The Challenge of Knowledge Soup

John F. Sowa

VivoMind Intelligence, Inc.

Abstract. People have a natural desire to organize, classify, label, and define the things, events, and patterns of their daily lives. But their best-laid plans are overwhelmed by the inevitable change, growth, innovation, progress, evolution, diversity, and entropy. When the Académie Française attempted to legislate the vocabulary and definitions of the French language, their efforts were undermined by uncontrollable developments: rapid growth of slang that is never sanctioned by the authorities, and wholesale borrowing of words from English, the world's fastest growing language. In Japan, the pace of innovation and borrowing has been so rapid that the older generation of Japanese can no longer read their daily newspapers. These changes, which create difficulties for people, are far more disruptive for the fragile databases and knowledge bases in computer systems. The term *knowledge soup* better characterizes the fluid, dynamically changing nature of the information that people learn, reason about, act upon, and communicate. This talk addresses the complexity of the knowledge soup, the problems it poses for intelligent systems, and the methods for managing it. The most important measure for any intelligent system is its flexibility in accommodating and making sense of the knowledge soup.

This report combines the speaker's slides from two different conferences:

- The PerMIS'04 Conference at NIST in Gaithersburg, Maryland, in August 2004.
- The epiSTEME-1 Conference in Goa, India, in December 2004.

Both lectures had the same title, and there was a large overlap in the material presented.
This report includes all the slides together with some additional comments and references.

Issues in Knowledge Representation

Outline of this lecture:

1. **Formal axioms and definitions**

   Essential for well-defined problems,
   But most problems aren't well defined.

2. **Knowledge Soup**

   "There are more things in heaven and earth, Horatio,
   Than are dreamt of in your philosophy."

   William Shakespeare

3. **Semiotics by Charles Sanders Peirce**

   Categories of signs,
   Cycle of cognition,
   Analogical Reasoning.

Aristotle's Syllogisms

System of logic based on four sentence patterns:

1. ***Universal affirmative*. Every employee is human.**
2. ***Particular affirmative*. Some employees are customers.**
3. ***Universal negative*. No employee is a competitor.**
4. ***Particular negative*. Some customers are not employees.**

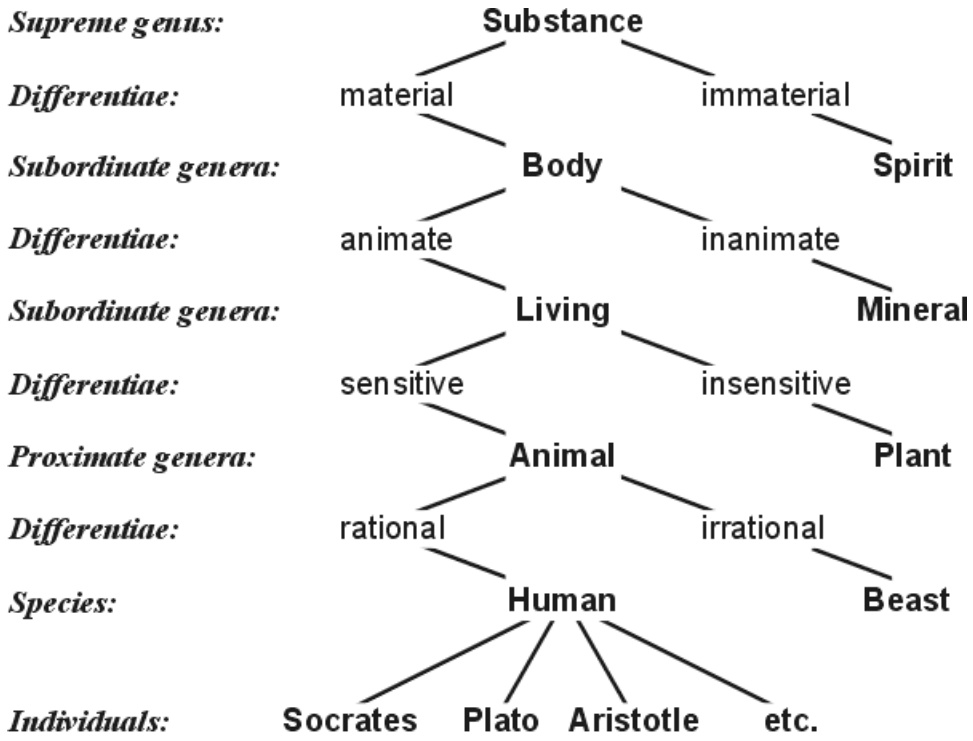
Affirmative patterns are used to define the type hierarchy and state inheritance.

Negative patterns are used to state constraints.

Note: Although Aristotle's syllogisms are the oldest system of formal logic, they are still an important version of logic. They form the core of modern *description logics* (DLs), such as OWL, which are widely used for defining ontologies. OWL and other DLs add important features, such as numeric-valued functions, in addition to Aristotle's monadic predicates. But for many applications, the Aristotelian subset is sufficient.

Tree of Porphyry

Shows inheritance of *differentiae* from genus to species:



Porphyry drew the first known tree diagram for organizing Aristotle's categories in the third century AD. This version was translated from the *Summulae Logicales* by Peter of Spain (1239). It shows that Body is defined as material Substance, and Human as rational Animal.

By following the path all the way to the top, the category Human would *inherit* all the differentiae along the way: rational, sensitive, animate, material Substance.

Similar diagrams are widely used today to represent hierarchies of concept types.

Gottfried Wilhelm Leibniz

Encoded Aristotle's categories as integers:

- **Prime numbers to encode primitive concepts,**
- **Products of primes for compound concepts.**
- **Concept X is a subtype of Y iff Y divides X.**
- **The result is a *lattice* with multiple inheritance.**

But he never realized his grand hope:

> ***The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say: Let us calculate, without further ado, in order to see who is right.***

Immanuel Kant

Proposed twelve categories as a replacement for Aristotle's:

Quantity	Quality	Relation	Modality
Unity	Reality	Inherence	Possibility
Plurality	Negation	Causality	Existence
Totality	Limitation	Community	Necessity

But he never realized his grand hope:

> ***If one has the original and primitive concepts, it is easy to add the derivative and subsidiary, and thus give a complete picture of the family tree of the pure understanding. Since at present, I am concerned not with the completeness of the system, but only with the principles to be followed, I leave this supplementary work for another occasion.***

Note the words in red: whenever a philosopher or a mathematician says that something is easy or obvious, that is a sure sign of difficulty.

Académie Française

Primary mission:

- **Défense de la langue française.**

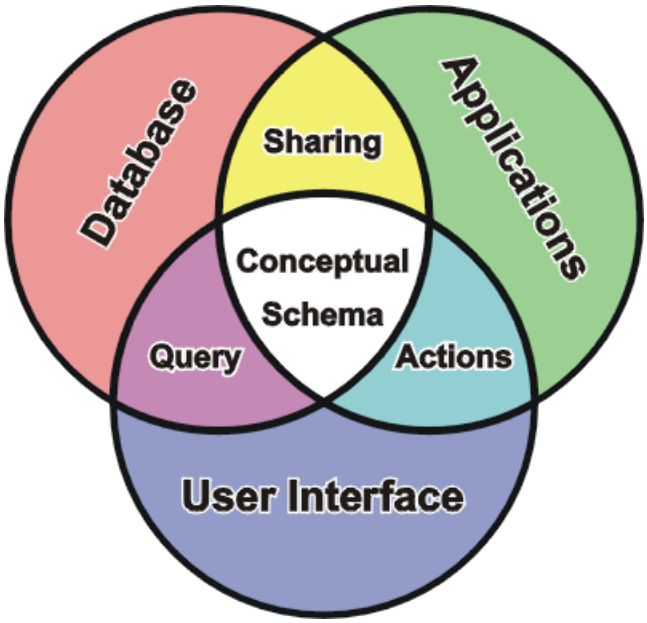
Strategy:

- **Create a dictionary that freezes the meaning of every French word.**

Result:

- **Uncontrollable growth of slang terms that never appear in the dictionary.**
- **Wholesale borrowing of new words from English.**

Conceptual Schema



ANSI SPARC, 1978.

ISO Standards Project, R.I.P. 1999.

Born again as the Semantic Web.

The above diagram illustrates the *three-schema approach* proposed by the ANSI SPARC committee. It would enable multiple applications to interoperate on a common semantic foundation represented in a *conceptual schema*. It was a good idea, but the achievements fell far short of the hopes. The currently popular term for a very similar idea is *ontology*.

World's Largest Ontology Project

Cyc project started in 1984 by Doug Lenat.

- **Name comes from the stressed syllable of *encyclopedia*.**
- **Goal: implement the commonsense knowledge of an average adult.**
- **After $70 million and 700 person-years of work,**

  600,000 categories
  defined by 2,000,000 axioms
  organized in 6,000 microtheories.

For more information about Cyc, see their web site:

http://www.cyc.com

Project Halo

Project for evaluating methods of knowledge representation.

Goal: Build an intelligent tutor.

Test case: Encode knowledge from a chemistry textbook in order to answer questions on a freshman chemistry exam.

Participants: Cycorp, OntoPrise, SRI International.

Results:

- **Average score: about 40% to 47% correct.**
- **Cost to encode knowledge: average about $10,000 per page from the textbook.**
- **Despite its large knowledge base, Cyc had the lowest score.**

With better tools and cheaper labor, the cost has been reduced to about $100 per page. But that is still too high for large-scale knowledge acquisition.

The Halo Project was funded by Microsoft cofounder Paul Allen as part of his goal to develop a "Digital Aristotle." Following is the official web site:

http://www.projecthalo.com/

For more reports about the Halo Project, type the keywords "Halo" and "chemistry" to your favorite search engine. To narrow the search, add more keywords, such as "Cyc" or "SRI."

Utterance by a 3-year-old Child

Sentence expressed by Laura at 34 months:

> ***When I was a little girl, I could go "geek, geek" like that;***
> ***but now I can go "This is a chair."***

Enormous logical complexity in one short passage:

- **Subordinate and coordinate clauses**
- **Tenses: Earlier time contrasted with "now"**
- **Modal auxiliaries: *can* and *could***
- **Quotations: "geek, geek" and "This is a chair"**
- **Metalanguage about her own linguistic abilities**
- **Contrast shown by *but***
- **Parallel stylistic structure**

Reference:

Limber, John (1973) "The genesis of complex sentences," in T. Moore, ed., *Cognitive Development and the Acquisition of Language*, Academic Press, New York, 169-186.

Available at http://pubpages.unh.edu/~jel/JLimber/Genesis_complex_sentences.pdf

Observations

The child has much less technical knowledge than Cyc.

But her learning ability is far more flexible and far more efficient:

- **Only three person-years of effort.**
- **No need for knowledge encoding at $10,000 or even $100 per page.**

Can our computer systems ever be as flexible?

Limitations of Current Approaches

The logics of the Semantic Web (RDF, OWL, and Rule ML) are useful for many applications, but there is nothing new:

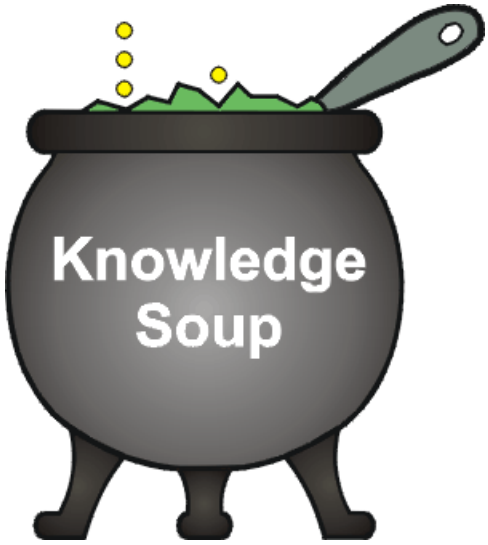
- **They're bracketed on the low end by Aristotle's syllogisms and on the high end by Cyc.**

$100 to encode one page from a textbook is a major barrier to widespread use.

In recent years, the Cyc knowledge base has expanded from 100,000 axioms to 2,000,000 axioms — but the cost of adding new knowledge has not gone down.

There's no evidence that an expansion from two million to two billion would make much, if any reduction in cost.

The Challenge



The fluid, loosely organized, dynamically changing contents of the human mind.

Examples of Knowledge Soup

- **Overgeneralizations: Birds fly.**

  But what about penguins? A day-old chick? A bird with a broken wing? A stuffed bird? A sleeping bird? A bird in a cage?

- **Abnormal conditions: If you have a car, you can drive from New York to Boston.**

  But what if the battery is dead? Your license has expired? There is a major snowstorm?

- **Incomplete definitions: An oil well is a hole drilled in the ground that produces oil.**

  But what about a dry hole? A hole that has been capped? A hole that used to produce oil? Are three holes linked to a single pipe one oil well or three?

- **Conflicting defaults: Quakers are pacifists, and Republicans are not.**

  But what about Richard Nixon, who was both a Quaker and a Republican? Was he or was he not a pacifist?

- **Unanticipated applications: The parts of the human body are described in anatomy books.**

  But is hair a part of the body? Hair implants? A wig? A wig made from a person's own hair? A hair in a braid that has broken off from its root? Fingernails? Plastic fingernail extender? A skin graft? Artificial skin used for emergency patches? A band-aid? A bone implant? An artificial implant in a bone? A heart transplant? An artificial heart? An artificial leg? Teeth? Fillings in the teeth? A porcelain crown? False teeth? Braces? A corneal transplant? Contact lenses? Eyeglasses? A tattoo? Make-up? Clothes?

Devil in the Details

Most banks offer similar services with similar terminology:

- **Checking, savings, loans, mortgages...**

Banks interoperate on electronic funds transfer.

But when two banks merge, they never merge their databases.

Two common strategies:

- **Keep running both databases indefinitely, or**
- **Close some or all accounts of one bank, and open new accounts in the database of the other bank.**

There are too many incompletely documented details.

Limits of Definability

- **Immanuel Kant:**

  "Since the synthesis of empirical concepts is not arbitrary but based on experience, and as such can never be complete (for in experience ever new characteristics of the concept can be discovered), empirical concepts cannot be defined.

  "Thus only arbitrarily made concepts can be defined synthetically. Such definitions... could also be called *declarations*, since in them one declares one's thoughts or renders account of what one understands by a word. This is the case with *mathematicians*."

- **Wittgenstein's *family resemblance*:**

  Empirical concepts cannot be defined by a fixed set of necessary and sufficient conditions. Instead, they can only be taught by giving a series of examples and saying "These things and everything that resembles them are instances of the concept."

- **Waismann's *open texture*:**

  For any proposed definition of empirical concepts, new instances will arise that "obviously" belong to the category but are excluded by the definition.

References:

Kant, Immanuel (1800) *Logik: Ein Handbuch zu Vorlesungen*, translated as *Logic* by R. S. Hartmann & W. Schwarz, Dover Publications, New York, 1988; also translated as *Lectures on Logic* by J. M. Young, Cambridge University Press, Cambridge, 1992.

Waismann, F. (1952) "Verifiability," in A. Flew, ed., *Logic and Language*, first series, Basil Blackwell, Oxford.

Wittgenstein, Ludwig (1953) *Philosophical Investigations*, Basil Blackwell, Oxford.

Limits of Logic

Alfred North Whitehead, *Modes of Thought*:

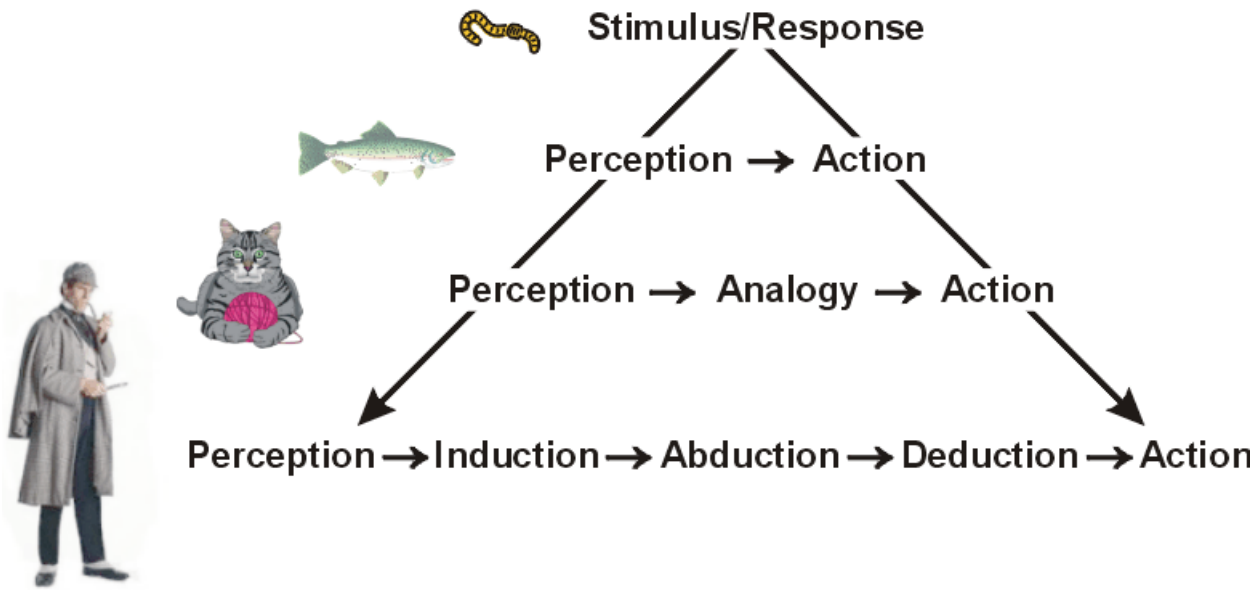
- **"Both in science and in logic, you have only to develop your argument sufficiently, and sooner or later you are bound to arrive at a contradiction, either internally within the argument, or externally in its reference to fact."**
- **"The topic of every science is an abstraction from the full concrete happenings of nature. But every abstraction neglects the influx of the factors omitted into the factors retained."**
- **"The premises are conceived in the simplicity of their individual isolation. But there can be no logical test for the possibility that deductive procedure, leading to the elaboration of compositions, may introduce into relevance considerations from which the primitive notions of the topic have been abstracted."**

Summary: "We must be systematic, but we should keep our systems open."

Reference:

Whitehead, Alfred North (1938) *Modes of Thought*, Macmillan Co., New York.

Evolution of Cognition



Every organism retains the capabilities of all earlier forms.

Peirce's Classification of Reasoning

Three methods of logic plus analogy:

1. **Deduction: Deriving implications from premises.**
2. **Induction: Deriving general principles from examples.**
3. **Abduction: Forming a hypothesis that must be tested by induction and deduction.**
4. **Analogy: "Besides these three types of reasoning there is a fourth, analogy, which combines the characters of the three, yet cannot be adequately represented as composite."**

Analogy is more primitive, but more flexible than logic.

The methods of logic are disciplined ways of using analogy.

For a summary of Peirce's relevance to modern cognitive science, see

http://www.jfsowa.com/pubs/csp21st.pdf

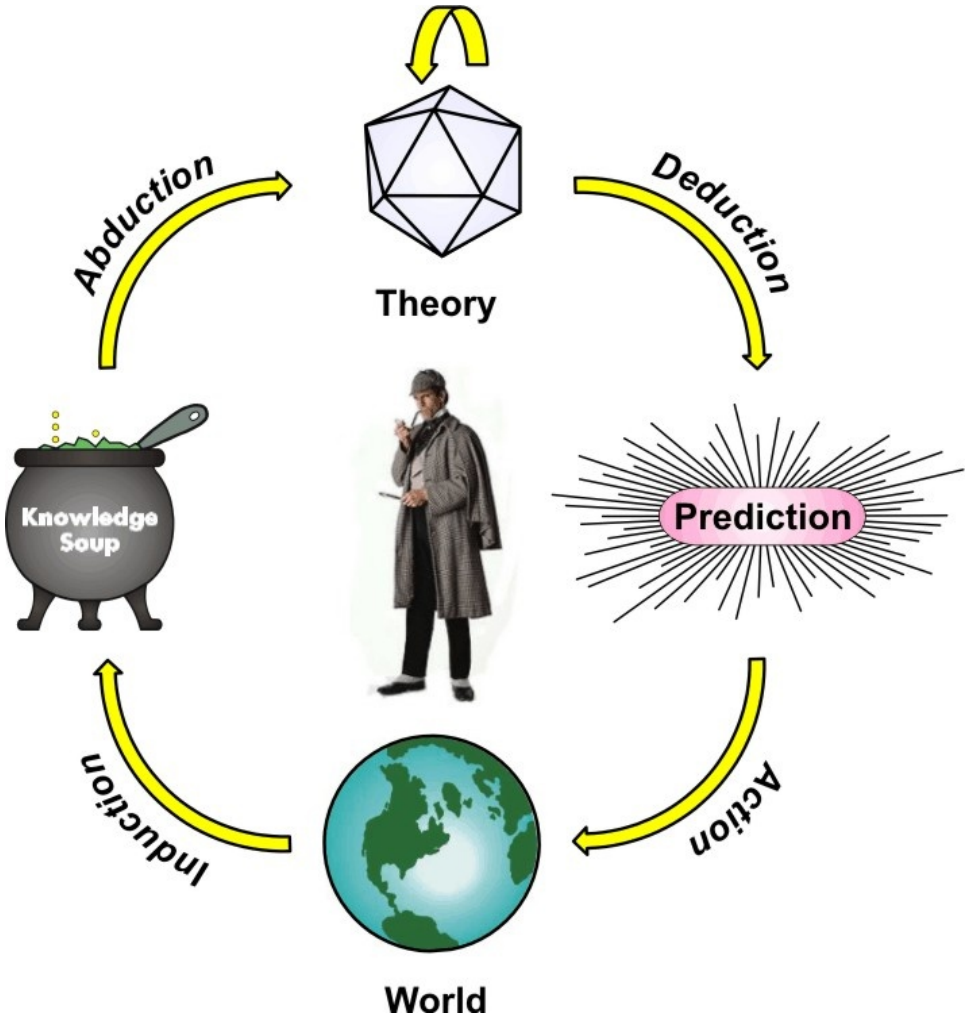
For Peirce's own writings about logic and related philosophy, see

Peirce, Charles Sanders (1898) *Reasoning and the Logic of Things*, The Cambridge Conferences Lectures of 1898, ed. by K. L. Ketner, Harvard University Press, Cambridge, MA, 1992.

Peirce, Charles Sanders (1903) *Pragmatism as a Principle and Method of Right Thinking*, The 1903 Lectures on Pragmatism, ed. by P. A. Turrisi, SUNY Press, Albany, 1997.

Peirce, Charles Sanders (EP) *The Essential Peirce*, ed. by N. Houser, C. Kloesel, and members of the Peirce Edition Project, 2 vols., Indiana University Press, Bloomington, 1991-1998.

Peirce's Cycle of Cognition

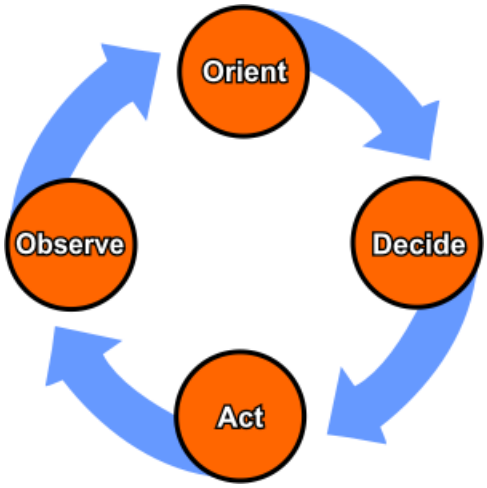


Peirce never drew such a diagram, but he described the stages in various writings.
See the recommended readings on the previous page.

A Continuum of Reasoning Processes

Peirce's cycle characterizes reasoning processes at every level of difficulty and for time periods of any length:

- **Real-time operations, as described by Boyd's OODA loop (Observe, Orient, Decide, Act), may happen in seconds or milliseconds.**
- **Problem-solving cycles may take minutes to days.**
- **Scientific research may take months to decades.**

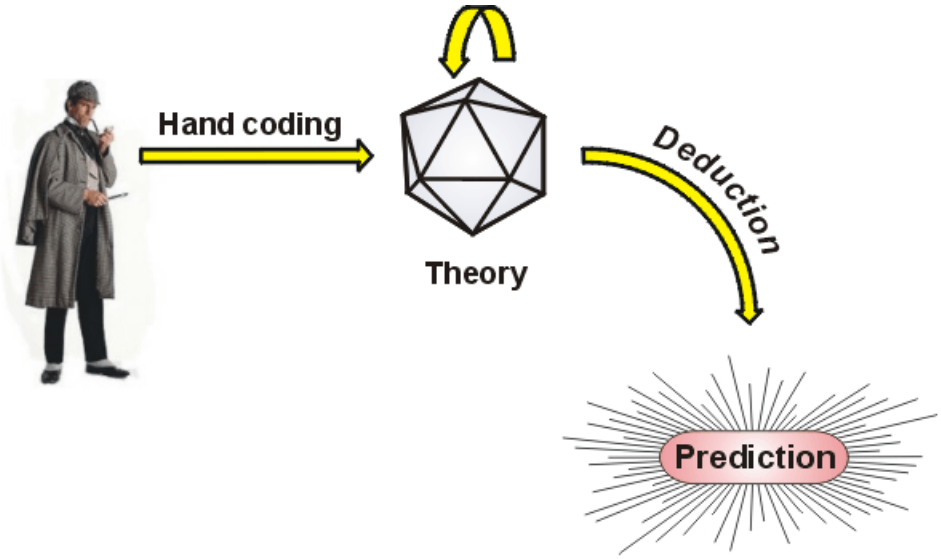


The central feature of Peirce's pragmatism is the grounding of the reasoning process in perception at one end and action at the other.

The OODA loop is a special case of Peirce's cognitive cycle.

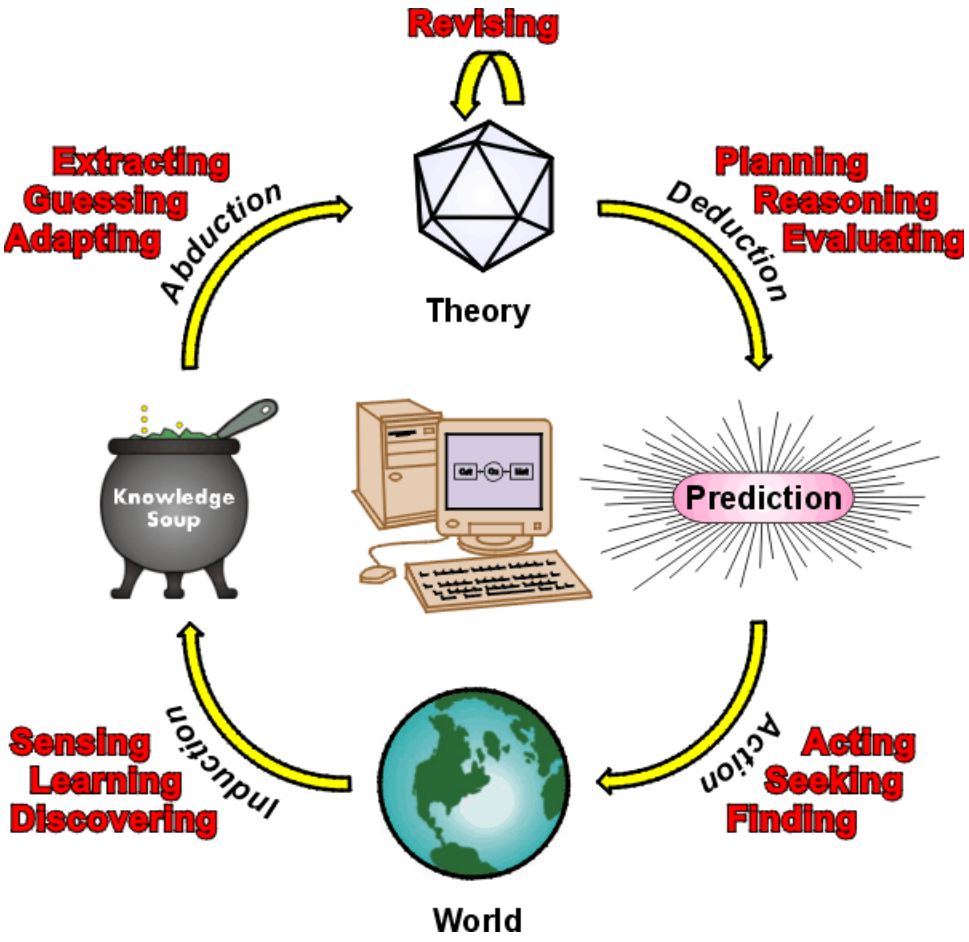
For further discussion and references, type "John Boyd" and "OODA loop" to your favorite search engine.

Cyc's Piece of the Pie

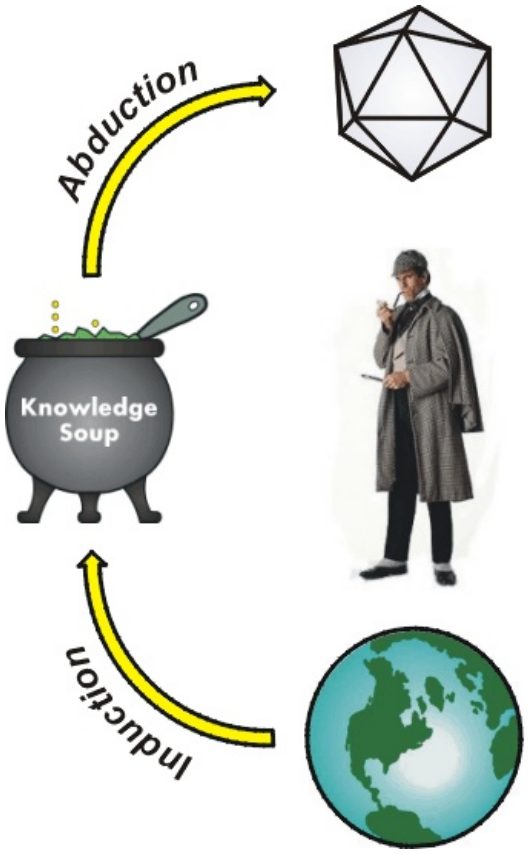


- **Cyc does not automate Sherlock Holmes.**
- **It requires people like him to write axioms.**
- **At a cost of $10,000 to encode one page from a textbook.**

Deduction is only 25% of the Cycle



The Challenge of Knowledge Soup



- **Computer systems are better at deduction than most people.**
- **But the greatest challenges and opportunities are on the other side.**
- **How is new knowledge added to the soup?**
- **How is structured knowledge derived from the unstructured soup?**
- **How is relevant knowledge found and used when needed?**
- **And how can those processes be automated?**

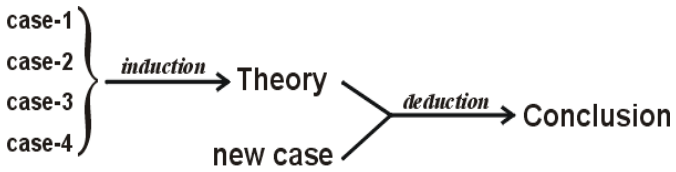
Ibn Taymiyya Contra Aristotle

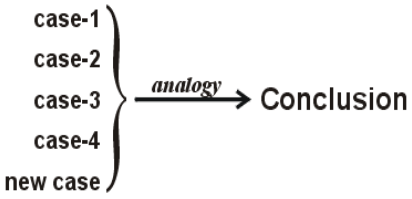
- **Fourteenth-century Islamic legal scholar.**
- **Admitted that deduction is necessary for pure mathematics.**
- **But for reasoning about the world, deduction is limited to the accuracy of the induction.**
- **Given the same data, analogy can replace induction + deduction.**

Reference:

Hallaq, Wael B. (1993) *Ibn Taymiyya Against the Greek Logicians*, Clarendon Press, Oxford.

Ibn Taymiyya's Argument





A theory can be very useful when available, as in mathematics, science, and engineering.

But in law, medicine, business, and everyday life, reliable theories are seldom available.

Those subjects depend on analogies to support *case-based reasoning* (CBR).

Structure Mapping

Mapping one conceptual structure to another can have four logical effects:

1. **Equivalence: $CS_1 \equiv CS_2$**
2. **Generalization: CS_1 implies CS_2**
3. **Specialization: CS_2 implies CS_1**
4. **Similarity: Neither one implies the other.**

Analogy uses all four kinds of mapping.

Logic uses only the first three kinds.

The same mechanisms, computational and neurophysiological, underlie both.

An analogy about analogy:

Logic is to analogy as dancing is to walking.

Dancing is a stylized form of walking that uses the same muscles and motions as walking, but in a more structured, disciplined form.

Logic is a stylized form of analogical reasoning that uses the same mental processes (whatever they may be), but in a more structured, disciplined form.

VivoMind Analogy Engine

Structure-mapping methods used in analogy:

1. ***Matching labels:***
    - **Compare type labels on conceptual graphs.**
2. ***Matching subgraphs:***
    - **Compare subgraphs independent of labels.**
3. ***Matching transformations:***
    - **Transform subgraphs.**

Methods #1 and #2 take (N log N) time.

Method #3 takes polynomial time (analogies of analogies).

Reference:

Sowa, John F., & Arun K. Majumdar (2003) "Analogical reasoning," in A. de Moor, W. Lex, & B. Ganter, eds., *Conceptual Structures for Knowledge Creation and Communication*, Proceedings of ICCS 2003, LNAI 2746, Springer-Verlag, Berlin, pp. 16-36.

Available at http://www.jfsowa.com/pubs/analog.htm

Intelligent Assessor

A textbook publisher wanted a method for grading free-form answers (one or two English sentences) to examination questions.

The test case was student explanations of algebra word problems.

Three companies proposed methods for addressing the task:

1. **One company recommended a deductive approach similar to Cyc.**
2. **Another company recommended Latent Semantic Analysis (LSA) for measuring the similarity of word choice between a student's answer and a correct answer.**
3. **VivoMind proposed the analogy engine for comparing student answers to a selection of correct and incorrect answers.**

Method #1 required too much knowledge representation by teachers who had no experience in KR, and method #2 could not distinguish correct answers from incorrect answers because they used similar selections of words.

Note: The three slides about the Intelligent Assessor were presented at the epiSTEME-1 Conference, but not at PerMIS'04.

Sample Data

The publisher evaluates exam questions with several classrooms of actual students.

For each question, they collect sample answers from about 6 teachers and about 50 students.

1. **Some answers are completely correct, but stated with various words and phrasing.**
2. **Some are partially correct, and a teacher wrote a comment to explain what is missing.**
3. **Some are wrong, and a teacher wrote a helpful comment.**
4. **The rest are blank or wrong, and no teacher wrote a comment.**

How can a computer system grade student answers and write comments that are comparable to those of a good teacher?

VivoMind Approach

Don’t try to reproduce all the reasoning done by the students or the teachers.

Just use the analogy engine to match any new answer to one of the sample answers:

1. **Translate all the sample answers from English to conceptual graphs (CGs).**
2. **Translate each new answer to a CG.**
3. **Use the VivoMind Analogy Engine (VAE) to find which of the sample CGs is the closest match to the new CG.**
4. **Print the teacher’s evaluation and comment associated with the matching CG.**

This method correctly graded all answers presented to it.

Unfortunately, the project manager died, and the budget was canceled.

Using VAE for Knowledge Fusion

Challenge:

- **A large corporation needed to analyze and re-engineer mainframe software and documentation dating back to 1962.**
- **Extract and combine knowledge from highly structured software and unstructured English.**
- **Cross-index all the sources.**
- **Detect inconsistencies.**
- **A major consulting firm estimated that it would take 40 people two years to do the analysis.**

Understanding unrestricted natural language is still an unsolved problem.

But if the structured data is processed first, the results can be used to interpret the unstructured English.

VAE was used to find and compare structures from different languages when translated to conceptual graphs.

Note: The three slides about knowledge fusion and legacy reengineering were presented at the PerMIS'04 Conference, but not at epiSTEME-1.

Legacy Re-engineering Task

Compare three different languages:

- **1.5 million lines of COBOL.**
- **Several hundred JCL scripts.**
- **100 megabytes of English documentation — text files, e-mails, Lotus Notes, HTML, and transcriptions of oral communications.**

Intellitex parser used a different grammar for each language:

- **First translate the structure declarations from COBOL and JCL to conceptual graphs.**
- **VAE was used to find and interpret relevant passages from the English documentation.**
- **Internally all information was represented in CGs.**
- **Output was translated to UML diagrams and English text.**

Results

Job finished in 8 weeks by two programmers, Arun Majumdar and André LeClerc.

- **Four weeks for customization:**
  - **Design and logistics.**
  - **Additional programming for I/O formats.**

- **Three weeks to run Intellitex + VAE + extensions:**
  - **24 hours a day on a 750 MHz Pentium III.**
  - **VAE handled matches with strong evidence.**
  - **Matches with weak evidence were confirmed or corrected by Majumdar and LeClerc.**

- **One week to produce a CD-ROM with integrated views of the results:**

  Glossary, data dictionary, UML diagrams.

Conclusions

Peirce's semiotic is important for analyzing and clarifying the relationships among different methods of reasoning:

1. **Although deduction is important, it is only one of the four methods of reasoning.**
2. **Induction, abduction, and analogy are at least as important, and they are necessary for learning or acquiring new knowledge.**
3. **Current computer systems such as Cyc come close to human ability in deduction.**
4. **But they are far inferior in learning, which depends heavily on the other three methods of reasoning.**

Peirce was also a superb teacher:

> **One math teacher claimed that some students could never learn mathematics. Peirce bet that he could teach the three worst students in the class. After he tutored them, all three became very good at math, and one of them became the best in the entire class.**

Peirce's insights may help revolutionize both cognitive science and methods of teaching.

Related Readings

For further analysis of the knowledge soup, see Chapter 6 of

> Sowa, John F. (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks/Cole Publishing Co., Pacific Grove, CA.

The published paper with the same title:

> The Challenge of Knowledge Soup
> http://www.jfsowa.com/pubs/challenge.pdf

A description of the VivoMind Analogy Engine and the Intellitex parser, co-authored with Arun Majumdar:

> Analogical Reasoning
> http://www.jfsowa.com/pubs/analog.htm

Abstracts and pointers to related topics:

> A Guided Tour of Ontology
> http://www.jfsowa.com/ontology/guided.htm

Model-theoretic foundation of logics with multiple metalevels and nested contexts:

> Laws, facts, and contexts: Foundations for multimodal reasoning
> http://www.jfsowa.com/pubs/laws.htm

Graphic and language interfaces to intelligent systems:

> Graphics and Languages for the Flexible Modular Framework
> http://www.jfsowa.com/pubs/gal4fmf.htm

A description of how the VivoMind Analogy Engine was used to support legacy re-engineering:

> LeClerc, André, & Arun Majumdar (2002) "Legacy revaluation and the making of LegacyWorks," *Distributed Enterprise Architecture* 5:9, Cutter Consortium, Arlington, MA.