

Foundations in Data Semantics

- Chapter 4

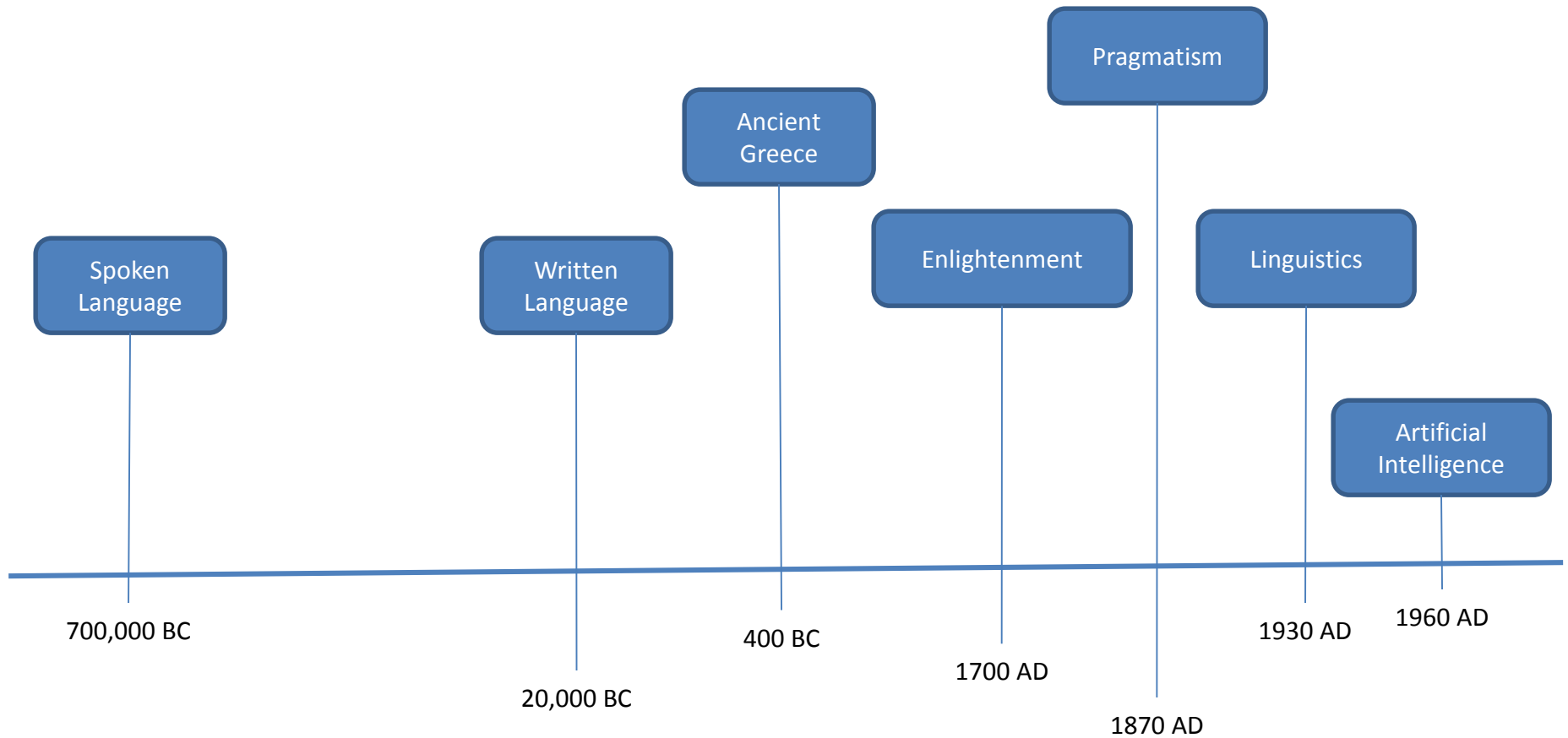
Introduction

- IT is inherently incapable of the analog processing the human brain is capable of. Why?
 - Digital structures consisting of 1s and 0s
 - Rule-based system on absolutes (either 1 or 0)
- Need to make semantics explicit, not leave it to the minds of the programmers
- Need to create “AI for middleware”
 - With optimized and decentralized understanding in communications
 - Without monolithic “intelligent” machines like HAL computer (“2001 Space Odyssey” film in 1968)
- Another bad example: OSI network stack. No semantics!

Introduction

- People and semantics: Is it something new? Have scientists dealt with it before? Where are the roots of semantics?
- Most influential thinkers of humanity contributed to a debate on truth, knowledge, wisdom, for 3000 years

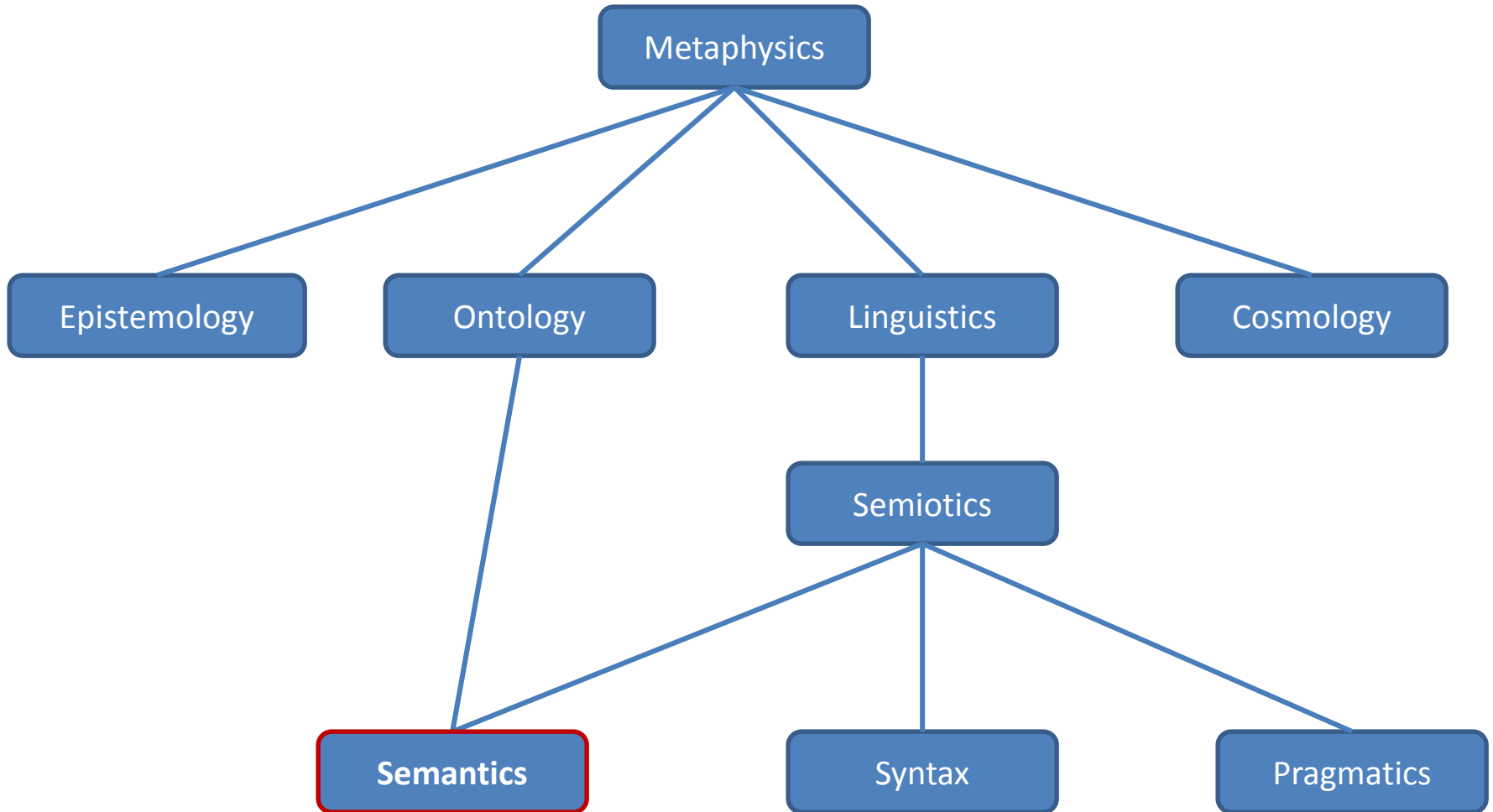
Brief History of Semantics



A Brief History of Semantics

- Spoken lang: from screaming to using words
- Written lang: communication with people who were not present
- Ancient Greece: Finding deeper meaning in words, inference, classification
- Enlightenment: Experimental verification
- Pragmatism: logic on semantics, deductive understanding of nature
- Linguistics: investigation of human languages (definition of category, or type)
- Artificial Intelligence: ontology, inference

The Place of Semantics



The Place of Semantics

Metaphysics: Explain the nature of everything, in particular the relationship of mind to matter

- **Ontology:** Structure, organization and classification of knowledge (but you need to know the meaning first)
 - **Semantics:** Study of meaning
- **Linguistics:** Study of language, sounds, etc.
 - **Semiotics:** Study of signs and symbols as used in language
 - **Semantics:** Study of meaning
 - **Syntax:** How to construct basic grammars (first order logic)
 - **Pragmatics:** Relationship between language (or signs) and context of people using it. Ex: What do we mean by a specific “business rule” in the business context

The Great Debate

- Many famous thinkers in humanity explored the great debate. Some of its topics:
 - Truth
 - Knowledge
 - Logic
 - Wisdom
 - Causality
 - Scientific method
 - Mathematics
 - Aesthetics
 - Physics
 - Relationships
 - Universal and particular
 - etc.

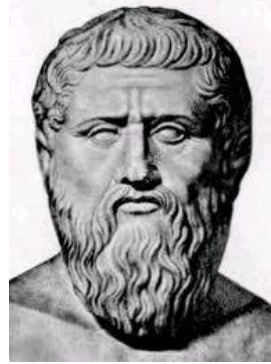
The Great Debate

- What do these topics have to do with information integration and computing?

Philosophical Topic	Implementation
Logic of truth in business rules	Methods and exceptions
Abstraction and (data) modeling	Software classes and objects
Communication of meaning	Interoperability of systems
Global and local subjectivity	Constants and variables
...	...

- ▶ Understanding philosophical context is first step in overcoming IT limitations

Plato (428BC – 348BC)



- Does the essence of reality and truth lie in the ideal or tangible realm?
- Theory of Forms (Ideas):
 - There exists an immaterial universe with perfect aspects of everyday things (table, birds, joy, action, etc.) called Forms or Ideas
 - The objects in our material world are mere shadows of these Forms
 - These Forms (e.g. of a chair) unite all instances of chair in the tangible, physical world

Plato's Cave

- **Reality/truth of the prisoners.** What is reality to the four prisoners?
- **Reality/truth within the cave.** What was the reaction of the freed prisoner when:
 - He saw the fire and the people holding the signs
 - He was told everything he knew before was illusion but now this is reality
 - What was real to him: shadows or surroundings?
- **Reality/truth outside the cave.** What was his reaction when he faced the new “above ground” reality?
 - First he saw the shadows more clearly
 - Then the reflections in the water
 - Then the objects themselves
 - The sun in its proper place

Plato's Cave

- Now he has seen the sun (knows the truth).
 - “Better to be the poorer servant of a poorer master, rather than think as they do...”
- He spread the truth...
- He was ridiculed...
- They would put him to death..
- It is the task of the enlightened to be willing to descend again...

It is the task of the enlightened not only to ascend to learning and to see the good but to be willing to descend again to those prisoners and to share their troubles and their honors, whether they are worth having or not. And this they must do even with the prospect of death.

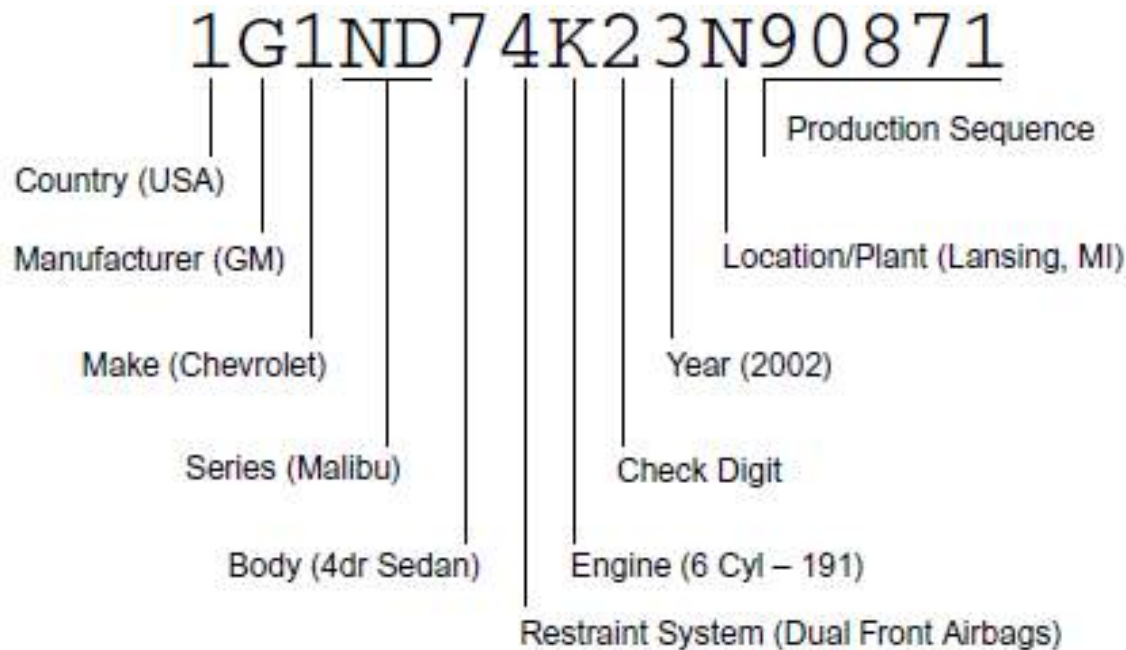
Connection to the Course?

- Plato argues that an Idea (Form) is the ultimate reality (truth), and objects in our material world are shadows of this reality
- Contrast this with objects in Object Oriented systems: A class is the Idea and objects that belong to the class are run-time instances of that class. See the analogy?
- On semantics: Two database tables are named “Customers” (for a grocery db) and “Patients” (for a hospital db). They both represent instances of the same entity/reality/truth (that of a client). However, we need to identify the semantics of these tables to use them correctly.

Aristotle (384BC-322BC)



- Truth and reality lie in what can be measured
- Aristotle's Universe
 - Individual objects (e.g. a falling rock) and systems (e.g. the motion of the planets) subordinate their behavior to an overall plan, or destiny
 - This is mostly apparent in living systems where component parts function in a cooperative way to achieve a final purpose, or end product
- Contrast with Plato:
 - Duality of “abstract” and “concrete”



- Where in software we have the duality of “abstract” and “concrete”?
 - The answer is in slide 14 (think first before you see it...)
- We need to model both
 - Abstract: Business rules, processes, workflow
 - Concrete: Inventory parts, customers, product assemblies

Immanuel Kant (1724 -1804)



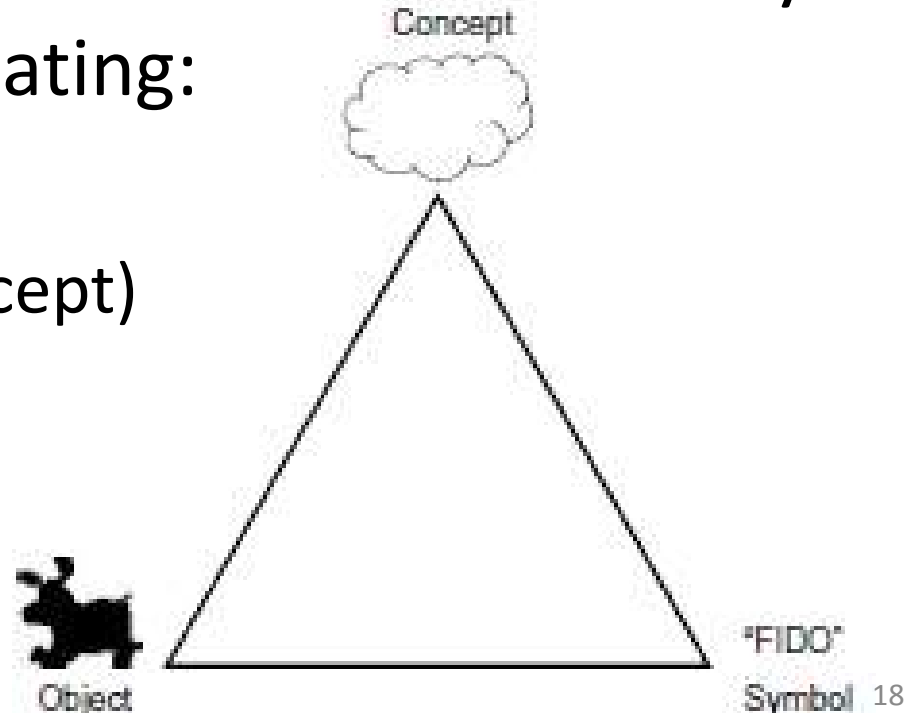
- Tried to answer “What can we know?”
- Father of Information Theory and separated
 - **Data** (sense) from
 - **Semantics** (understanding)
- However, both are needed for complete information
- Information is context dependent and varies from person to person

Charles S. Peirce (1839-1914)



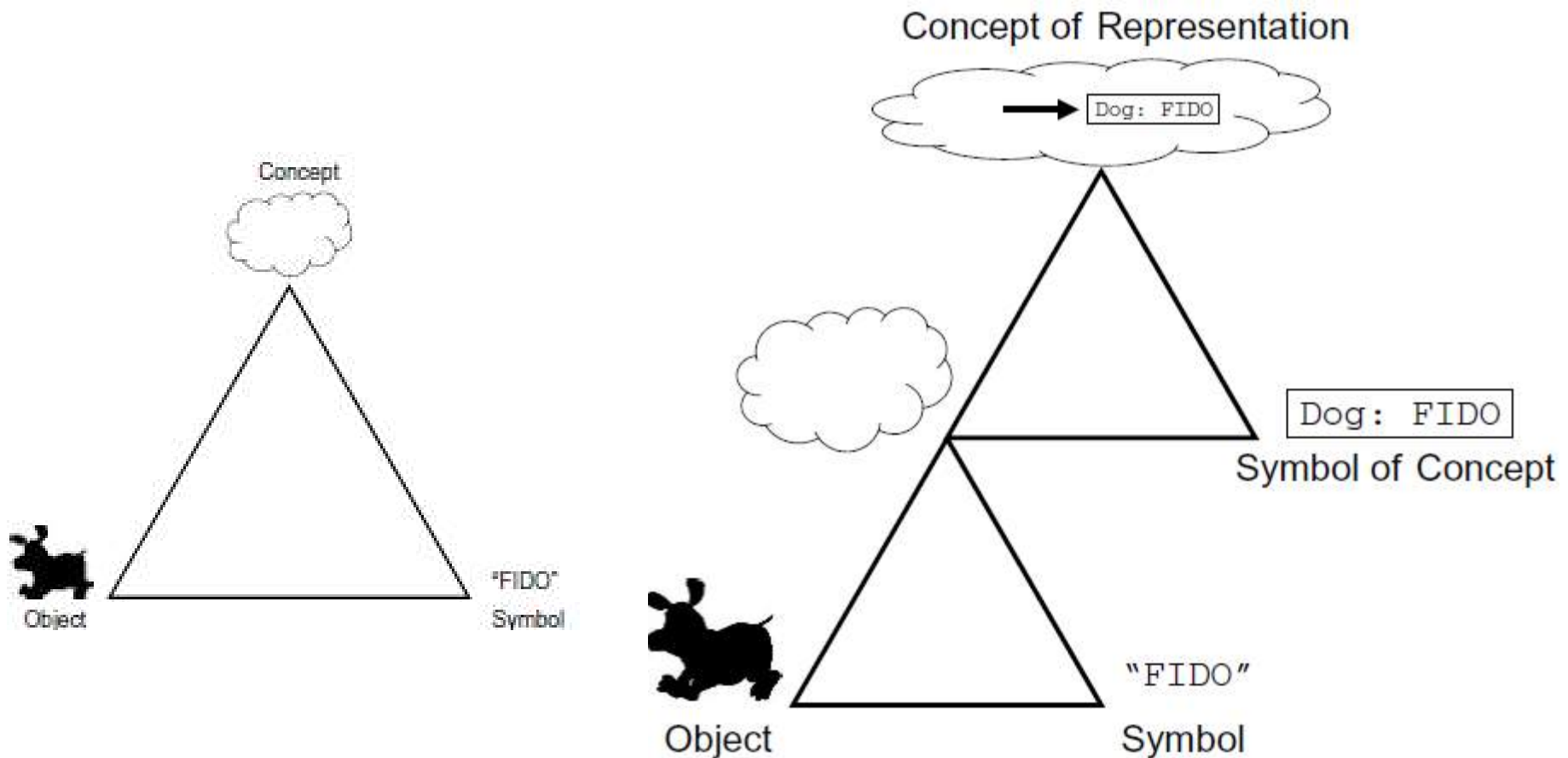
- One of the greatest logicians ever lived
- Pioneered areas in semiotics (study of signs)
- His greatest contributions were to the study of meaning, by triangulating:

- Object
- Representation (concept)
- Referent



Charles S. Peirce (1839-1914)

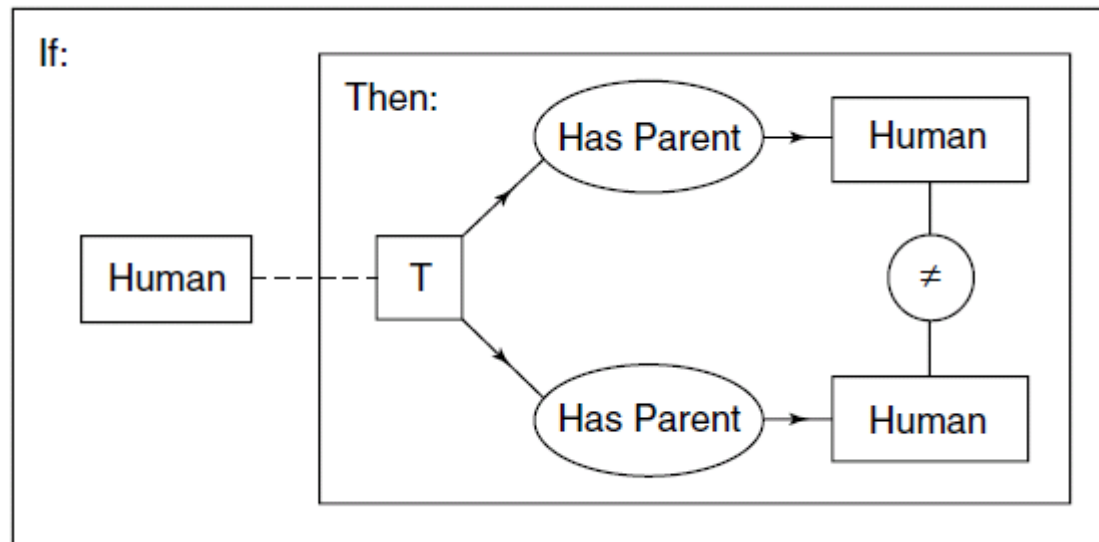
- Nested levels of meaning



John Sowa



- Invented conceptual graphs (CG) based on semantic networks and on Peirce's existential graphs
- Ex. CG: If there is a human s/he has two distinct p



Natural Language and Meaning

- Human mind draws conclusions and connections through means not fully understood by science
- We know we compute, but don't know how...
- Digital communications does not have:
 - Neural or analog processing
 - Adaptability of human mind
 - Ability to handle information overload
- However, computers are capable of utilizing First Order Logic (FOL)

First Order Logic (FOL)

- Science of symbolic logic (logic with symbols) containing:
 - Primitive symbols
 - Axioms
 - Combinations of the above
 - Rules of inference
- Most “semantic technologies” have their roots in FOL
- However, most IT practitioners lack understanding and appreciation of FOL’s importance in IT systems

We Stalled...

- Do you see why there is a gap between natural language meaning and IT systems?



Fuzzy Language

- Analyze this: “John, Bill, and Tom killed each other”
- Natural language is imprecise and context dependent
 - Polysemy (related meanings)
 - Ex: open – unfold, expand, reveal, make openings
 - Homonymy (unrelated meanings – same sound)
 - Ex: bark – by a dog, or surrounding a tree
 - Categorical ambiguity (unrelated meanings – different syntax)
 - Ex: sink – noun (the sink), verb (to sink)

Context and Meaning

- Domain Context
 - Community Culture (folklore, jargon)
 - Business Processes (workflow affects data's meaning)
 - Business Rules (their change affects data's meaning)
 - Data Usage Scenarios (diff apps convey differences)
 - Application functions (data can change meaning as it moves from function to function)
 - Reporting formats (reports are interpreted differently by different people)
 - User interface (influences the meaning of data)


Context and Meaning

- Local Context
 - RDBMS tables (catalog information indicates data's context)
 - Markup tags (reveal context)
 - Data-layer design elements (depth of data hierarchies reveals context)
 - Application-layer design elements (use of inheritance and encapsulation reveals context)
- Understanding the role of context is crucial to understanding semantics in digital systems
- Date: in Procurement vs. in Sales

Data Semantics

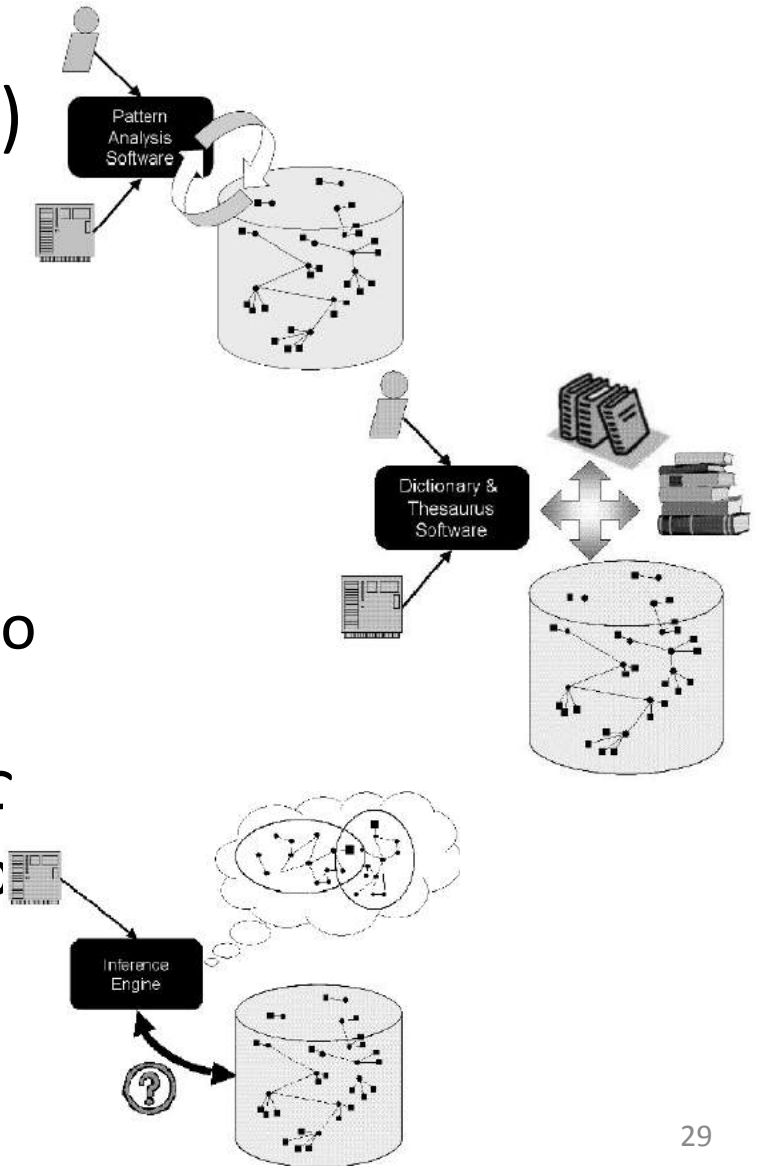
- Definition
- The meaning of data. Meaning is subjective, constrained by the interpreter's context. Semantics are real-time, all the time. Data semantics are implicit, but must be made explicit for data processing. Techniques to make them explicit include
 - Pattern analysis
 - Dictionaries/thesauri
 - Inference
 - Semantic mapping
 - Conceptual graphing

Make Semantics Explicit

- Problem is NOT how to insert semantics in software (programmers do it all the time)
- Problem is how to make them explicit so the underlying knowledge  is available to others who have not participated in the programming process

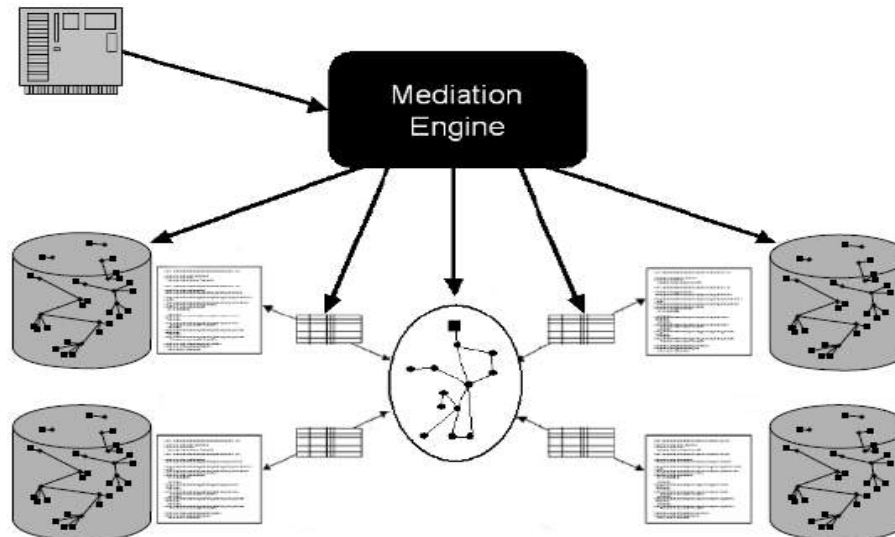
Approaches for Explicit Semantics

- Pattern analysis (data mining)
 - Statistical analysis
 - Artificial intelligence
 - Machine learning
- Definition and synonym relationships
 - Use synonyms and antonyms to infer relationships
- Inference and deductive logic
 - Techniques to find relationships that are not explicit



Approaches for Explicit Semantics

- Context-aware schema mappings
 - Sophisticated routines to make data relationships explicit
 - Ex: A db may be denormalized to gain performance
 - Identify relationships to form schema to schema mappings between denormalized data



What is Information in IT?

$$\text{Information} = \text{Data} + \left(\begin{array}{l} \textit{Semantics} \left(\begin{array}{l} \textit{patterns} + \textit{references} + \textit{relationships} + \\ \textit{logic}(\textit{inductive}, \textit{deductive}, \textit{abductive}) \end{array} \right) \\ \hline \textit{Context} (\textit{local cues} + \textit{domain cues}) \end{array} \right)$$

Logic

- **Abductive:** a is an explanation of b
 - Arriving at a hypothetical explanation a from observation
 - Well-calculated guessing
- **Deductive:** infer b from a
 - b is always a formal consequence of a
- **Inductive:** infer b from a
 - b is not always a consequence of a

Representing Knowledge

- Knowledge representation (KR) encompasses many disciplines and has a broad definition of a schema describing that, which something knows about. It plays many roles:
 - KR is a surrogate – something external to ourselves
 - KR is a set of ontological commitments – how and what we see in the world
 - KR is a medium for efficient computing – to represent what we compute
 - KR is a medium of human expression – affected by personal view

Representing Ontology

- Another broad and fuzzy subject
- Historically, a concept of metaphysical philosophy
- In the computer world, is the study of how to represent knowledge for computing, an “explicit specification of a conceptualization”
- What does this mean?
 - A technically constrained and processable set of data about a collection of concepts describing the world within a given context

Understanding Ontology

- An ontology, then, is an active model that contains a variety of data structures and some way of propagating changes through itself. It can comprise a host of things:
 - Taxonomies of data objects;
 - Taxonomies of relationships or typed links (often expressed as verb phrases), from “is associated with” to “is a kind of” to “contains” or “produces” or “consumes” or even “enjoys” or “prefers” or “burns.”
 - Those relationships can usually be modeled or represented by combining other more elemental components, or through applications that implement (for example) all the things that a customer can do or can have done to her and her account.
- (Another example: Burning is a specific kind of destruction; it is also a chemical process. Which representation you use depends on the context.)

Ontology Types

- **Interface Ontology**— as in service and API interface descriptions
- **Process Ontology**— both fine-grained and coarse-grained procedural descriptions
- **Policy Ontology**— access, privilege, security and constraint rule descriptions
- **Information Ontology**— all things about business contents
- **Industry Ontology**— domain concept descriptions
- **Social/Organizational Ontology**— organizational and social networks
- **Metadata Ontology**—published content descriptions
- **Common Sense Ontology**—general facts about life
- **Representational Ontology**—meta-information about classifications
- **Task Ontology**— term and task matching descriptions

Summary

- The foundations of semantics lie in a 3000+-year debate of philosophy, scientific method, and mathematics
- Understanding meaning is inherently fuzzy, paradoxical, and context dependent
- Semantics in digital systems can be discovered through multiple avenues including pattern analysis, thesauri, inference, semantic mapping, and data nets
- Most information technology contains inherent, but only implicit, semantics
- Semantics are evolutionary—data meanings change over time

End of Chapter 4