An Introduction to

Natural Language Processing
and
Machine Learning

Karthik Sankar
Department of CSE
NIT Trichy

Natural Language Processing

A lot of human communication is by means of natural language

So computers could be a ton more useful if they could read our email, do our library research, chat to us, do all of these things involve dealing with natural language

They're pretty good at dealing with machine languages that are made for them, but human languages, not so.

“Look. The computer just can't deal with the kind of stuff that humans produce, and how they naturally interact”

We’re exploiting human cleverness rather than working out how to have computer cleverness.

Definition

NLP is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages

Categories

❖ Phonology	-	study of speech sounds
❖ Morphology	-	study of meaningful components of words
❖ Syntax	-	study of structural relationships between words
❖ Semantics	-	study of meaning

Phonology

Modeling the pronunciation of a word as a string of symbols – PHONES

Articulatory Phonetics: How phones are produced as the various organs in the mouth, throat and nose modify the airflow from the lungs.

Can
Chair
Coach

Syllables

Morphology

Identification, analysis and description of the structure of words.

Inflections

Number
Tense
Case
Gender
Person

dog/dogs : goose/geese
hunt – hunted
his - hers

Word Formation

mother in law
hot dog

Finite State Machines
Finite State Transducers

Syntax

Part of Speech Tagging

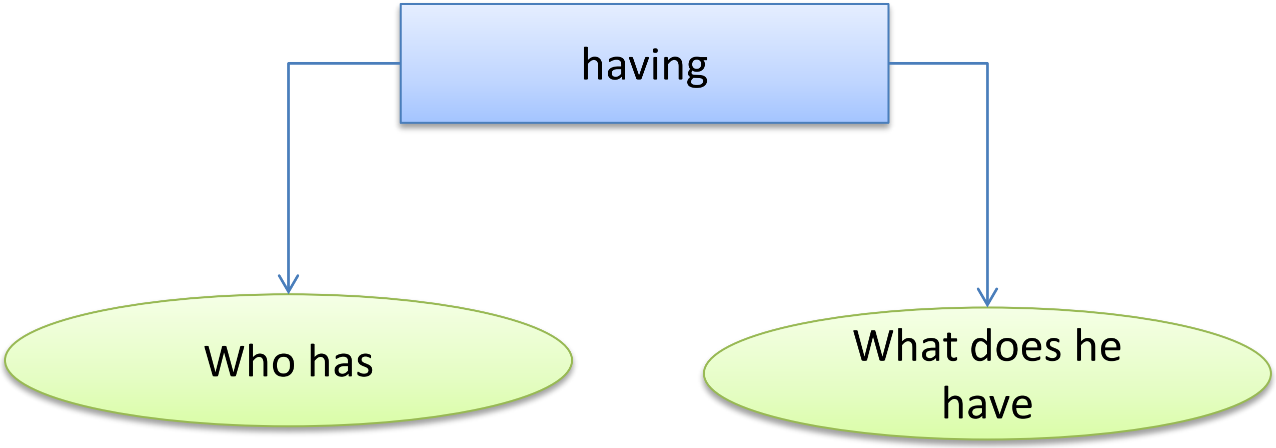
Noun
Verb
Adjective …

I can write – aux. verb **OR** verb **OR** noun

Context Free Grammars

Semantics

Understanding and representing the meaning



First Order Predicate Calculus

Has(Ram, book)

Ambiguity

Adjective: the adjectives are associated with which of the two nouns ?
"pretty little girls' school"

Pronoun: which noun does 'they' relate to ?
We gave the monkeys the bananas because *they* were hungry.
We gave the monkeys the bananas because *they* were over-ripe.

Emphasis: notice the change in meaning due to the change in stress
I never said she stole my money
I **never** said she stole my money
I never **said** she stole my money
I never said **she** stole my money
I never said she **stole** my money
I never said she stole **my** money
I never said she stole my **money**

Ambiguity - contd

Fed raises interest rates half a percent in effort to control inflation

She rates highly
Our water rates are high

Japanese movies interest me
The interest rate is 8 percent

Fed raises
The raises we received was small

Resolving Ambiguity

- ✓ Part of Speech Tagging
- ✓ Word Sense Disambiguation
- ✓ Probabilistic Parsing
- ✓ Speech Act Interpretation

Perceptions

Perception provides agents with information about the world they inhabit.

Perception is initiated by sensors.

A sensor is anything that can record some aspect of the environment and pass it as input to an agent program.

The sensor could be as simple as a one-bit sensor that detects whether a switch is on or off or as complex as the retina of the human eye, which contains more than a hundred million photosensitive elements

- Image processing
- Computer Vision
- Speech recognition
- Facial recognition
- Object recognition

Applications

- ❖ Information retrieval & Web Search

  Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching databases and the World Wide Web.

- ❖ Information Extraction

  Information extraction (IE) is a type of information retrieval whose goal is to automatically extract structured information, i.e. categorized and contextually and semantically well-defined data from a certain domain, from unstructured machine-readable documents

- ❖ Question Answering

  Type in keywords to Asking Questions in Natural Language.
  Response from documents to extracted or generated answer

- ❖ Text Summarization

  Process of distilling most important information from a source to produce an abridged version

- ❖ Machine Translation

  use of computer software to translate text or speech from one natural language to another.

Applications

- Speech - recognition & synthesis

  Deriving a textual representation of a spoken utterance

- Natural Language understanding and generation

  NLG system is like a translator that converts a computer based representation into a natural language representation.

- Human - Computer Conversation

  Dialogue between humans and computers using natural language.

- Text Generation

  A method for generating sentences from “keywords” or “headwords”.

- Hand writing recognition

  Ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices

Machine Learning

Machine Learning

The ability to learn

- Learning something new
- Learning something new about something you already knew
- Learning how to do something better, either more efficiently or with more accuracy

A system can improve its problem solving accuracy (and possibly efficiency) by learning how to do something better

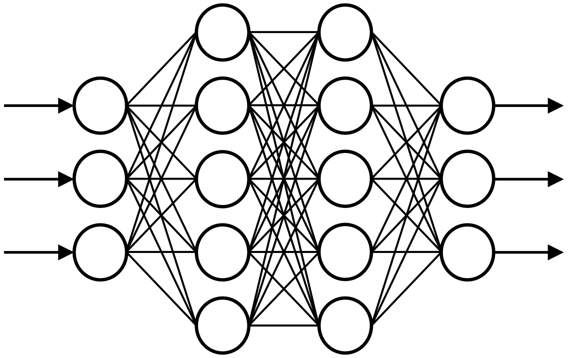
Types of Machine Learning - 1

Symbolic

Explicitly represented Domain knowledge

Sub-Symbolic or Connectionist Networks

- Neural Networks
- simulate the structure and/or functional aspects of biological neural networks
- Simple processing elements (neurons), which can exhibit complex global behaviour, determined by the connections between the processing elements and element parameters



Genetic and Evolutionary Learning

Learning through adaptation

Types of Machine Learning - 2 - “is there a teacher ???”

Supervised

Training data is available

Unsupervised

Training data is not available. Self learning process

Reinforcement

how an agent ought to take actions in an environment so as to maximize some notion of long-term reward

Types of Machine Learning - 3

Knowledge acquisition

Learning through problem solving

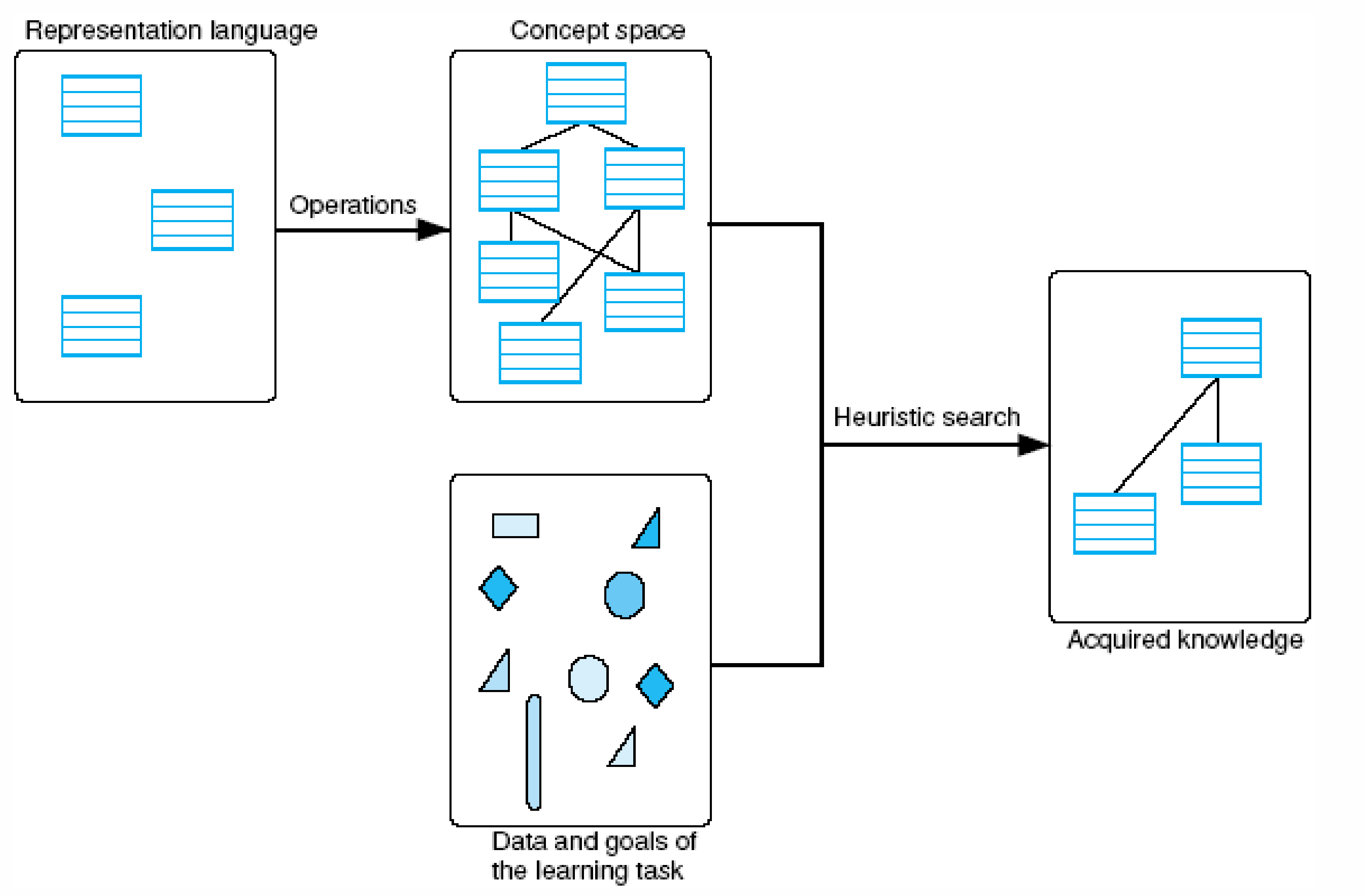
Explanation based learning

Analogy

Framework for Symbol Based Learning

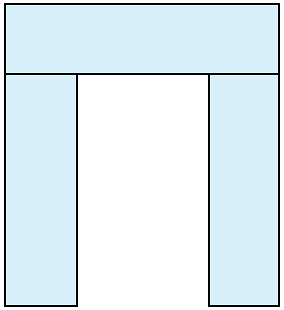
- Data and the goals of the learning task
- The representation of Learned Language
- A set of operations
- The concept space
- Heuristic Search

Framework for Symbol Based Learning

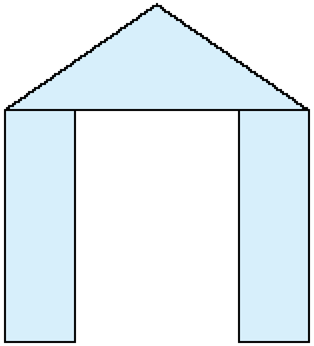


Example - the goal is to build an arch

positive



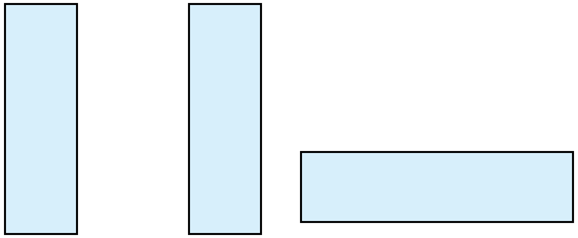
Arch



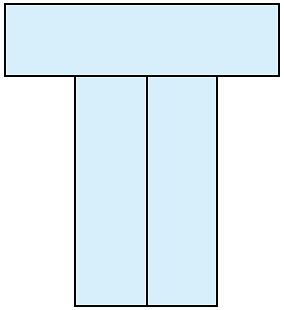
Arch

positive

negative



Near miss

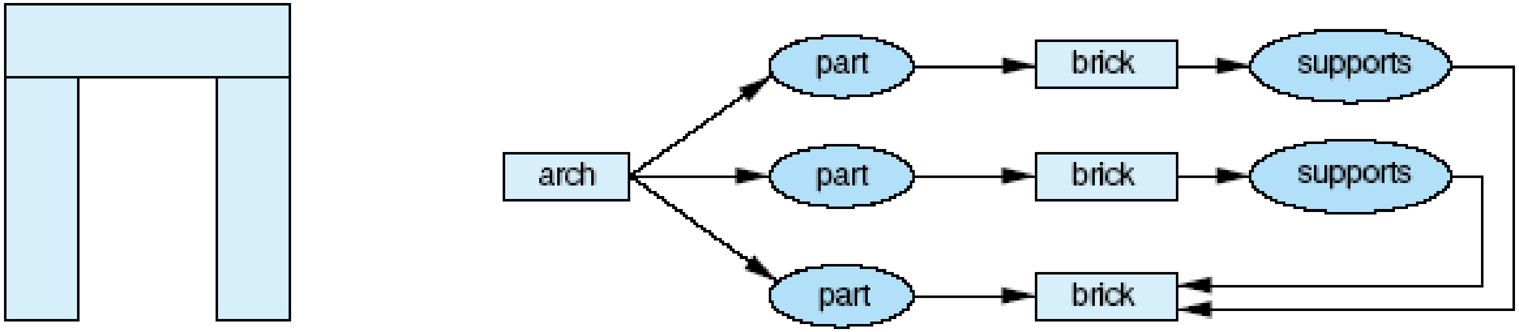


Near miss

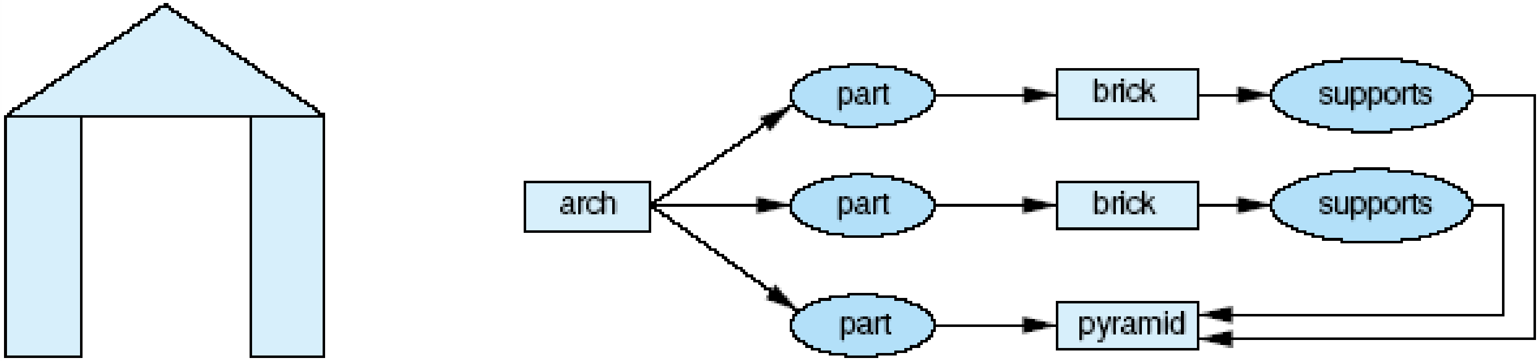
negative

Example

a. An example of an arch and its network description

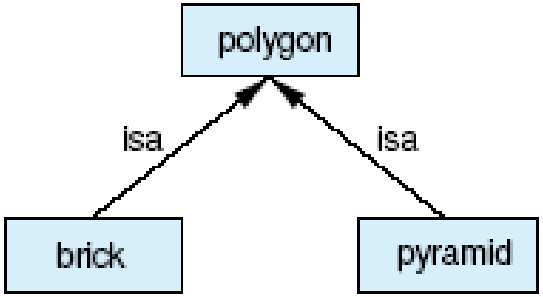


b. An example of another arch and its network description

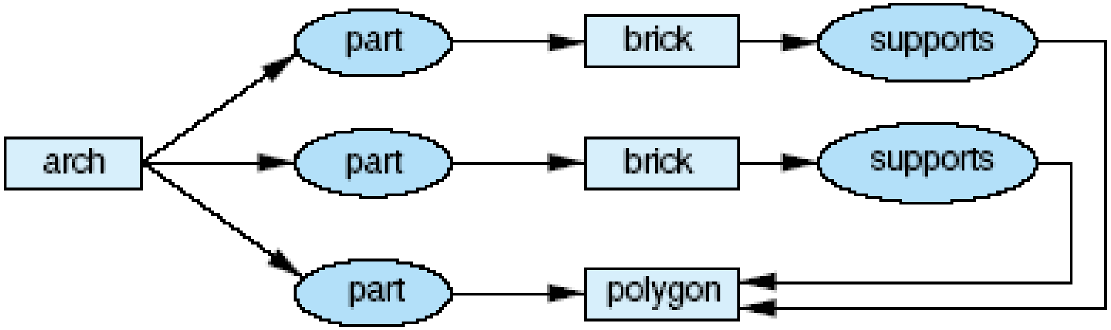


Example

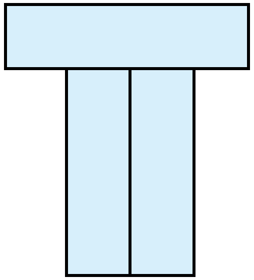
c. Given background knowledge that bricks and pyramids are both types of polygons

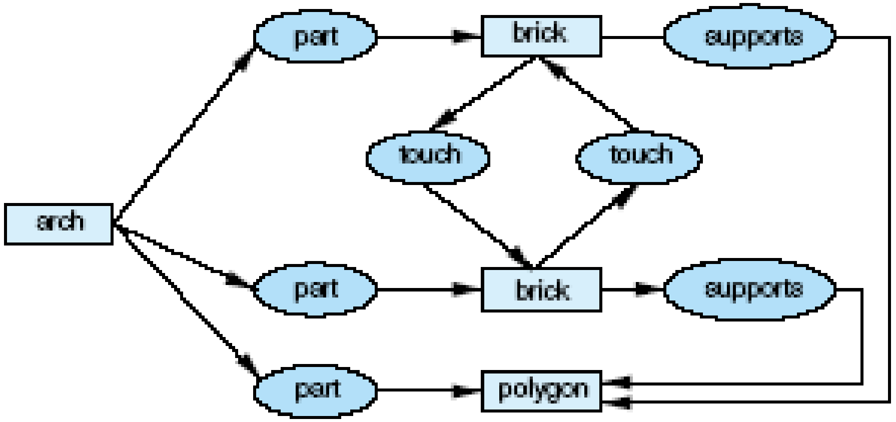


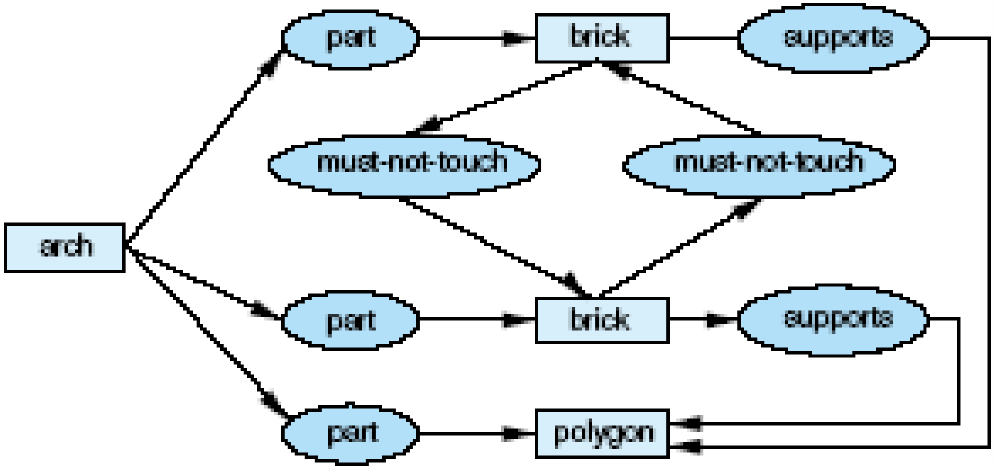
d. Generalization that includes both examples



Example

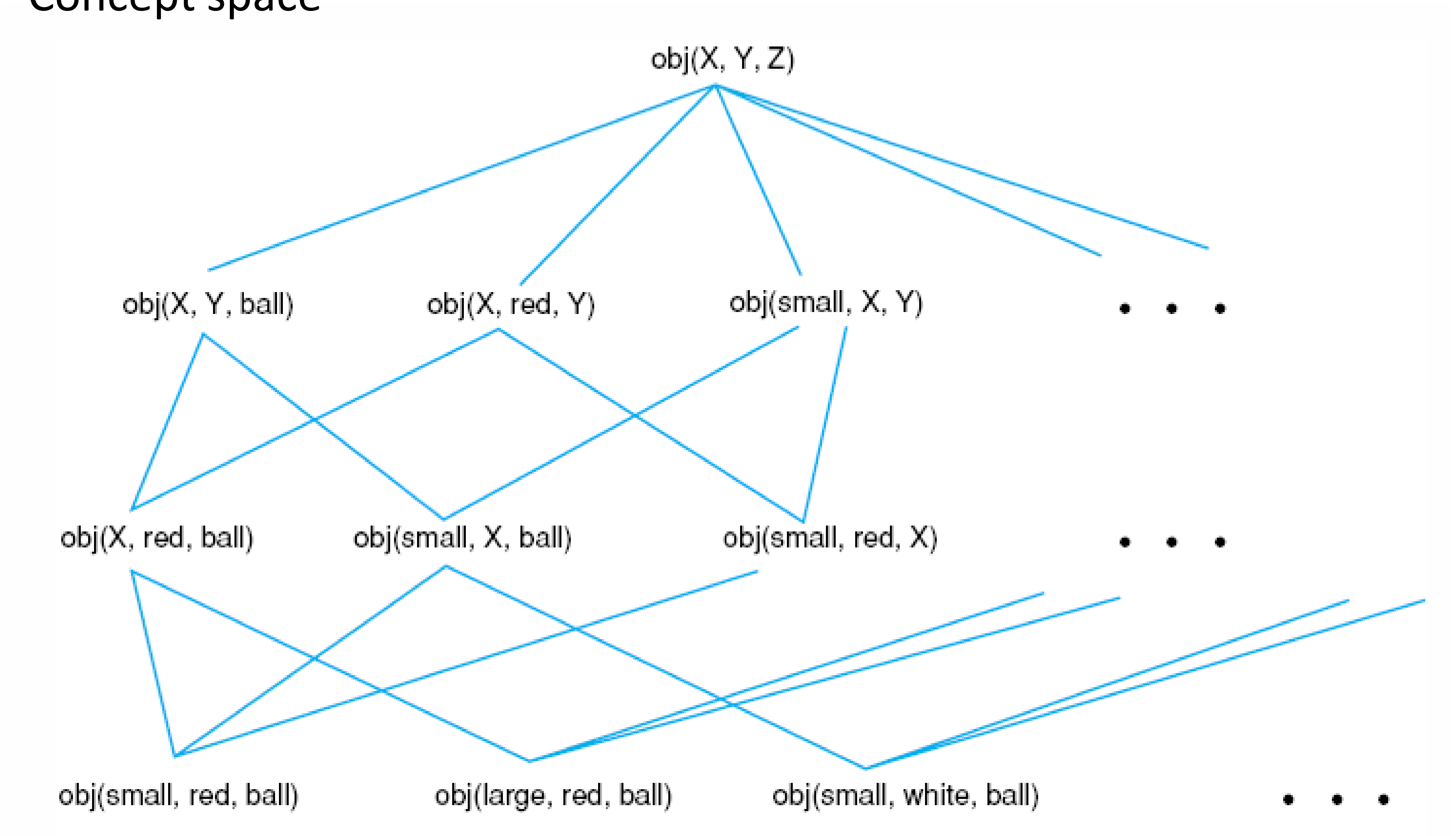



part
brick
supports
touch
touch
arch
part
brick
supports
part
polygon


part
brick
supports
must-not-touch
must-not-touch
arch
part
brick
supports
part
polygon

Version Space Search

Concept space



Version Space Search

Generalization Operations

Color(ball, red)
generalizes to Color(X, red)

Shape(X, round) ^ Size(X, small) ^ Color(X, red)
generalizes to Shape(X, round) ^ Size(X, small)

Covering

p *covers* q

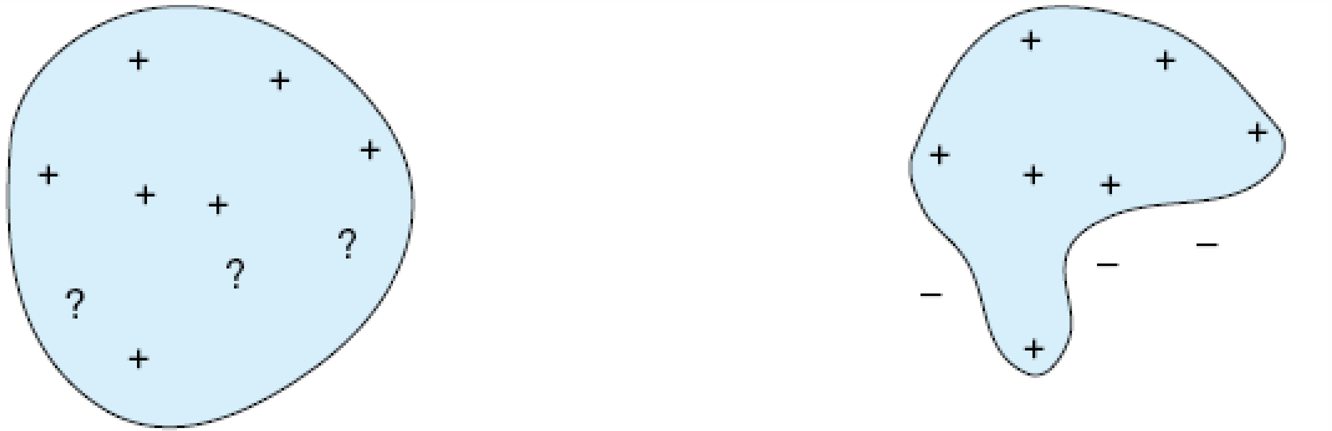
Version Space Search

Candidate Elimination Algorithm

- Specific to general direction
- General to specific direction
- Bi-directional

Version Space Search

Role of negative examples



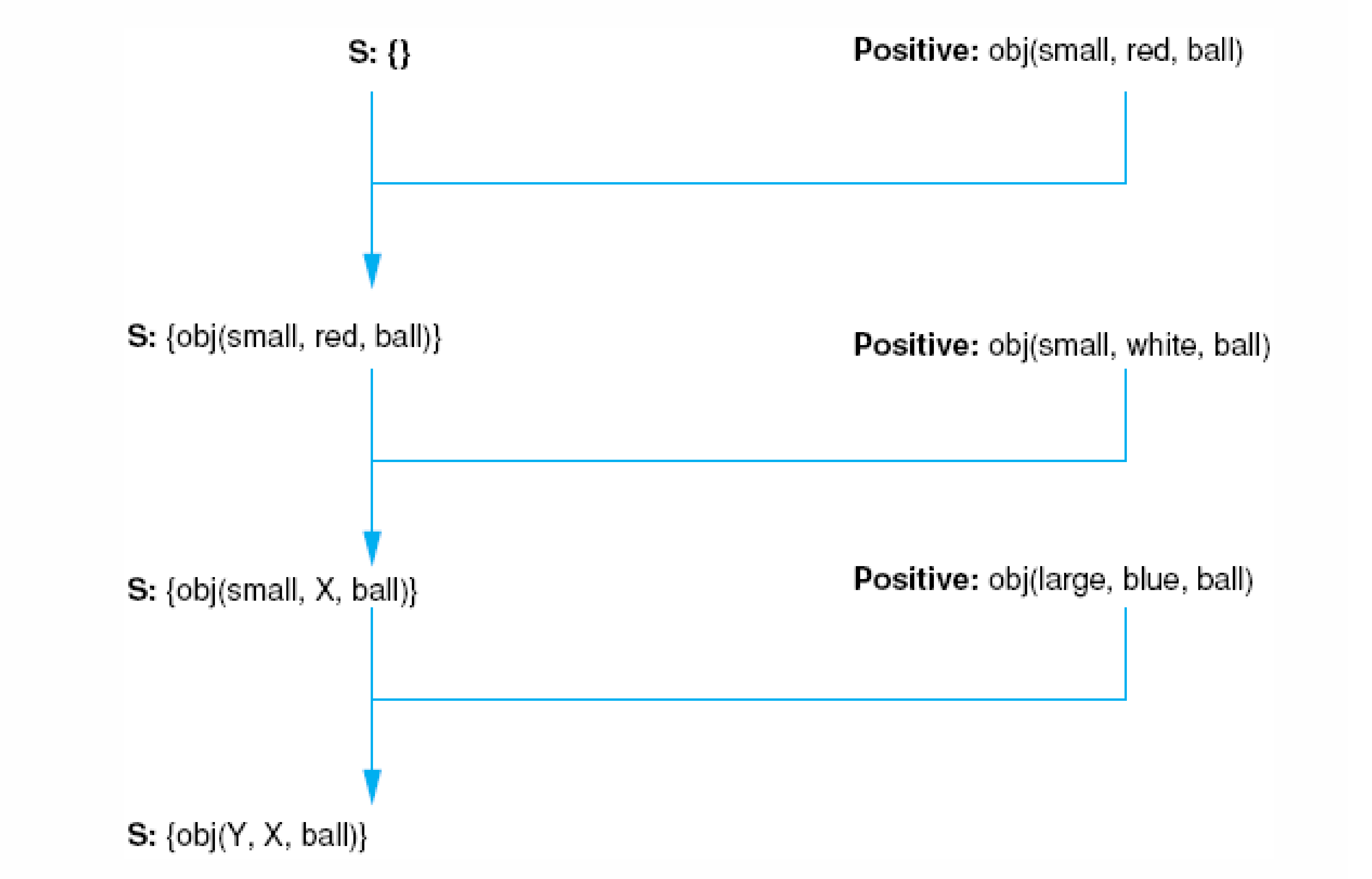
Concept induced from positive examples only

Concept induced from positive and negative examples

Version Space Search - Specific to general direction

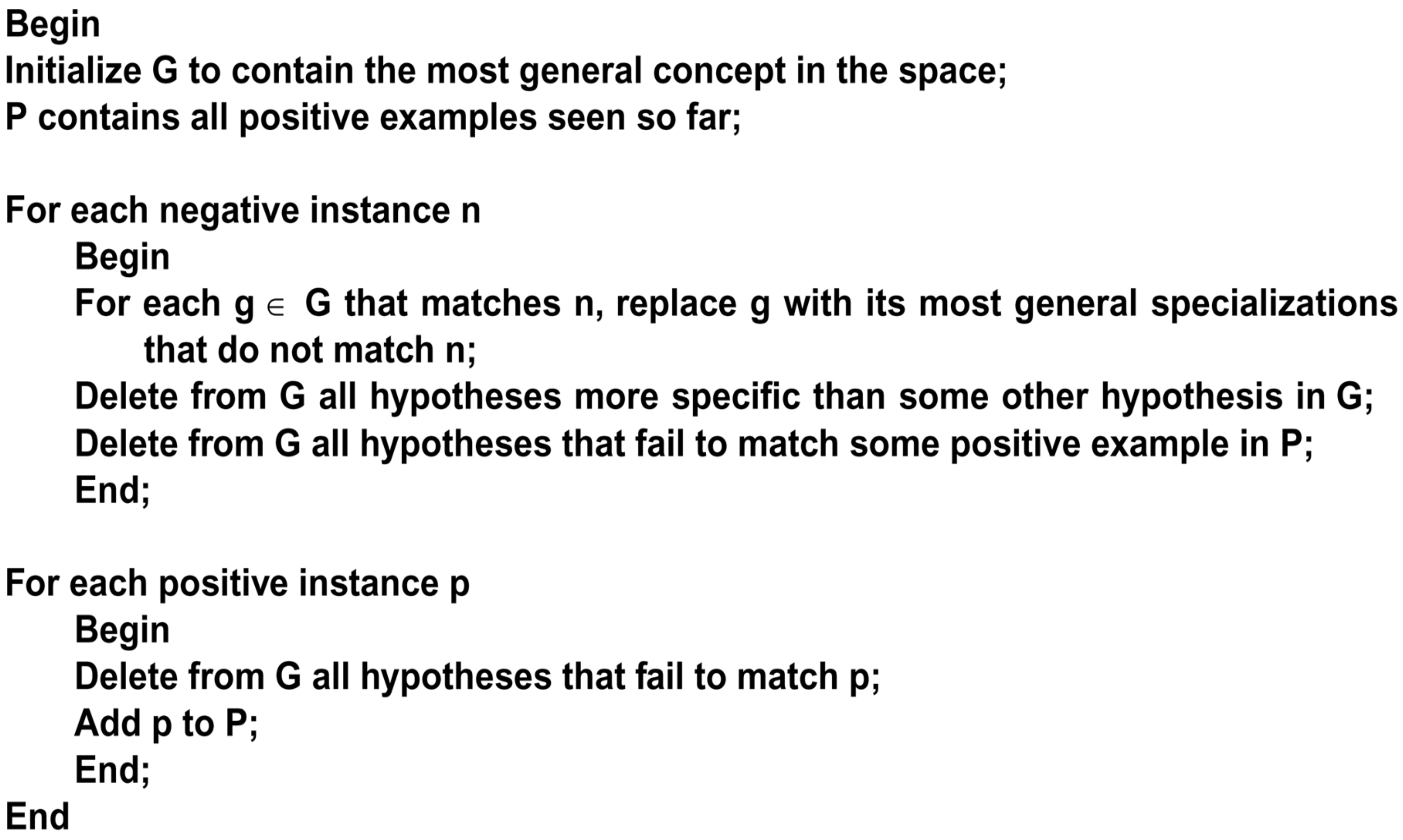
```
Begin
Initialize S to the first positive training instance;
N is the set of all negative instances seen so far;

For each positive instance p
    Begin
    For every s ∈ S, if s does not match p, replace s with its most specific
        generalization that matchs p;
    Delete from S all hypotheses more general than some other hypothesis in S;
    Delete from S all hypotheses that match a previously observed negative
        instance in N;
    End;
For every negative instance n
    Begin
    Delete all members of S that match n;
    Add n to N to check future hypotheses for overgeneralization;
    End;
End
```

Version Space Search - Specific to general direction - example



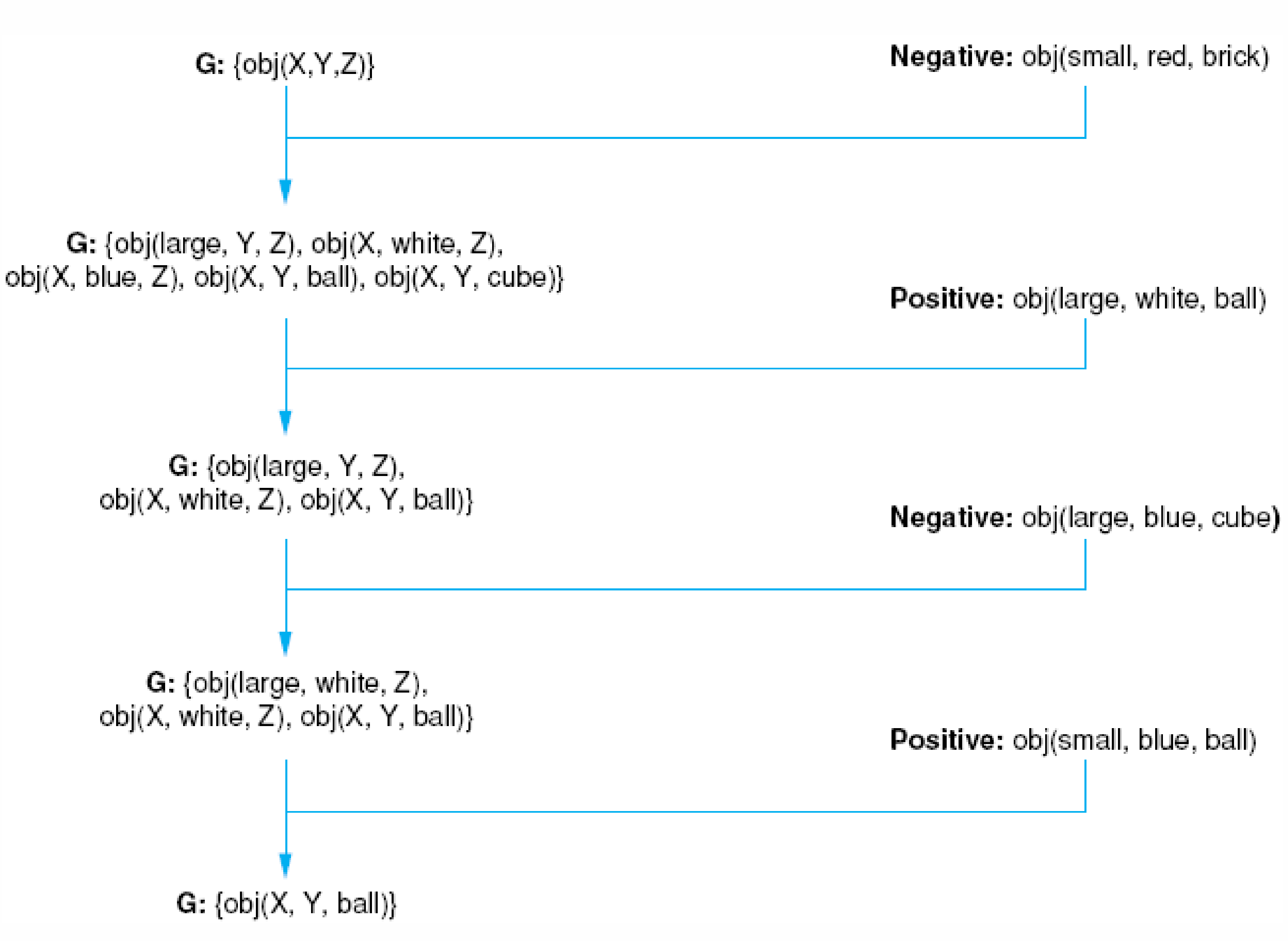
S: {}
Positive: obj(small, red, ball)
S: {obj(small, red, ball)}
Positive: obj(small, white, ball)
S: {obj(small, X, ball)}
Positive: obj(large, blue, ball)
S: {obj(Y, X, ball)}

Version Space Search - General to specific direction



```
Begin
Initialize G to contain the most general concept in the space;
P contains all positive examples seen so far;

For each negative instance n
    Begin
    For each g ∈ G that matches n, replace g with its most general specializations
        that do not match n;
    Delete from G all hypotheses more specific than some other hypothesis in G;
    Delete from G all hypotheses that fail to match some positive example in P;
    End;

For each positive instance p
    Begin
    Delete from G all hypotheses that fail to match p;
    Add p to P;
    End;
End
```

Version Space Search - General to specific direction - example



G: {obj(X,Y,Z)}

Negative: obj(small, red, brick)

G: {obj(large, Y, Z), obj(X, white, Z), obj(X, blue, Z), obj(X, Y, ball), obj(X, Y, cube)}

Positive: obj(large, white, ball)

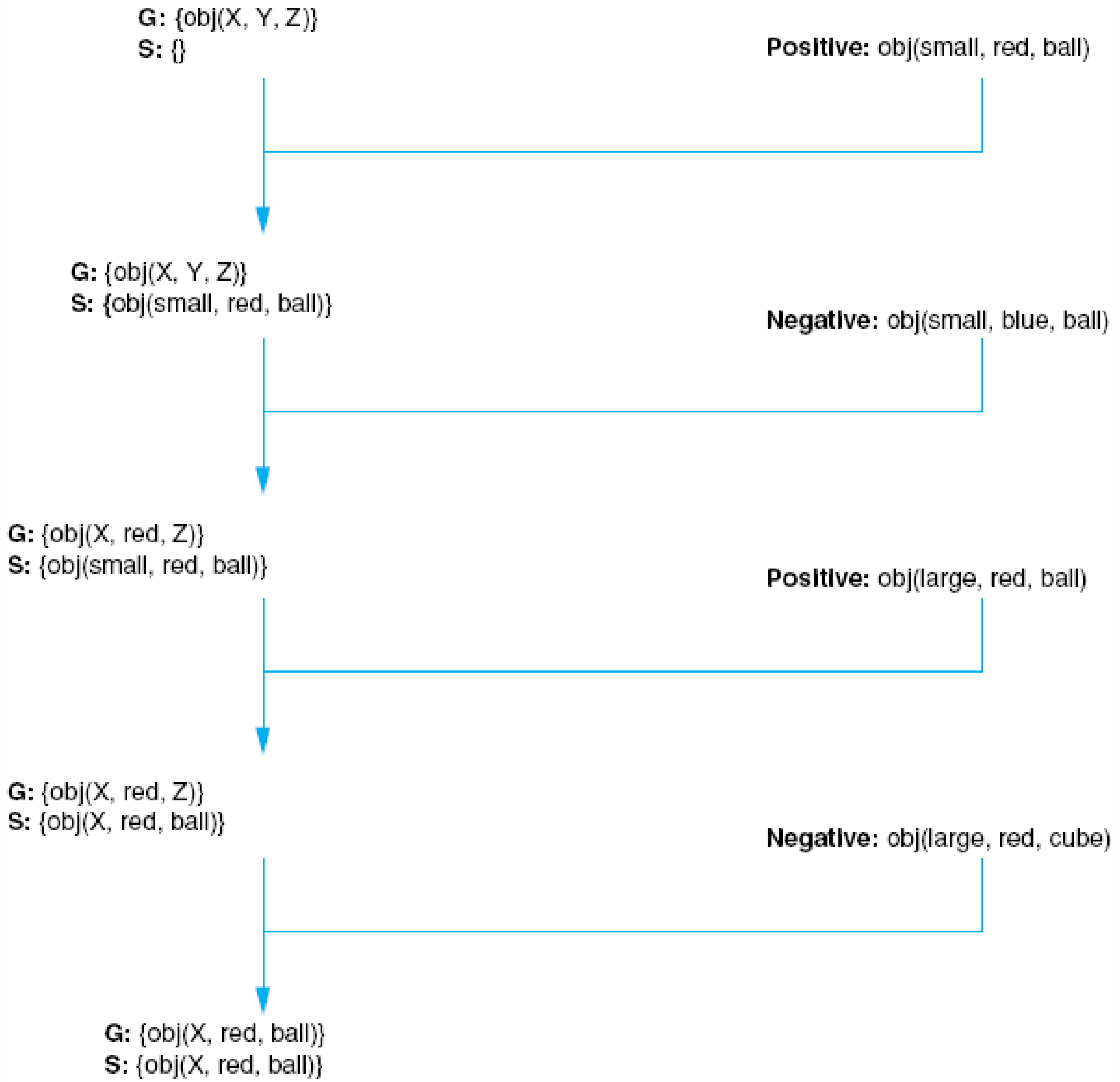
G: {obj(large, Y, Z), obj(X, white, Z), obj(X, Y, ball)}

Negative: obj(large, blue, cube)

G: {obj(large, white, Z), obj(X, white, Z), obj(X, Y, ball)}

Positive: obj(small, blue, ball)

G: {obj(X, Y, ball)}



G: {obj(X, Y, Z)}
S: {}

Positive: obj(small, red, ball)

G: {obj(X, Y, Z)}
S: {obj(small, red, ball)}

Negative: obj(small, blue, ball)

G: {obj(X, red, Z)}
S: {obj(small, red, ball)}

Positive: obj(large, red, ball)

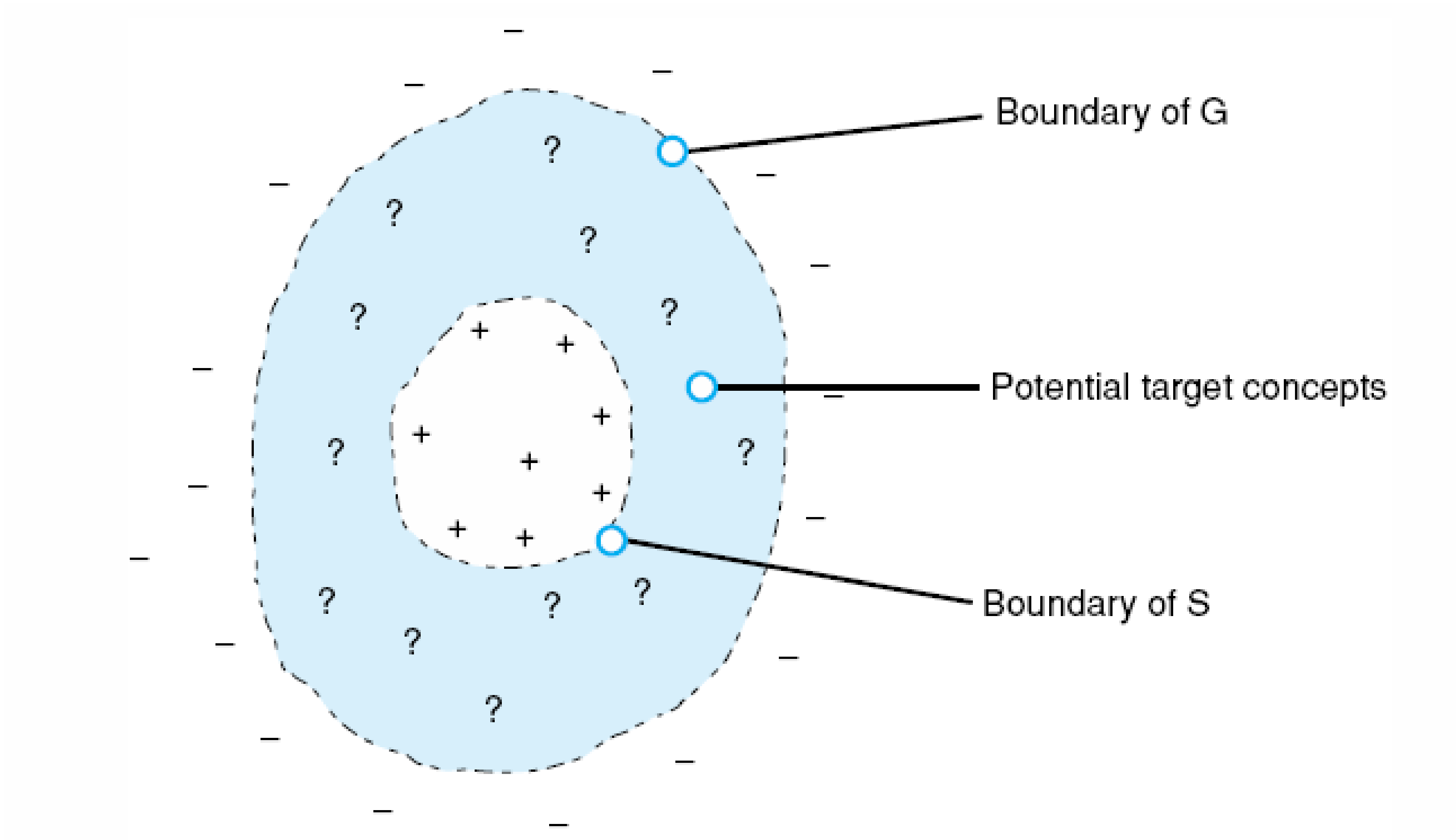
G: {obj(X, red, Z)}
S: {obj(X, red, ball)}

Negative: obj(large, red, cube)

G: {obj(X, red, ball)}
S: {obj(X, red, ball)}

Version Space Search

How the algorithm works



Thank you