidence.

Unlike areas of humanities such as history (5.13.0), science depends on the replicability of its effects.

9.2.2. Scientific Knowledge: Properties, Laws, and Models

There are many ways we know things, such as listening to others but science supports a more systematic investigation. Perhaps we also develop models of the world (4.4.4). Science generally tries to develop conceptual models of the phenomenon under investigation. A wide variety of phenomena are under study. For instance, medical science models complex processes and systems. On the other hand, systematic science-like analysis is applied by detectives. A paradigm is a body of scientific knowledge, not just one model. Standard theory for science. Science as developing new frames.

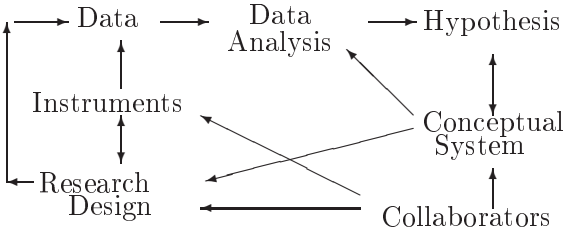

Figure 9.9: Many factors contribute to effective scientific inference (adapted from[79]). (check permission)

Scientific Laws and Models

Science as developing descriptions about entity classes. This often means developing models. Models are not reality. Scientific explanations. Discrete or continuous models. Entities/Systems and Interactions/Processes.

Laws are regularities with a broad application profile.

Accepted scientific explanations may change. Isaac Newton's models of gravity were replaced by Einstein's (Fig. 9.11) as described below.

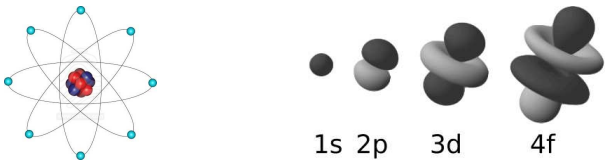

Figure 9.10: Two conceptual models of atoms: The Bohr model (left) and the quantum model (right) which is derived from quantum equations (redraw) (check permission)

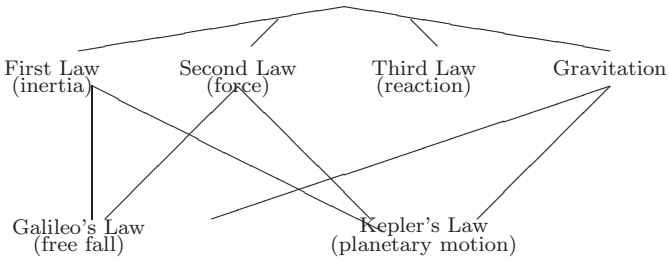

Figure 9.11: Explanatory coherence for Newton's Laws (from[78]). The set of Laws neatly covers commonly observed phenomena. (check permission)

Models may provide conceptual frameworks and science depends on developing conceptual models. These are often causal models. Models are also be useful for formulate new hypotheses (Fig. **??**).

Describe the entity classes and the relationships among them. Theories, Laws, and Models. Conjecture, Hypothesis, Theory, Law, Constants. Not all research is hypothesis testing. Because they are at the highest level, scientific laws (constants) do not have an explanation; rather, they are assumed to be true. Pre-causal models.

Equations and logic as models. For instance, Newton's F=MA

Process Models for Science The development of new scientific models is a creative activity. Cognition and the development of models such as reasoning by analogy (4.3.4). There has to be a base of knowledge and experience in a field. In other cases, models develop from a systematic search for evidence.

Reductionism proposes that complex phenomena can be described by appeal to low-level underlying processes. A reductionist would explain biology entirely in terms underlying chemical and physical processes. Reductionism avoids appeals to emergent phenomena. Minimize the number of assumptions.

Science attempts to develop more detailed and also more general models. Science as sense-making. Increasingly, rich and inter-related models are preferred over separate, equally adequate models. This is a principle known as Occam's Razor. However, there may be disagreement about what is superfluous. Indeed, scientific paradigms are often composed of complex interlocking assumptions and there may be an ongoing debate about them.

Scientific Explanations

Explanation versus modeling versus causation. Explanation (6.3.4).

Abstractions and generalization. Role of causality in exaplanations. Causation versus mechanisms. Science explanations as building models for abstractions. Different from explanations of history (11.3.3).

Explanations based on laws.

Functional explanations.

How do we decide what is pseudoscience?.

Is causation needed for scientific explanation? Problem of explanation and quantum theory.

Semantic technologies are also relevant. Importance of how entities are defined and how they relate to models. This is seen directly in taxonomy. Analogical models can be helpful (4.3.4).

Hypothesis generation. Undiscovered public knowledge (9.1.2). Mining scholarly literature text and citation linkages.

Causation in Science

9.2.3. Evidence for Science

What is adequate evidence to accept a theory?

Examples of accidental discoveries. Even more subtly, we might try to develop metrics for the quality of science.

Experimentation. Antropology. Astronomy,

Research designs versus research methods. Scientific research methods are often aligned with given disciplines (8.13.2).

Scientific research reports as creating a research space and staking a claim [75]. Science as deduction vs as argumentation (6.3.5). There are aspects of both. Operational definition. Testable hypotheses. Science is based on evidence and potentially every scientific generalization is disconfirmable[58]. Negative results. Fasifiability. How much evidence is needed to disconfirm especially if a finding does not match a popular model. Acceptance requires replicability.

Thought experiment. However, counterfactuals can get convoluted. For the purposes of developing models. Reasoning about expectations. Retrodiction.

What is a plausible relationship between data and a hypothesis. Depends on the paradigm. Handling anomalies. Fig. 9.12

Figure 9.12: Causal model with observations. The observations do not perfectly reflect the actual concepts.

Scientific method. Scientific induction. Induction and abduction (-A.7.1)/ Abduction, for instance about the fossil record. Statistical tests of hypotheses.

Bacon and experiments. Fisher and randomized control.

Volunteers collecting observations and networking the results of those observations. Citizen cyber-science.

Science is a human activity, as such it is susceptible to human frailty. Scientific fraud. Systematic attacks science by publishing and promoting dubious research.

Scientific explanations. Could something happen versus does it explain a given situation. Amount of variance accounted for by an effect

Science is never settled. What does it mean that a model is accepted. How much proof is required. Acceptance of some scientific theories is often a matter of consensus, sometimes by scientific review panels, but in other cases it simply the judgment of individuals. Indeed, the norms of science encourage the minimization of social gamesmanship in the interest of letting data stand on its own (9.2.0). For instance, the use of honorary titles may be discouraged.

Using the Scientific Literature

This is an aspect of scientist information behavior. Science and information systems. Tacit knowledge in science (7.3.4). Collaboration among scientists (9.2.3). Scientist's use of journals (9.2.3). Scholars and scientists are demanding users of information resources. As we have noted, scholarship and science are interactive activities. Many scientists spend hours reading[77]. Keeping up with field. Examples of information use (1.4.3). Science-related information tasks [?]. Current awareness (2.5.5). There are marked differences across disciplines in the patterns of using information.

More broadly, supporting discover.

Bench science. Laboratory life[44]. Collaboration in scientific groups as a community of practice (5.8.2). Workflow in science. Scientific process automation. Virtual laboratories typically provide platforms for access to data sets and tools often associated with those the domain which the data represent.

d

9.2.4. Scientific Collaboration and Communities

Science is a social activity in several senses. This is reflected in multi-author and often, multi-institution, publications. Joint authorship in citations. Co-citation networks may show schools of thought (9.1.2). Scientific communities are a type of scholarly community (9.1.1). Social structure of science research. Distributed research teams. Aspects of science communication[45]. Scientists from various fields differ considerably in their willingness to share data with colleagues.

Scientific Communities

Disciplines. Interdisciplinary interfaces.

Conceptual revolutions are generally changes of conceptual frameworks. There often seems to be a tipping point where many scientists change the conceptual models they apply. Such tipping points have been termed paradigm shifts. However, it remain unclear whether such paradigm shifts are the result of a rational consensus, or generational changes. Indeed, are many cases in which paradigm shifts are led by older scientists.

Scholarship evolves as new ideas are introduced. In some cases, that evolution seems to be the result of a paradigm shift[41]. We can see the transitions in the visualization of citation networks. One way to analyze the patterns is to analyze how scholarship is done. Scientific thinking evolves. Sometimes rather than gradual changes, there are dramatic paradigm shifts[41]. The shift in physics from the Bohr atom to the quantum model. Or the shift of psychology from field theory to computation models of cognition (4.5.2). Bernoulli principle and airplane flight. Thinking in physics changed dramatically from Newtonian determinism with the acceptance of quantum mechanics which is probabilistic. Causes of ulcers.

Scientific models follow trends in information processing[25].

Scientific Communication and Publishing

The primary goal of scientific publishing is the dissemination of findings. Science writing should present a clear description of the claims persuasion but maybe this should be viewed as persuasion by logic.

All scholarly publications have standardized discourse structures, but these are especially pronounced for scientific publications. Discourse in science articles (6.3.2). The most common Genre Template is Introduction, Methods, Results, Discussion (IMRD)[75]. Scientific argumentation versus deduction and hypothesis testing. Discourse in academic lectures. Research genres.

Functional units of language and tasks CARS model for introduction to scientific articles [?]. Applied by the author to persuade the reader. Linking to entities and to models. Module reuse in scientific publishing. Model-oriented scientific research reports[20].

Coordinating vocabularies across disciplines.

Scientist's workbenches. Tools to understand research fronts. Exploratory search.

Collaboratories

Research groups can be tightly interwoven collaborative teams. Indeed, the tacit knowledge of procedures in often difficult to transplant. It is often most efficient for a member of the team to move from to a new laboratory. Exchange of samples, methods, and data among researchers. However, there is also competition.

Teams may collaborate on complex scientific problems. Facilities which support collaboration are sometimes called collaboratories (Fig. 9.13).

Figure 9.13: Astrophysics Simulation Collaboratory Portal.

Citizen Science

Citizen data collection. Crowdsourcing science. Concern with data quality.s

Increased public information and computational resources means that significant data analysis can be done by citizens. Moreover, individuals may also have strong motivation to conduct research (for instance a parent may have a child with a disease). This gets to the question of who is qualified to do science. There are issues of the reliability of data and sophistication of the analyses.

9.3. Measurements and Instrumentation

Recall that in science, it's common to assume that most valid observations are based on our senses. This is not to say that our senses can't be fooled

For science and commerce. Testing laboratories. Certification of testing laboratories.

9.3.1. Measuring Things (Metrology)

Metrology is the study of making measurements. Accurate measurement is important in areas such as business and science. When you buy gasoline or meat at a butcher you rely on the scales.

Scaling (-A.11.2). Accuracy and precision.

Sensors (-A.19.0). Data collection from instruments and calibration of those instruments to ensure accuracy. Measurements are inherently inexact.

Standard weights and measures.

Tracability.

Measurement theory and scaling for social science.

9.3.2. Instruments

Technology has a great effect on the advancement of science. Instruments for basic measurements. Should be consistent across environments. easy to use. Thermometer. Record keeping for calibrations and settings. Social science instruments such as surveys.

Role of networking in coordinating data from different instruments.

The availability of instruments to collectn data in specific ways helps to shape the focus of research and, thus, the nature of knowledge. Highly distributed sensors as a new type of measurment instrument.

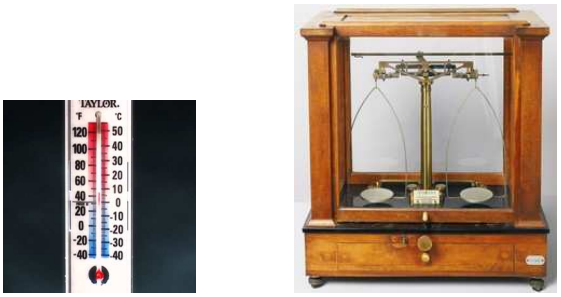

Figure 9.14: Thermometer (left) and beam balance (right). (check permission)

Science often requires measuring unusual properties in unusual conditions So, special instruments need to be developed.

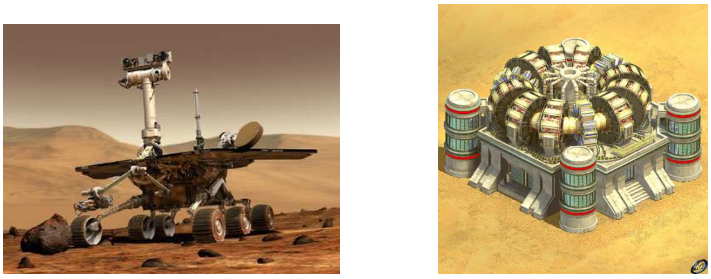

Figure 9.15: Advances in instrumentation often lead to advances in science: Mars rover (left) and supercollider (right). (check permission)

Automating exploration of a large space of parameters.

Sensors as instruments.

9.3.3. Laboratory Notebooks and Integrated Research Data Management

Laboratory notebooks. Notebooks and open notebooks. Tools to help scientists manage their data (9.6.0). Records, audits, and preservation. Ensuring patent right. Laboratory notebooks as part of an ecology of information genres for scientists.

Scientific workflow provenance.

Open science. Open notebook science Fig. **??**. Laboratory notebooks as blogs.

9.4. Science, Technology, and Society

Science and technology interact with society. Among other things, science often studies things which are socially important. Development of socially important knowledge.

9.4.1. Science and Society

Science generally serves social needs. Science is a social activity [28] [45]. Science and technology have transformed society. Boundary between politics and science. Science education (5.11.5). More than a socio-technical system.

Social construction of scientific knowledge.

Within scientific disciplines ad hoc groups may emerge to address more specific issues. These are effectively communities of practice (5.8.2). they emerge, develop structure, assign roles, and use language through stages [**?**]. Medicine.

What social structures lead to the a science-oriented society. The Age of Reason. The Enlightenment.

Broader social value of science and of the other hand, the problem of not having data available. How social values may determine scientific frameworks [?]

Society and social science.

Using unsettled science not only for policy but also to sway public opinion.

Because science can be highly effective it can be persuasive. Indeed, we have many calls for "evidence-based" However, the results of science may take a while to be settled and social science results are often particularly contentious. Moreover, individual scientific results may be grabbed by the media as though the were settled and widely accepted.

9.4.2. Science and Technology Policy

Science and technology are so integral to society that science policy is developed. Encouraging commercialization. Ideally, patents (8.2.2) encourage development of commercially valuable technologies. Incubators. Innovation.

Images of computing technology. Technology and humanity.

Funding for science

Funding agencies.

Report possible conflict of interest which might bais research outcomes. There are a variety of mechanisms for science funding and feedback into societal needs.

There are large differences in funding for different areas of science. Presumably the funding is prioritized to social needs such as eliminating disease. The funding model should require that the results should be independent of the results obtained. Otherwise, there is a bias and science loses its value. Some funding models present challenges to academic freedom.

Scientific debates often go on for years without resolution. However, funders would like evidence that their money is being spent on areas where progress is being made and where it is able to be made. Science metrics (scimetrics) attempts to determine the interaction of science and society. Measures such as bibliometric impact factors (9.1.3) and the number of patents to determine the quality of a research program and thus to set overall funding.

ICTD (8.9.1). Disruptive technologies.

On one hand. science is useful when it helps to resolve policy debates. On the other hand, it can often easily be distorted when caught in the middle of a political discussion.

Political and scientific controversies and what research gets done. Lysenko. There are also many examples of scientists who ignore social pressure.

Science is a social activity [46]. Scientists communicate scientific ideas to the public. Many scientific areas overlap with policy. Increasingly, scientists are being expected to make complex assessments of risk [54].

While many early scientists worked alone with modest funding, some areas of science require substantial funding. There are still many areas in which individuals new ideas and clear thinking can have a big effect, but there are also areas where large infrastructure is required. Ideally, science would be independent of social pressure but that often does not happen. Big science. Large projects with many people involved. IRB.

Social needs for research often implies support by public. Government funding of science both directly

and indirectly. Difficulty of government funding research. However, because government is a fundamentally political, there can be difficulties in government funding science. Science funding by industry. For instance, drug companies may fund evaluation of the drugs they produce.

Research institutions (8.12.3). Pure research versus applied research (Fig. 9.16)[73].

Evidence-based initiatives.

	Useful	
Understanding	**No**	**Yes**
No	-	Edison
Yes	Bohr	Pasteur

Figure 9.16: Types of scientific research with examples of practitioners of each. "Pasteur's quadrant" is applied research[73].

9.4.3. Public Understanding of Science

Public understanding of science (PUS)[15]. Dissemination of scientific results. Society has become increasingly dependent on science but the public is poor at evaluating scientific controversies.

By the traditional values of science, a scientist should shun publicity because it might corrupt the neutral stance toward data. Increasingly, however, scientists are engaging on policy issues and some are actively seeking publicity.

Popularized social blogs. Science is the press. Science journalism. Public outreach by science.

As with other literacies (5.12.2), we would like citizens to know the basics of science and be able to apply basic science models in everyday contexts.

There are many complications in trying to base policy on scientific research that is not settled.

Many controversies are played out in the press. People often believe results presented in the form of science although the quality controls such as peer review may not be followed. Pseudo-Science. The tendency to believe causal chains (4.5.0). Belief in UFOs[62].

Science Controversies and the Public. The scientific community usually eventually reaches consensus about whether to accept evidence. However, there are many scientific controversies which are being evaluated in the court of public opinion. These range from evolution, to the causes of global warming.

Examples of manipulation of science on both sides of the global warming debate.

9.4.4. Engineering, Technology, and Society

Technology is the application of scientific knowledge to practical problems. Technology made dramatic changes in the living conditions in the Industrial Revolution and that rate of change continues with the developments in information technology.

Engineering

Engineering is the application of technologies for soling problems. Technology and engineering are very different from the scientific knowledge generation. Engineering as problem solving. Engineering is, typically, very pragmatic and outcome oriented with an emphasis on cost-minimization. Technology is the application of knowledge to practical needs. Applied science is often combined with engineering (8.12.3). Engineering and design. R&D. Concurrent engineering. Information engineering.

Difficulty of engineering very complex systems.

Large-scale engineering projects such as the design and development of a new appliance or a new building.

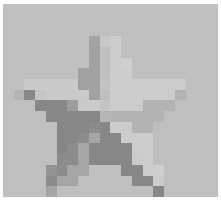

Figure 9.17: The Imperial Hotel in Tokyo survived a major earthquake because it incorporated adequate knowledge.

Figure 9.18: While Interstate highways have brought many benefits to society, they have also created new problems. In some places the highways divided neighborhoods. The "BigDig" tried to undo the effects of a major highway dividing the center of Boston.

Technology

Infrastructure as technology. Developing partial production methods based on scientific findings. Technology has provided society with a remarkable string of accomplishments. Technology is the application of scientific knowledge. Technology management. Technology and design (3.8.0). Technology transfer. Technology and social policies. Institutions and procedures as technologies. Socio-technical systems.

Technology shapes society. It is both agent and recipient of practical action.

Discourse around technologies. "Progress is our most important product".

The relationship between technology and culture. From the perspective of culture, technology is that which disrupts culture and tradition. There are certainly social changes related to technology and, of course, technology does not always improve people lives.

Pasteur's quadrant.

Technology has sometimes been portrayed as leading to an unalloyed benefit notion of progress. While science and technology have clearly allowed the human population to increase and for substantial advances in human welfare such as improved health, it not improved living standards for all people.

Technology as a double-edged sword.

Technology transfer to business.

Polygraph[14]. How much social shaping of technology is there?

History and Economics of Technology Technology and economics (8.8.2). Technological developed is greatly affected by the economic infrastructure. Clearly a system of intellectual property protection and encouraging entrepreneurship facilitates the development of technology. Patents are intended as incentives for technological innovation (8.2.2).

Technology development. Innovation (8.14.0).

9.5. Simulations

Large scale computer-driven simulations are increasingly common. They are found in sciences such as astronomy but also in policy arenas such economics and climate modeling. Indeed, we have encountered models throughout this text (e.g., (1.1.2). Models have representations. Conceptual models (4.4.1).

Simulations as information resources. Simulation documentation rather than metadata.

Here, we focus on the process of complex modeling with computer programs. A model of land use may forecast the extent of urban sprawl over a certain period given certain variables, such as growth rates, predicted economic changes, etc. An individual in the speculative real estate business may find such information useful, should they believe the model to be accurate. Therein lies the difficulty of predictive modeling: we tend to not need predictive models to forecast simple, or easily predictable events; they are most useful for complex situations that, by their very nature, are difficult or impossible to predict, whether or not we have a modeling program designed to do just that. A simulation then, is the output of the runs with one of the models. Business process models (8.11.2). Simulation for the purpose of education (serious games) versus entertainment. Models are difficult to validate across a broad range of cond

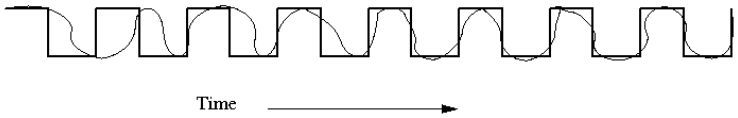

Figure 9.19: An adaptive dynamic model. Imagine a system that re-produces a regular pattern (dark line) (repeat figure)

Simulation is generative reproduction. It acts on a model to generate output.

Simulation as a thought experiment suggesting what is possible.

Simulated interactive environments as related to games.

9.5.1. Types of Modeling and Simulation

Different types of simulations are used for different purposes. Some types that differentiate simulations are shown in Fig. 9.20. Models are particularly useful as a representation for information systems (1.1.2). Modeling attempts to develop an accurate description of complex plans. That model is often used in simulation.

This sort of modeling enables a designer to see the overall picture of a project, and whether or not it satisfies its set of requirements before embarking on the time-consuming task of actually implementing the design. Mathematical models are more abstract examples of regular models, in which mathematical formulas are constructed to represent what sort of outcomes a particular set of input variables will produce (-A.10.0). Modeling for decision support services are particularly complex, as they typically involve many variables in a complex, or probabilistic scenario (3.4.2).

System dynamics (-A.10.2).

Representations. Mathematical models (-A.10.0).

Dimension	Description
interactive / non-interactive	Is there user interaction beyond setting the initial conditions?
deterministic / stochastic events	Does the simulation follow a set course or is it probabilistic?
discrete-event / continuous systems	Does the simulation run from beginning to end or does it pause for input following each event?

Figure 9.20: Some dimensions of simulations.

Finite-element analysis for simulation. Deterministic models. Models that show learning.

Simulations languages and tools. Domain specific.

Non-interactive animations could be implemented as simulations. Fig. 9.21 shows a frame from a computer generated animation in which the fluid motion of the clothing on a dancer is simulated.

Figure 9.21: An animation showing a simulation of clothing and hair moving while the character is dancing[4]. (check permission)

Some stochastic models, such as climate change models, predicting earthquakes, and turbulent flow are recognized to be extremely difficult to simulate (9.5.4). In part, this depends on the resolution of the calculations, and the probabilistic nature of the system being simulated. These systems cannot be "solved," and therefore a simulation is evaluated by how closely it conforms to the analogous process in the real world. These are said to be model-based simulations; that is, they are based on an existing model, though the details and effects of simulations run on that model may be unclear. The model becomes a metaphor for the simulated system.

Simulation-rendering - Alice.

Neural networks (-A.11.4).

Procedural Modeling

Agent-Based Simulations

Computer simulations can be constructed with independent agents. Understanding that many systems to be modeled involve various discrete elements that interact to create an overall effect, agent-based simulations attempt to re-create this scenario by constructing multiple "agents" that act according to pre-defined instructions in response to various stimuli; the net effect of all individual agents produces the desired simulation. These are autonomous agents, and their actions are not coordinated by a central control.

Multiple-agent, aggregate effect simulation can be applied to individual agents as well. (7.7.8).

9.5.2. Practicalities of Simulations

Composable and Multiscale Simulations

Ideally, simulations would be scalable. As for systems biology. Systems dynamics for supporting multiscale simulation. Many difficulties of composable simulations. Creating modular simulations that work together. Because the formatting and language is the same, "composable simulations" allow modules to be reused [?] and save designers time CAD (8.12.3).

Divisible modularity. At different levels, different processes dominate.

Re-use has been emphasized throughout. When models need to be re-used they need to be assembled.

Reuse and interoperability. Interoperability of simulations at different levels of details. Composable simulations can consist of hierarchical components. It is difficult to develop cross functionality between

simulations if the simulations or events to be modeled are of a very different variety.

The evaluation of interactive simulations is more complex.

Parameters and disruptive events outside the model.

Interactive Simulations

Simulations are often a part of interactive virtual environments (11.10.2), and can greatly enhance their plausibility. Simulations, to be effective, often need to be scalable, in the sense of added information. New information — whether it be user actions, updated environmental data, or completely new variables or limits — have an effect on the simulation, and if agents are participating, will have an effect on that outcome. Interactivity adds an external factor – the actions of the user — to simulations.

This interactivity can be added to simulations in two ways: through feedback from environmental sensors, and through user interaction.

Simulations can be produce information artifacts. Another simulation support tool allows the addition of annotations to existing simulations.

9.5.3. Example: Weather and Climate Models

9.5.4. Evaluating Simulations

It is often helpful to have model for complex processes. These produce predictions of complex behavior. Malthus (Fig. 9.22).

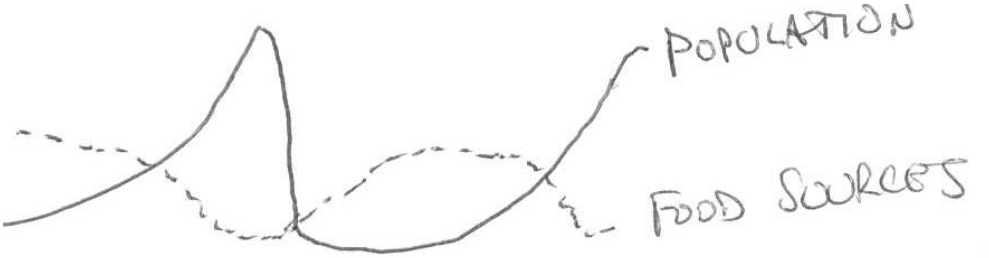

Figure 9.22: Models have been used to make predictions about complex systems such as the use of resources. Malthus predicted that a population of animals would grow geometrically in the presence of adequate food. However, a limited food supply may be exhausted and the population will crash. (redraw)

However, minute changes in data and only minor miscalculations can lead to massive irregularities and inaccuracy when their effects are spread across and entire system. These models predicted the depletion of resources. Many of these predictions have proven to be unreliable. Complex systems such as the economy and the climate are notoriously difficult to predict. One notable example is the Club of Rome report on the effects of over-population[49]. They do not include the full complexity of these very complex systems.

Indeed, a nearly infinite number of models can be developed by adding parameters. Thus, we prefer models with fewer parameters. Free parameters. (-A.10.0). However, we also need to consider the generality of simulations. One strategy to improve accuracy in domain-specific modeling. Over-simplification by incompletely modeling the complexity of the system. Failure to consider critical but hidden factors. There is an ongoing debate about whether climate models incorrectly estimate feedback processes in the same way.

Climate models are often evaluated by testing seeing how well they predict historical climate changes. It is often helpful to couple climate models. That is, to run two different models simultaneously and to combine their predictions. While one model may be getting of track, it can be stabilized by the other model.

Some types of phenomena are famously difficult to model. The Butterfly Effect (9.5.4). illustrates how unpredictable weather can be. Cellular automata. Weather prediction. Time-step predictions.

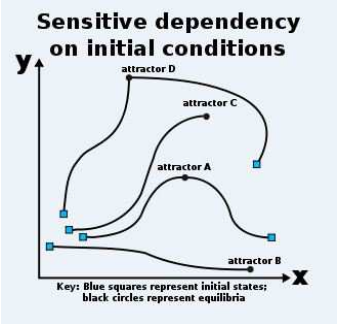

Figure 9.23: Butterfly Effect. (with permission)

Some models are so complex, that it's difficult to analyze these directly. Sensitivity analysis for empirically determining the impact of various parameters. How the output of the model is affects by different values in the inputs to that model. Fig. 9.24. Sensitivity analysis and risk analysis.

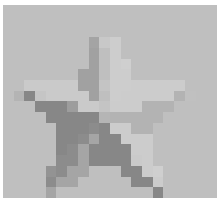

Figure 9.24: Sensitivity analysis.

9.6. Scholarly and Scientific Data Sets

A lot of data is generated by science, but there are also massive data sets from many other sources such as the government, corporations (7.4.1, 7.4.4), and from data such as personal medical records (9.9.0). Big data sets and mass personalization (4.10.3). Beyond data sets created specifically for scholarship, many other data sets are used in scholarship.

This surge in data is generated by a increasing array of sensors which measure attributes such as light, sound, temperature, and even magnetic fields. These are coupled with the development of data centers and networking for processing and communicating the data. A data set is a collection of related data.

Just as we describe APIs as a standard interface for one program to interact with another, we can say there is an API for data bases. If this is one the Web we might call it a Web Service. An API can be considered a type of representation for the component.

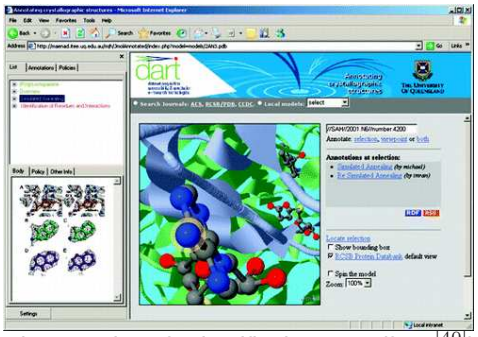

Figure 9.25: Visualization and annotation of scientific data sets (from[40]). (check permission)

Primary data. Is it reproducible or not.

Big data. Could also include big metadata.

Standardized metadata systems. Communities which define their metadata standards. Earthcube.

Data data model: collection, deliverable unit, manifestation, file. For data, preservation can be defined as the ability to be reuse.

Moreover, data sets may be linked into larger scale data-sets. Tools to support data analysis. Visual analytics.

High-performance information processing. Solving certain large scale problems.

9.6.1. Collections of Data

A data set in set of observations. It is similar to a work (2.4.3). Not only do they have a consistent structure but they are managed much like library content. Should we trust the data in the dataset?

Interoperability among different data sets. Difficulty of interoperability also has policy and social origins.

Live data sets may keep changing. Difficulties of dynamic records. Do we have to keep all changes and be able to reproduce the exact data set on which a given calculation was completed.

Data management institutions: FBI and fingerprints[6].

Figure 9.26: Fingerprint data set.

Data grid (7.7.2). Infrastructure for grid.

Figure 9.27: UN data.

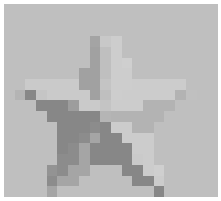

Figure 9.28: Sloan Sky Survey.

Gigaword corpus, Google Books. Social media data sets.

Data sets for decision making versus data sets for science.

eScience has an emphasis on large collections of data. Linking data sets to literature. We considered the conceptual foundations of science earlier in the chapter. Here, we examine data sets which are collected for science. Separation from hypothesis testing.

Standard data sets. Data for future reference - just in case it's interesting or because it shows an effect we know to be of interest. Reference works (3.3.2). New access for data sets such as the Library of Life

[8]. Data sets which are high quality may be cited by scientists to help validate a claim based on them.

Coordination and preservation of complex data sets.

Many sciences are becoming highly data driven. Collect a lot of data and then analyze it. Record keeping (7.4.1).

44000 terabytes of NOAA data.

Materiality.

Reference collections (7.6.2). Data authors as a type of information professional (5.12.4).

Scientists annotation of their own data sets.

Scientific publishing and data sets. Active publishing. For example, highly fragmented least-publishable units. Versioning.

Observing scientists and their use of metadata. It is not clear who will be doing this work.

Mashups for coordinating data from different data sets. This can be viewed as a type of data fusion.

Entity authority file for names of objects in a database.

Data sets in the real world are often messy.

Data cleaning. Data releases. FRBR and data sets.

Requires data specialists and extensive domain knowledge.

Annotation of data sets. Annotation with SWAN.

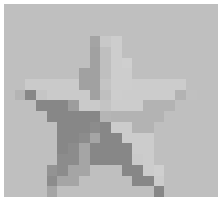

Figure 9.29: SWAN annotation ontology. (check permission)

Scientific Data Sets

Collecting some data without an immediate reason. Extensive Data sets have been fundamental to discoveries such as Darwin's theory of evolution. Instrumentation which obtains a lot of data needs to be followed my systems for the management of that data. Tyco Brahe was able to develop predictive equations for planetary motion by using the data tables developed by Kepler (Fig. 9.30).

Management of scientific data. Record keeping is essential for science.

Sensors and the ability to collect massive data sets makes systematic hypothesis testing less important and it fundamentally changes the nature of science. Still, the key is the interpretation of data rather than the collection of data.

Data retrieval from large datasets can be available on the grid. Mining data for research. Science and information systems are becoming more inter-twined. Open notebook science for confirming the procedure used to collect the data sets. Shared data which can be useful to everyone. History of how a data set has been processed. Are two data sets related?

FRBR for data sets.

Figure 9.30: Kepler (left) collected extensive data tables about the motion of the planets in the sky which Tyco Brahe (right) used to formulate equations. (check permission)

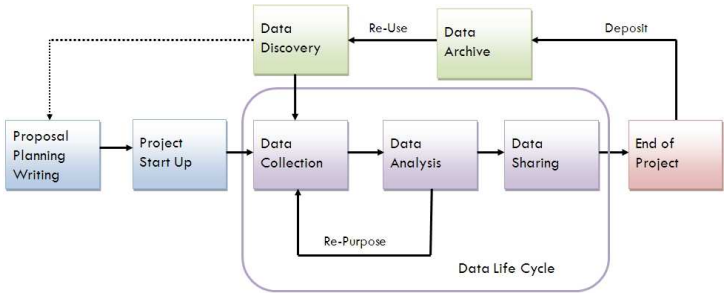

Figure 9.31: Research data lifecycle. (redraw) (check permission)

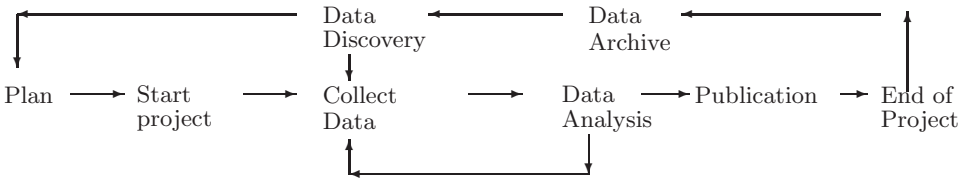

Figure 9.32: Research data lifecycle. (vs OAIS) Increasingly, also need to support publishing and reuse. (check permission)

Baseline collections such as large sets of shells which in science museums against which to gauge ecological changes.

Interoperability of data sets.

Automated filtering of data sets for low probability events.

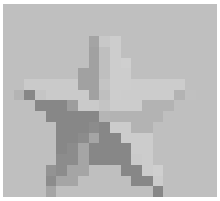

Figure 9.33: O-Ring failures. Challenger accident.

9.6.2. Applied Data Sets

Earth monitoring. Medical data sets.

9.6.3. Frameworks for Managing Data Sets and Data Curation

There are many similarities between the management of text libraries and data libraries. Data libraries are related to data warehouses (7.4.4). Data archiving. Linking to publications, DRYAD. Analogous to

the approaches for archives. Data management planning. Coordinating and reconciling metadata and field names.

Just as policy was critical for libraries and archives, policy is also central for frameworks for data management.

The collection and management of data sets such as census, economic statistics, labor statistics, geographic data, and sometimes scientific data is a significant government function.

At the least, the data should be machine processable. This often means it is structured with XML.

Data Sets

Data set as a work.

What makes a dataset citable. Corpus linguistics.

Errors in Data Sets Outliers. Missing cells. Limits to resolution in measurement accuracy. Data cleaning.

Human-to-human communication around data. Storytelling about data. Data communities analogous to document communities (5.8.2).

Data Curation

Scientific data use often doesn't match the relational model.

Data reuse.

Trusted data. Data provenance[65].

The intermediate data products may be captured.

Difficulties of data analyses across different levels of a model. Such multilevel models are common in complex systems.

Many information systems need to manage large amounts of numerical data. There's a wide range of data sets. Data sets need to be managed as though they are libraries. Data curation is still fundamental. Coordination across data sets.

Data custodians and data managers.

Similar to many of the issues for libraries and archives (7.5.1). Trusted repositories. Risk analysis for content. Provides a cost-benefit advantage.

Data curation and data management[53]. Collection management issues are like those for digital libraries. Quality of the data. Identify new data used in an analysis. Should partial results be kept. In laboratory research, there is a lot of control over the quality of the data.

Curating data sets is similar in some ways to collection management (7.2.2). This includes consistent organizational commitment and continuity. Audit procedures (7.10.2). Security, information assurance, history of access and updates. Audit trails for determining the provenance and quality of data. tracking any changes that are made.

Narrative journalism.

Some of this is specific to science datasets but other aspects apply to all large datasets. Messy natural data sets. (3.9.2). In the sense used here, the curator is like a steward.

Semantic Web and ontologies are widely used for description of data sets. (2.2.2).

Representation information

Data Management Systems and Services

Managing petabytes (7.7.3) and scale. Some datasets are so large the the cannot be run across distributed networks. current data networks cannot handle them. Large scale data flows[21] (Fig. 9.34). We need systems for accessing large data sets across the network. The Data Grid combines features of libraries, archives, and knowledge management systems. As we saw for those applications, they can be described by the services they offer. (7.0.0). Federated data. XML format sports data. Many of the issues we discussed for cultural archives (7.5.1). also apply for preservation of data sets. Sneaker net. Moving data vs. moving computation to the data. Science DMZ.

Data

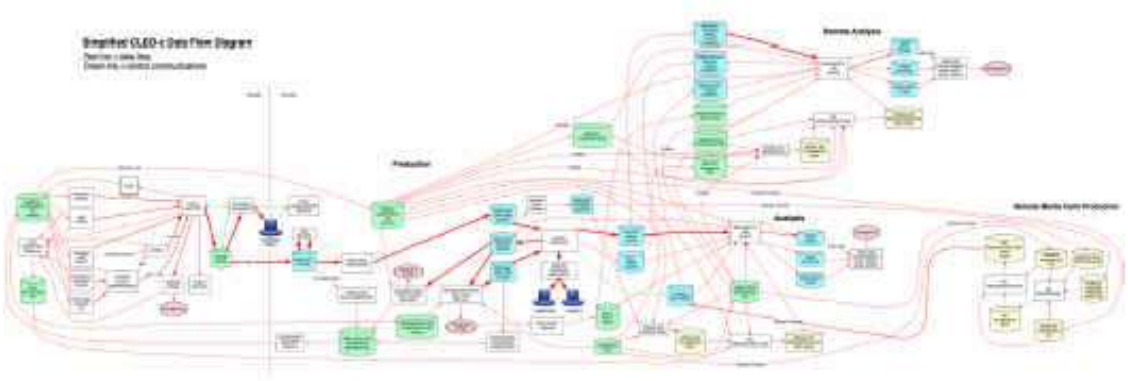

Figure 9.34: Here is a data management map for CLEO which is a high-energy physics project based at Cornell. It includes hundreds of people and instruments. One of the problems of distributed data management is maintaining consistency when doing updates[21]. A workflow management system can be used to keep track of complex processes.

Derived data products.

Representing and reasoning about provenance. Grid transactions. Events in the user interface. Process documentation. p-query.

Federation of data grids. Coordination of policies for data grids. Data grid chaining.

Difficulty of moving large amounts of data around the data gird. Collecting data from instruments in remote locations. Data centers for scientific data.

Merging results from very different fields requires both coordination of data sets and also ultimately, coordination of the organizations which produce the data and analyses. Cyberinfrastructure for science. Implementing policies for science repositories. For instance, this might include IRB approval[51].

Data synchronization.

Describing Data Sets

Metadata and Semantic Web for describing the content of scientific data sets.

Metadata for statistical data. Standard descriptors for similar concepts across implementations.

Interoperability.

Classification of statistical data sets.

Common information model.

Gene ontology, (2.2.2). In some cases, some or all metadata about a digital object may be lost.

Interconnecting data sets with workflows.

This is particularly important since we need to know exactly what is in the data sets.

Machine-processable metadata.

Self-describing datasets.

Data format definition language.

Metadata for data libraries. Potentially a very large amount of metadata. Effects of scientific data sets on the way science gets done. Scientific metadata evolves.

Linking datasets and publications.

Data Preservation

Large collections of data may be maintained as data libraries. S.O.A.P. (Selection, Organization, Access Persistence) principles should also apply here. Each domain will have its own specific data structures but the background principles should be the same.

Loss of data sets can be very expensive and can set back research. It is expensive and sometimes impossible to recreate scientific data sets. Collecting and preservation data which might be useful sometime in the future. It is especially worth preserving scientific data when that data is unique or when it is difficult to reconstruct. Scientific datasets allow the user to go back to old data to see what has been recorded and what it shows. As with other types of digital preservation, formats and metadata are major issues (7.5.5). Preservation institutions and infrastructure. Importance of keeping data in the original format (Fig. 9.33).

Some data sets about phenomena such as records of earthquakes and volcanoes are not reproducible. Particular preservation of at risk data or data which would be particularly difficult to replace.

Trusted data. Data provenance[30].

Economic models for data collection and storage.

Figure 9.35: Frame from a reconstructed NASA video of the first step on the moon. The original was lost. (repeat)

Preservation data requires that the protocols used to generate the data are described. Importance of guaranteeing data quality.

Importance of information needs for data libraries. For instance, the astronomers need to be able to compare observations across time.

Federated data. Repositories of government data.

Structured browsing of the content of data libraries. For instance, a collection of data about ship wrecks could include data sets organized by the components of a ship.

. Data documentation initiative. Recent versions include details about the lifecycle and workflow. Lifecycle management for datasets. Archiving data sets. Data preservation. Sets of related materials for a certain objective. Text documents, images, data. Unlike some archives, only the values in the data set are important, the original appearance and format are generally not important. Data liefcycle issues in DDI-3.

Some data sets such as those generated for relational databases (3.9.0). Data cleaning.

Preserving Workflows and Contexts

9.6.4. Data Rights and Policies

Policies

As with other collections of information resources, there need to be both legal and institutional policies governing their use. One claim is that data from research which is funded by government grants should generally be made publicly available. Indeed, increasingly this is part of the terms of awarding a grant. Embargo for use of data collected by one researcher. We have considered the basics of records for commerce (7.4.1) but that is often focused on satisfying legal obligations. Here we consider scientific and scholarly records where there would be different types of issues. For instance, sharing data across a research community is often a major consideration.

Recognition for the creators of data sets. Issues around open data sets. Confidentiality in some social science data. Open data sets.‘ Open-source Project for a Network Data Access Protocol standard interfaces for visualization software. There's a cost to not having open data. Civic data.

Data cannot be copyrighted; it describes facts. Rights and responsibilities for data management. Language for specifying constraints on the use of data.

Beyond retaining data from individual experiments. Science is an accumulation interlocking results. Data progresses from collections of measurements, to edited databases and handbooks, to derived physical constants. The periodic table effectively summarizes a large amount of data from many observations. It is a very effective representation.

Aggregation of records from many sources. Re-identification of records in supposedly anonymized data sets.

Data Privacy. Corporate records. Security for managing privacy. Anonymize data to facilitate privacy.

Substantial privacy issues for some data sets.

Data Sharing and Open Data

Shared raw data sets (9.6.3) can supply more information about publications, especially when linking is provided between data sets and journal literature. Norms across disciplines for data sharing. Cost and business models for data sharing. Time-embargoed on data. Facilitating interaction among researchers. Open microscopy environment[12]. Tool sets and desktops (3.5.4). Data sharing and scientific collaboration (5.3.3). This is similar to the unwillingness of sharing data in organizations which leads to silos (7.3.6). Perhaps even mandatory posting of data. Collecting and managing the collection. Who controls data? Open data.

Policies to encourage access especially of government data. There could be considerable value in making data available. Data practices are human routines organized around data. Sometimes there is rapid sharing or data and in other cases, the data is hoarded. However, there are substantial differences in disciplines about this. Not just open data but follow the open data by open analysis.

9.6.5. Data Mining, Data Visualization, and Visual Analytics

The activity of science consists of a set of systematic workflows. Science data flow models. Design of laboratory notebooks should be based on scientists use. The model becomes an artifact showing the dataflow used in a given analysis and can be managed like other information resources. Indeed, the reliability of laboratory notes needs to be comparable to other trusted archival materials. Flow of the routine data analysis. Workflow. Providing an index of content, a friendly interface, and security. Workflow for the analysis of data. For instance, Kepler workflow (Fig. 9.36).

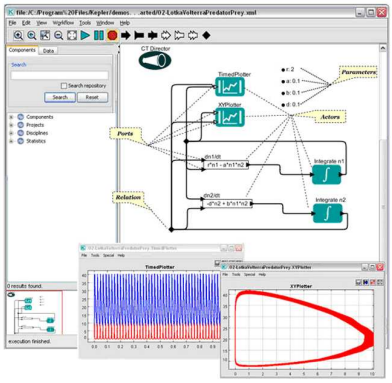

Figure 9.36: The Kepler data management system has a GUI dataflow for science and can provide provenance for derived data. (Workflow only. check permission)

Fig. 9.37[16].

Figure 9.37: Many-Eyes community data analysis. (blogging based around visualization tools)

Complex decision support (3.4.2).

Exploratory data analysis.

Large data sets can be extremely rich sources of information. Thus, their exploration and extraction of information from them is actively pursued.

Determining semantic structure of data sets. For instance, "data detectors".

Information visualization (11.2.5) emphasizes quantitative relationships in data. However, we can also consider displays of quantitative, numerical relationships. Because this is often associated with scientific research, this is often called "scientific visualization". This often allows for controlling perspective and coordinates. Data may also be presented with animated visualization. For instance, particle flurries identify on a sample of exemplary particles. The motion of typical particles can illustrate complex behavior such as flow through a blood vessel (Fig. 9.38). Data analysis environments versus workflows. These could, for instance, support decision making with DSSs (3.4.2).

Data analysis from secured data sets. For instance, maintain privacy.

Data analysis environments and workflow. Limitations of visualization.

Data behaviors as a type of information behavior. Query previews (3.9.2).

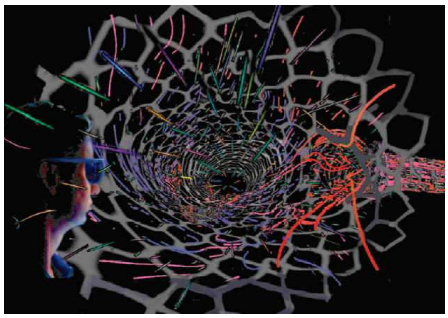

Figure 9.38: The flow of red and white blood cells can be illustrated with animations of particle flurries[13]. (check permission)

There are many possible applications: Visualization and statistical analyses. Clustering. Data warehousing. Information fusion (9.6.5). Visual analytics uses visualization to support analyses. Visualization data in a schematic of the environment from which the data was drawn.

Collaborative visualization. A visualization can serve as a conversational artifact. Virtual observatories.

However, there is mixed evidence about whether visualization is actually superior to other interfaces for information presentation.

Information fusion is combining evidence from many sources. For instance, from sensors, reports by humans, and historical knowledge.

Data mining. Privacy preserving data mining.

Data narratives.

Sensors in the name of science. Changes the way people do science and greater value in the analysis than in the collection of data. Need to reward collection.

9.7. Mathematics, Statisitics, and Logic

Although mathematics is generally paired with science, it is quite different. It can be quite abstract and in that sense, it seems far removed from typical information systems. The criterion for acceptance of a mathematical principle is consistency with the rest of the system of expressions. Mathematics is useful for science because mathematics facilitates the abstraction of some types of processes. Math can be thought of as a collection of languages. Mathematical knowledge and axioms[42]. Working with math[43]. Godel[36]. Teaching math (5.11.5).

Numeracy. Cognition and numbers[31]. Reasoning about probability (4.3.4). Clearly, math needs to be broadly defined to include statistics and probabilities.

Logic (-A.7.0)

9.7.1. Mathematics and Logic as Frameworks for Representations

Mathematics has been described as exploring the "relationship of relationships" [55]. Abstract procedures. State models and discrete math. The nature of mathematical proof.

Math as abstraction. Branches of mathematics: Topology, Graph Theory, Algebras.

Library of mathematics[29]. Coherency of mathematics. Proof development systems. Mathematical proof and certainty of knowledge.

Geometric objects as ideal forms.

Collaborative mathematical environments.

9.7.2. Structure of Mathematical Expressions and Mathematics Markup

Languages of mathematics. Geometry as a visual language. Equations in text such as $(a+b)^2$ can be presented with an extension of XML known as MathML[10]. As shown in Fig. 9.39, the format can be separated from semantics. Parsing equations.

```
<msup>                  <apply>
  <mfenced>               <power/>
    <mi>a</mi>            <apply>
    <mo>+</mo>              <plus/>
    <mi>b</mi>              <ci>a</ci>
  </mfenced>                <ci>b</ci>
  <mn>2</mn>              </apply>
</msup>                   <cn>2</cn>
                        </apply>
```

Figure 9.39: MathML presentation format specification (left) and content format (right) for $(a+b)^2$.

9.7.3. Mathematical Argumentation

9.7.4. Automated Theorem Proving

Logical operations.

Proof verification.

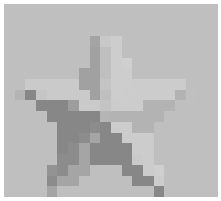

Figure 9.40: Automated theorem proving.

9.7.5. Other Math Processing Tasks

Structured math documents such as math practice sheets.

Math searching[11]. Can I get the derivative of $\frac{1}{e^x}$ Math descriptive metadata.

9.7.6. Statistics

Exploratory Data Analysis.

9.8. Science and Informatics

9.8.1. Biology as an Information Science

"...modern biology has evolved into a science of information," – David Baltimore

There are several senses in which biology is an information science. Taxonomic classification has been an important theme for many years. Taxonomy based on evolution rather than, say, on structure.

There is a greater awareness of DNA as a representation. Biological systems have representations which model they world. Their representations are biological processes. While DNA represents the information necessary for making life is clear, the processes by which that representation gets expressed are also integral. Information polymers: DNA, RNA, proteins, and carbohydrates.

Biologically inspired systems. Genetic algorithms (-A.11.6).

Issues of representation are essential for biology. Bioinformatics. How do organisms grow and reproduce themselves. Partly the DNA but the entire context. Data management of biological data. Encyclopedia of life. iPlant. Brain science (-A.12.2). Spore is a game which simulates biological processes (Fig. 9.41). Intellectual property issues for genetic modifications (8.2.2).

Figure 9.41: Natural computing. (check permission)

Naming and Classifying Living Things

Since the time of the great Swedish taxonomist, Linnaeus, taxonomies and classification have been fundamental to biological [70]. Categorization and classification (2.1.2). As with other classification systems, there are problems with the definition of the attributes on which the classification is based. In the case of biological classification one question is what exactly is a species [82].

Biological taxonomies. Versus, say, agricultural taxonomies.

Originally based on physical structures and now similarity of DNA and, for that, models of likely evolution.

SuperTrees [9] (Fig. 9.42).

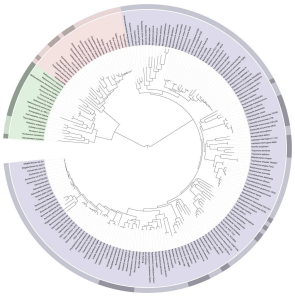

Figure 9.42: SuperTrees are cluster diagrams which try to link a very large number species[9].

Genomics and Proteomics

DNA carries the information essential for life. In effect, DNA is the representation of the organism though complex processes are required for expression of that representation.

Genome sequencing. Human DNA is 99.9% similar; 0.1% difference.

Diseases determined by mutations. Analyzing the function of gene sequences. HMMs for gene sequences.

Gene ontology. Dimensions of gene activity. Function, Gene annotation and entities [47]. This is an example of a collaborative information resource. Evidence codes. Re-sequencing.

Interestingly, the sequence of codes in DNA chains and proteins can be examined with some of the same techniques that we use for text retrieval. For instance, we may want to match a fragment of DNA

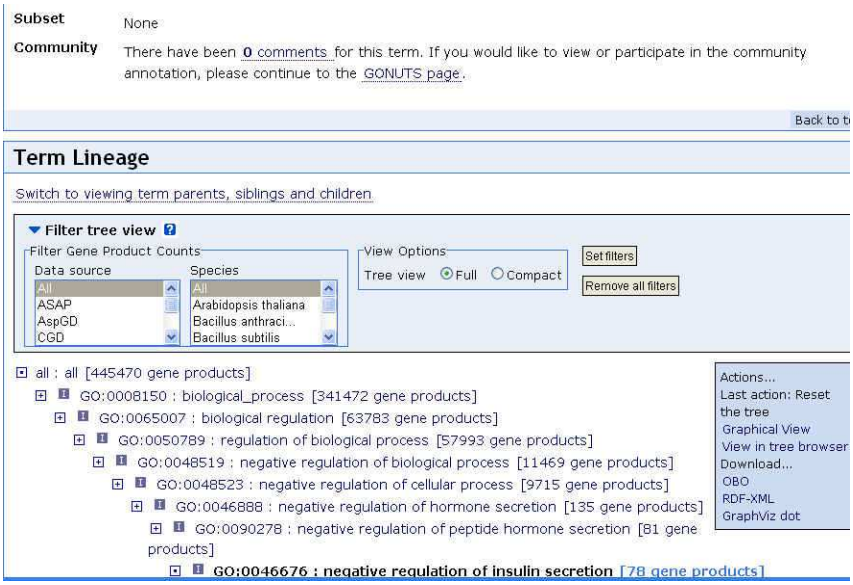

Figure 9.43: Gene ontology. (repeat figure)

to find the sequence where it belongs. Also, uses for proteins analysis. From DNA sequences to the 3-D structure of a protein is difficult. Protein folding.

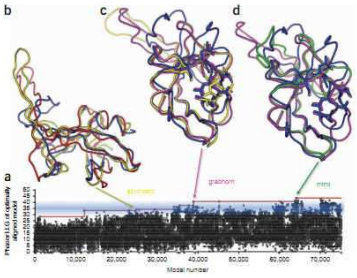

Figure 9.44: Fold-it is a crowd-sourced puzzle for solving the complexitiies of protein folding (from [?]) (check permission)

Searching Sequences with BLAST (Basic Local Alignment Search Tool) search aligns sequences. This makes DNA sequencing possible. Matching amino acid sequences in proteins.

MKLLQRGVALALLTTFTKASETALA

Figure 9.45: Fragment of an amino acid sequence in a protein.

Protein shape searching in drug development.

Translational medicine - from science to clinical impact.

Patents may be given, for genetically, modified crops.

Control of Gene Expression Design control in biological systems.

Personalized medical treatment by understanding individual gene sequences.

Genetic testing and personal genomics. Personal DNA screening. Genetic testing is a predictor of a

person's health. It can lead to personalized medicines, health insurance, Personal knowledge about genetic conditions. There can be difficult decisions about having children with knowledge of genetic conditions. These tests raise ethical and privacy issues[5] Large data sets of personal genetic information are needed for research but are prone to privacy problems. (8.3.1).

Models of Biological Processes and Learning in Biological Systems

Systems biology explores how the components work together. Organisms can be considered as the complex combination of many interacting components. We can consider them as systems. Thus, some of the same tools we use for modeling information systems may be applied to modeling biological systems. For instance, biological systems may be modeled with UML (3.10.2)[33] (Fig. 9.46). Furthermore, we can use visualization to explore these models. This model could then be used to do the drug personalization. Semantic systems biology.

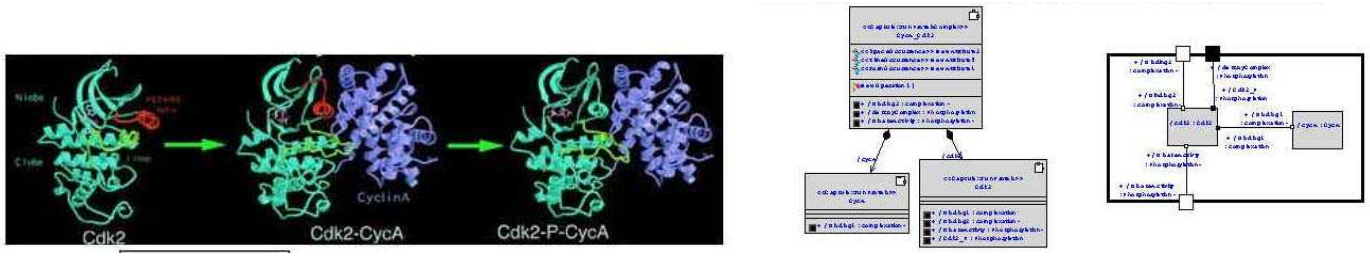

Figure 9.46: UML for systems biology. The transitions between molecules in a chemical pathway (left) can be modeled as states in a statechart (right). Specifically, the pathways of cyclin-dependent Cdk2 kinase are shown [?]. (check permission)

Evolutionary game theory.

Simulations and systems biology. In some cases, UML (3.10.2) is used for these simulations.

Learning is an aspect of information systems and we see learning in biological systems. The primary representation is DNA. Survival of the fittest. The species is an adaptive unit. Evolution of altruism. Evolution and game theory[56]. Cost-benefit ratio times the number of links.

$$B/C * K \tag{9.2}$$

Genetic algorithms (-A.11.6) are a type of machine learning. Natural Computing. Evolutionary systems.

9.8.2. Cheminformatics

Just as classification has been a cornerstone of biology, it has also been central to chemistry. The development of the periodical table of the elements was a major advance. Increasingly coordinated with biological informatics.

9.8.3. Other Sciences

Geology. Astronomy. Economics. Archaeology as a spatial scientific and cultural science.

9.9. Medical, Clinical, and Health-Care Informatics

Medicine and health-care are information and inference intensive. These issues range from principles of medicine, to the management of the entire health system, to the clinical treatment of individuals and patient involvement in decision making and Medicine employs many of the information management techniques we have explored. Many of these points echo discussions we have had elsewhere. Health care is information intensive. Thus, there are issues related to decision support, semantics, and data management.

Medicine would seem to be an area where progress is hard to argue with. Life spans have been increasing steadily. Translational medicine attempts technology transfer for medical discoveries. That is, it attempts to facilitate moving good ideas from the laboratory to practice.

9.9.1. Medical Concepts, Research, Scholarship, and Teaching

Scholarly publishing (9.1.1) for medicine. Medical scholarship. PubMed. Disease database. Low recall for physician information seeking[24].

Automatic monitoring of patients with semi-intelligent devices.

Open-Source Medical Research.

Over the years, many different medical description languages have been developed. Recently, many of the components of those languages have been unified into a single system, the Unified Medical Language System (UMLS). UMLS has attempted to create a umbrella which encompasses the different domains of medicine but it is unclear if it accomplishes that. Ontologies (2.2.2). UMLS is composed of three main parts: a semantic network, a lexicon, and a thesaurus, which combine knowledge representation with subject classification systems. From these components, UMLS has developed a framework describing the medical literature based on the concepts it contains. While medical publications are relatively formal, as with any natural-language descriptions, the concepts may be described with several different words. Indeed, each of those lexical units uses different string units. Fig. 9.47 illustrates the relationship of some of these concepts. UMLS has many applications; for instance, it is a type of query reformulation (10.7.2) with different terms. This is a relatively well-defined ontology but as we have noted for other attempts to exactly specify concepts (1.1.4) and to develop controlled vocabularies (2.5.3), here are still many vague terms.

Medical Searching

Use of medical literature (9.9.2). Medical question answering. The patients other have a lot of anxiety about their medical problems. Medical searching search engine as a specialized search. Evidence-based medicine (9.9.2). Clinical outcomes research. Characteristics of different searches.

Semantic types in UMLS. Text processing of medical content.

Difficulties of consistency in UMLS [**?**].

String (SUI)	Lexical (LUI)	Concept (CUI)
S0016668 Atrial fibrillation	L0004238 Atrial fibrillation Artrial fibrillations	C0004238 Atrial fibrillation Artrial fibrillations
S0016669 Artrial fibrillations		
S0016899 Auricular fibrillations	L0004327 Auricular fibrillation Auricular fibrillations	Auricular fibrillation Auricular fibrillation
S0016900 Auricular fibrillations		

Figure 9.47: There are different levels of description. The Unified Medical Language System (UMLS) defines String variants (SUI), Lexical variants (LUI), and Conceptual Units (CUI). One concept, (C0004238), is expressed with several terms (L0004238 and L0004237) and each of those terms has several strings (S0016668, S0016669, S0016899, S0016900)[39]. (check permission)

Visualizations such as particle flurries (11.2.5). Augmented realities.

X-ray image processing. Indexing cell lines. Organizing information to protect meat supplies.

Medical training and medical simulation. Fig. 9.48 Humanoid robots may also be used for medical simulations.

Figure 9.48: An example of simulation for medical education. (check permission)

9.9.2. Clinical Health-Care Informatics

Information systems are being applied in medicine in many ways.

Decompose the complex activities into workflows. Many types of information are needed to make effective medical decisions. Information about the patients' condition and prior medical interventions. However, this is often not available.

Small search tasks in everyday activities. Select from many alternatives. Many reports.

Hospital. Doctors rarely have all of the information they need about patients when interviewing them. Major information management constraints. Siloed information organization. Supporting diagnosis and decision making. Supporting collaboration in medical settings.

Surgical workflows.

Safety and error prevention.

Care assistant.

Medical Diagnosis and Decision Making

Diagnosis is part classification and part causal modeling (4.4.5). Doctors[7]. Medical diagnosis can be particularly challenging to because of the complexity of the human body. There many different ways that medical symptoms can manifest themselves and there is the possibility that there can be multiple diseases.

Making decisions about individual cases. Medical inference systems. Diagnosis and treatment decisions about individuals.

Evaluation with positive predictive value which is similar to the notion of precision for evaluating search performance. Fig. 9.49

		Actual Condition		
		Yes	**No**	**Total**
Diagnosis	**Yes**	10	10	20
	No	20	60	80
	Total	30	70	100

Figure 9.49: The positive predictive value is the value of a diagnostic test.

Nosology is the classification of diseases. Especially contentious with respect to concepts related to mental health.

LOOK-DECIDE-DO There are many devices for information delivery such as PDAs. Consider the

information needs of health professionals in an emergency room. Complex data and a stressful environment.

Medical decision support systems for making medical decisions. Clinical decision support systems. As with other decision support systems, the data presentation should be easy to understand. Computer-supported diagnosis helps to minimize errors from cognitive load. Medical diagnosis as cultural phenomenon. Low cost wide-spready medical testing. Diagnosis for all. Networked communication to support simple medical diagnostics. ICTD. Best practices for health care in a crisis.

Managing treatment. However, devices may affect physician workflow and this becomes an obstacle to their acceptance. Current versions often fail to take contextual factors into consideration.

Evidence in Medicine and Evidence-Based Medicine

There is often a substantial gap between research findings and applications of medical treatments. Medical treatments should be based on the latest finding but that often doesn't happen; that is, we should have Evidence-Based Medicine (EBM). There are several reasons for this gap. Doctors don't know all the latest research in every field and even if they do research on those finding, the research results are not readily accessible. Speed the time from literature to practice, a type of technology transfer. Making the literature clearer with narrative and systematic reviews. Also with structured abstracts which highlight dimensions which are particularly relevant to medical treatments. Rely less on human judgment which is often fallible.

Stepping further back, we can ask what is evidence in medicine. There is a lot of medical research but there also seem to be a lot of contradictory results. It is relatively easy to obtain correlational data but it is harder to do experiments. Randomized clinical trials (RCTs) provide stronger evidence than correlational studies. Generates large data sets.

Designing Medical Work Environments

Collaboration of a medical team in the trauma center. Social structure and communication of medical delivery teams.

Information in Health-Care Delivery

Information is essential in the health-care delivery environment. Hospitals are at once highly complex, highly traditional, and data-rich. Hospitals are among information institutions invented. Physician-order entry systems – for managing treatment. Management of costs. Effective outcomes. RFID. Sensor monitoring of health-care procedures and treatments. Information for preventative medicine.

Health gaming.

Medical images (e.g., Fig. 9.50) are very different from pictures of your vacation. Radiological images are often difficult to decipher And may require pattern recognition approaches such neural networks have been used for detection of tumors. Medical image modeling. Computational anatomy[3].

Teleradiology. Collections of digital images will have high space and network requirements.

Patient Information Behavior

Self-diagnosis. Hypercondira. Cyber-condria. Consumer health vocabularies.

Issues for decision making. It is often difficult for a patient to understand all the implications of various treatment options. In some cases, it may be easier to have a physician or social worker talk through the choices and provide advice. Shared decision making (SDM) between patient and physician. Like other decisions, this will include information collection. Patient decision making.

Problem of patient conclusions. Quackery. Laetrile.

Online patient support group.

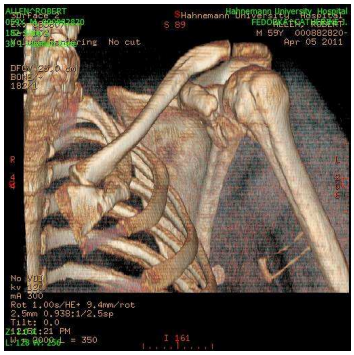

Figure 9.50: CT scan image.

Increasingly, patients are accessing online information resources to try understand their own medical conditions. Consumer health information. How citizens acquire medical information[22].

Information for Patients

Doctor-patient interaction (6.4.3). Patients will use colloquial language and may attend to unimportant details about their medical condition. Patient's descriptions of their symptoms and complaints. Telling the story of your medical history.

One way of structuring information about a patient is with "PICO" questions. These are: Patient Problem, Intervention, Comparison, Outcome.

So called, information prescriptions ask the patient to explore the implications of a condition or range of treatment options. Making health materials easy for patients to understand. Personalized health care benefits from extensive personal health records.

Tools to support doctor's communicating of information to patients. Information for hospital patients (Fig. 9.51).

Figure 9.51: Prototype display of patient notices[52]. (check permission)

Decisions about Treatments

Like other types of medical decision making there must be estimates of risks and risk management (7.10.3). Patient decisions about treatment options, about lifestyle choices.

Social networks and the diffusion of medical information (5.1.3). Social links seem to affect medically significant behavior.

Supporting Physican Activities

Coordination in hospital team work. Complications of a continually changing situation. Support for patient health care management. Management of chronic diseases.

9.9.3. Medical Records

Hospitals have a very complex organizational structure. There are many formalized roles. Encouraging transparency about health-care costs. We have already seen a wide range of information system interaction in institutional dialogs (6.4.3). These are socio-technical system. Related to other electronic records (7.4.1).

Personal and Patient Health Records

The potentially flexibility and accuracy of medical records. Improving health-care with information technology. Potentially, there would be a great deal of benefit from being able to access complete records about a patient. This is a special case of the management of electronic records (7.4.1). Keeping patient records have the potential to improve the quality of health-care delivery. However, the massive conversion costs are substantial. The requirements are difficult such as handling insurance claims. Universal electronic patient records. Many solutions have been proposed. Perhaps simple records are better than overly complex records. A more limited version is the records we find in an organization. Patient records in a hospital. Health insurance forms. Google healthVault. Personal recording of information. Paper chart for traditional medical records. Conversion of paper-based clinical medical records.

Electronic health records (EHR) and personal health records (PHR). Health Level 7 is an XML-based standard for exchange of health information. There are many different layers of records which could be incorporated such as lab reports, prescriptions, treatment records. Conversion of paper records to electronic records. Medical data codes involve classification of medical conditions (2.1.2)which can be useful for public health.

Beyond patients: Staff and stuff. Logistics (8.12.1). Equipment, business end, insurance/billing. Electronic Medical Records (EMR). Usability of health records. Adoption of health information systems by hospitals. EMR data centers and hospital information systems.

There are many difficulties in implementing them. It is important for people to have records of their medical condition. Doctors information needs are relatively well served by existing approaches. Medical privacy issues are often an extreme concern (8.3.1). However, there are many different types of information and strong requirements for privacy that developing a standard has proven extremely difficult. Broader ethical questions such as releasing information about people's genetic dispositions. There are also, many complications to allowing medical records to be disseminated. This includes establishing policies and workflows for managing the data. Medical information policy. Such as HIPAA privacy policy. Implementing HIPAA policies. These records have severe privacy issues. Difficulties of data mining the results given the privacy constraints. Legal issues for records, not just privacy. Liability and medical records. Security breaches. Ease of access.

Mining the data has the potential to greatly improve health. What's more, combing records of family members, or by census data has the potential to detect important health trends. Social determinants of health. Social networks and risky health activities.

Standardized indexing and organizing information are crucial to the effective records. Controlled vocabularies for specification of patient information. Use of UMLS and patient health records (PHR). While medical concepts can be relatively cleanly defined in the research literature, it is far more difficult to do that in the everyday diagnosis.

Letting people view and correct their own health records.

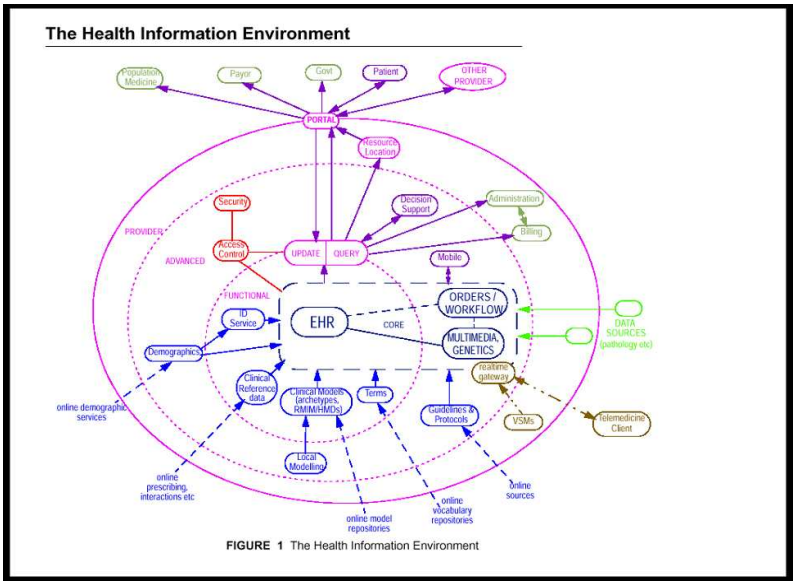

Figure 9.52: HL7 health information environment which is centered around the Electronic Health Record (EHR)[1]. (check permission) (simplify) (redraw)

9.9.4. Public Health Information and Public Awareness of Health Information

Analysis of population trends and precautions. How do people adjust to changing recommendations about treatments. Promoting health. Encouraging health literacy. Health care information access by the public. Developing a system that would be useful in detecting influenza outbreaks based on search characteristics.

Public health data sets.

Data collections from clinical sources. Disease informatics. Epidemiology such as for SARS and bird flu. Data analysis. Models of infectious disease spread.

Flu Trends. Predicting flu epidemics from search engine trend data. Privacy issues (8.3.1).

Deployment of Health-Care Information Systems

There are many promises of information systems in medicine but there are also many difficulties for that. Given the complexity of hospitals and other health-care situations it shouldn't be surprising that they have many of the problems we described earlier for other types of organizations.

Several distinct users. Physicians, Nurses, Patients, Administrators. Use cases and the difficulties of the distinct user groups. For instance, physicians generally do not like to enter information into structured forms.

Cost of health-care information systems.

Failures of health-care information systems. [63].

9.10. Spatio-Temporal Information Systems

Representations and data models for time and space. Spatial information systems describe the properties of objects in space such as geospatial relationships representing 3-D spaces, 3-D objects and environments (11.8.1). In these systems, one can interact with physical objects and their representation in space. The primary role of the representations is to refer to the location of objects in the physical world. One can observe the relative locations of parts of the body.

Spatial analysis. Optimization for spatial data. For instance, where are best places to position fire stations. Coverage models.

Locative media. Incorporating location information into other services.

9.10.1. Varieties of Spatial Information Systems

The most typical spatial information systems are geographic (or geo-spatial) information systems (GIS), which show spatial relationships of the earth. However, there are other types of spatial information and these are often available from information systems. It is possible to explore virtual environments for architecture. This allows architects and their clients to understand what a building will look like before it is built. Urban modeling such as knowing the layout of buildings.

There are many different applications beyond traditional geographic information, including astronomical star charts. And there are other systems in which geo-referenced data may be included; oceanographic and atmospheric systems are data libraries (9.6.3). Systems also differ in how they present space and in the level of interactivity they allow. Models may have metric versus non-metric spatial relationships and they may assume 2D GIS models or 3-D GIS models. Spatial environmental models (Fig. 9.53). Standardized metadata systems.

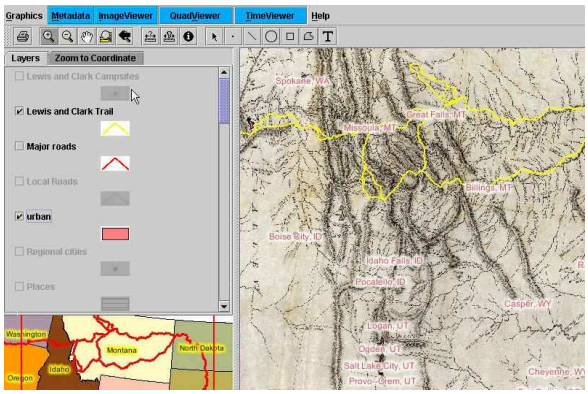

Figure 9.53: Maps as data sets[60]. Here, we see Lewis and Clark's route overlaid on a historical map. (check permission)

Geocoding. Map projections and consistent frameworks for matching points.

Overlaying GIS data sets with maps[60]. Data sets (9.6.0). Maps as representations, indeed visualizations, of schematics. Spatial humanities.

Inferring 3-D from photos. PhotoCity.

Panoramic views of streets.

GIS and epidemiology and criminal justice.

Location technologies and location analytics. Each map search tells a great deal about that person about the person Who does the search. Similarly, knowing the coordinates of a user's mobile device tells a lot. Landscape modeling. Stories and memories associated with land and the physical environment.

Coordinating GIS search with general knowledge about a user's interest.

3-D reconstructions from historical videos of places could be useful as a component of history education.

Stories (6.3.6)related to locations.

Location models. Roles of maps in games. Mirror worlds and the metaverse.

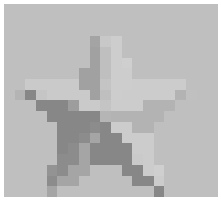

Figure 9.54: Policy map.

9.10.2. Geospatial Data

Data management issues (9.6.3). Historic real estate maps. The ambiguity of geo-spatial objects. Geographic data science.

Incorporating other models into the GIS. For instance, hydrological modeling of water flow combined with landuse.

Geography data sets can serve as a structure around which other data can be organized; GIS data can be used for visualization (11.2.5). There are many examples of such geo-referenced data sets. Other databases are intertwined with geographic information; for instance, the census is broken down by state.

Environmental information can also be geo-referenced, as can video[69]. Many aspects, legal issues. Closely connected to mapping. These maps can be used to biodiversity (Fig. **??**)[61].

Describing Geographic Objects

Information model for GIS objects. This is another example of a Knowledge Organizing System (2.2.0).

"Metadata" descriptions of spatial objects (Fig. 9.55). [38]. FGDC. There are many difficulties in representing political debates if an administrative unit is challenged. OGIS.

Relationship	**Example**
Administrative part of	Mumbai :: India
Administrative seat of	Mexico City :: Mexico
Sub-feature of	Hudson Bay :: Arctic Ocean
Physical Containment	Lake Victoria :: Africa

Figure 9.55: Some relationships among geographic entities.

Describing geospatial objects and the problem of describing physical objects in general. Point objects and vector objects. Spatial synonymy. There is a difficulty identifying place — and place may have two separate names. For instance, a place may be identified by administrative boundaries and one identified by geospatial features. Proper names of geographic objects are also often very ambiguous; for one thing, the names of geographic areas may change over time. Gazetteers list place names and clarify exact locations[38]. Gazetteer details help distinguish Pittsburgh, Pennsylvania from Pittsburg, Kansas or Pittsburg, California (Fig. 9.56). As with most other categories, geographic categories, are social constructs and generally reflect the orientation of the speakers. From a European perspective, Europe is considered a separate continent although it is continuous with Asia. whereas North and South American are considered one continent although they are connected by only a narrow strip of land.

Location	**Description**
Pittsburg	1 city W. Calif. NE of Oakland on San Joaquin River, *pop. 19,062* 2 city SE Kans., *pop. 18,678*
Pittsburgh	city SW Pa., *pop. 604,332*

Figure 9.56: Geographic gazetteer entries for Pittsburg and Pittsburgh[17].

Geographic meta-models.

North Korea economy watch. Satellite images and determining social structure (Fig. 9.57). Surveillance (8.3.3).

Figure 9.57: Detail of a region in North Korea which has been identified by Web-based observers as a compound for the elite Party members (North Korean Economy Watch). (check permission)

GIS Queries and Interfaces

Spatial relationships imply different types of queries than we saw for text queries. Many spatial query relationships are relative to other objects: left, right, above, near, facing, and among/between are some of the terms used to describe them. Relationships like "between" are sometimes ambiguous because the positions of the objects are sometimes complex (Fig. 9.58). Range queries ask "Is a spatial object within a given area".

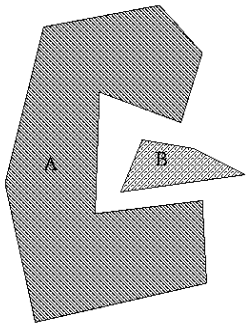

Figure 9.58: This illustration shows one of the difficulties of range queries. A query might ask whether city B in region A.

Queries to GISs can be complex and require considerable inference about geographic relationships. For instance, the query "what towns are in flood plains in the Pittsburgh area?" would require knowledge of both towns and flood plains. GIS queries and interfaces support tasks and decision-making. As with text queries, the "query semantics" should match the data model.

9.10.3. Searching Geospatial Information

Coordinating search with locations. For instance, finding pizza close to your current location. Alternative to yellow pages. GPS. Geospatial objects include natural objects such as mountains and rivers, human constructions such as highways, and abstract concepts such as boundaries and property lines. In ordinary use, the definitions of geospatial objects can be ambiguous; for instance, a "street" may be confused with a "road" or a "highway". However, it is possible to give a more exact definition.

Some spatial objects are "point" spatial objects such as a city center, others are "vector" spatial objects, such as a road. Still others are best described with a bounding box or perhaps more irregular polygonal footprints. Sometimes a system must deal with disjoint entities, such as the locations of all branches of a bank.

Spatial indexing organizes objects by locations. Data structures for spatial indexing such as bin-trees and R*-trees. The familiar longitude and latitude lines can be the basis of indexing locations in geographic space; other systems of coordinates such as the Universal Transverse Mercator (UTM), are also used. Geo-referencing is one example of coordinating data sets. The most common representation of space in cartography but other representations are possible.

9.10.4. Spatial Cognition, Orientation, and Navigation

Spatial Cognition

People learn about space by moving, crawling, and walking. How do people understand, remember, and use concepts of space? The representation and inferences people make about spatial relationships are known as "spatial cognition". There are probably multiple, interlocking cognitive representations for space. As with other inferences (4.3.4), people may make incorrect inferences about spatial relationships Some people are confused about whether San Francisco or Los Angeles is closer to Hawaii. As you can see in Fig. 9.59, San Francisco is closer although it is farther north. Presumably, people reason that LA is south of SF and Hawaii is even further south. They don't also realize that San Francisco is much further west than Los Angeles.

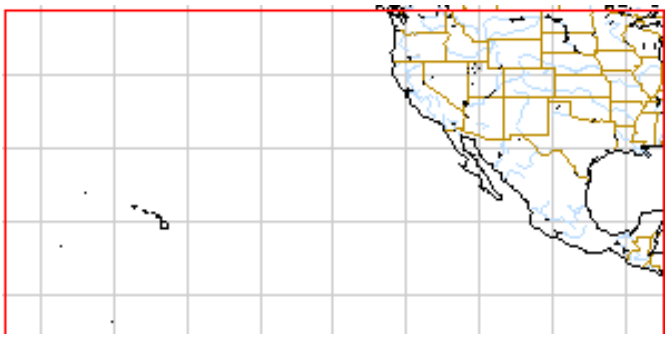

Figure 9.59: Without seeing a map, most people believe that Hawaii is closer to Los Angeles than to San Francisco. (check permission)

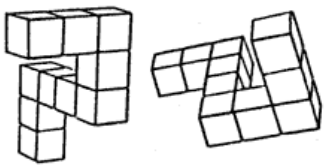

Figure 9.60: Mental rotation task. Are these two shapes congruent? (check permission)

9.10.5. Maps

Maps are schematics that support orientation and navigation; they are a graphical symbolic representation of space. Maps typically go beyond pictorial presentation of symbolic information. Maps can be used to find locations corresponding to physical reality, and to compare various properties of physical objects. Fig. 9.61 is a well-known drawing that blends pictorial and map-like information. Maps employ a specialized visual language (11.2.4) that represents space. When correlated with the world, they have a "referential semantics" (6.2.3). Place is emergent. Maps help people claim territory. Cartographic representation.

Spatial history. Cultural mapping.

Migration and cultural changes. Economic activity, trade, and commodities.

Place is emergent often related to cultural stories. Re-naming of place is often related to cultural transitions.

Figure 9.61: A drawing which suggests the typical mental map a New York City resident has of the western part of the United States. The focus is the area around 9^{th} Avenue. Beyond that, the view is not linear but distorted like a "fisheye" and only a few salient locations in the distance are included[72]. (redraw-K) (check permission)

A map is a document that identifies objects and shows their spatial relationships. As with any document, it reflects the author's sense of what will be most useful and most effectively understood by the viewer. Maps in the tradition of the Western culture often show things that can be enumerated, such as distance and population. Although we think of maps as primarily visual, geospatial information can also be conveyed in other ways. For instance, poems and songs may provide a cultural memory for spatial relationships. Places in digital environments. Spatial brain cells (-A.12.2).

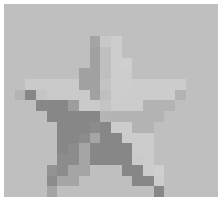

Figure 9.62: Map.

Usability of maps. Orienting oneself with a map. First-person perspective. Similar concerns arise in virtual reality. Tethered perspective. Absolute perspective. This may also include focus+context visualization as we saw earlier (11.2.5).

Metadata and Cataloging Maps

As with other complex information resources, there are catalogs of map collections. Even seemingly straightforward attributes, such as the date, can be crucial, since geography is dynamic. Several metadata standards for maps are built on Dublin Core (2.4.4).

Spatial narrative. Spatial history. Spatial animations. Personal testimony.

Interactive Maps and Spatial Presentations

Interactivity allows users to explore details. We have already seen lenses to allow users to explore maps (11.2.5). Maps can be used to display responses to GIS queries, and as query interfaces for georeferenced systems (9.10.2). Maps may also be used to orient people to other types of processes. or instance, SimCity uses spatial representation as the basis for games.

Maps may be elements in virtual or augmented reality. Indeed, they are an aspect of many games.

Spatial Modeling

3-D modeling.

9.10.6. GIS-Based Services and Neogeography

Many services have a GIS component. Location as a context. It can be useful in, for example, services to do with crisis management (8.6.4), law enforcement, the census, or the environment, GIS libraries. Map services.

KML files.

Mirror world. Metaverse (11.10.1).

Political disputes about boundaries reflected in the map. Layer of commentary and opinions over the spatial descriptions. This is sometimes termed "neogeography". sNorth Map-based mashups with GIS.

Spatial Orientation and Navigation

People use their knowledge of space to determine their location and to move from one place to another. Orientation is locating oneself within an established framework – especially in space. To use a map, a person must understand how the map relates to the physical space. Maps may be egocentric or exocentric navigation – that is they may be based on an absolute position or they may be relative to an individual's position. How do you figure out where you are in a virtual landscape? Context is vital to orientation, which is partly a recognition problem. Spatial navigation is called "wayfinding".

Signage should be based on scenarios of use.

Figure 9.63: Signage is often essential for wayfinding. Each path a person may take through the physical environment and their spatial information needs should be considered. (check permission)

"Landmarks" are points that are easily recognized. Identification of the landmark allows orientation. Landmarks are important aids to knowing where you are and guiding navigation. They generally are familiar and highly visible, and have distinctive features. Longitude and latitude indexing can provide orientation as can be an alternative to using landmarks for navigation. Still other methods such as Western and Polynesian methods for navigation.

Figure 9.64: Polynesia navigation stickchart. (check permission)

Sensations such as smell, sounds, warmth, and wind often provide orienting cues. But, these are frequently not available in computer-generated virtual worlds (11.10.2). Pop-up navigation in 3D-spaces and virtual worlds (11.10.2).

Navigation is movement through a series of stages toward a goal (2.6.3). Polynesian navigation. While we may navigate an interface but navigation is regarded as moving through space. Landmarks and sign posts are among the clues that can be used to support navigation.

The explanation of directions illustrates some of the ways in which groups develop a shared understanding. The process of negotiating meaning during conversation (6.4.0) is particularly apparent when giving directions[74]. Sketching can add a visual component to the directions

Some people do navigation by a grid such as following a spatial grid such as city blocks.

9.10.7. Events and Time

Timelines and temporal histories. Temporal information retrieval. History (5.13.0).

Relevance changes over time. Episodic retrieval.

Processes.

Temporal landmarks.

Exercises

Short Definitions:

Academic freedom	Occam's razor	Scientific methods
Bibliometrics	Primary literature	
Collaboratory	Scholarship	

Review Questions:

1. Scholarship. (9.0.0)
2. How is science different from technology? (9.4.4)
3. Give an example of workflow in science. (9.3.3)
4. What Web site lists the format used for US Census data? (9.6.3)

Short-Essays and Hand-Worked Problems:

1. Calculate linkages in the citation lists of several related articles. (9.1.2)
2. What are the difficulties of using bibliometrics as quality indicators for scholarship? (9.1.3).
3. What is the impact of a scholarly journal that publishes 32 articles per year and each of them receive a total of 200 citations. How does that compare with a journal that publishes 110 articles per year that receives a total of 330 citations? (9.1.3)
4. Sometimes the number of citations to a scholarly paper is taken as an indication of quality. What are some of the problems with that? (9.1.3)
5. Contrast the goals of "text retrieval" from "knowledge discovery". (9.2.2, 10.9.0).
6. What type of metadata might be appropriate for a data set collected for a space probe sent to Mars? (9.6.3)
7. Genomics. (9.8.1)
8. We have defined information as that which reduces uncertainty. Do biological representations reduce uncertainty? ((sec:infodefinition), 9.8.1, -A.1.0)
9. GIS. (9.10.0)
10. Explain the differences between first-person and third-person perception from maps. (9.10.4)
11. It is said that people from older cities in the U.S. (e.g., Boston) navigate by landmarks while people from the newer cities (e.g., Phoenix) navigate by compass directions. Why might this be true? (9.10.4)
12. Describe the characteristics that make a map effective. (9.10.5)
13. Observe how people use maps. How could their use be better supported by interactive maps? (9.10.5)
14. Ask someone to give you directions. Note what they say. Record the questions the ask and describe the assumptions they are making. (9.10.6)

Practicum:

1. Data libraries,

Going Beyond:

1. Is scholarship a largely academic, ivory tower exercise or is it relevant to social issues and problems? (9.0.0)
2. Should a scholarly paper that has been found to plagiarize a previous work be deleted from a library? (5.12.3, 9.1.1)
3. Describe the pros and cons of peer reviewing and an evaluation for the value of ideas. (9.1.1)
4. Create a citation map of a field. (9.1.2)
5. Are citations good measures of quality for scholarly research? (9.1.2)
6. Are scientific laws created or discovered? (9.2.2)
7. What is the proper role of the government in science? (9.4.0)
8. How does the notion of a "collection" apply to a data library? Is a database a data library? (7.2.2, 9.6.3)
9. Specify data management procedures for a database of clinical trials. (9.6.3)
10. Describe a task for which the particle flurries would be a distinct advantage over other visualization techniques. (9.6.5)
11. Develop a MathML specification for $\frac{1}{\frac{1}{2}^2}$ (9.7.2)
12. If Biology is an information science rather than a physical science, what sort of generalizations can we make about it? (9.8.1)
13. Should personal genetic information affect insurance rates? (9.8.1)
14. Is there a net cost savings from implementing health-care information systems? (9.9.2)
15. Describe technologies what would be needed to match pictures with text documents. (9.10.5)
16. Describe how OCR techniques might be applied to automatic map analysis. (9.10.5, 10.1.5)

Teaching Notes

Objectives and Skills:

Instructor Strategies:

Related Books

- Baker, S. *The Numerati.* Houghton-Mifflin, New York, 2008.
- Boot, M. *War Made New: Technology, Warfare, and the Course of History, 1500 to Today.* Council for Foreign Relations, 2006.
- Borgman, C. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet.* MIT Press, Cambridge MA, 2007.
- Bowker, G.C. *Memory Practices in the Sciences.* MIT Press, Cambridge MA, 2006.
- Brown, M.B. *Science in Democracy: Expertise, Institutions, and Representation* MIT Press, Cambridge MA, 2009.
- Bunker, W.E., Bal, R., and Hendirks, R. *The Paradox of Scientific Authority: The Role of Scientific Advice in Democracies.* MIT Press, Cambridge MA, 2009.
- Cole, S. *Making Science.* Harvard University Press, Cambridge MA, 1992.
- Deutsch, D. *The Fabric of Reality.* Penguin Press, New York, 1997.
- Giere, R.N., Bickle, J., and Mauldin, R.F. *Understanding Scientific Reasoning.* Thomson-Wadsworth, Belmont CA, 5^{th} ed., 2006.
- Green, M.A., and Bowie, M.J. *Essentials of Health Information Management.*
- Gilovich, T. *How Do we Know What Isn't So? The Fallibility of Human Reason in Everyday Life.* Free Press, New York, 1991.
- Heuer, R.J. *Psychology of Intelligence Analysis.* Novkinka Books, New York, 2006.
- Latour, B. and Woolgar, S. *Laboratory Life: The Construction of Scientific Facts.* Princeton University Press, Princeton NJ, 1986.
- Owens, W.A., Dam, K.W., and Lin, H. *Technology, Policy, Law, and Ethics Regarding U.S. Acquisition and Use of Cyberattack Capabilities.* NAP, Washington DC, 2009.
- MacEachren, A.M., *How Maps Work: Representation, Visualization, and Design.* Guilford, New York, 1995.
- Meadow, C.
- Merton, R.K. *The Sociology of Science.* University Chicago Press, Chicago, 1973.
- Nagl, J. *Learning to Eat Soup with a Knife: Counterinsurgency Lessons from Malaya and Vietnam.* Praeger, Westport CT, 2002.
- NAP *Ensuring the Integrity, Accessibility, and Stewardship of Data in the Digital Age.* National Academies Press, 2009.

- Pearl, J. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge UK, 2000.
- Simon, H.A. *Sciences of the Artificial.* 3^{rd} ed. MIT Press, Cambridge MA, 1996.
- Singer, P.W. *Wired for War: The Robotics Revolution and Conflict in the* 21^{st} *Century.* Penguin, New York, 2009.
- Stokes, D. *Pasteur's Quadrant: Basic Science and Technological Innovation.* Brookings Institution, Washington DC, 1997.
- Tenner, E. *Our Own Devices: How Technology Remakes Humanity.* Vintage, New York, 2003.
- Thagard, P. *Conceptual Revolutions.* Princeton University Press, Princeton NJ, 1992.
- Von Baeyer, H.C. *Information: The New Language of Science.* Harvard University Press, Cambridge MA, 2004.
- Yoon, C.K. *Naming Nature: The Clash Between Instinct and Science* Norton, New York, 2009.