# What is Data?

## Data is how we express **observations** in reusable form.

## Human observations are collections of objects, their qualities and their connections. Our observations may be about fine art, scientific experiments, government policies or food.

## Machine representations of human observations – also called data representations – are lists of entities, their properties and their relationships.

## Data representations are inscribed in documents using a notation. A specific notation is a data language - a system of signs and syntax for encoding and decoding data.

## A data language records facts – data in some context. The data language is composed of sentences. Each sentence represents a single **observation**.

## Sentences in data documents are “flattened” using a serialization format that can be shared over an electronic network.

## When combined, the sentences in a shared data document express knowledge.

## With the basic unit of data expression in place - the sentence – we next ask: how to compose a sentence?

## The key components are the subject, predicate, and object. For data representations, these "parts of speech" are also known as the entity, attribute, and value - or - the subject, verb, and object.

# Sentence Construction

## To create a sentence, we need a signing mechanism to identify the subject, predicate, and object.

## At an early age, humans are given tools to identify those terms and use them in the proper context. A machine also needs tools to lookup ("de-reference") and interpret the signs.

## With signs in place, sentences need to be arranged in a systematic order (e.g., subject > predicate > object), according to the rules, or grammar, of a specific data language.

## With signs and grammatical arrangement in place, we need to know the role played by the subject, the predicate, and the object.

## For example, the subject role is the focal point in a sentence. The subject possesses some discernible characteristic (or attribute).

## The object role represents the value of the subject’s characteristic (or attribute).

## The predicate role expresses the intersection between the subject role and the object role.

## Collectively, each of the roles above expresses the meaning of the sentence – the semantics of an **observation**.

## In a data language, the roles express the nature of a data document: the type of connections (a relation) and its membership (the collection of sentences that represent entities that are related in a specific way).

# Properties of Relationship Types

## Two or more entities can be related (connected) in a variety of ways. Relationship types could be reflexive, symmetric, and/or transitive, among others. Here are simple examples:

#### A "knows" relationship type is defined as one that can only have a Person as the subject and another Person as the object.

#### An "employed by" relationship type is defined as one that can only have a Person as the subject and an Organization as the object.

#### A "mother of" relationship type is defined as one that can only have a Female Person as the subject and a Person as the object.

# Challenges for making data machine-readable

## Humans are taught to use text and numbers to communicate knowledge in human-readable form.

## Machines also need to be taught to share and apply knowledge – knowledge encoded in a machine-readable form. But there are challenges that need to be solved to make that possible. The challenges are:

#### Open Standards for digital sentence representation - e.g., the RDF Language from the W3C

#### Open Standards for signs that function like names - e.g., HTTP URIs (hyperlinks) from IETF

#### Open Standards for representing relationship types - e.g., RDF Schema and OWL. These are collections of RDF statements (packaged as a vocabulary or ontology) that describe the nature of different entity-relationship types in machine-comprehensible form.

#### Open Standards-based Query Language for interacting with relations represented as fine-grained sentence-graphs, rather than as coarse-grained records in a table.

#### Database Management Systems that can manage data represented as sentence-graphs, using the open standards above.

# Conclusion

## Data lies at the root of Language and is effectively mankind's most powerful tool.

## Our next challenge is getting computers to augment our existing ability to communicate via sentences and statements.

## The use of **very basic sentences** would escalate the power of computing and overall human productivity to new levels, as already demonstrated by what's happened thus far on the World Wide Web. There are plenty of words on the Web, but no implicit sense of whether they are nouns, verbs, or adjectives, or whether they form sentences.

## Ultimately, the “semantics of relations” is the key to machine augmented knowledge. These “semantics of relations” are how we encode and decode data.

## When properly applied we can put data into perspective – a **perspective** context.

## The next step is to create machine readable knowledge by putting that same data into a **cognitive** context.