Knowledge-Based Agents

Chapter 7.1-7.3

Some material adopted from notes by Andreas Geyer-Schulz and Chuck Dyer

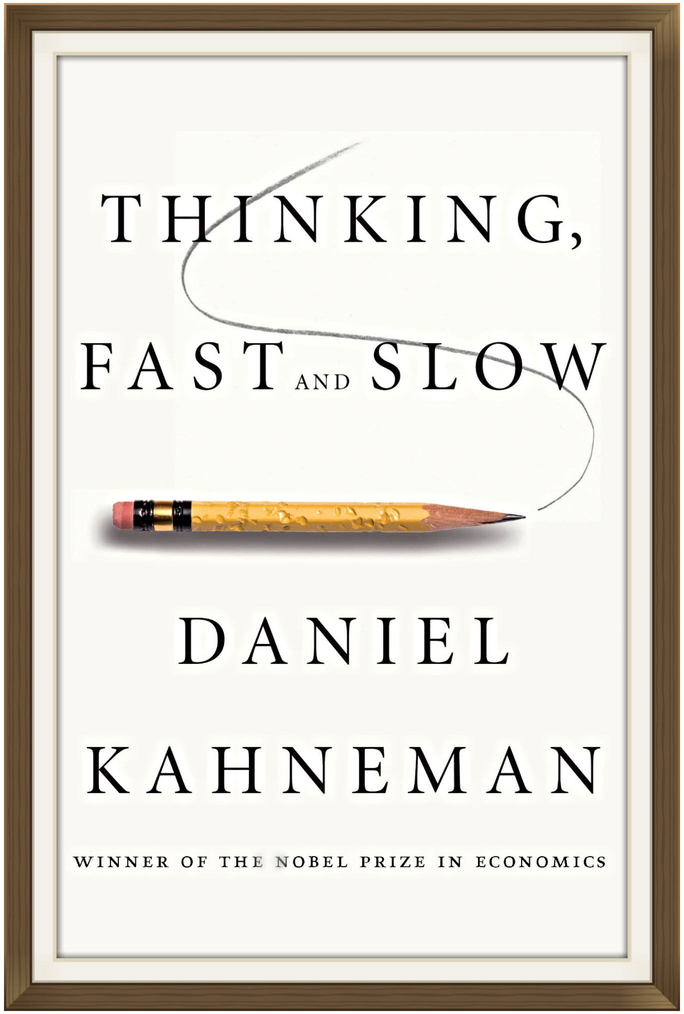
Big Idea

- Drawing reasonable conclusions from a set of data (observations, beliefs, etc.) seems key to intelligence
- Logic is a powerful and well-developed approach to this & highly regarded by people
- Logic is also a strong formal system that computers can use (cf. John McCarthy’s work)
- We can solve some AI problems by representing them in logic and applying standard proof techniques to generate solutions

Inference in People

- People can do logical inference, but are not always very good at it
- Reasoning with negation and disjunction seems particularly difficult
- But people seem to employ many kinds of reasoning strategies, most of which are neither *complete* nor *sound*

Thinking Fast and Slow



- A popular 2011 book by a Nobel prize winning author
- His model is we have two different types of reasoning facilities
- **System 1** operates automatically and quickly, with little or no effort and no sense of voluntary control
- **System 2** allocates attention to effortful mental activities that demand it, including complex computations (e.g., logic, arithmetic, writing software, etc.)

SYSTEM 1

Intuition & instinct

95%

Unconscious
Fast
Associative
Automatic pilot

SYSTEM 2

Rational thinking

5%

Takes effort
Slow
Logical
Lazy
Indecisive

Source: Daniel Kahneman

Question #1

Here is a simple puzzle

Don’t try to solve it -- listen to your intuition

Question #1

Here is a simple puzzle

Don’t try to solve it -- listen to your intuition

- A bat and ball cost $1.10
- The bat costs one dollar more than the ball
- How much does the ball cost?

Question #1

Here is a simple puzzle

Don’t try to solve it -- listen to your intuition and type an answer into the chat

- A bat and ball cost $1.10
- The bat costs one dollar more than the ball
- How much does the ball cost?

The ball costs $0.05

Question #2

Try to determine, as quickly as you can, if the argument is logically valid. Does the conclusion follow the premises?

Question #2

Try to determine, as quickly as you can, if the argument is logically valid. Does the conclusion follow the premises?

- **All roses are flowers**
- **Some flowers fade quickly**
- **Therefore, some roses fade quickly**

Question #2

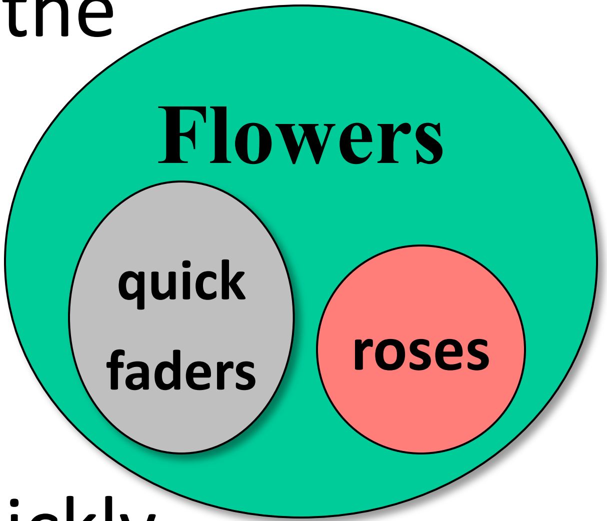
Try to determine, as quickly as you can, if the argument is logically valid. Does the conclusion follow the premises?

- All roses are flowers
- Some flowers fade quickly
- Therefore, some roses fade quickly

It is possible that there are no roses among the flowers that fade quickly

Question #2

Try to determine, as quickly as you can, if the argument is logically valid. Does the conclusion follow the premises?



- All roses are flowers
- Some flowers fade quickly
- Therefore, some roses fade quickly

It is possible that there are no roses among the flowers that fade quickly

Question #3

It takes 5 machines 5 minutes to make 5 widgets

How long would it take 100 machines to make 100 widgets?

Question #3

It takes 5 machines 5 minutes to make 5 widgets

How long would it take 100 machines to make 100 widgets?

- **100 minutes or 5 minutes?**

Question #3

It takes 5 machines 5 minutes to make 5 widgets

How long would it take 100 machines to make 100 widgets?

- 100 minutes or 5 minutes?

5 minutes

Wason Selection Task

- I have a pack of cards; each has a *letter* written on one side and a *number* on the other
- I claim the following rule is true:

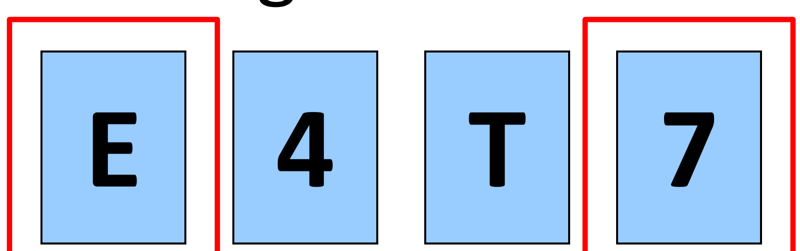
  If a card has a *vowel* on one side, then it has an *even number* on the other
- Which cards should you turn over in order to decide whether the rule is true or false?

E	4	T	7

Wikipedia

Wason Selection Task

- Wason (1966) showed people are bad at this task
- To disprove rule P=>Q, find a situation in which P is true but Q is false, i.e., show P^~Q
- To disprove **vowel => even**, find a card with a vowel and an odd number
- Thus, turn over the cards showing **vowels** and those showing **odd numbers**



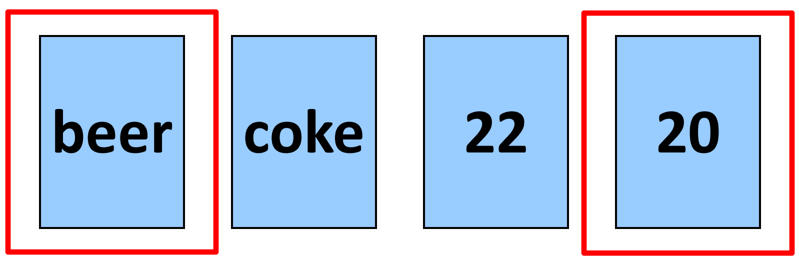
Wason Selection Task

- This version is easier for people, as shown by Griggs & Cox, 1982
- You are the bouncer in a bar; which of these people do you card given the rule: *You must be 21 or older to drink beer.*

beer	**coke**	**22**	**20**

Wason Selection Task

- This version is easier for people, as shown by [Griggs & Cox, 1982]
- You are the bouncer in a bar; which of these people do you card given the rule: *You must be 21 or older to drink beer.*



Perhaps easier because it's more familiar or because people have special strategies to reason about certain situations, such as cheating in a social situation

Negation in Natural Language

- We often model the meaning of natural language sentences as a logic statements
- This maps these into equivalent statements
  - All elephants are gray
  - No elephant are not gray
- Double negation is common in informal language: th*at won't do you no good*

As a way to state a negative more strongly

Negation in Natural Language

- It's not just informal language actually
- What does this mean:

  we cannot underestimate the importance of logic

- Does is mean logic is important or not?

- See the LanguageLog blog misnegation archive for lots of real-world examples

Logic as a Methodology

Even if people don't use formal logical reasoning for solving a problem, logic might be a good approach for AI for many reasons

- Airplanes don't need to flap their wings
- Logic may be a good implementation strategy
- Solution in a formal system offers other benefits, e.g., letting us prove properties of the approach

- See neats vs. scruffies

Knowledge-based agents

- Knowledge-based agents have a knowledge base (KB) and an inference system
- KB: a set of representations of facts believed true
- Each individual representation is called a **sentence**
- Sentences are expressed in a **knowledge representation language**
- The agent operates as follows:
    1. It **TELL**s the KB what it perceives
    2. It **ASK**s the KB what action it should perform
    3. It performs the chosen action

Architecture of a KB agent

- **Knowledge Level**
  - Most abstract: describe agent by what it knows
  - Ex: Autonomous vehicle knows Golden Gate Bridge connects San Francisco with the Marin County
- **Logical Level**
  - Level where knowledge is encoded into *sentences*
  - Ex: **links**(GoldenGateBridge, SanFran, MarinCounty)
- **Implementation Level**
  - Software representation of sentences, e.g. `(links,goldengatebridge,sanfran,marincounty)`

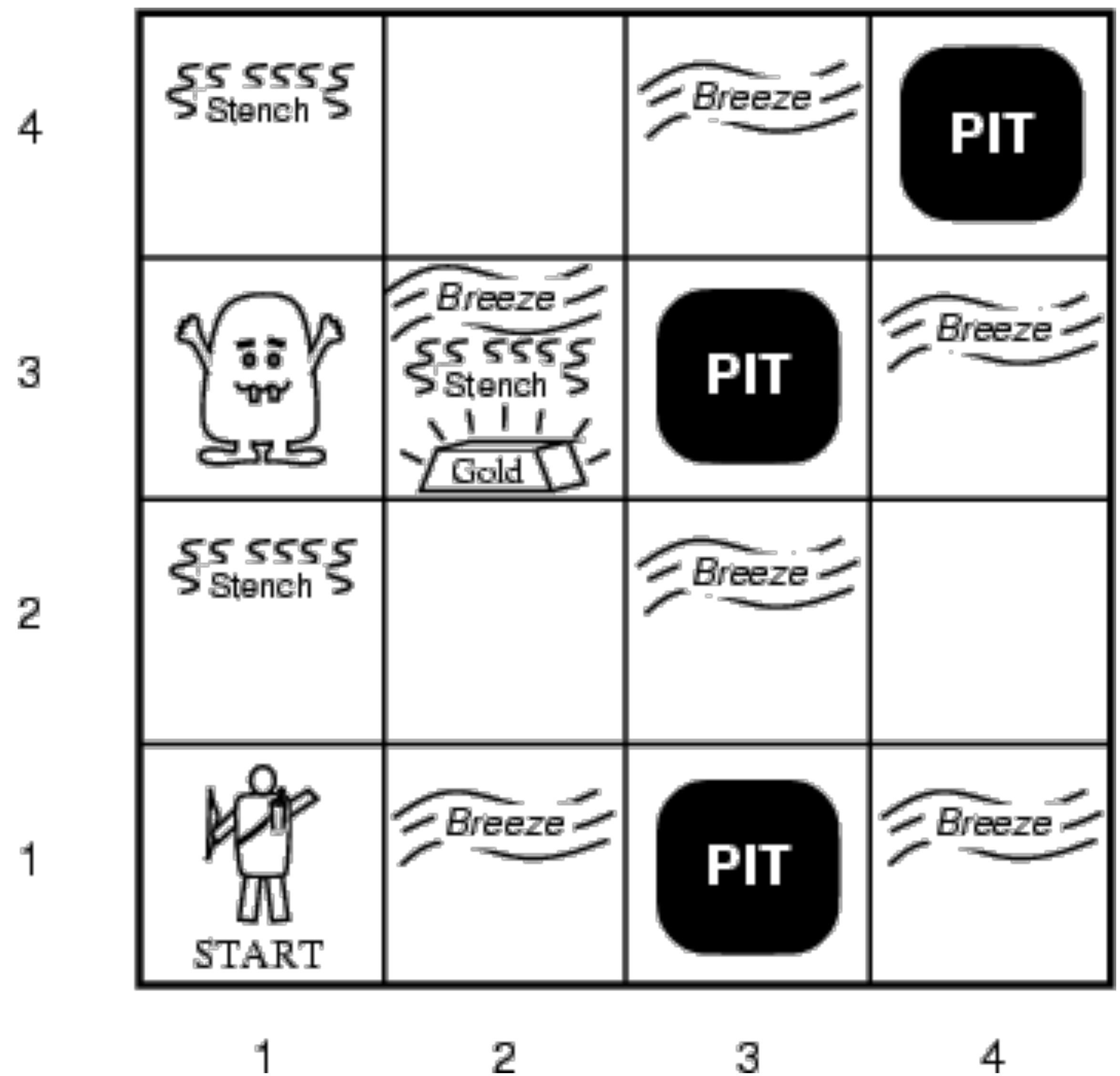
Wumpus World environment

- Based on Hunt the Wumpus computer game from 1972
- Agent explores cave of rooms connected by passageways
- Lurking in a room is the *Wumpus*, a beast that eats any agent that enters its room
- Some rooms have *bottomless pits* that trap any agent that wanders into the room
- Somewhere is a heap of gold in a room
- Goal: collect gold & exit w/o being eaten

AIMA's Wumpus World

The agent always starts in the field [1,1]

Agent's task is to find the gold, return to the field [1,1] and climb out of the cave



Agent in a Wumpus world: Percepts

- The agent perceives
  - **stench** in square containing Wumpus and in adjacent squares (not diagonally)
  - **breeze** in squares adjacent to a pit
  - **glitter** in the square where the gold is
  - **bump**, if it walks into a wall
  - Woeful **scream** everywhere in cave, if Wumpus killed
- Percepts given as five-tuple, e.g., if stench and breeze, but no glitter, bump or scream:

  (Stench, Breeze, None, None, None)
- Agent cannot perceive its location, e.g., (2,2)

Wumpus World Actions

- **go forward**
- **turn right** 90 degrees
- **turn left** 90 degrees
- **grab**: Pick up object in same square as agent
- **shoot**: Fire arrow in direction agent faces. It continues until it hits & kills Wumpus or hits outer wall. Agent has one arrow, so only first shoot action has effect
- **climb:** leave cave, only effective in start square
- **die:** automatically and irretrievably happens if agent enters square with pit or living Wumpus

Wumpus World Goal

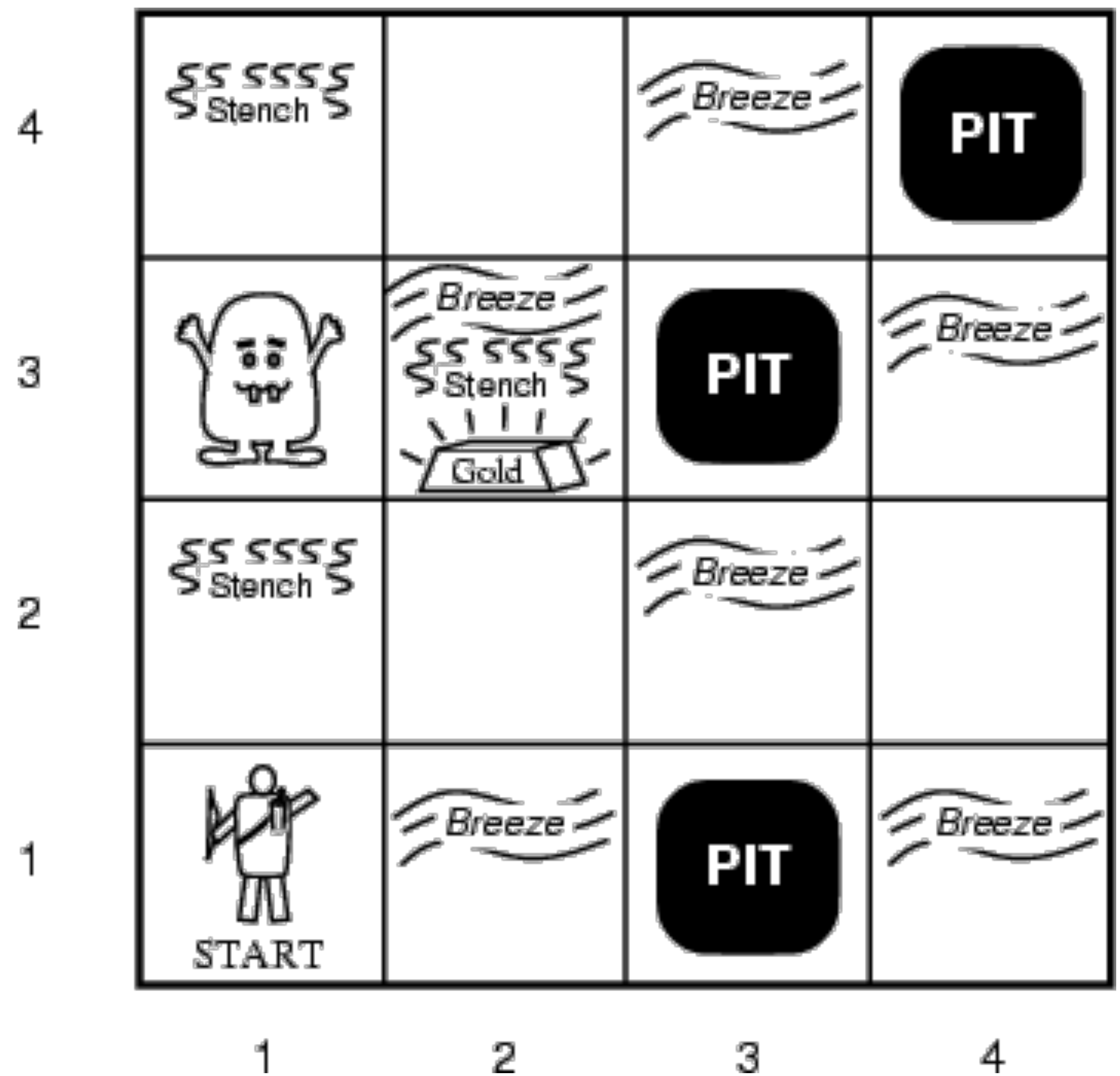
Agent's goal is to find the gold and bring it back to the start square as quickly as possible, without getting killed

- 1,000 point reward for climbing out of cave with gold
- 1 point deducted for every action taken
- 10,000 point penalty for getting killed

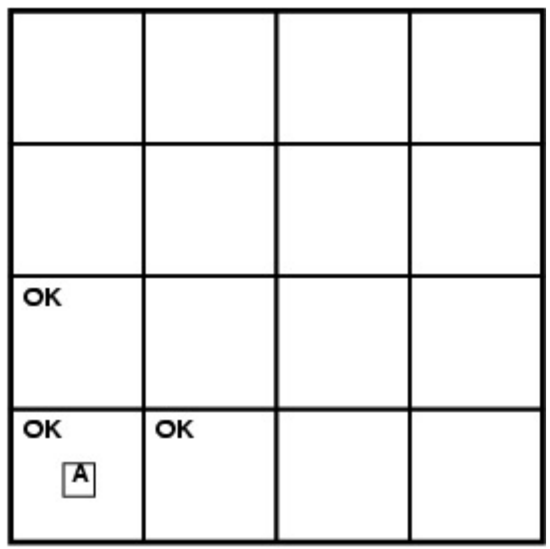
AIMA's Wumpus World

The agent always starts in the field [1,1]

Agent's task is to find the gold, return to the field [1,1] and climb out of the cave



Exploring a wumpus world



label	fact
A	agent
B	breeze
G	glitter
OK	safe cell
P	pit
S	stench
W	wumpus

We label cells with facts agent learns about them as it moves through world

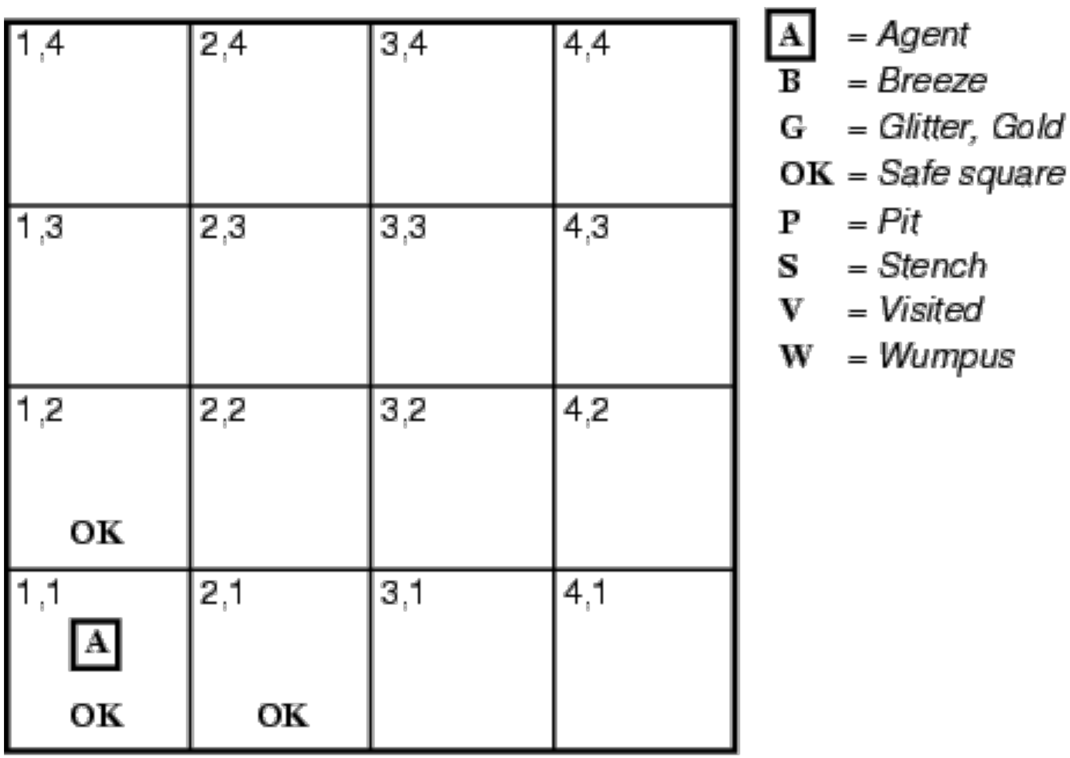
The Hunter’s first steps

(a) (b)

Since agent is alive and perceives neither breeze nor stench at (1,1), it **knows** (1,1) and its neighbors are OK

Moving to (2,1) is a **safe move** that reveals a breeze but no stench, **implying** that Wumpus isn’t adjacent but one or more pits are

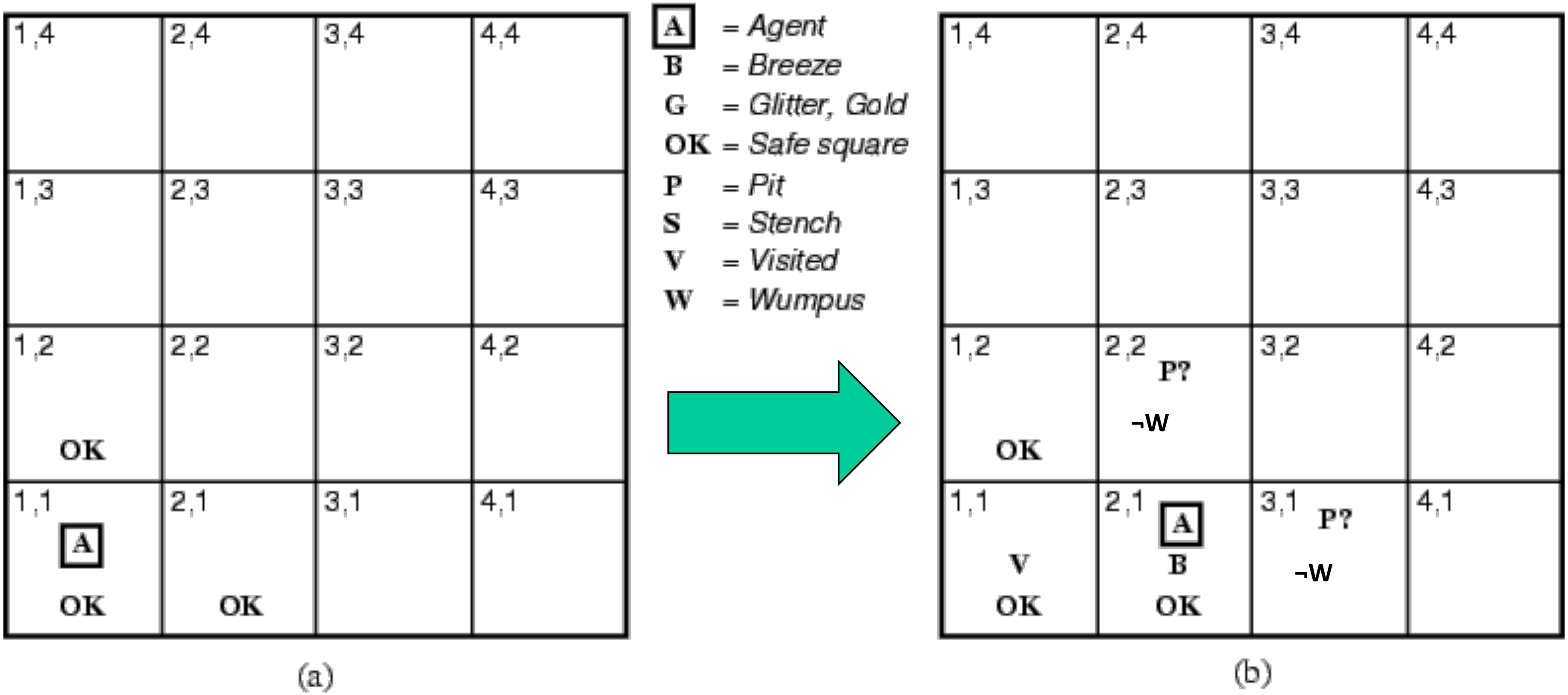
The Hunter's first steps



(a)

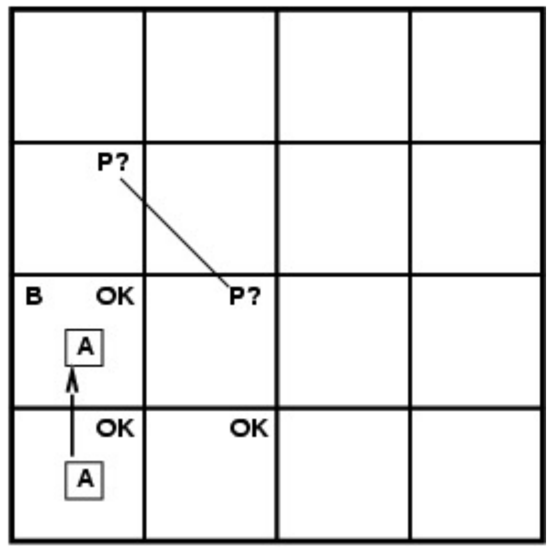
Since agent is alive and perceives neither breeze nor stench at (1,1), it **knows** (1,1) and its neighbors are OK

The Hunter's first steps



Moving to (2,1) is a **safe move** that reveals a breeze but no stench, **implying** that Wumpus isn't adjacent but one or more pits are

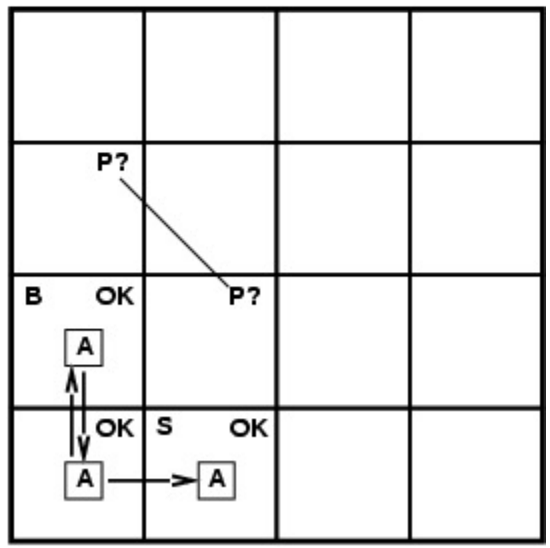
Exploring a wumpus world



A	agent
B	breeze
G	glitter
OK	safe cell
P	pit
S	stench
W	wumpus

Let's start over: assume the agent moves to (1,2) and detects a Breeze. A pit must be in (1,3) or (2,2). What should the agent do next?

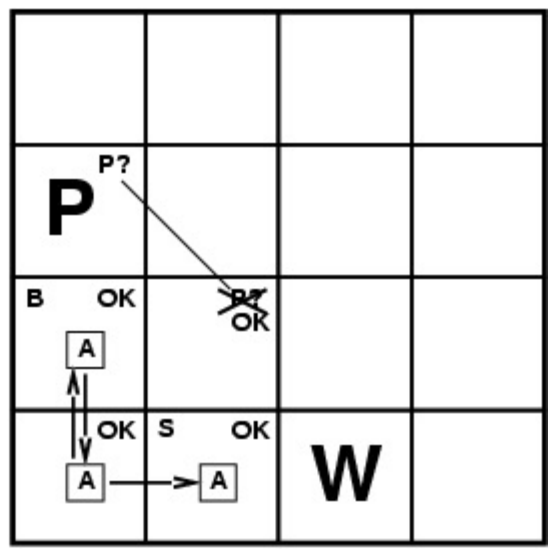
Exploring a wumpus world



A	agent
B	breeze
G	glitter
OK	safe cell
P	pit
S	stench
W	wumpus

Returning to (1,1), and then going to (2,1) is a safe move. Always prefer a safe move to a risky one. If the agent perceives a stench but no breeze in (2,1), what can it conclude?

Exploring a wumpus world

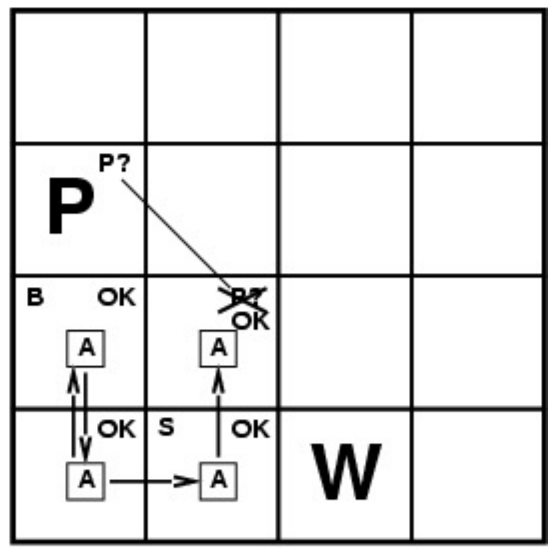


A	agent
B	breeze
G	glitter
OK	safe cell
P	pit
S	stench
W	wumpus

No stench in (1,2) => Wumpus not in (2,2) => Wumpus in (1,3)

No breeze in (2,1) => no pit in (2,2) => pit in (1,3)

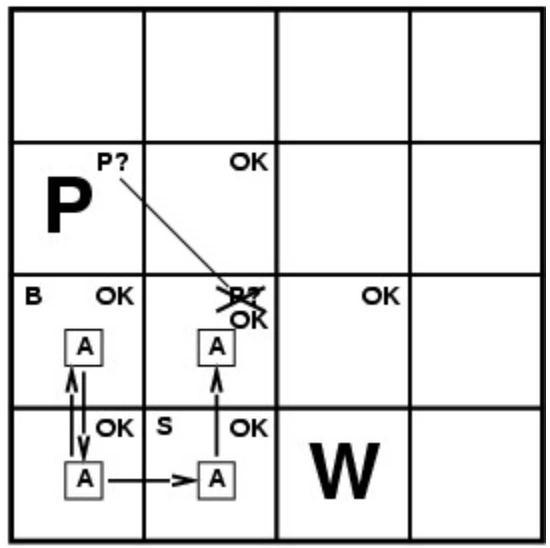
Exploring a wumpus world



A	agent
B	breeze
G	glitter
OK	safe cell
P	pit
S	stench
W	wumpus

The agent goes to (2,2) since it's safe

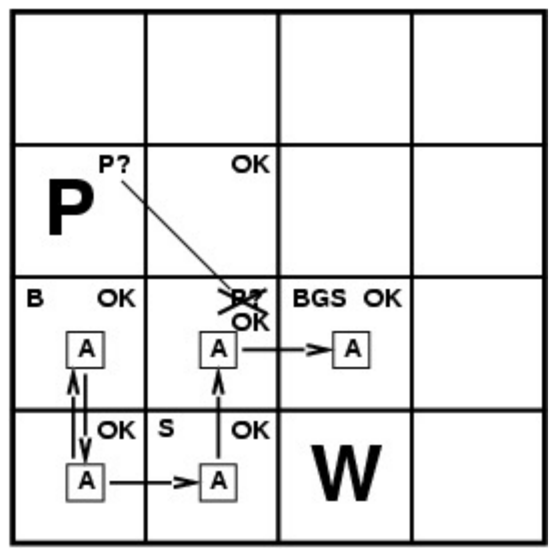
Exploring a wumpus world



A	agent
B	breeze
G	glitter
OK	safe cell
P	pit
S	stench
W	wumpus

Detecting neither a stench or breeze means that (2,3) and (3,2) are safe

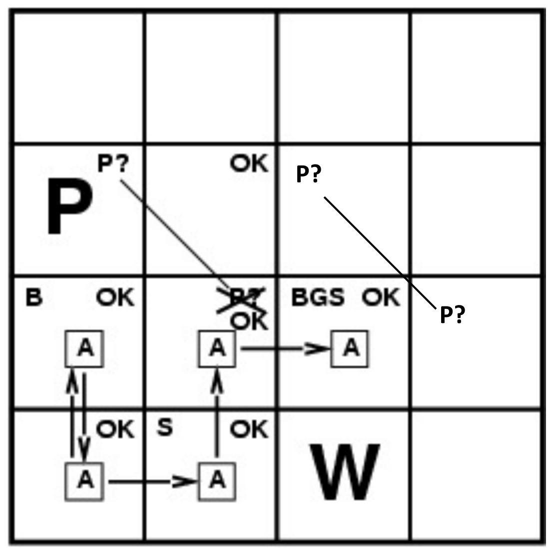
Exploring a wumpus world



A	agent
B	breeze
G	glitter
OK	safe cell
P	pit
S	stench
W	wumpus

We pick one of the safe moves, (2,3), and detect a breeze, stench and glitter.

Exploring a wumpus world



A	agent
B	breeze
G	glitter
OK	safe cell
P	pit
S	stench
W	wumpus

Found gold! Now it must find way back to (1,1). Hopefully, it has been remembering it's moves so can quickly return to (1,1) via safe moves

Logical reasoning

- As we'll see, the agent can represent
  - Its knowledge about the world in general, e.g., you can smell the Wumpus in the next cave
  - New facts it learns, e.g., no smell in 1,1
- And then draw conclusions, e.g., no Wumpus in 1,2 or in 2,1

Logic in general

- **Logics** are formal languages for representing information so that conclusions can be drawn
- **Syntax** defines the sentences in the language
- **Semantics** define the "meaning" of sentences
  - i.e., define **truth** of a sentence in a world

E.g., the language of arithmetic

- $x+2 \geq y$ is a sentence; $x2+y > \{\}$ is not a sentence
- $x+2 \geq y$ is true iff the number $x+2$ is no less than the number y
- $x+2 \geq y$ is true in a world where $x = 7, y = 1$
- $x+2 \geq y$ is false in a world where $x = 0, y = 6$
- $x+1 > x$ is true for all numbers x

Entailment

- Entailment: one thing follows from another
- KB ╞ α
- Knowledge base *KB* entails sentence α iff α is true in *all possible worlds* where *KB* is true
- A **possible world where KB is true** can contain additional facts as long as they don't contradict anything in the KB

  E.g.: 'what's known today' + 'there's life on Mars'

Entailment

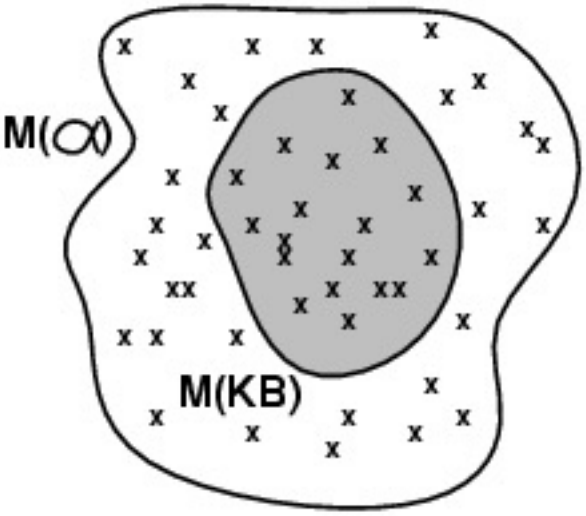
- Entailment: one thing follows from others
- KB $\models$ α
- Knowledge base *KB* entails sentence α iff α is true in *all possible worlds* where *KB* is true
  - E.g., the KB containing "UMBC won" and "JHU won" entails "Either UMBC won or JHU won"
  - E.g., x+y = 4 entails x = 4 - y
  - Entailment is a relationship between (sets of) sentences (i.e., syntax) that is based on **semantics**

Models

- Logicians talk of models: formally structured worlds w.r.t which truth can be evaluated
- *m* is a model of sentence α if α is true in *m*

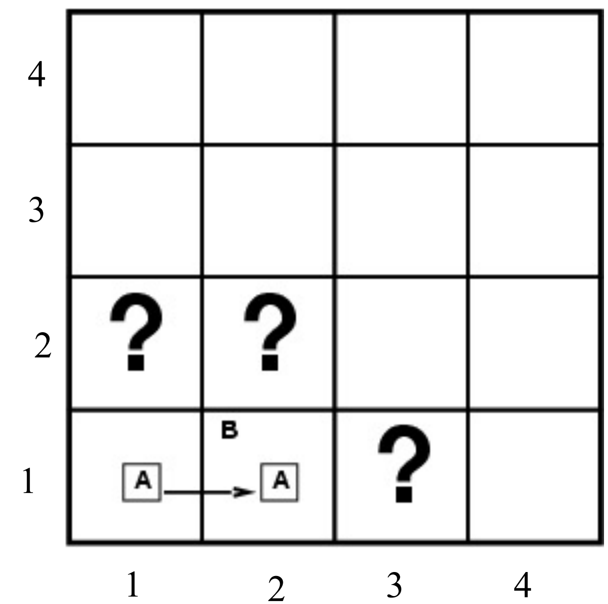
  Lots of other things might or might not be true or might be unknown in *m*

- *M(α)* is the set of all models of α
- Then KB $\models$ α iff $M(KB) \subseteq M(\alpha)$
  - *KB* = UMBC and JHU won
  - α = UMBC won
  - Then KB $\models$ α



Entailment in the Wumpus World

- Situation after detecting nothing in [1,1], move right, breeze in [2,1]
- Possible models for *KB* assuming only pits and restricting cells to {(1,3)(2,1)(2,2)}
- Two observations: ~B11, B12
- Three more propositional variables variables: P13, P21, P22
- ⇒ 8 possible models consistent with observations

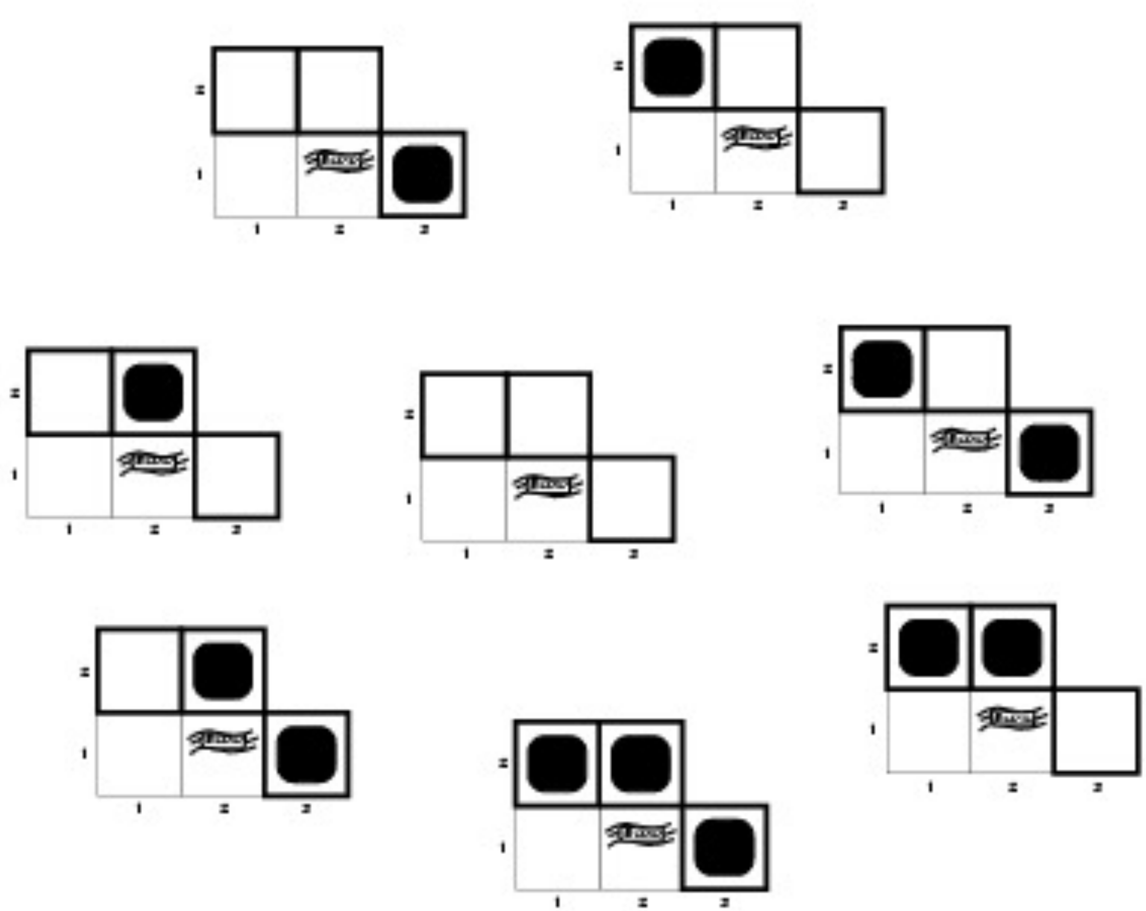


B11: breeze in (1,1)
P13: pit in (1,3)

Wumpus models

P13	P21	P22
F	F	F
F	F	T
F	T	F
F	T	T
T	F	F
T	F	T
T	T	F
T	T	T

Each row is a potential world



Some of these are inconsistent with the observed facts

Wumpus World Rules (1)

- **If a cell has a pit, then a breeze is observable in every adjacent cell**
- In propositional calculus we can not have rules with variables (e.g., forall X…)

  P11 => B21

  P11 => B12

  P21 => B11

  P21 => B22 …

If a pit in (1,1) then a breeze in (2,1), …

Wumpus World Rules (1)

- **If a cell has a pit, then a breeze is observable in every adjacent cell**
- In propositional calculus we can not have rules with variables (e.g., forall X…)

P11 => B21

P11 => B12

P21 => B11

P21 => B22 …

. . .

If a pit in (1,1) then a breeze in (2,1), …

these also follow

~B21 => ~P11

~B12 => ~P11

~B11 => ~P21

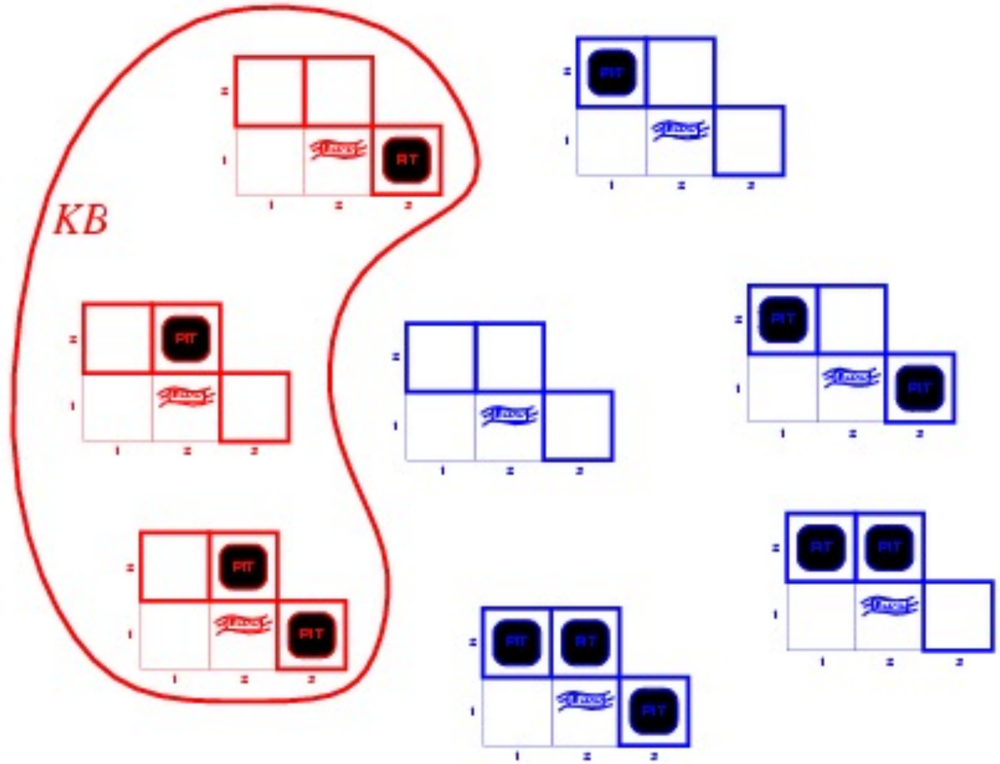
~B22 => ~P21

…

Wumpus models

P13	P21	P22
~~F~~	~~F~~	~~F~~
F	F	T
~~F~~	~~T~~	~~F~~
~~F~~	~~T~~	~~T~~
T	F	F
T	F	T
~~T~~	~~T~~	~~F~~
~~T~~	~~T~~	~~T~~

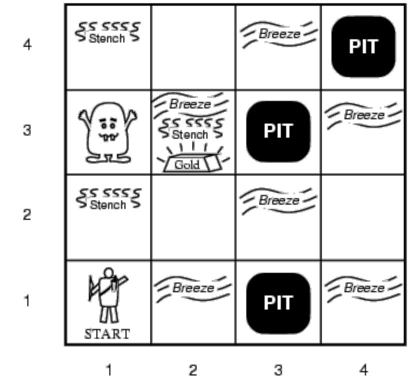
(P13 ∨ P22) ∧ ~P21



KB **= wumpus-world rules + observations**

- Only **three** of the possible models are consistent with what's known
- Any might be the way the world really is

Wumpus World Rules (2)



- **Cell safe if it has neither a pit nor wumpus**

  OK11 => ~P11 ∧ ~W11

  OK12 => ~P12 ∧ ~W12 …

OK11: (1,1) is safe
W11: Wumpus in (1,1)

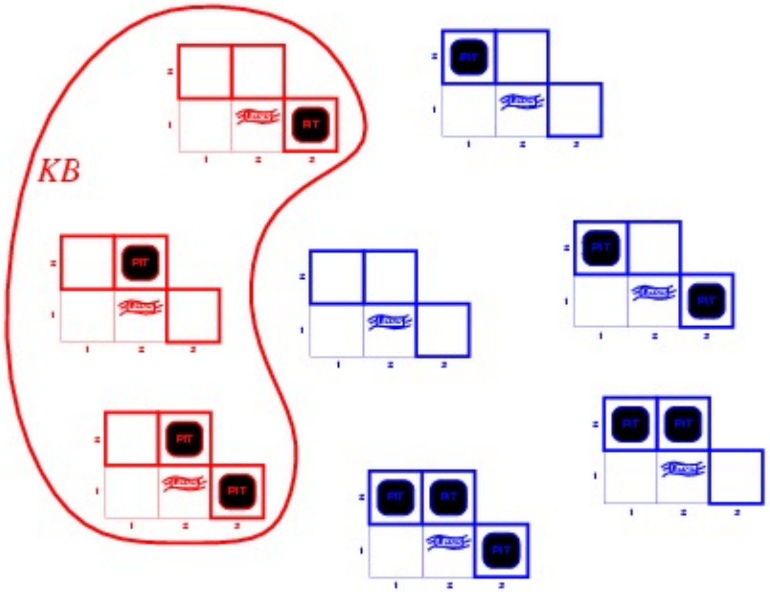
- From which we can derive the more useful “rules”

  P11 ∨ W11 => ~OK11

  P11 => ~OK11

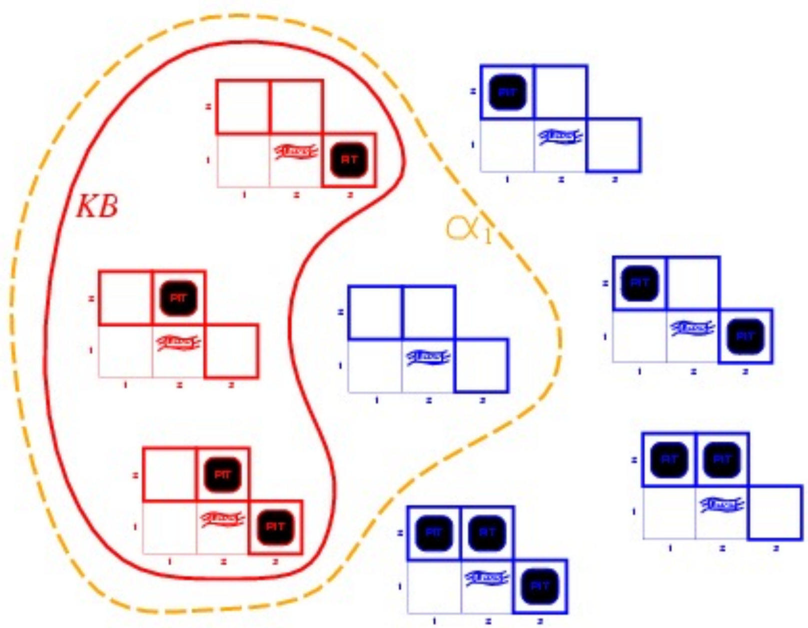
  W11 => ~OK11 …

Wumpus models



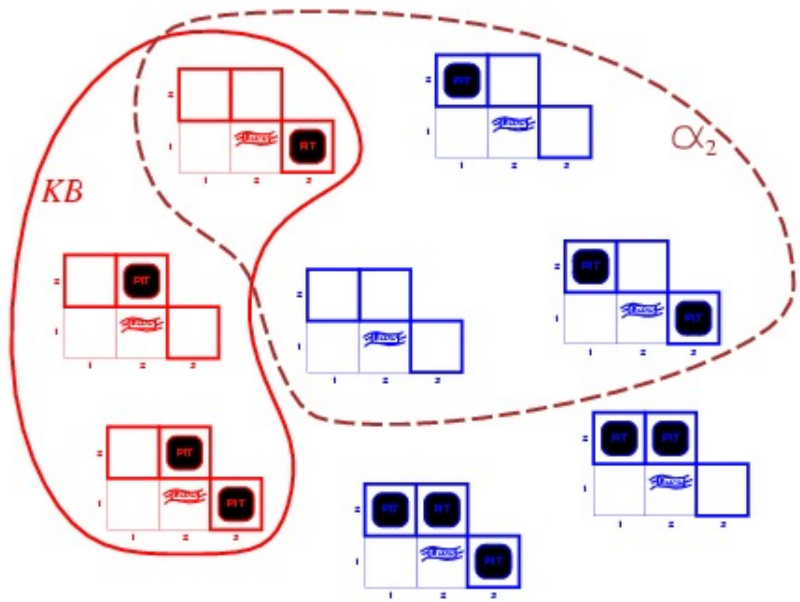
KB = wumpus-world rules + observations

Is (1,2) Safe? Yes!



- *KB* = wumpus-world rules + observations
- α_1 = "[1,2] is safe"
- *Since all models include* α_1
- *KB* $\models \alpha_1$, proved by **model checking**

Is (2,2) Safe? Maybe, Maybe Not!



- *KB* = wumpus-world rules + observations
- α_2 = "[2,2] is safe"
- Since some models don't include α_2, $KB \not\models \alpha_2$
- We cannot prove OK22; it might be true or false

Inference, Soundness, Completeness

- *KB* $\vdash_i$ α : sentence α can be derived (inferred) from *KB* by procedure *i*
- **Soundness:** *i* is sound if whenever *KB* $\vdash_i$ α, it is also true that *KB* $\models$ α
- **Completeness:** *i* is complete if whenever *KB* $\models$ α, it is also true that *KB* $\vdash_i$ α

- Preview: **first-order logic** is expressive enough to say almost anything of interest and has a **sound** and **complete** inference procedure

Soundness and completeness

- A *sound* inference method derives only entailed sentences
- A complete inference method can (eventually) derive any entailed sentence

- Analogous to the property of *soundness* and *completeness* in search

Summary

- Intelligent agents need knowledge about world for good decisions
- Agent’s knowledge stored in a knowledge base (KB) as **sentences** in a knowledge representation (KR) language
- Knowledge-based agents needs a **KB** & **inference mechanism**. They store sentences in KB, infer new sentences & use them to **deduce** which actions to take
- A **representation language** defined by its syntax & semantics, which specify structure of sentences & how they relate to facts of the world
- **Interpretation** of a sentence is fact to which it refers. If fact is part of the actual world, then the sentence is true

Fin