

Explainable AI Past, Present, and Future: A Scientific Modeling Approach



User with
a Task

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

William J. Clancey, PhD

Senior Research Scientist

Florida Institute for Human & Machine Cognition

Pensacola, FL

Explaining Expert System Behaviors in 1970s

“INTERACTIVE EXPLANATION CAPABILITIES:

Another important component of the system permits the consultation program to explain its knowledge of infectious disease therapy and to justify specific therapeutic recommendations.”

MYCIN Project Proposal, Stanford University, October 1973

**** WHY DIDN'T YOU NEED TO KNOW WHETHER ORGANISM-1 IS A CONTAMINANT?**

During the preceding consultation, I never had reason to try to find out whether ORGANISM-1 is a contaminant. It would have been necessary to know this information in order to use any of the following: RULE006, RULE007, RULE106, RULE108, RULE109, RULE159. However, none of these succeeded in the context of ORGANISM-1. If you would like an explanation for why any of these rules failed, please enter their numbers:

**** 159**

Rule159 was tried in the context of ORGANISM-1, but it failed due to clause 1 ["it is suspected that the identity of the organism is corynebacterium-non-diphtheriae"]

Sample Questions about a Consultation

- 1) what is <parm> of <cntxt>
TO WHAT CLASS DOES ORGANISM-1 BELONG?
IS ORGANISM-1 CORYNEBACTERIUM-NON-DIPHThERIAE?
- 2) how do you know the value of <parm> of <cntxt>
HOW DO YOU KNOW THAT CULTURE-1 WAS FROM A STERILE SOURCE?
DID YOU CONSIDER THAT ORGANISM-1 MIGHT BE A BACTEROIDES?
WHY DON'T YOU THINK THAT THE SITE OF CULTURE-1 IS URINE?
WHY DID YOU RULE OUT STREPTOCOCCUS AS A POSSIBILITY FOR ORGANISM-1?
- 3) how did you use <parm> of <cntxt>
DID YOU CONSIDER THE FACT THAT PATIENT-1 IS A COMPROMISED HOST?
HOW DID YOU USE THE AEROBICITY OF ORGANISM-1?
- 4) why didn't you find out about <parm> of <cntxt>
DID YOU FIND OUT ABOUT THE CBC ASSOCIATED WITH CULTURE-1?
WHY DIDN'T YOU NEED TO KNOW WHETHER ORGANISM-1 IS A CONTAMINANT?
- 5) what did <rule> tell you about <cntxt>
HOW WAS RULE 178 HELPFUL WHEN YOU WERE CONSIDERING ORGANISM-1?
DID RULE 116 TELL YOU ANYTHING ABOUT INFECTION-1?
WHY DIDN'T YOU USE RULE 189 FOR ORGANISM-2?

“Why Not” Questions in MYCIN (1976)

**** WHY DID YOU RULE OUT STREPTOCOCCUS AS A POSSIBILITY FOR ORGANISM-1?**

The following rules could have been used to determine that the identity of ORGANISM-1 is streptococcus: RULE033. However, none of these succeeded in the context of ORGANISM-1. If you would like an explanation for why any of these rules failed, please enter their numbers:

**** 33**

Clause 2 of rule033 (“the morphology of the organism is coccus”) was already known to be false for ORGANISM-1, so the rule was never tried.

**** WHY DON'T YOU THINK THAT THE MORPHOLOGY OF ORGANISM-1 IS COCCUS?**

It is definite that the morphology of ORGANISM-1 is rod. Knowing this with certainty rules out all other values for the morphology of ORGANISM-1, including coccus.

MYCIN Consultation & Typical Rule

31) Has Pt538 ever undergone any type of surgery?
** YES
32) Does Pt538 have a history of NEUROSURGERY?
** NO
33) Does Pt538 live in a crowded environment?
** NO
34) Do you suspect recent alcoholic history in Pt538?
** NO
35) Is meningitis a hospital-acquired infection?
** YES
36) Is Pt538's clinical history consistent with EPIGLOTTITIS?
** NO
37) Is Pt538's clinical history consistent with OTITIS-MEDIA?
** NO
38) Has Pt538 ever undergone splenectomy?
** NO
39) Is Pt538 a burn patient?
** YES

If: 1) The infection which requires therapy is meningitis,
2) Only circumstantial evidence is available for this case,
3) The type of meningitis is bacterial,

4) The age of the patient is greater than 17 years, and
5) The patient is an alcoholic,

Then: There is evidence that the organisms which might be causing the infection are
diplococcus-pneumoniae (.3) or e.coli (.2)

Design of MYCIN's Rule Set

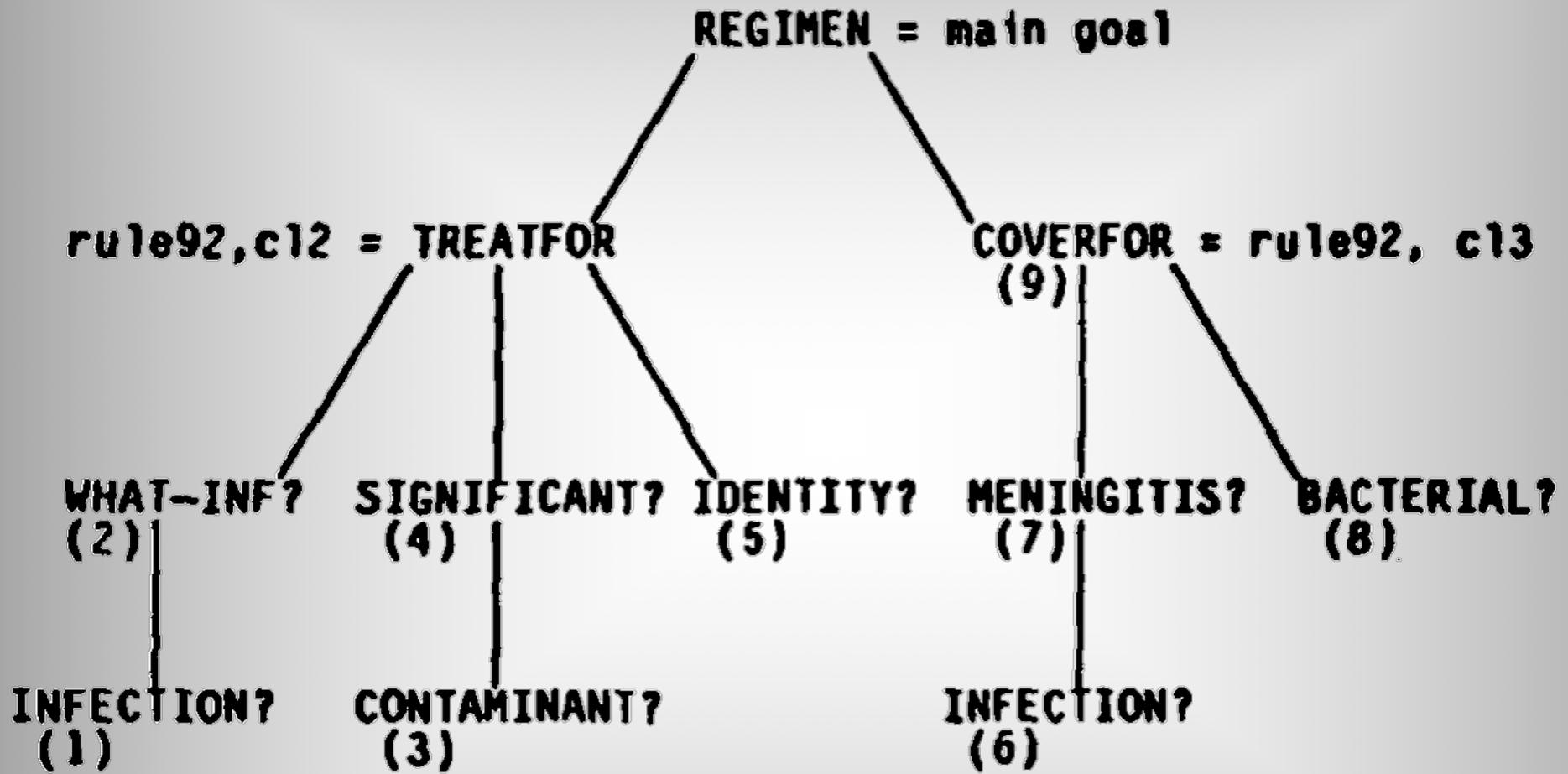


Figure 14. Portion of MYCIN's inference structure
(Numbers give the order in which non-place-holder goals are achieved by the depth-first interpreter.)

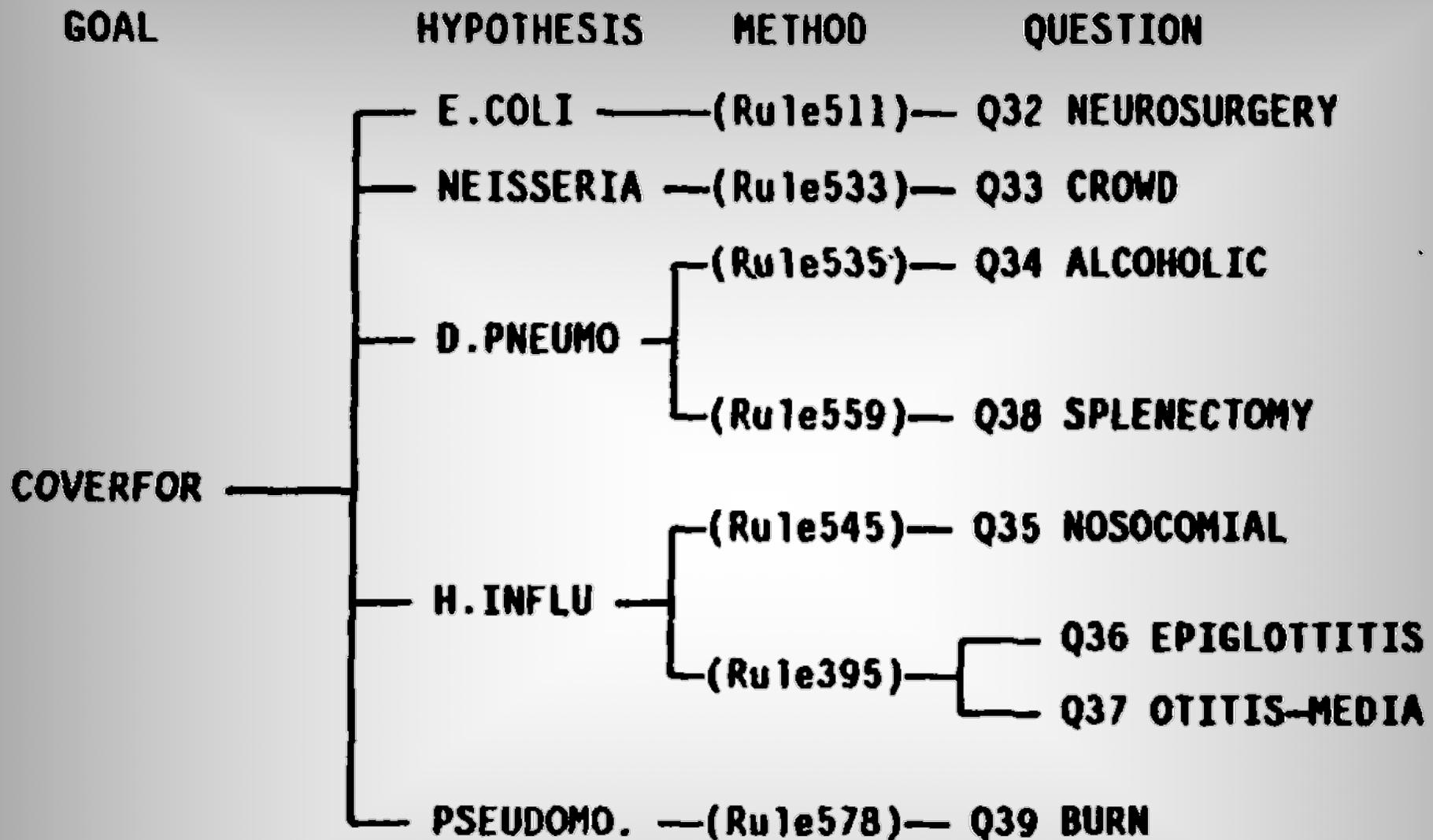


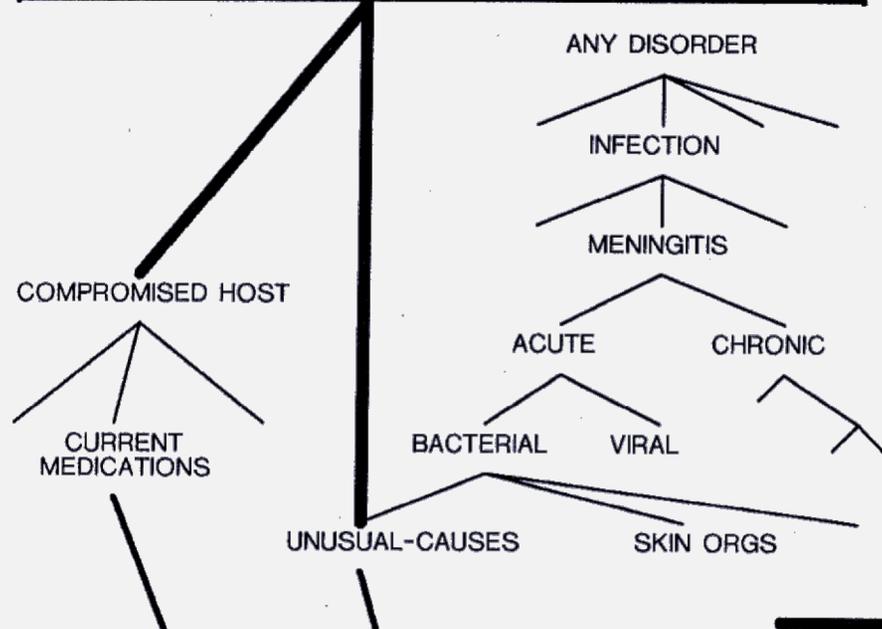
Figure 12. Portion of the AND/OR tree corresponding to the questions shown in Figure 11 (reorganized according to the hypothesis each rule supports).

Explanation Framework for Rule-based Diagnostic Systems

ESTABLISH HYPOTHESIS SPACE:
CONSIDER DIFFERENTIAL-BROADENING FACTORS

STRATEGY

(RULE MODEL)
IN BACTERIAL MENINGITIS, COMPROMISED HOST
RISK FACTORS SUGGEST UNUSUAL ORGANISMS



STRUCTURE

if STEROIDS then GRAM-NEGATIVE ROD ORGS

INFERENCE RULE

STERIODS IMPAIR IMMUNO-RESPONSE
MAKING PATIENT SUSCEPTIBLE TO
INFECTION BY ENTEROBACTERIACEAE,
NORMALLY FOUND IN THE BODY

SUPPORT

“Knowledge for indexing, justifying, and invoking a MYCIN rule.”

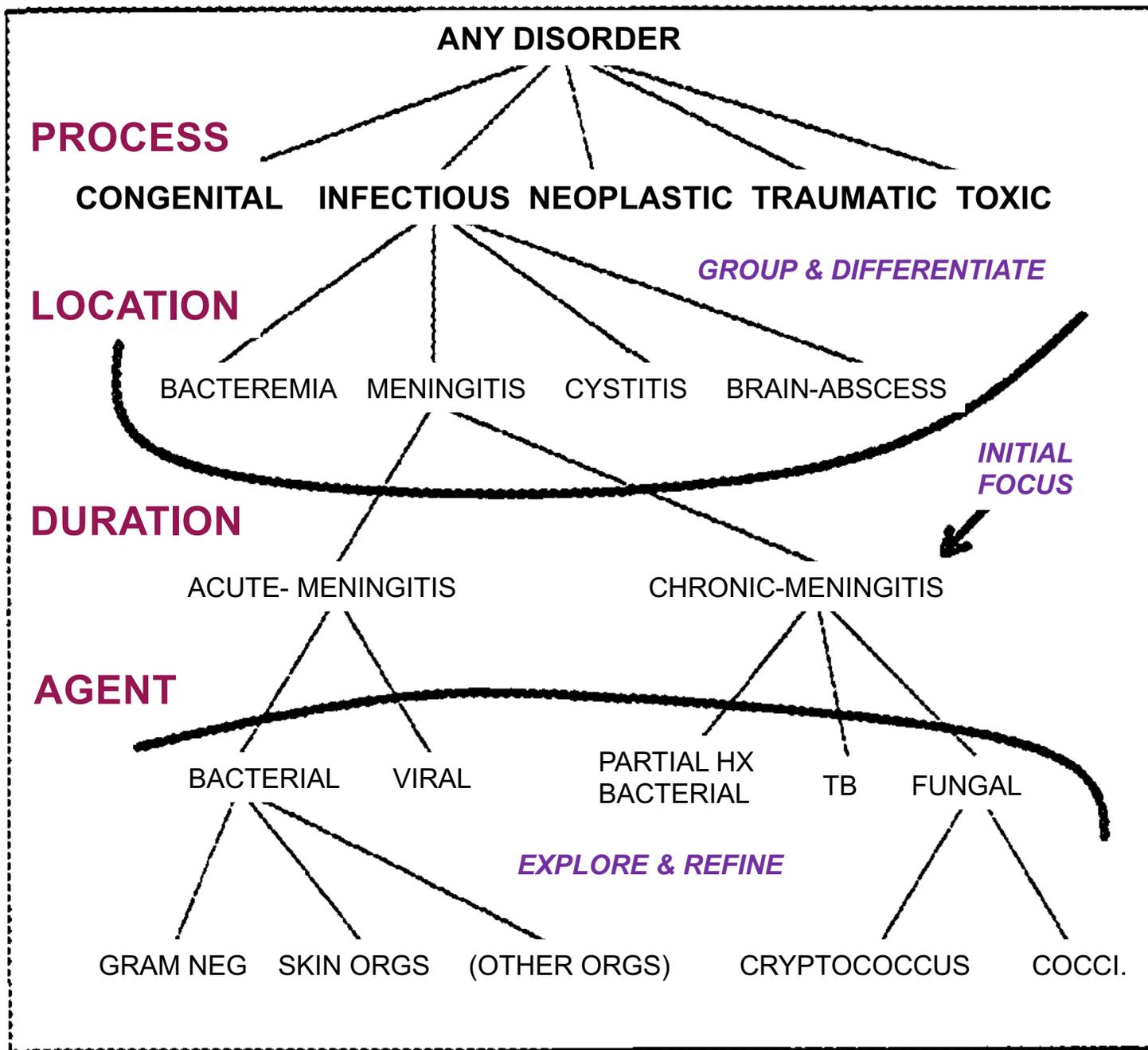


Figure 11. Looking Up and Looking Down in Diagnostic Search. Disease knowledge is represented as a taxonomy of processes. At the highest level are internal aber-

HEURISTIC

Compromised Host \Rightarrow Gram-Negative Infection

GENERALIZATION

SUBTYPE

Immunosuppressed

E.coli Infection

GENERALIZATION

Leukopenia

Heuristic Classification Pattern in MYCIN

DEFINITIONAL

HEURISTIC MATCH

Patient Abstractions \Rightarrow Disease Classes

Low WBC

DATA
ABSTRACTION

REFINEMENT

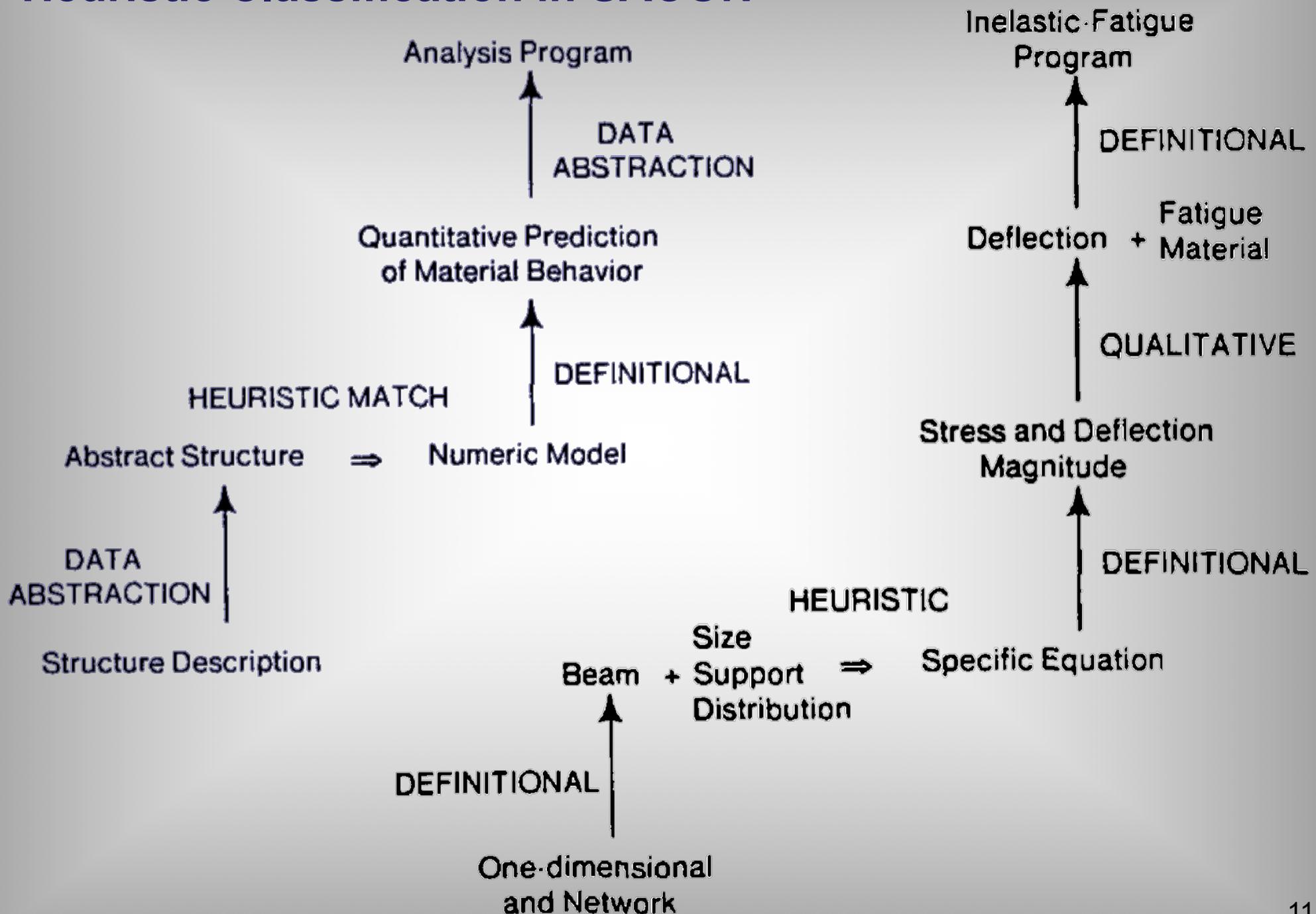
QUALITATIVE

Patient Data

Diseases

WBC < 2.5

Heuristic Classification in SACON



CASTER Consultation

MYCIN → EMYCIN → SACON :: Neomycin → HERACLES → Caster

16-May-85 22:43:56

-----CASTING-4-----

Please enter information about the casting

1) Casting Type:

IRON

2) Please describe the chief complaints:

** SHRINKAGE-CAVITIES

**

Caster begins by applying the heuristic rule suggesting inadequate supply.

DIFFERENTIAL: (INADEQUATE-SUPPLY 800)

Caster is exploring and refining the differential.

Inadequate-supply has been replaced by metal-leak, feed-is-shut-off, and no-metal-to-feed. Metal-leak suggests a broken-mold, so Caster inquires whether there is evidence of a runout.

3) Does Casting-4 have a runout?

** N

Caster has looked at feed-is-shut-off, which suggests that fillets may be too small, so the system asks if the shrink is at a corner, which would be strong evidence for fillets-too-small.

4) Does Casting-4 have a shrink at a corner?

** Y

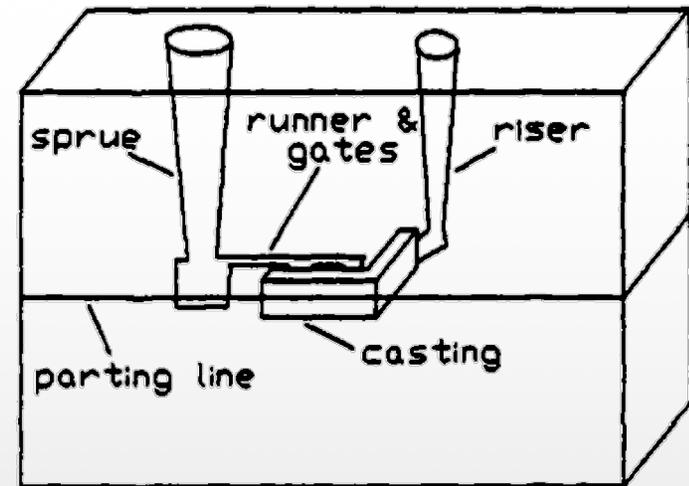
Finally, Caster looks for evidence that there is no metal available to feed, namely that there is a high reject ratio.

5) Does Casting-4 have a high reject ratio?

** N

DIFFERENTIAL: (METAL-LEAK 100) (FEED-IS-SHUT-OFF 100)

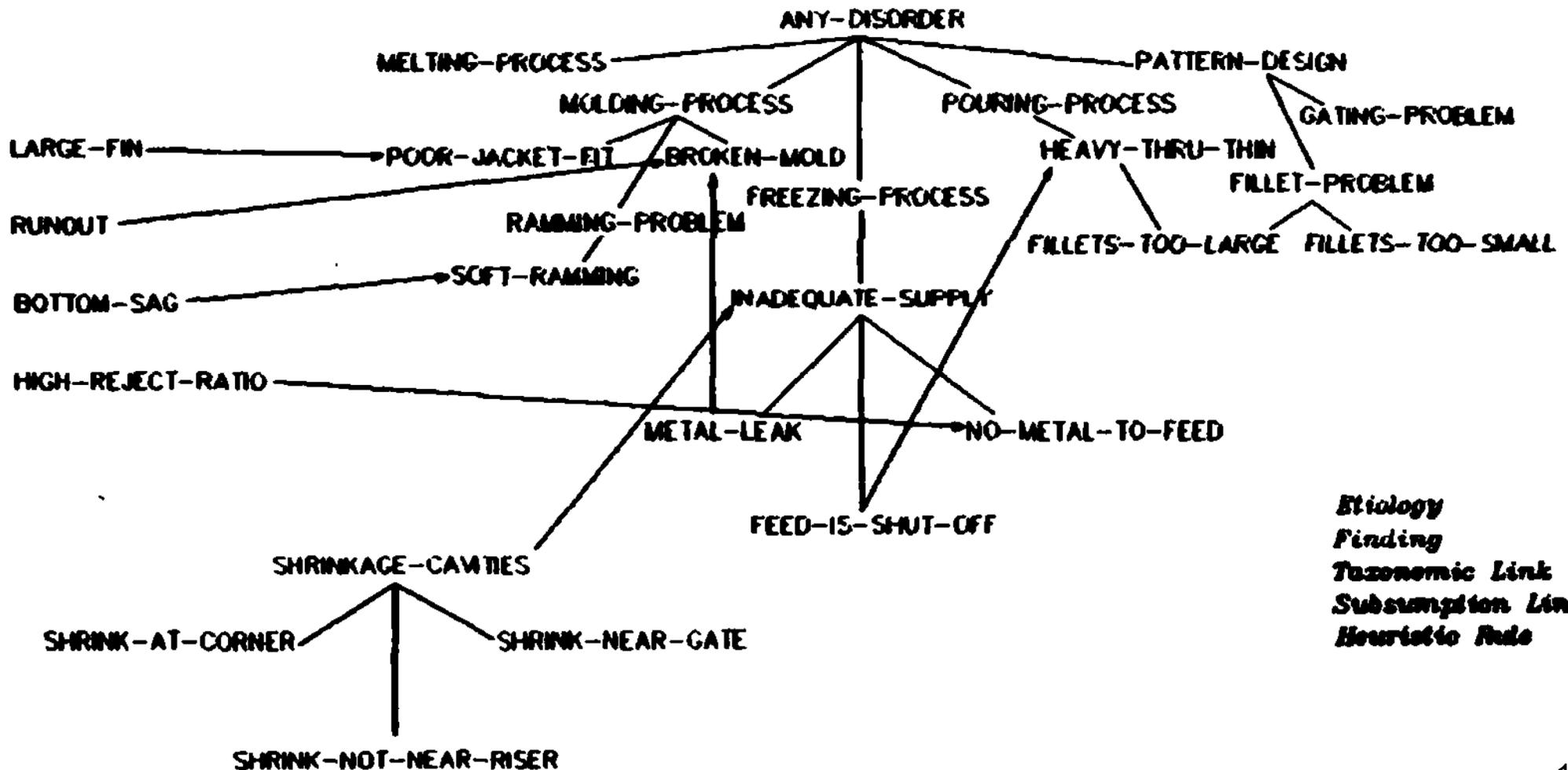
(FILLETS-TOO-SMALL 700)



Scientific approach:

- Abstract all processes
- Represent in general systems-modeling framework
- Create domain-general frameworks (consultation, tutoring, explanation) to manipulate situation-specific models.

Caster's Disorder Taxonomy of Sand-Casting Processes



Therapy Explanation in MYCIN

**** WHY DID YOU GIVE CHLORAMPHENICOL FOR E.COLI IN REC-1?**

CHLORAMPHENICOL was prescribed for ITEM-2 in RECOMMENDATION-1:

Since

- CHLORAMPHENICOL is a treatment of choice for e.coli in meningitis
- ITEM-2 is sensitive to CHLORAMPHENICOL
- there were no contraindications for it

CHLORAMPHENICOL was prescribed because it was part of the recommendation that covers for all of the items, using the fewest number of drugs.

Replacing original therapy rules by table-driven generate & test method made the therapy selection process formal & perspicuous, significantly improving maintainability, and provided a new capability for explanation and instruction.

“Why Not?” Therapy Explanation in MYCIN

**** WHY DIDN'T YOU SUGGEST PENICILLIN IN REC-1 FOR STAPH-COAG + ?**

PENICILLIN was not prescribed for ITEM-1 in RECOMMENDATION-1:

PENICILLIN was discounted for ITEM-1 because it is NOT DEFINITE that the item is sensitive to this drug. There are other potential therapies under consideration which are much more desirable, viz., current therapies or drugs to which the item is definitely sensitive.

Would you like to see some details? ** YES

The drugs to which the staphylococcus-coag-pos is sensitive are: cephalothin (1.0) vancomycin (1.0) gentamycin (1.0) tobramycin (1.0) erythromycin-and-tetracycline (1.0) chloramphenicol-and-erythromycin (1.0) [RULE098 RULE445]

Would you like to know about the history of PENICILLIN in the decision process up to this point? ** YES

-- PENICILLIN is a treatment of choice for staphylococcus-coag-pos in meningitis. But as explained above, PENICILLIN was discounted.

Critiquing Method for Explanation

(Developed for MYCIN/Guidon c. 1978)

**** WHY DIDN'T YOU GIVE AMPICILLIN-AND-GENTAMICIN FOR E.COLI IN REC-1?**

AMPICILLIN-AND-GENTAMICIN was not prescribed for ITEM-2 in RECOMMENDATION-1:

AMPICILLIN-AND-GENTAMICIN is a plausible choice for e.coli in meningitis, and was not explicitly rejected for use against ITEM-2 in RECOMMENDATION-1. However, the best therapy did not include AMPICILLIN-AND-GENTAMICIN.

If you would like to suggest therapy which includes AMPICILLIN-AND-GENTAMICIN, your regimen will be compared to MYCIN's. Would you like to do this? **** YES**

Critiquing Method (continued)

For each item in turn, enter the drug you would have prescribed in RECOMMENDATION-1.

In order to minimize the number of drugs in your recommendation, you may not want to prescribe therapy for every item. Items which represent the most likely organisms are indicated with a plus sign (+).

- + ITEM-1 -- the staphylococcus-coag-pos ** **GENTAMICIN**
- + ITEM-2 -- the e.coli ** **AMPICILLIN GENTAMICIN**
- + ITEM-3 -- the klebsiella-pneumoniae ** **GENTAMICIN**
- + ITEM-4 -- the pseudomonas-aeruginose **

[Checking for contraindications . . .]

[Considering AMPICILLIN-AND-GENTAMICIN for use against INFECTION-1 . . .]

[No contraindications found . . .]

[Now comparing your prescription to MYCIN's . . .]

Critiquing Method (continued)

Perhaps you did not realize that one of the drugs you prescribed, GENTAMICIN, will cover for ITEM-4, an item for which you did not prescribe therapy. I have changed your prescription accordingly.

| ORGANISMS | Your regimen Drug -- Choice | MYCIN's regimen Drug -- Choice |
|---------------|--------------------------------------|-------------------------------------------|
| "most likely" | | |
| ITEM-3 | GENTAMICIN -- 3rd | CHLORAMPHENICOL-AND- GENTAMICIN -- 1st |
| ITEM-2 | AMPICILLIN-AND- GENTAMICIN -- 1st | CHLORAMPHENICOL -- 1st |
| ITEM-1 | GENTAMICIN -- 2nd | GENTAMICIN -- 2nd |
| "less likely" | | |
| ITEM-4 | GENTAMICIN -- 2nd | GENTAMICIN -- 2nd |

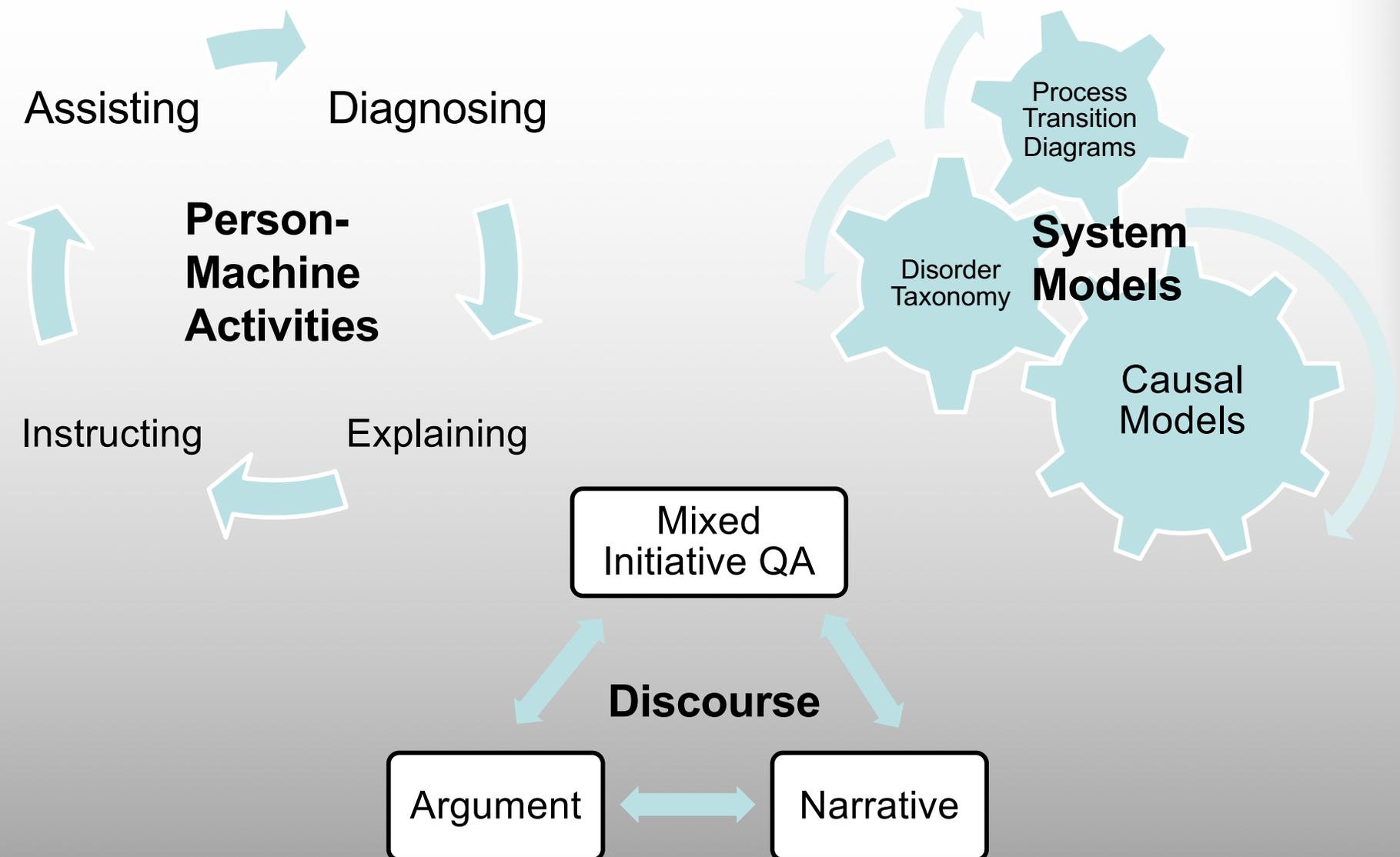
(The desirability of a drug is defined to be its lowest ranking for the items it covers.)

Both prescriptions include fewer than 3 drugs, so we must look at how highly ranked each prescription is for the most likely organism(s).

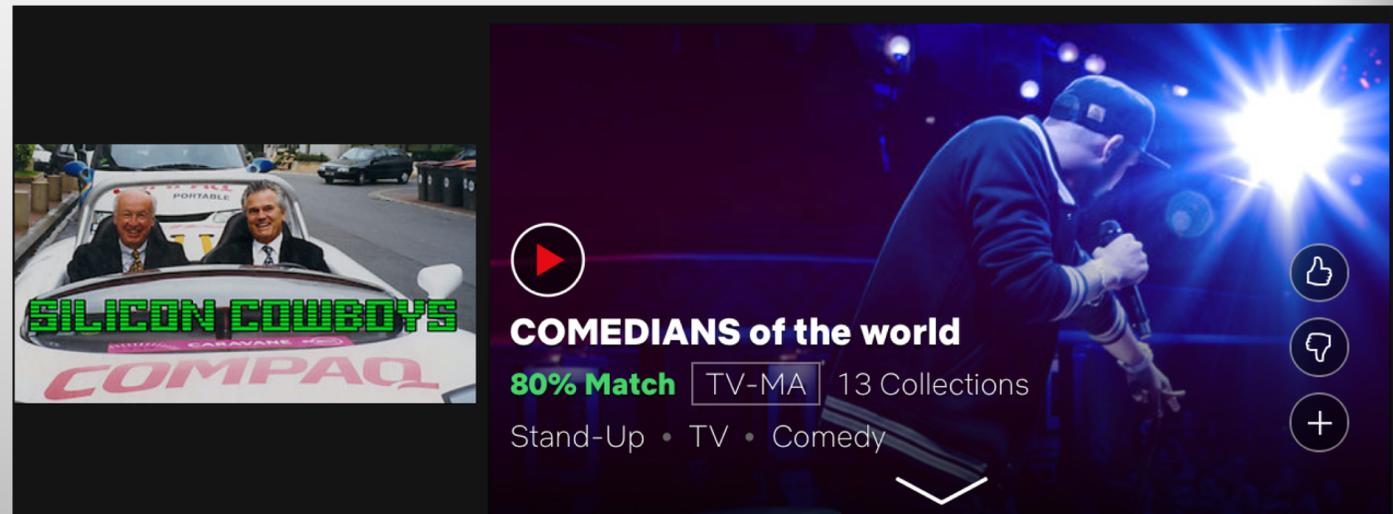
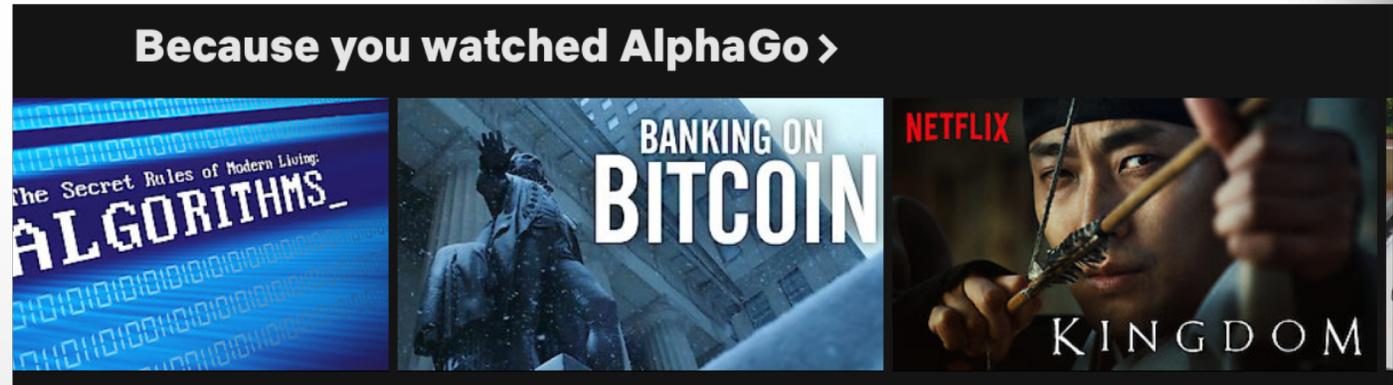
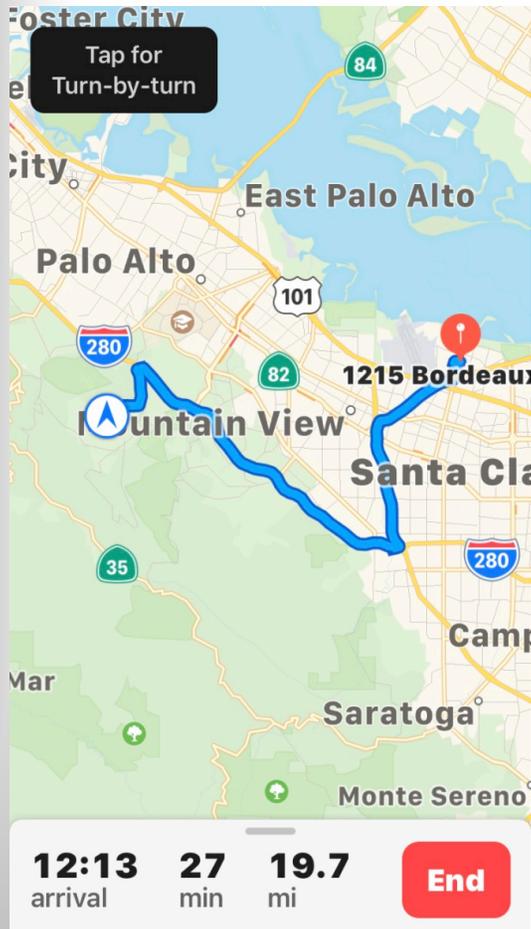
Your prescription of 1 first choice drug (AMPICILLIN for ITEM-2) and 1 third choice drug (GENTAMICIN for ITEM-3) is not as good as MYCIN's prescription of 1 first choice drug (CHLORAMPHENICOL for ITEM-2 and Item-3) and 1 second choice drug (GENTAMICIN for ITEM-1).

[You may refer to your regimen as RECOMMENDATION-2 in later questions.]

Explanation Research was an Integral Part of AI Applications in 1970s & 1980s



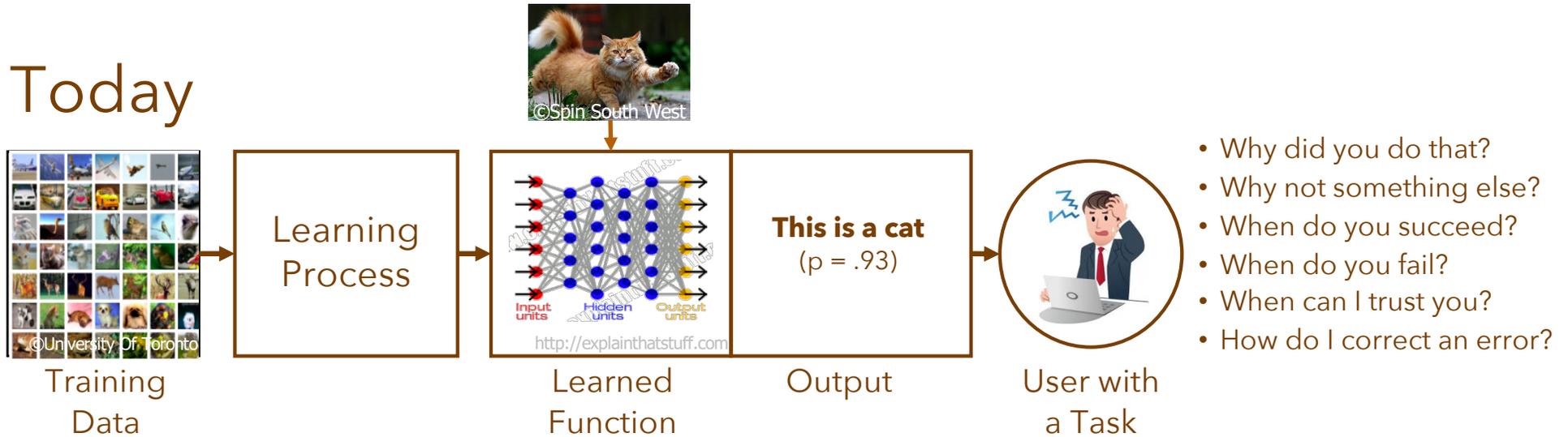
Common Interactive Systems Today Cannot Explain Behavior or Advice



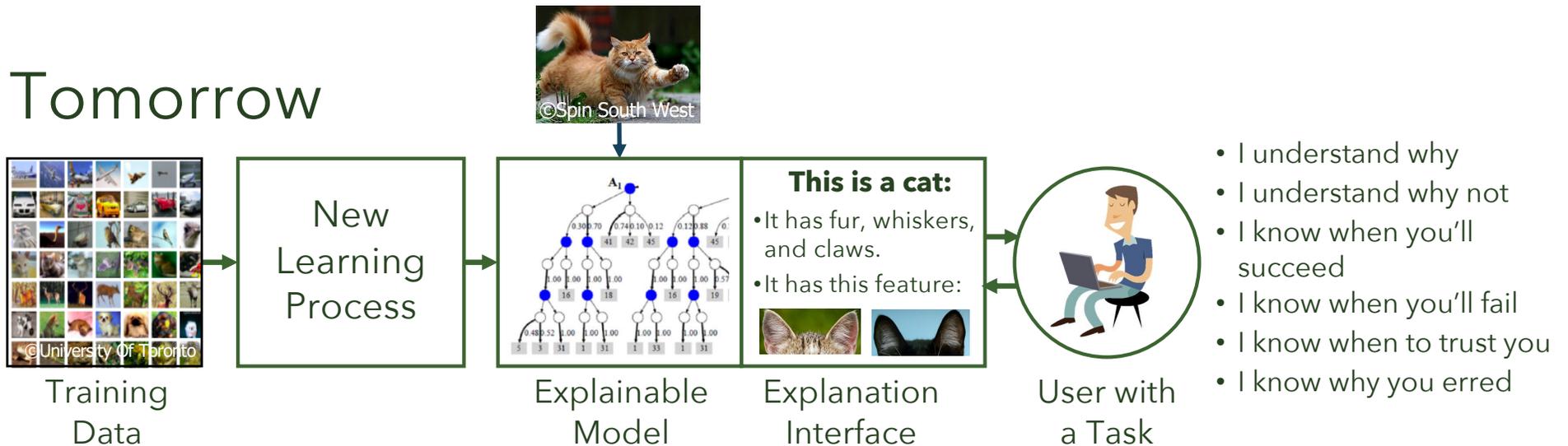
Why not Page Mill & 101?
What comes after 280?

How are these Netflix suggestions related to AlphaGo?
Show me the highest “match” ones first...

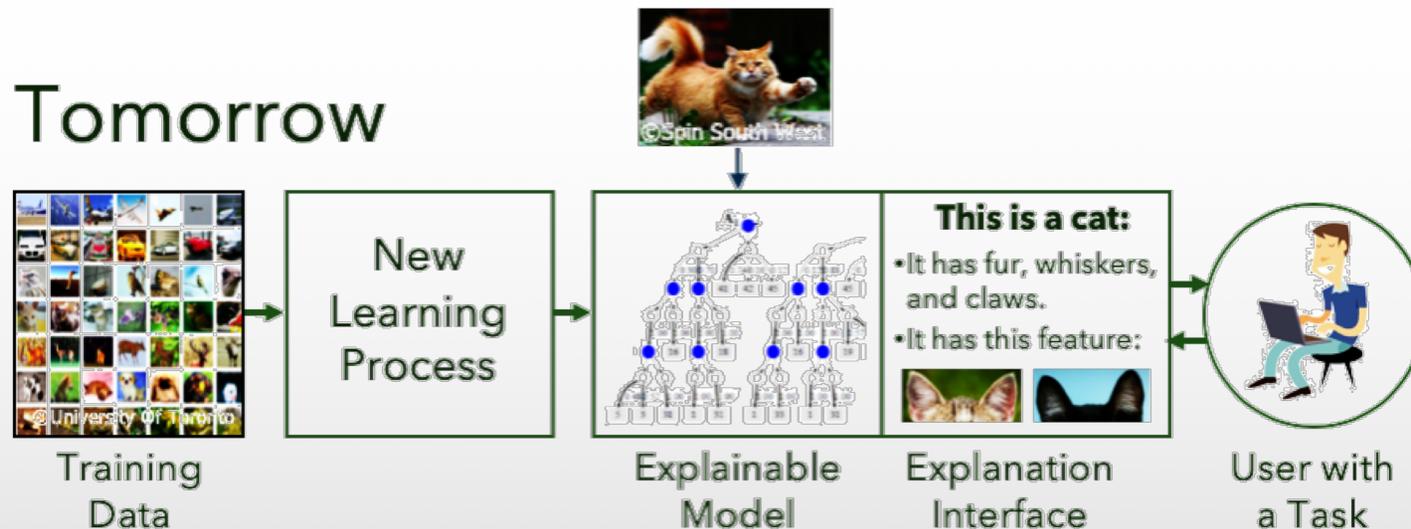
Today



Tomorrow



Future Research for Technology People Understand & Trust



- Operations are dynamic and interactive in a system of people, technology, and the environment – not linear & placeless.
- Explanation involves discourse, follow-up and mutual learning; requires a “user model” of interests & knowledge.
- “Explanation” is not a module – rather it drives the design process; needs are empirically discovered in prototype experiments.



Shift to Systems Thinking:

Work Practice Constrains Explanation Requirements

- Nature of the work system activity (e.g., diagnosing, predicting, controlling, configuring, planning)
- Operational setting (e.g., safety critical, extreme environment, business office)
- The role and capabilities of the worker
- Interactions with other tools and people in practice
- The dynamics of interactions during work activities (especially time available)



Work System Design Methodology: *Empirical requirements analysis, Participatory design, Participant observation, Incrementally prototyping useful tools, Experiments in authentic work contexts*



Explanation Might Address Shortcomings of Symbolic AI



- Identifying domain representations as “knowledge” obscured system-modeling methods & hence the domain-general scientific accomplishment
→ *abstract the modeling frameworks & operational tasks*
- Ongoing tuning and extension required “Knowledge Engineers”
→ *e.g., need principles for appropriate/sufficient data sets*
- Brittle – boundaries not tested; system not reflective
→ *recognize operating outside designed & tested situations*
- Not integrated with legacy systems & work practice
→ *develop tools for work systems; iterative experiments with prototypes*

For more information...

- Buchanan & Shortliffe, 1984. ***Rule-Based Expert Systems: The MYCIN Experiments of the Heuristic Programming Project.*** Available: www.Shortliffe.net.
- Clancey, 1985. **Heuristic classification.** *Artificial Intelligence* 27:289-350.
- Thompson & Clancey, 1986. **A qualitative modeling shell for process diagnosis.** *IEEE Software* 3(2) 6-15.
- Clancey, 1987. ***Knowledge-Based Tutoring.*** MIT Press. (1979 PhD dissertation)
- Clancey, 1992. **Model construction operators.** *Artificial Intelligence* 53:1–115.
- Clancey, 2012. ***Working on Mars: Voyages of Discovery with the Mars Exploration Rovers.*** MIT Press.
- Arnold, (1959) 2016. ***Creative Engineering: Promoting Innovation by Thinking Differently.*** Introduction by WJ Clancey. See <http://Bill.Clancey.name>.

