Artificial Intelligence 1

Roman Barták
Department of Theoretical Computer Science and Mathematical Logic

Knowledge Representation: Propositional Logic

Today

Starting today we will design agents that can form **representations** of a complex world, use a process of **inference** to derive new information about the world, and use that information to **deduce** what to do.

They are called **knowledge-based agents** – combine and recombine information about the world with current observations to uncover hidden aspects of the world and use them for action selection.

We need to know:

– how to represent **knowledge?**

– how to **reason** over that knowledge?

Knowledge-based agent

A knowledge-based agent uses a **knowledge base** – a set of sentences expressed in a given language – that can be updated by the operation TELL and can be queried about what is known using the operation ASK.

Answers to queries may involve **inference** – that is deriving new sentences from old sentences (inserted using the TELL operations).

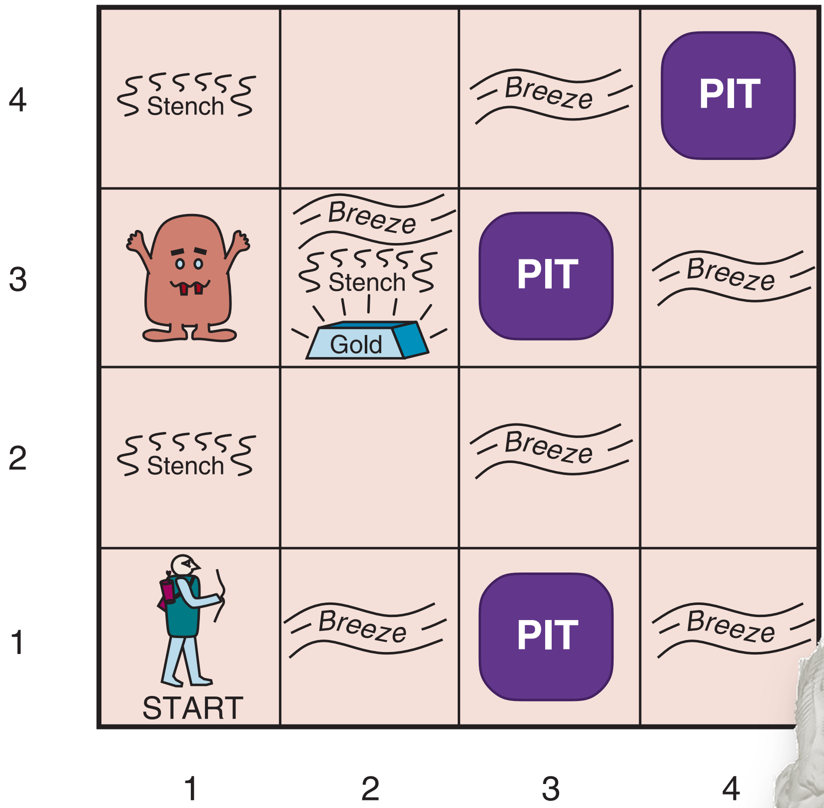
```
function KB-AGENT(percept) returns an action
  persistent: KB, a knowledge base
              t, a counter, initially 0, indicating time

  TELL(KB, MAKE-PERCEPT-SENTENCE(percept, t))
  action ← ASK(KB, MAKE-ACTION-QUERY(t))
  TELL(KB, MAKE-ACTION-SENTENCE(action, t))
  t ← t + 1
  return action
```

knowledge base contains information about observations as well as about own actions

inference will help the agent to select an action even if information about the world is incomplete

The Wumpus world: a running example

A cave consisting of rooms connected by passageways, inhabited by the terrible **Wumpus**, a beast that eats anyone who enters its room, containing rooms with bottomless **pits** that will trap anyone, and a room with a heap of **gold**.



- The agent will perceive a **Stench** in the directly (not diagonally) adjacent squares to the square containing the Wumpus.
- In the squares directly adjacent to a pit, the agent will perceive a **Breeze**.
- In the square where the gold is, the agent will perceive a **Glitter**.
- When an agent walks into a wall, it will perceive a **Bump**.
- The Wumpus can be shot by an agent, but the agent has only one arrow.
  - Killed Wumpus emits a woeful **Scream** that can be perceived anywhere in the cave.

Performance measure

- +1000 points for climbing out of the cave with the gold
- -1000 for falling into a pit or being eaten by the Wumpus
- -1 for each action taken
- -10 for using up the arrow

Environment

- 4×4 grid of rooms, the agent starts at [1,1] facing to the right

Sensors

- Stench, Breeze, Glitter, Bump, Scream

Actuators

- MoveForward, TurnLeft, TurnRight
- Grab, Shoot, Climb

Fully observable?

- NO, the agent perceives just its direct neighbour (partially observable)

Deterministic?

- YES, the result of action is given

Episodic?

- NO, the order of actions is important (sequential)

Static?

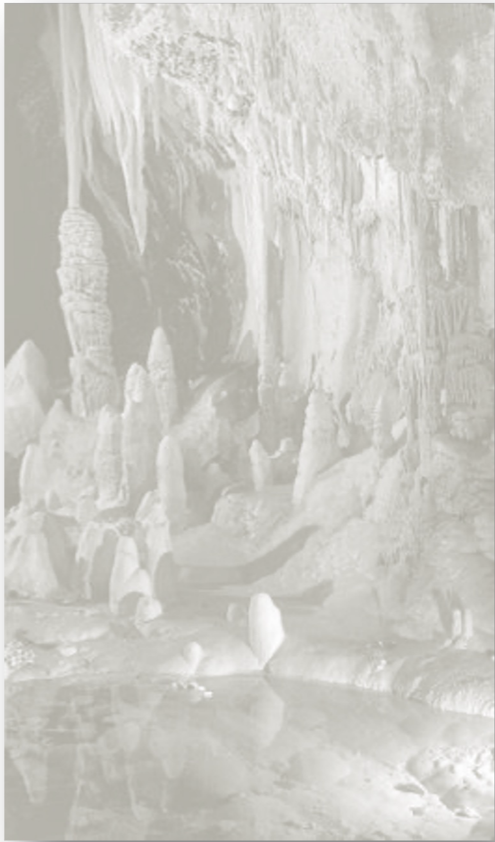
- YES, the Wumpus and pits do not move

Discrete?

- YES

One agent?

- YES, the Wumpus does not act as an agent, it is merely a property of environment

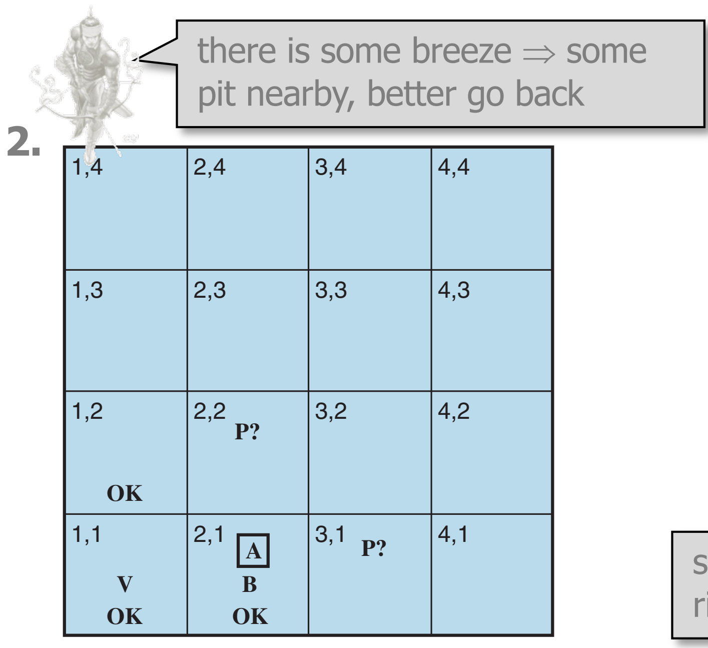


The Wumpus world: the quest for gold

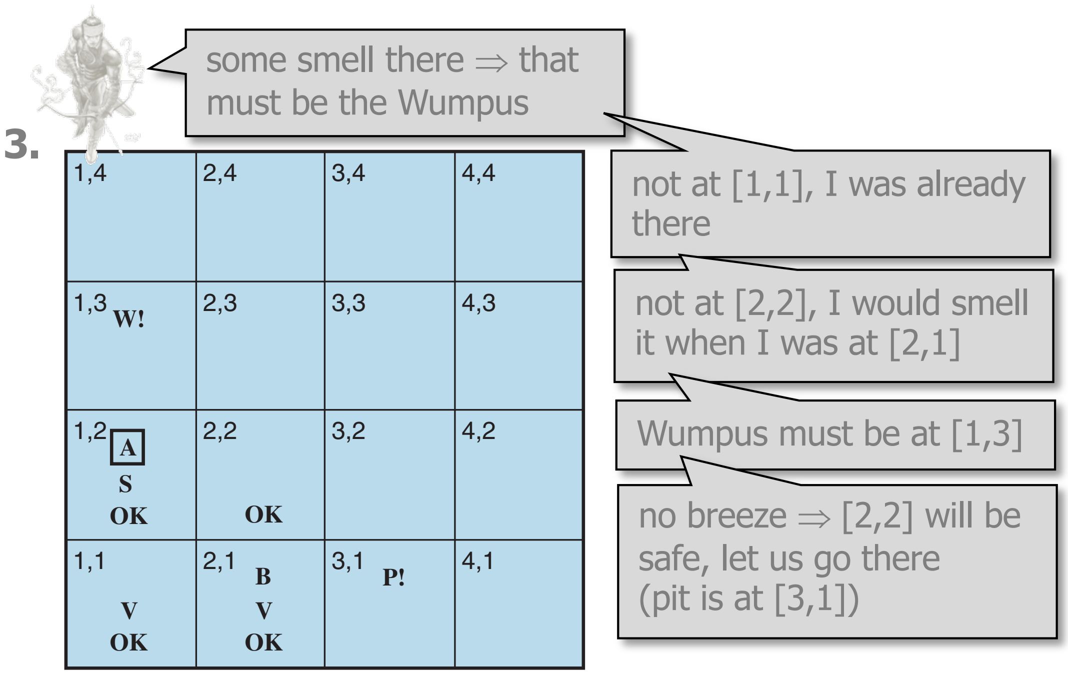
1.

- **A** = *Agent*
- **B** = *Breeze*
- **G** = *Glitter, Gold*
- **OK** = *Safe square*
- **P** = *Pit*
- **S** = *Stench*
- **V** = *Visited*
- **W** = *Wumpus*

2.

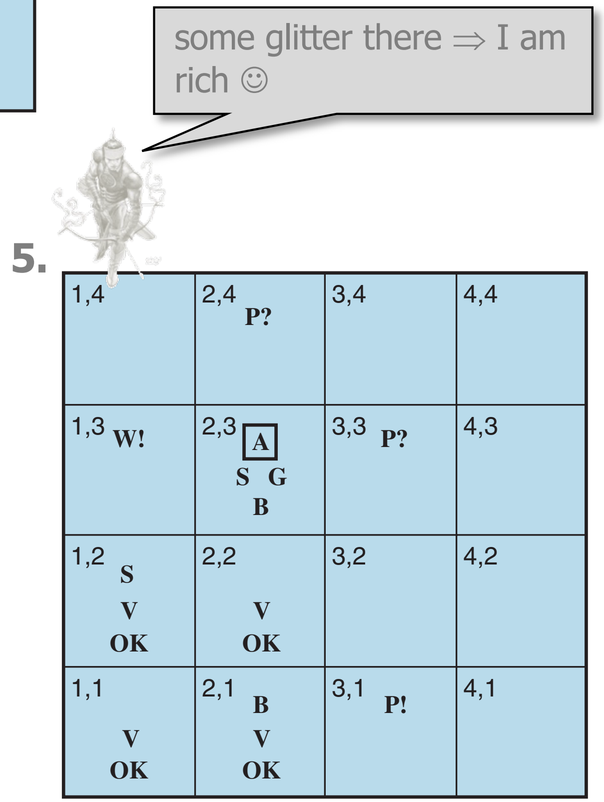


3.

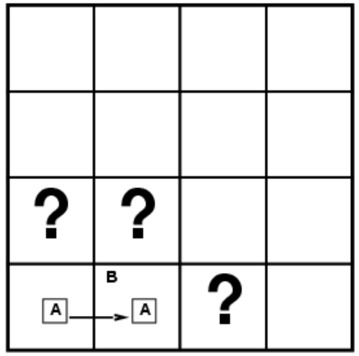


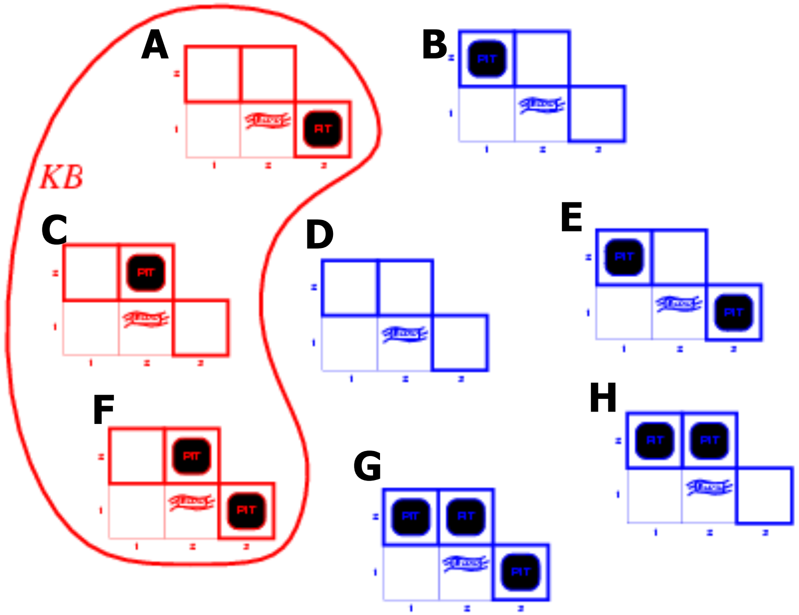
...

5.



Assume a situation when there is no percept at [1,1], we went right to [2,1] and feel Breeze there.





– For pit detection we have 8 (=2^3) possible models (states of the neighbouring world).
– Only three of these models correspond to our knowledge base, the other models conflict the observations:
  - no percept at [1,1]
  - Breeze at [2,1]

The Wumpus world: some consequences

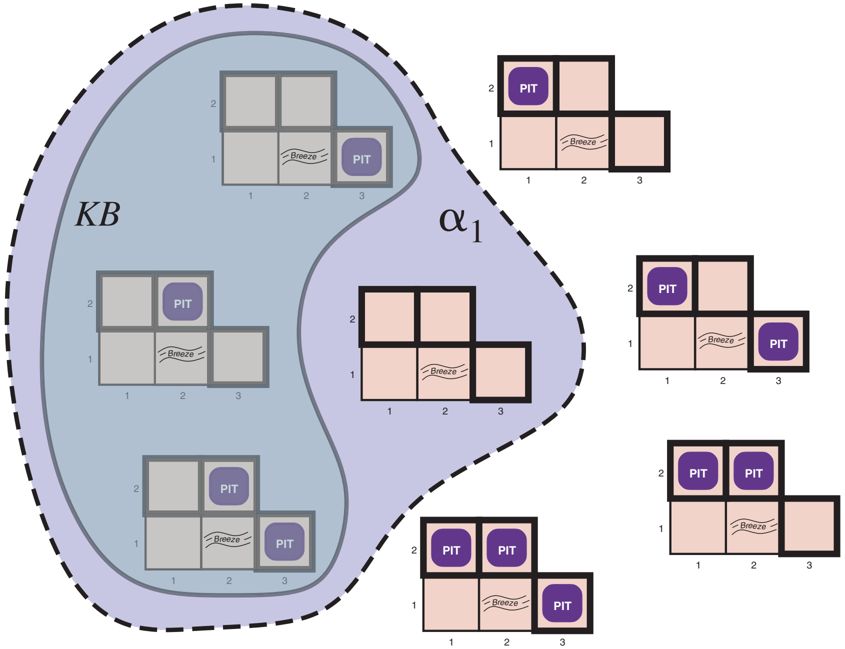
Let us ask whether the room [1,2] is safe.

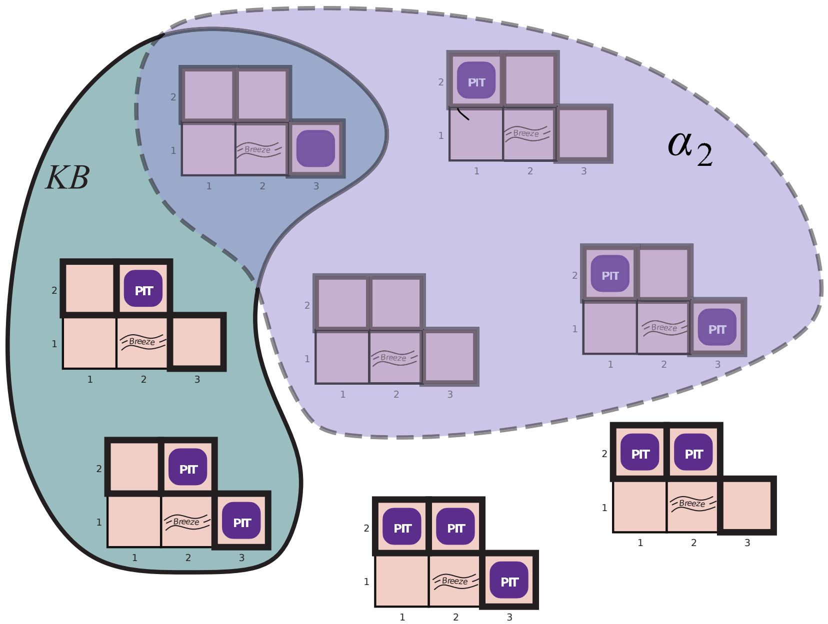
Is information α_1 = "[1,2] is safe" entailed by our representation?

- we compare models for KB and for α_1
- every model of KB is also a model for α_1 so α_1 is entailed by KB

And what about the room [2,2]?

- we compare models for KB and for α_2
- some models of KB are not models of α_2
- α_2 is not entailed by KB and we do not know for sure if room [2,2] is safe





How to implement inference in general?

We will use **propositional logic**. Sentences are propositional expressions and a knowledge base is a conjunction of these expressions.

- **Propositional variables** describe the properties of the world
  - **$P_{i,j}$ = true** if there is a pit at [i, j]
  - **$B_{i,j}$ = true** if the agent perceives Breeze at [i, j]
- **Propositional formulas** describe
  - known information about the world
    - $\neg P_{1,1}$ no pit at [1, 1] (we are there)
  - general knowledge about the world (for example, Breeze means a pit in some neighbouring room)
    - $B_{1,1} \Leftrightarrow (P_{1,2} \vee P_{2,1})$
    - $B_{2,1} \Leftrightarrow (P_{1,1} \vee P_{2,2} \vee P_{3,1})$
    - …
  - observations
    - $\neg B_{1,1}$ no Breeze at [1, 1]
    - $B_{2,1}$ Breeze at [2, 1]
- We will be using **inference** for propositional logic.

Syntax defines the allowable sentences.

- a propositional variable (and constants true and false) is an (atomic) sentence
- two sentences can be connected via logical connectives $\neg$, $\wedge$, $\vee$, $\Rightarrow$, $\Leftrightarrow$ to get a (complex) sentence

Semantics defines the rules for determining the truth of a sentence with respect to a particular model.

- **model** is an assignment of truth values to all propositional variables
- an atomic sentence P is true in any model containing P=true
- semantics of complex sentences is given by the truth table

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \Rightarrow Q$	$P \Leftrightarrow Q$
false	*false*	*true*	*false*	*false*	*true*	*true*
false	*true*	*true*	*false*	*true*	*true*	*false*
true	*false*	*false*	*false*	*true*	*false*	*false*
true	*true*	*false*	*true*	*true*	*true*	*true*

M is a **model** of sentence α, if α is true in M.

– The set of models for α is denoted M(α).

Entailment: KB $\models \alpha$ „KB platí v α"

means that α is a logical consequence of KB

– KB entails α iff $M(KB) \subseteq M(\alpha)$

We are interested in **inference methods**, that can find/verify consequences of KB.

– KB $\vdash_i \alpha$ means that algorithm i infers sentence α from KB

– the algorithm is **sound** iff KB $\vdash_i \alpha$ implies KB $\models \alpha$

– the algorithm is **complete** iff KB $\models \alpha$ implies KB $\vdash_i \alpha$

There are basically two classes of inference algorithms.

- **model checking**
  - based on enumeration of a truth table
  - Davis-Putnam-Logemann-Loveland (DPLL)
  - local search (minimization of conflicts)
- **inference rules**
  - theorem proving by applying inference rules
  - a resolution algorithm

```
function TT-ENTAILS?(KB, α) returns true or false
   inputs: KB, the knowledge base, a sentence in propositional logic
           α, the query, a sentence in propositional logic

   symbols ← a list of the proposition symbols in KB and α
   return TT-CHECK-ALL(KB, α, symbols, { })

function TT-CHECK-ALL(KB, α, symbols, model) returns true or false
   if EMPTY?(symbols) then
      if PL-TRUE?(KB, model) then return PL-TRUE?(α, model)
      else return true        // when KB is false, always return true
   else
      P ← FIRST(symbols)
      rest ← REST(symbols)
      return (TT-CHECK-ALL(KB, α, rest, model ∪ {P = true})
             and
             TT-CHECK-ALL(KB, α, rest, model ∪ {P = false }))
```

The Wumpus world
α_1 = "[1,2] is safe" = "$\neg P_{1,2}$" is entailed by KB, as $P_{1,2}$ is always false for models of KB and hence there is no pit at [1,2]

- We simply explore all the models using the **generate and test** method.
- Each model of KB must be also a model for α.

$B_{1,1}$	$B_{2,1}$	$P_{1,1}$	$P_{1,2}$	$P_{2,1}$	$P_{2,2}$	$P_{3,1}$	KB	α_1
false	*false*	*false*	*false*	*false*	*false*	*false*	*false*	*true*
false	*false*	*false*	*false*	*false*	*false*	*true*	*false*	*true*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
false	*true*	*false*	*false*	*false*	*false*	*false*	*false*	*true*
false	*true*	*false*	*false*	*false*	*false*	*true*	*true*	*true*
false	*true*	*false*	*false*	*false*	*true*	*false*	*true*	*true*
false	*true*	*false*	*false*	*false*	*true*	*true*	*true*	*true*
false	*true*	*false*	*false*	*true*	*false*	*false*	*false*	*true*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
true	*true*	*true*	*true*	*true*	*true*	*true*	*false*	*false*

nikde nesmí být false, true ←

Sentence (formula) is **satisfiable** if it is true in, or satisfied by, *some* model.

Example: $A \vee B$, C

Sentence (formula) is **unsatisfiable** if it is not true in *any* model.

Example: $A \wedge \neg A$

Entailment can then be implemented as checking satisfiability as follows:
KB $\models \alpha$ if and only if (KB $\wedge \neg\alpha$) is unsatisfiable.

- proof by **refutation**
- proof by **contradiction**

Verifying if α is entailed by KB can be implemented as the satisfiability problem for the formula (KB $\wedge \neg\alpha$).

Usually the formulas are in a **conjunctive normal form** (CNF)

- **literal** is an atomic variable or its negation
- **clause** is a disjunction of literals
- **formula** in CNF is a conjunction of clauses

Example: $(A \vee \neg B) \wedge (B \vee \neg C \vee \neg D)$

Each propositional sentence (formula) can be represented in CNF.

$B_{1,1} \Leftrightarrow (P_{1,2} \vee P_{2,1})$

$(B_{1,1} \Rightarrow (P_{1,2} \vee P_{2,1})) \wedge ((P_{1,2} \vee P_{2,1}) \Rightarrow B_{1,1})$

$(\neg B_{1,1} \vee P_{1,2} \vee P_{2,1}) \wedge (\neg(P_{1,2} \vee P_{2,1}) \vee B_{1,1})$

$(\neg B_{1,1} \vee P_{1,2} \vee P_{2,1}) \wedge ((\neg P_{1,2} \wedge \neg P_{2,1}) \vee B_{1,1})$

$(\neg B_{1,1} \vee P_{1,2} \vee P_{2,1}) \wedge (\neg P_{1,2} \vee B_{1,1}) \wedge (\neg P_{2,1} \vee B_{1,1})$

Davis, Putnam, Logemann, Loveland

– a sound and complete algorithm for verifying satisfiability of formulas in a CNF (finds its model)

```
function DPLL-SATISFIABLE?(s) returns true or false
  inputs: s, a sentence in propositional logic

  clauses ← the set of clauses in the CNF representation of s
  symbols ← a list of the proposition symbols in s
  return DPLL(clauses, symbols, { })

function DPLL(clauses, symbols, model) returns true or false

  if every clause in clauses is true in model then return true
  if some clause in clauses is false in model then return false
  P, value ← FIND-PURE-SYMBOL(symbols, clauses, model)
  if P is non-null then return DPLL(clauses, symbols – P, model ∪ {P=value})
  P, value ← FIND-UNIT-CLAUSE(clauses, model)
  if P is non-null then return DPLL(clauses, symbols – P, model ∪ {P=value})
  P ← FIRST(symbols); rest ← REST(symbols)
  return DPLL(clauses, rest, model ∪ {P=true}) or
         DPLL(clauses, rest, model ∪ {P=false}))
```

Early termination for partial models

- clause is true if any of its literals is true
- formula is not true if any of its clauses is not true

Pure symbol heuristics

- a pure symbol always appears with the same " sign" in all clauses
- the corresponding literal is set to true

Unit clause heuristics

- a unit clause is a clause with just one literal
- the corresponding literal is set to true

branching for backtracking

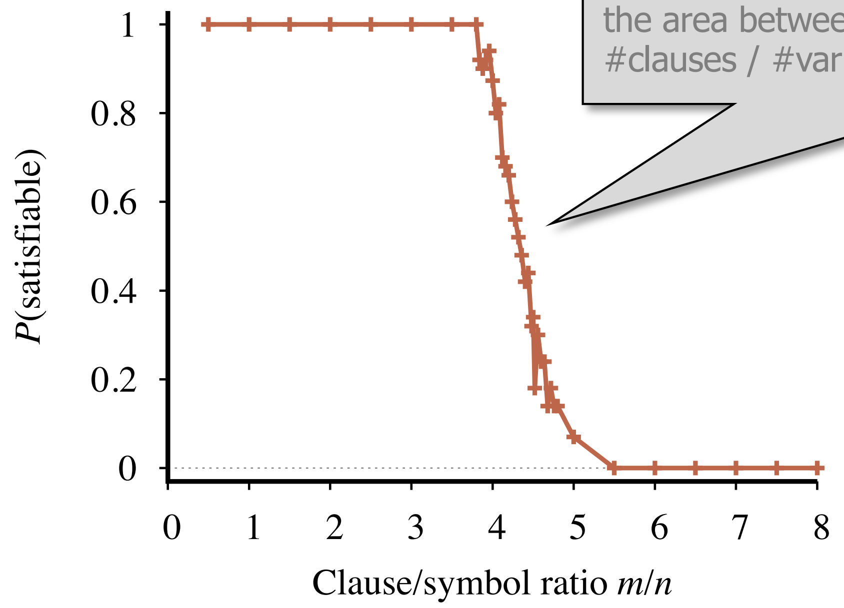
Hill climbing merged with random walk

- minimizing the number of conflict (false) clauses
- one local step corresponds to swapping a value of the selected variable
- **sound, but incomplete algorithm**

function WALKSAT(*clauses*, *p*, *max_flips*) **returns** a satisfying model or *failure*
  inputs: *clauses*, a set of clauses in propositional logic
    p, the probability of choosing to do a "random walk" move, typically around 0.5
    max_flips, number of value flips allowed before giving up

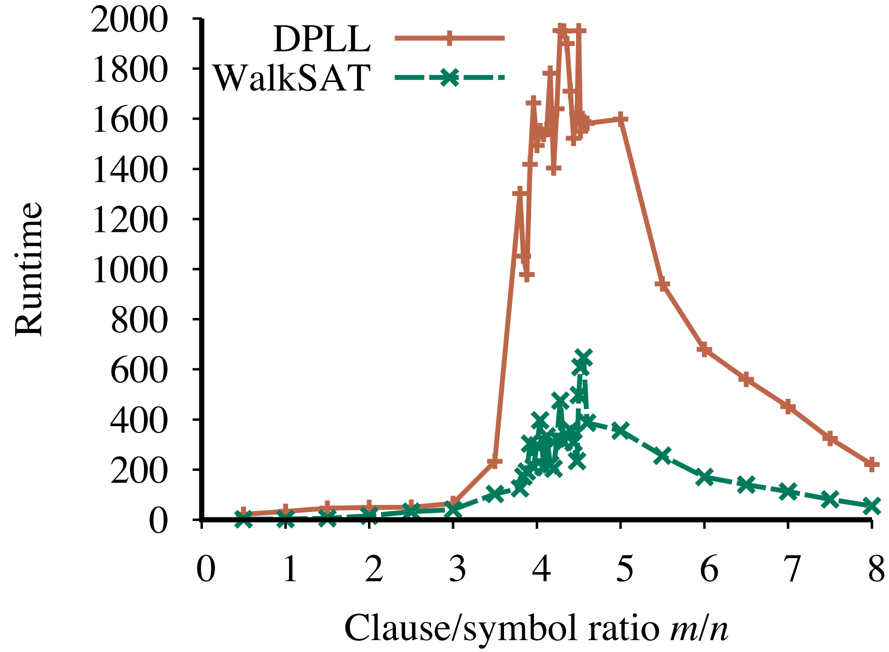
  model ← a random assignment of *true*/*false* to the symbols in *clauses*
  for each $i = 1$ **to** *max_flips* **do**
    if *model* satisfies *clauses* **then return** *model*
    clause ← a randomly selected clause from *clauses* that is false in *model*
    if RANDOM$(0, 1) \leq p$ **then**
      flip the value in *model* of a randomly selected symbol from *clause*
    else flip whichever symbol in *clause* maximizes the number of satisfied clauses
  return *failure*

The phase transition
the area between satisfiable and unsatisfiable formulas
#clauses / #variables= 4.3



Random 3-SAT problem with 50 variables

- each clause consists of three different variables
- probability of using a negated symbol is 50%



The graph shows medians of runtime necessary to solve the problems (for 100 problems)

- DPLL is pretty efficient
- WalkSAT is even faster

The resolution algorithm proves unsatisfiability of the formula (KB $\wedge \neg\alpha$) converted to a CNF. It uses a **resolution rule** that resolves two clauses with complementary literals (P and $\neg$P) to produce a new clause:

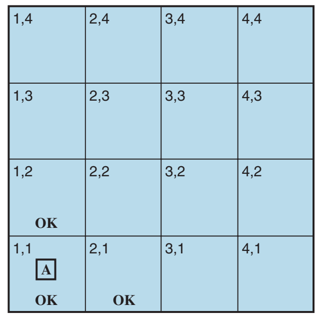
$$\frac{l_1 \vee \ldots \vee l_k \qquad\qquad m_1 \vee \ldots \vee m_n}{l_1 \vee \ldots \vee l_{i-1} \vee l_{i+1} \vee \ldots \vee l_k \vee m_1 \vee \ldots \vee m_{j-1} \vee m_{j+1} \vee \ldots \vee m_n}$$

where l_i and m_j are the complementary literals

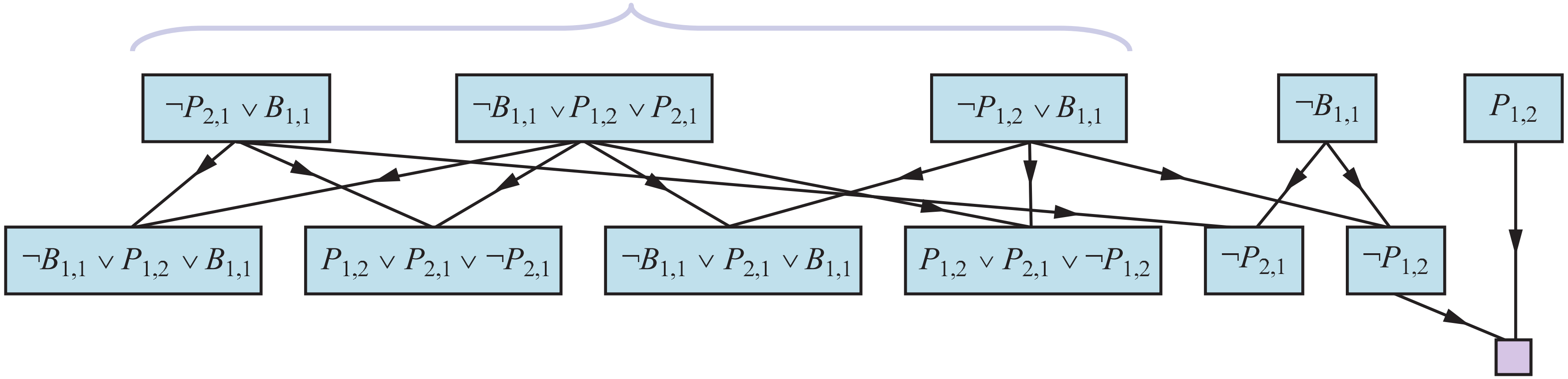
The algorithm stops when

- no other clause can be derived (then $\neg$ KB $\models \alpha$)
- an empty clause was obtained (then KB $\models \alpha$)

Sound and complete algorithm



$\mathbf{B_{1,1} \Leftrightarrow (P_{1,2} \vee P_{2,1})}$



Resolution algorithm

function PL-RESOLUTION(KB, α) **returns** *true* or *false*
  inputs: KB, the knowledge base, a sentence in propositional logic
    α, the query, a sentence in propositional logic

  $clauses \leftarrow$ the set of clauses in the CNF representation of $KB \wedge \neg\alpha$
  $new \leftarrow \{\}$
  while *true* **do**
    for each pair of clauses C_i, C_j **in** *clauses* **do**
      $resolvents \leftarrow$ PL-RESOLVE(C_i, C_j)
      if *resolvents* contains the empty clause **then return** *true*
      $new \leftarrow new \cup resolvents$
    if $new \subseteq clauses$ **then return** *false*
    $clauses \leftarrow clauses \cup new$

For each pair of clauses with complementary literals produce all possible resolvents. They are added to KB for next resolution.

an empty clause corresponds to false (an empty disjunction)
  → the formula is **unsatisfiable**

we reached a fixed point (no new clauses added)
  → formula is **satisfiable** and we can find its model
How to **find a model?**
  take the symbols P_i one be one
    if there is a clause with ¬ P_i such that the other literals are false with the current instantiation of $P_1,\ldots,P_{i-1}$, then P_i = false
    otherwise P_i = true

Many knowledge bases contain clauses of a special form – so called **Horn clauses**.

- Horn clause is a disjunction of literals of which at most one is positive

Example: $C \wedge (\neg B \vee A) \wedge (\neg C \vee \neg D \vee B)$

- Such clauses are typically used in knowledge bases with sentences in the form of an implication (for example Prolog without variables)

Example: $C \wedge (B \Rightarrow A) \wedge (C \wedge D \Rightarrow B)$

We will solve the problem if a given propositional symbol – **query** – can be derived from the knowledge base consisting of Horn clauses only.

- we can use a special variant of the resolution algorithm running in linear time with respect to the size of KB
- **forward chaining** (from facts to conclusions)
- **backward chaining** (from a query to facts)

From the known facts we derive all possible consequences using the Horn clauses until there are no new facts or we prove the query.

$(A \wedge B) \Rightarrow C$

This is a **data-driven method**.

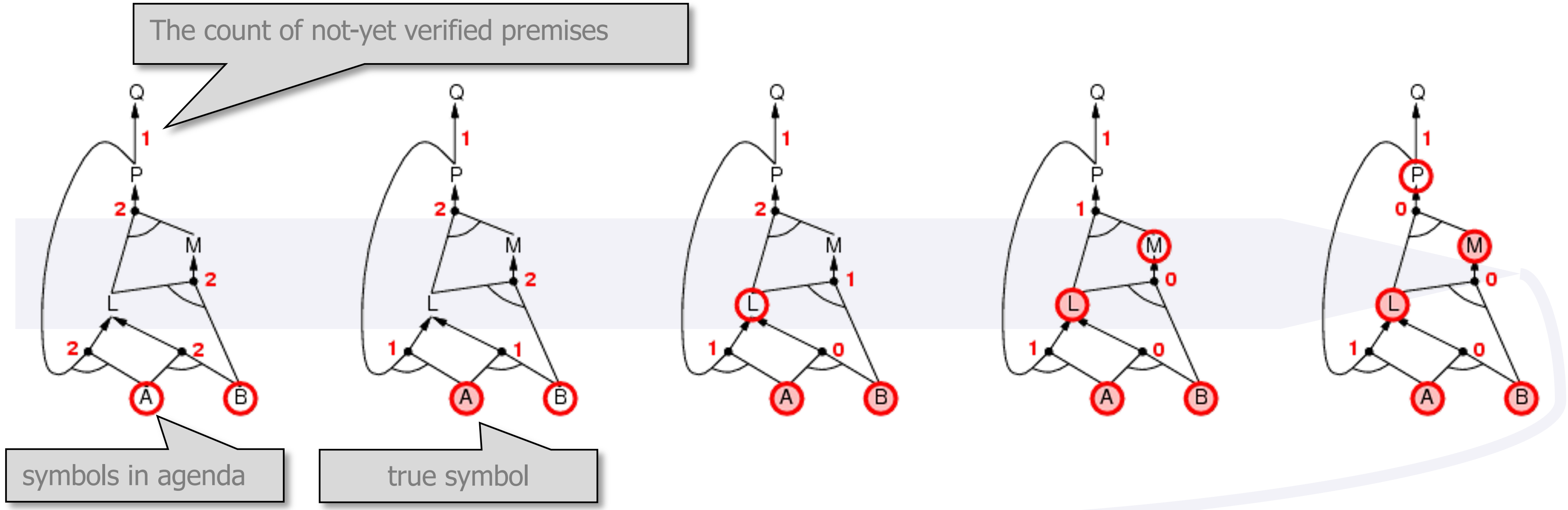
↳ postupně odvozuji fakta. Jakmile mám A a B, můžu odvodit i C.

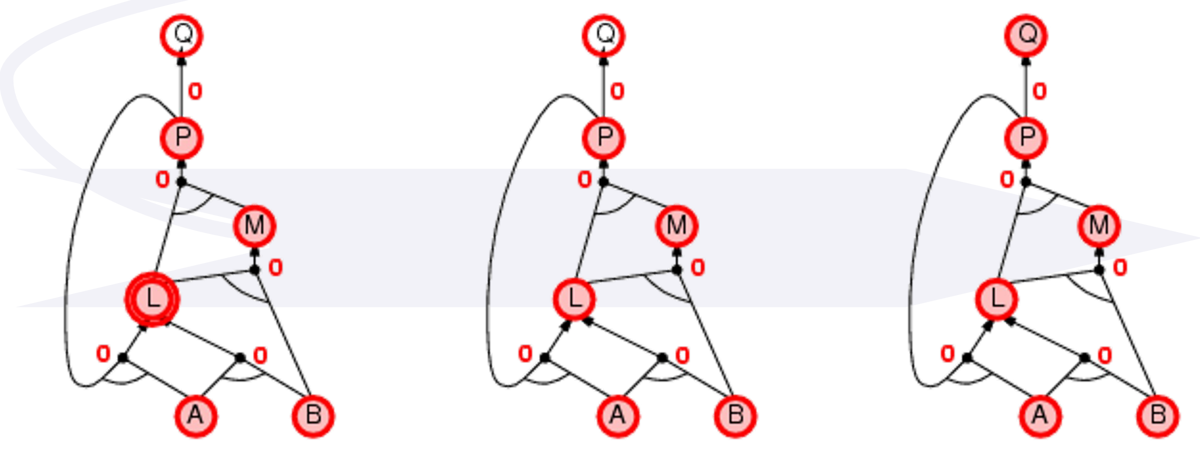
```
function PL-FC-ENTAILS?(KB, q) returns true or false
  inputs: KB, the knowledge base, a set of propositional definite clauses
          q, the query, a proposition symbol
  count ← a table, where count[c] is initially the number of symbols in clause c's premise
  inferred ← a table, where inferred[s] is initially false for all symbols
  queue ← a queue of symbols, initially symbols known to be true in KB

  while queue is not empty do
     p ← POP(queue)
     if p = q then return true
     if inferred[p] = false then
        inferred[p] ← true
        for each clause c in KB where p is in c.PREMISE do
           decrement count[c]
           if count[c] = 0 then add c.CONCLUSION to queue
  return false
```

For each clause we keep the number of not yet verified premises that is decreased when we infer a new fact. The clause with zero unverified premises gives a new fact (from the head of the clause).

- **sound and complete** algorithm for Horn clauses
- **linear time complexity** in the size of knowledge base

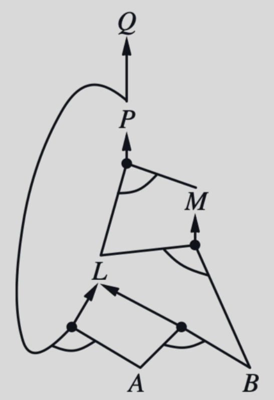
Forward chaining in example





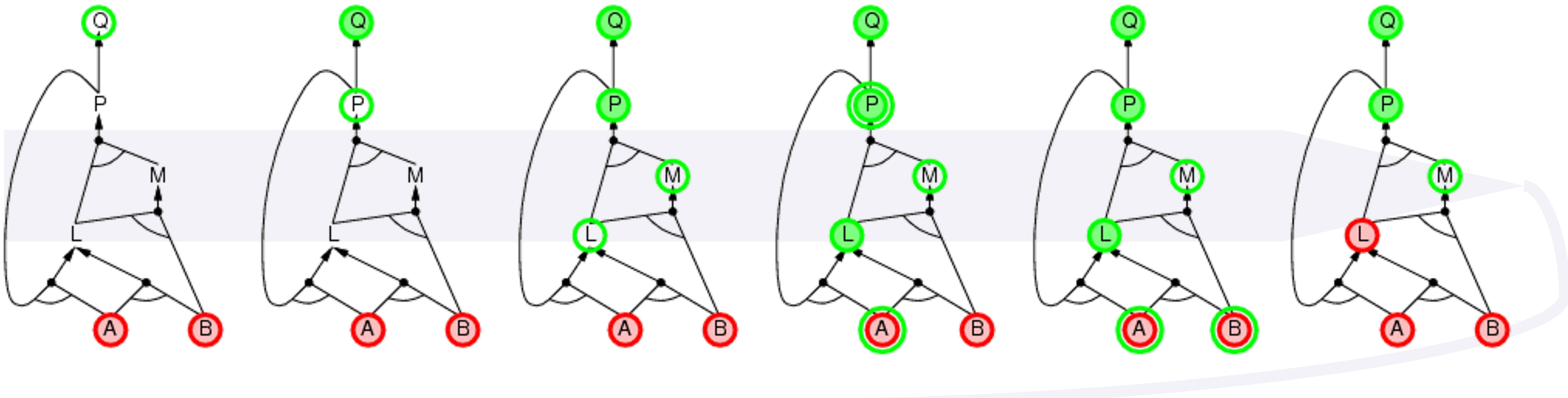
Knowledge base with a graphical representation

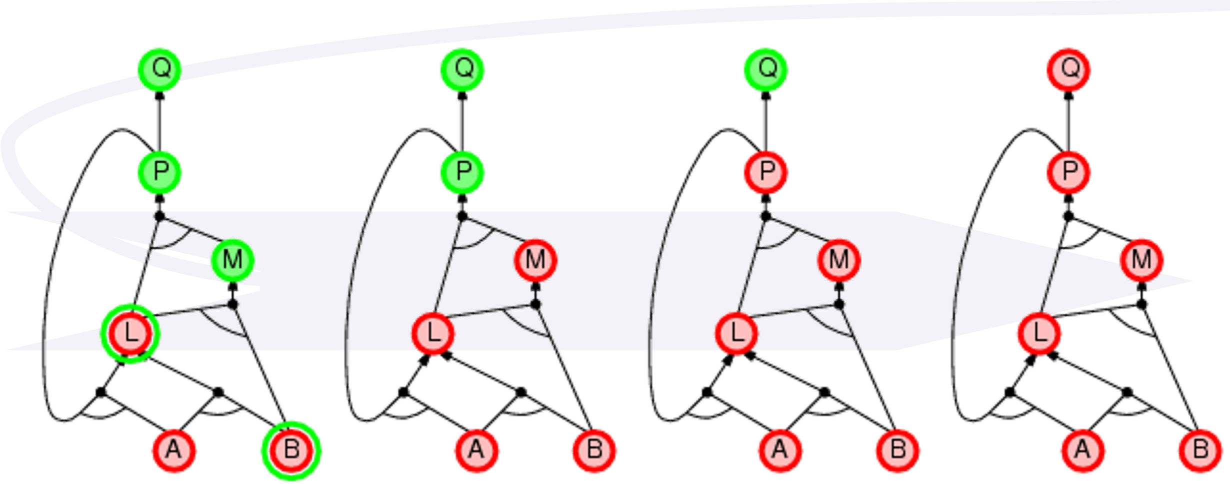
$$P \Rightarrow Q$$
$$L \wedge M \Rightarrow P$$
$$B \wedge L \Rightarrow M$$
$$A \wedge P \Rightarrow L$$
$$A \wedge B \Rightarrow L$$
$$A$$
$$B$$



The query is decomposed (via the Horn clause) to sub-queries until the facts from KB are obtained.

Goal-driven reasoning.





Knowledge base with a graphical representation

$P \Rightarrow Q$

$L \wedge M \Rightarrow P$

$B \wedge L \Rightarrow M$

$A \wedge P \Rightarrow L$

$A \wedge B \Rightarrow L$

A

B

For simplicity we will represent only the "physics" of the Wumpus world.

- we know that
  - $\neg P_{1,1}$
  - $\neg W_{1,1}$
- we also know why and where breeze appears
  - $B_{x,y} \Leftrightarrow (P_{x,y+1} \vee P_{x,y-1} \vee P_{x+1,y} \vee P_{x-1,y})$
- and why a smell is generated
  - $S_{x,y} \Leftrightarrow (W_{x,y+1} \vee W_{x,y-1} \vee W_{x+1,y} \vee W_{x-1,y})$
- and finally one "hidden" information that there is a single Wumpus in the world
  - $W_{1,1} \vee W_{1,2} \vee \ldots \vee W_{4,4}$
  - $\neg W_{1,1} \vee \neg W_{1,2}$
  - $\neg W_{1,1} \vee \neg W_{1,3}$
  - …

We can also include information about the **agent**.

- where the agent is
  - $L_{1,1}$
  - $FacingRight^1$
- and what happens when agent performs actions
  - $L^t_{x,y} \wedge FacingRight^t \wedge Forward^t \Rightarrow L^{t+1}_{x+1,y}$
  - we need an upper bound for the number of steps and still it will lead to a huge number of formulas

The Wumpus world: a hybrid agent

function HYBRID-WUMPUS-AGENT(*percept*) **returns** an *action*
  inputs: *percept*, a list, [*stench*,*breeze*,*glitter*,*bump*,*scream*]
  persistent: *KB*, a knowledge base, initially the atemporal "wumpus physics"
    t, a counter, initially 0, indicating time
    plan, an action sequence, initially empty

  TELL(*KB*, MAKE-PERCEPT-SENTENCE(*percept*, *t*))
  TELL the *KB* the temporal "physics" sentences for time *t*
  $safe \leftarrow \{[x, y] : \text{ASK}(KB, OK^t_{x,y}) = true\}$
  if ASK(*KB*, $Glitter^t$) = *true* **then**
    plan ← [*Grab*] + PLAN-ROUTE(*current*, {[1,1]}, *safe*) + [*Climb*]
  if *plan* is empty **then**
    $unvisited \leftarrow \{[x, y] : \text{ASK}(KB, L^{t'}_{x,y}) = false \text{ for all } t' \leq t\}$
    plan ← PLAN-ROUTE(*current*, *unvisited* ∩ *safe*, *safe*)
  if *plan* is empty and ASK(*KB*, $HaveArrow^t$) = *true* **then**
    $possible\_wumpus \leftarrow \{[x, y] : \text{ASK}(KB, \neg W_{x,y}) = false\}$
    plan ← PLAN-SHOT(*current*, *possible_wumpus*, *safe*)
  if *plan* is empty **then**   *// no choice but to take a risk*
    $not\_unsafe \leftarrow \{[x, y] : \text{ASK}(KB, \neg OK^t_{x,y}) = false\}$
    plan ← PLAN-ROUTE(*current*, *unvisited* ∩ *not_unsafe*, *safe*)
  if *plan* is empty **then**
    plan ← PLAN-ROUTE(*current*, {[1, 1]}, *safe*) + [*Climb*]
  action ← POP(*plan*)
  TELL(*KB*, MAKE-ACTION-SENTENCE(*action*, *t*))
  t ← *t* + 1
  return *action*

Add information about current observation

Find provably safe (no danger there) rooms.

Gold found, grab it and escape.

Explore the area – find a safe way to some frontier room.

No safe exploration, try to shoot Wumpus.

Explore the area with some risk (not provably safe).

OK, no way to gold (without being killed), escape the cave.

Department of Theoretical Computer Science and Mathematical Logic
bartak@ktiml.mff.cuni.cz