Differential Diagnosis

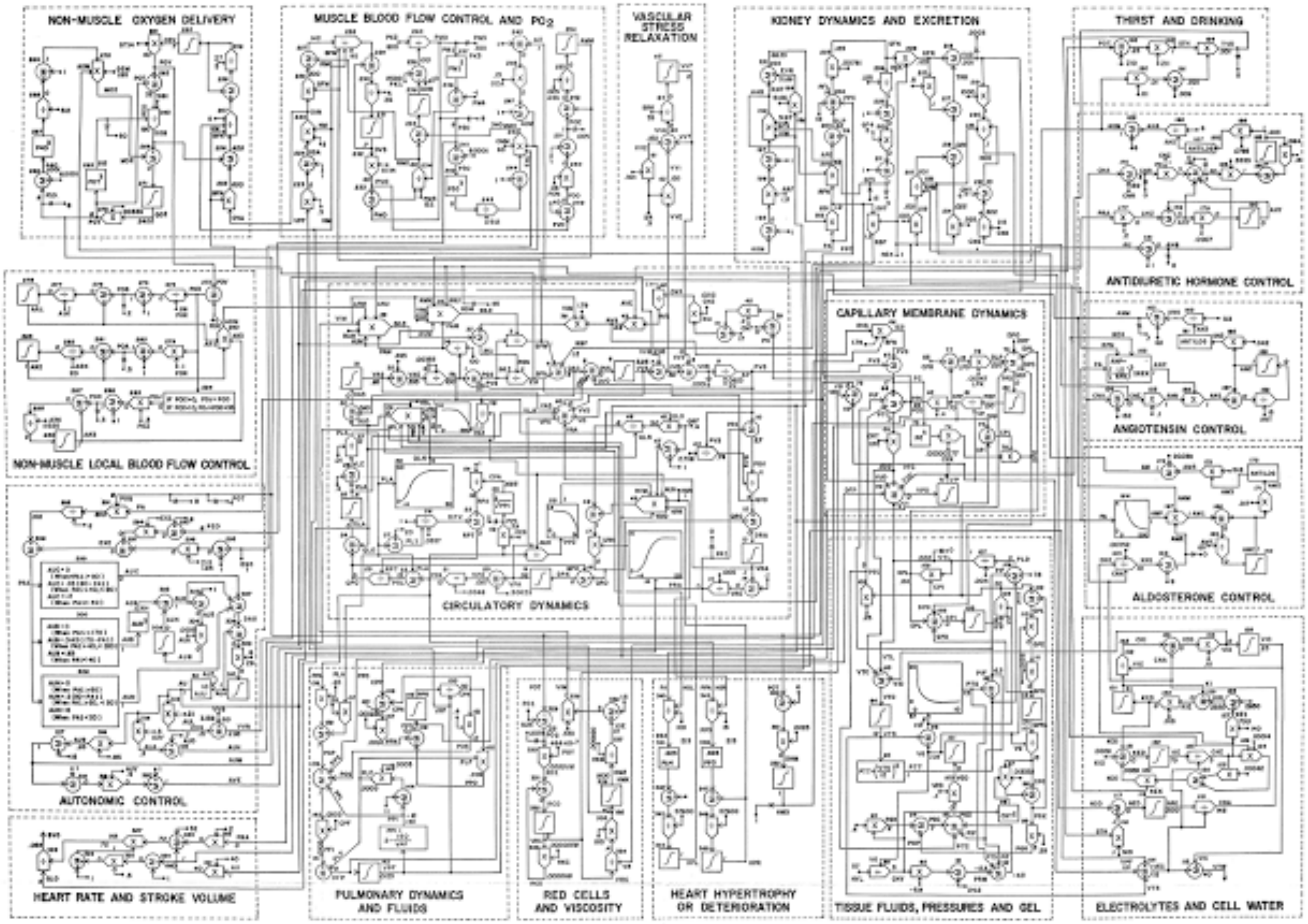
May 5, 2020

“**Diagnosis** is the identification of the nature and cause of a certain phenomenon”
“**differential diagnosis** is the distinguishing of a particular disease or condition from others that present similar clinical features”

—Wikipedia

Massachusetts Institute of Technology

Guyton's Model of Cardio-vascular Dynamics



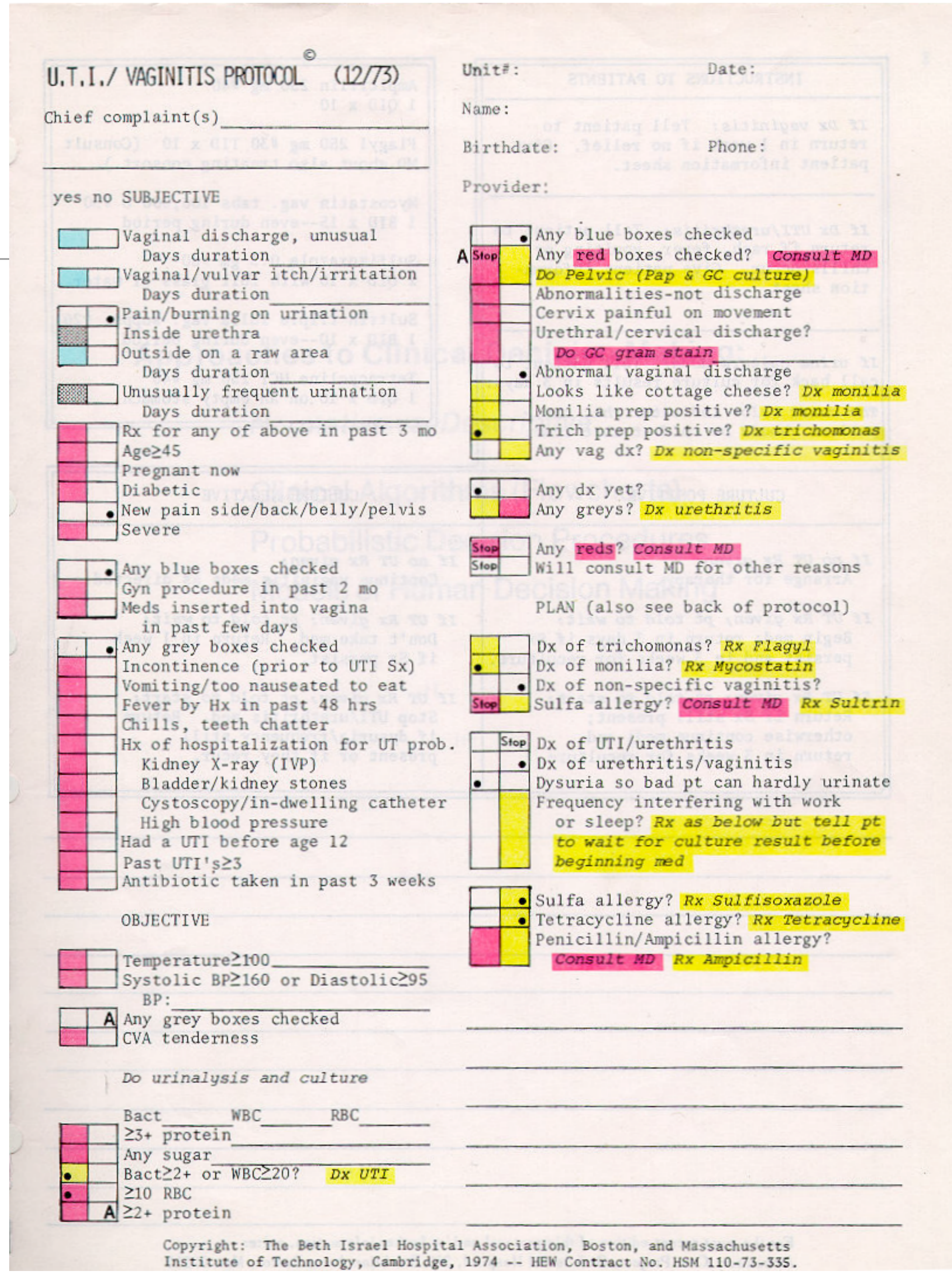
Models for Diagnostic Reasoning

- Flowcharts
- Based on associations between diseases and {signs, symptoms}
  - *"manifestations"* covers all observables, including lab tests, bedside measurements, …
- Single disease vs. multiple diseases
- Probabilistic vs. categorical
- Utility theoretic
- Rule-based
- Pattern matching

Sign: Any objective evidence of disease, as opposed to a **symptom**, which is, by nature, subjective. For example, gross blood in the stool is a sign of disease; it is evidence that can be recognized by the patient, physician, nurse, or someone else. Abdominal pain is a symptom; it is something only the patient can perceive.
https://www.medicinenet.com/script/main/art.asp?articlekey=5493

Flowchart

- BI/Lincoln Labs Clinical Protocols



©

U.T.I./ VAGINITIS PROTOCOL (12/73)

Chief complaint(s) ________

Unit#: Date:

Name:

Birthdate: Phone:

Provider:

yes no SUBJECTIVE

Vaginal discharge, unusual
Days duration ______
Vaginal/vulvar itch/irritation
Days duration ______
Pain/burning on urination
Inside urethra
Outside on a raw area
Days duration ______
Unusually frequent urination
Days duration ______
Rx for any of above in past 3 mo
Age≥45
Pregnant now
Diabetic
New pain side/back/belly/pelvis
Severe

Any blue boxes checked
Gyn procedure in past 2 mo
Meds inserted into vagina
in past few days
Any grey boxes checked
Incontinence (prior to UTI Sx)
Vomiting/too nauseated to eat
Fever by Hx in past 48 hrs
Chills, teeth chatter
Hx of hospitalization for UT prob.
Kidney X-ray (IVP)
Bladder/kidney stones
Cystoscopy/in-dwelling catheter
High blood pressure
Had a UTI before age 12
Past UTI's≥3
Antibiotic taken in past 3 weeks

OBJECTIVE

Temperature≥100 ______
Systolic BP≥160 or Diastolic≥95
BP: ______
A Any grey boxes checked
CVA tenderness

Do urinalysis and culture

Bact ___ WBC ___ RBC ___
≥3+ protein
Any sugar
Bact≥2+ or WBC≥20? *Dx UTI*
≥10 RBC
A ≥2+ protein

Any blue boxes checked
A Stop Any red boxes checked? *Consult MD*
Do Pelvic (Pap & GC culture)
Abnormalities-not discharge
Cervix painful on movement
Urethral/cervical discharge?
Do GC gram stain
Abnormal vaginal discharge
Looks like cottage cheese? *Dx monilia*
Monilia prep positive? *Dx monilia*
Trich prep positive? *Dx trichomonas*
Any vag dx? *Dx non-specific vaginitis*

Any dx yet?
Any greys? *Dx urethritis*

Stop Any reds? *Consult MD*
Stop Will consult MD for other reasons

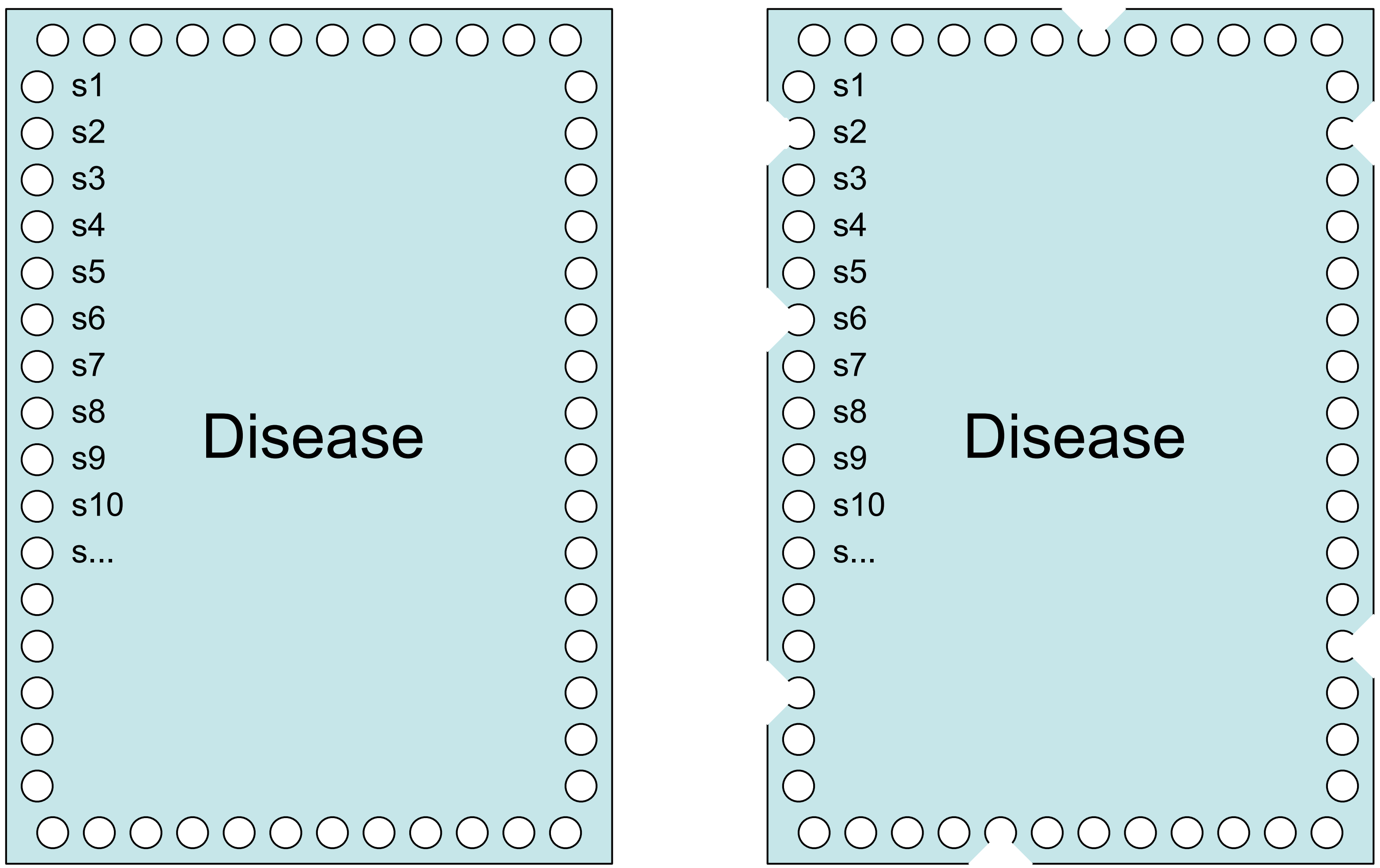
PLAN (also see back of protocol)

Dx of trichomonas? *Rx Flagyl*
Dx of monilia? *Rx Mycostatin*
Dx of non-specific vaginitis?
Stop Sulfa allergy? *Consult MD* *Rx Sultrin*

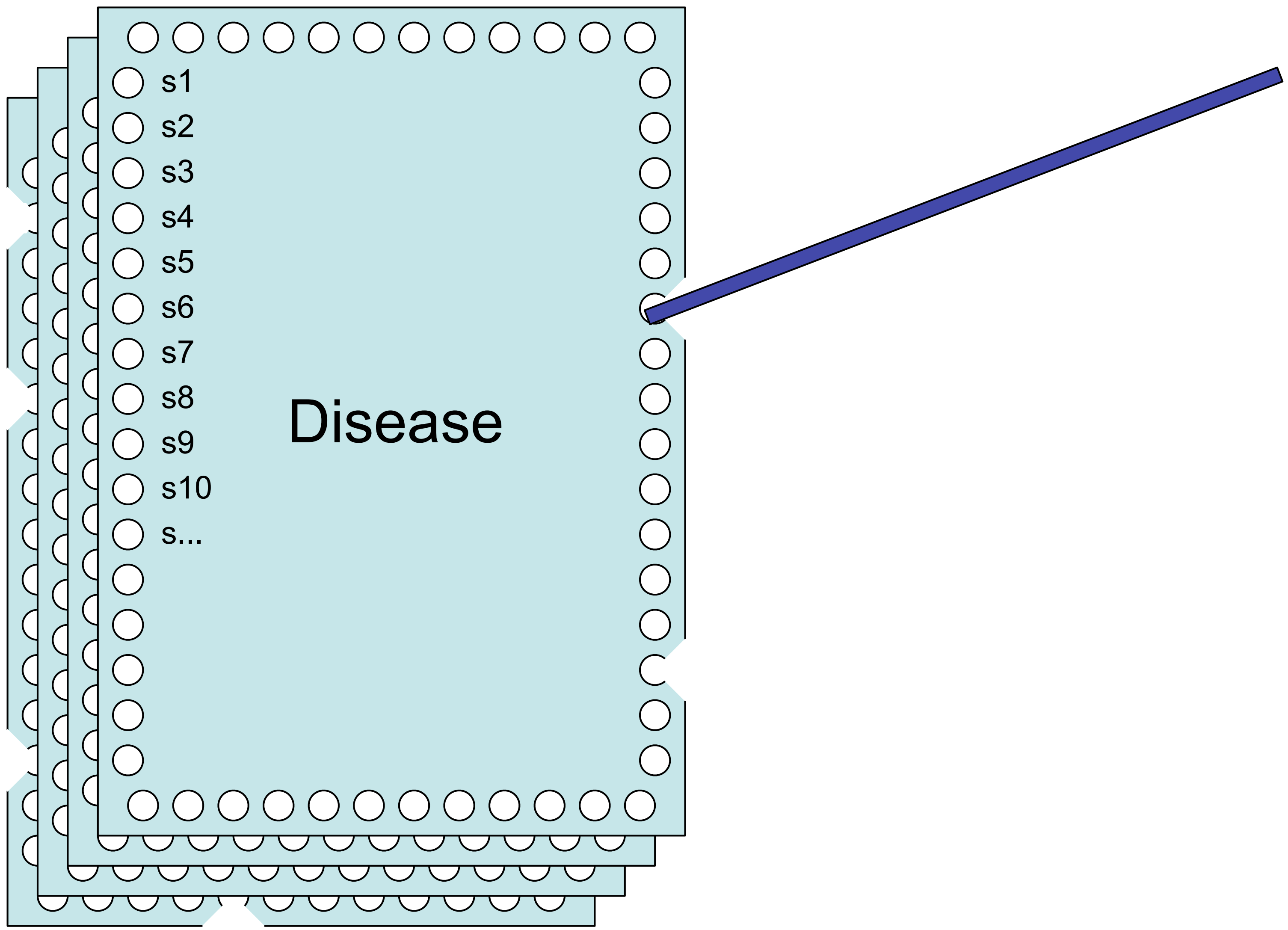
Stop Dx of UTI/urethritis
Dx of urethritis/vaginitis
Dysuria so bad pt can hardly urinate
Frequency interfering with work
or sleep? *Rx as below but tell pt to wait for culture result before beginning med*

Sulfa allergy? *Rx Sulfisoxazole*
Tetracycline allergy? *Rx Tetracycline*
Penicillin/Ampicillin allergy?
Consult MD *Rx Ampicillin*

Disease = {signs & symptoms}

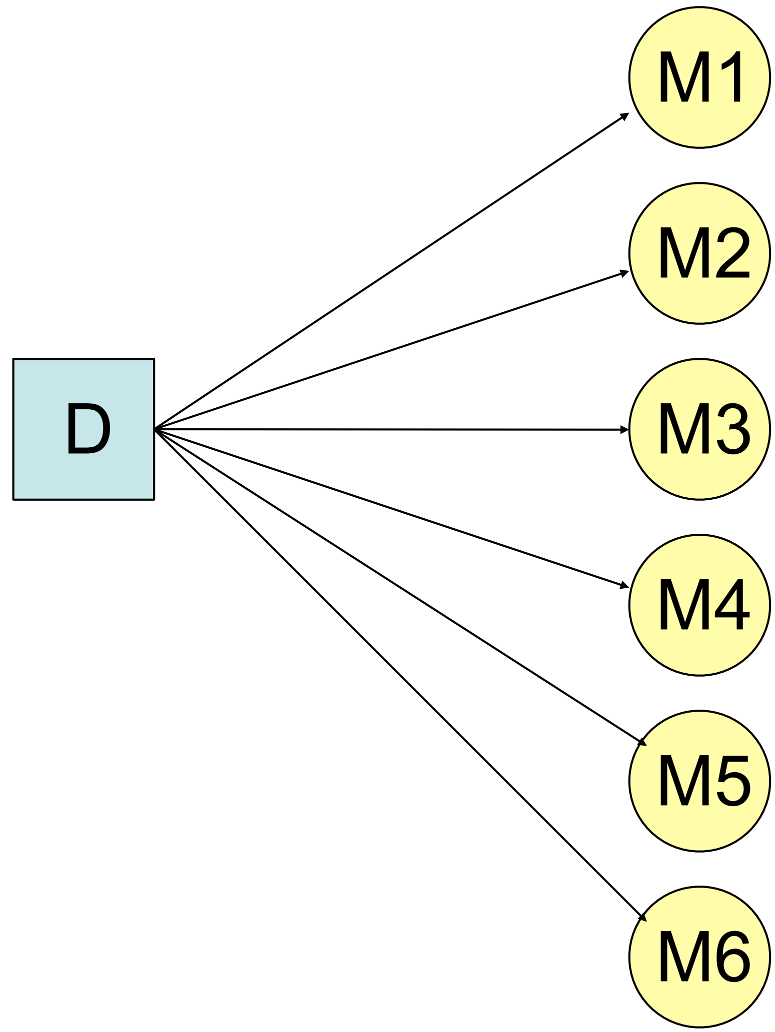


Diagnosis by Card Selection

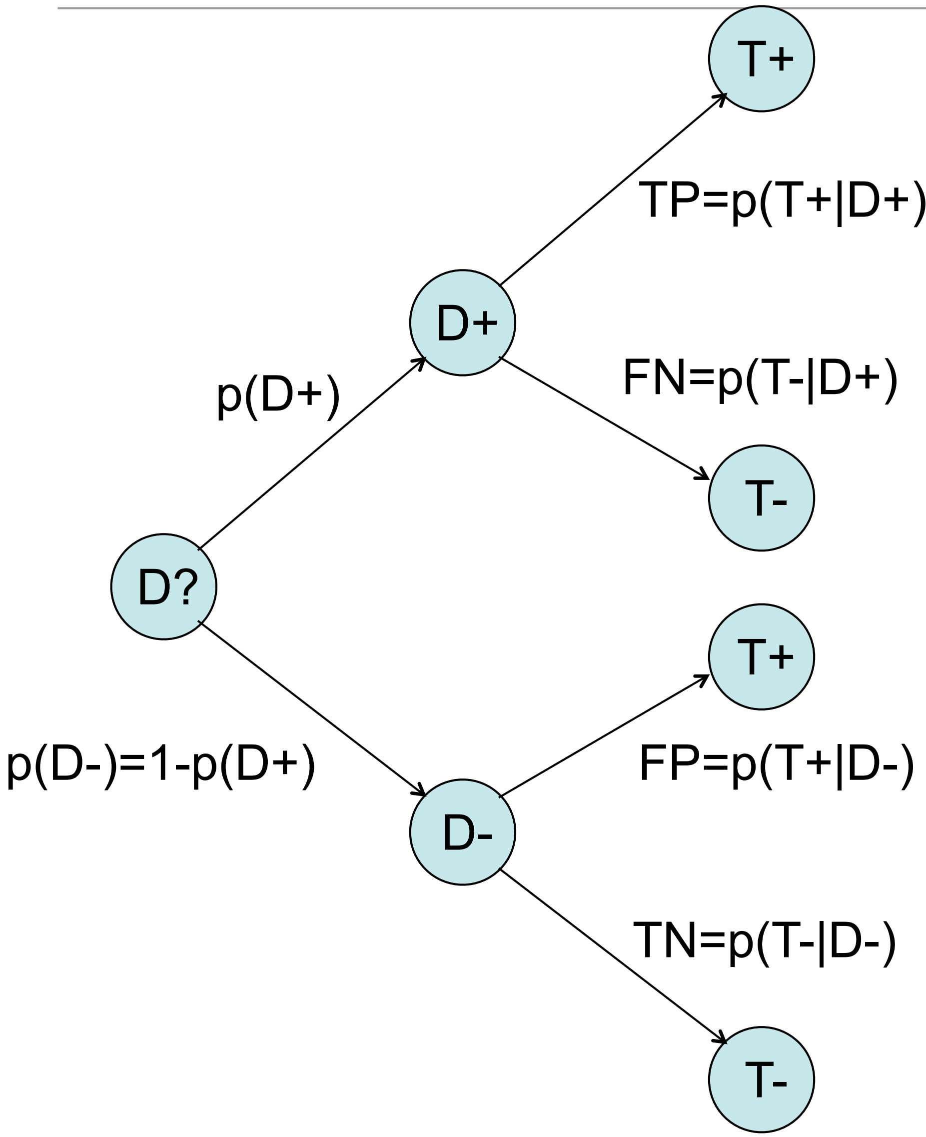


Naïve Bayes

- **Exhaustive** and **Mutually Exclusive** disease hypotheses (1 and only 1)
- **Conditionally independent** observables (manifestations)
- $P(D_i)$, $P(M_{ij}|D_i)$



How certain are we after a test?



Imagine P(D+) = .001 (it's a rare disease)
Accuracy of test = P(T+|D+) = P(T-|D-) = .95

Bayes' Rule:

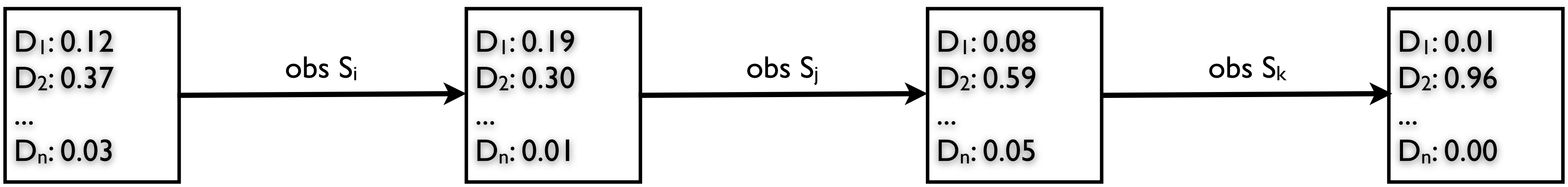
$$P_{i+1}(D_j) = \frac{P_i(D_j)P(S|D_j)}{\sum_{k=1}^{n} P_i(D_k)P(S|D_k)}$$

Diagnostic Reasoning with Naive Bayes

- Exploit assumption of conditional independence among symptoms

$P(S_1, S_2, \ldots, S_n | D_i) = P(S_1|D_i)P(S_2|D_i)P(S_n|D_i)$

- Sequence of observations of symptoms, S_i, each revise the distribution via Bayes' Rule



- After the j-th observation,

$$P^j(D_i|S_1, \ldots, S_j) = \frac{P^{j-1}(D_i)P(S_j|D_i)}{P^{j-1}(S_j)} = \frac{P^{j-1}(D_i)P(S_j|D_i)}{\sum_{i=0}^{n} P^{j-1}(D_i)P(S_j|D_i)}$$

Odds-Likelihood

- In gambling, “3-to-1” odds means 75% chance of success

$O = P/(1 - P) = P/\neg P$

- P = 0.5 means O=1
- Likelihood ratio
- Odds-likelihood form of Bayes rule

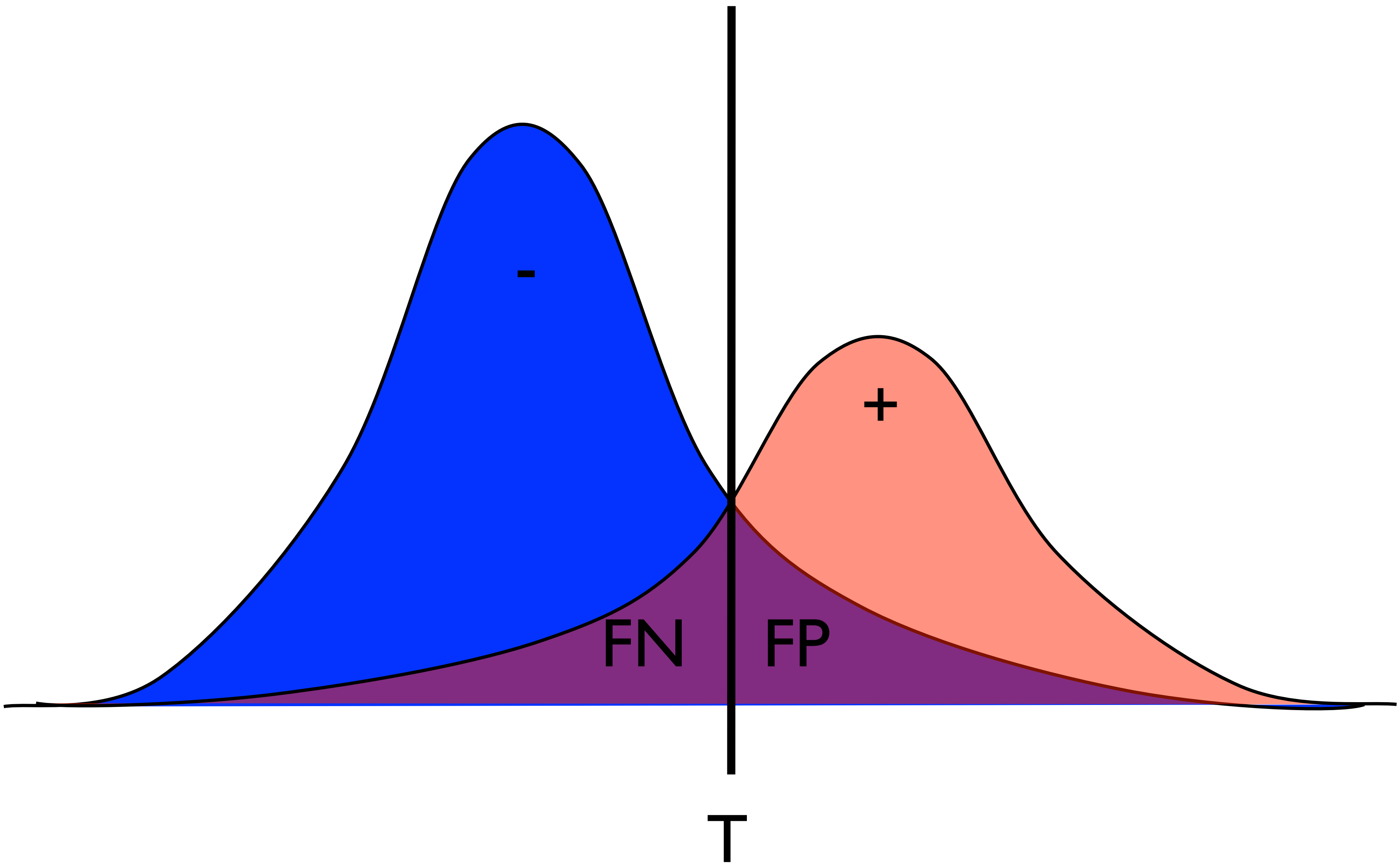
$L(S|D) = P(S|D)/P(S|\neg D)$

- Log transform

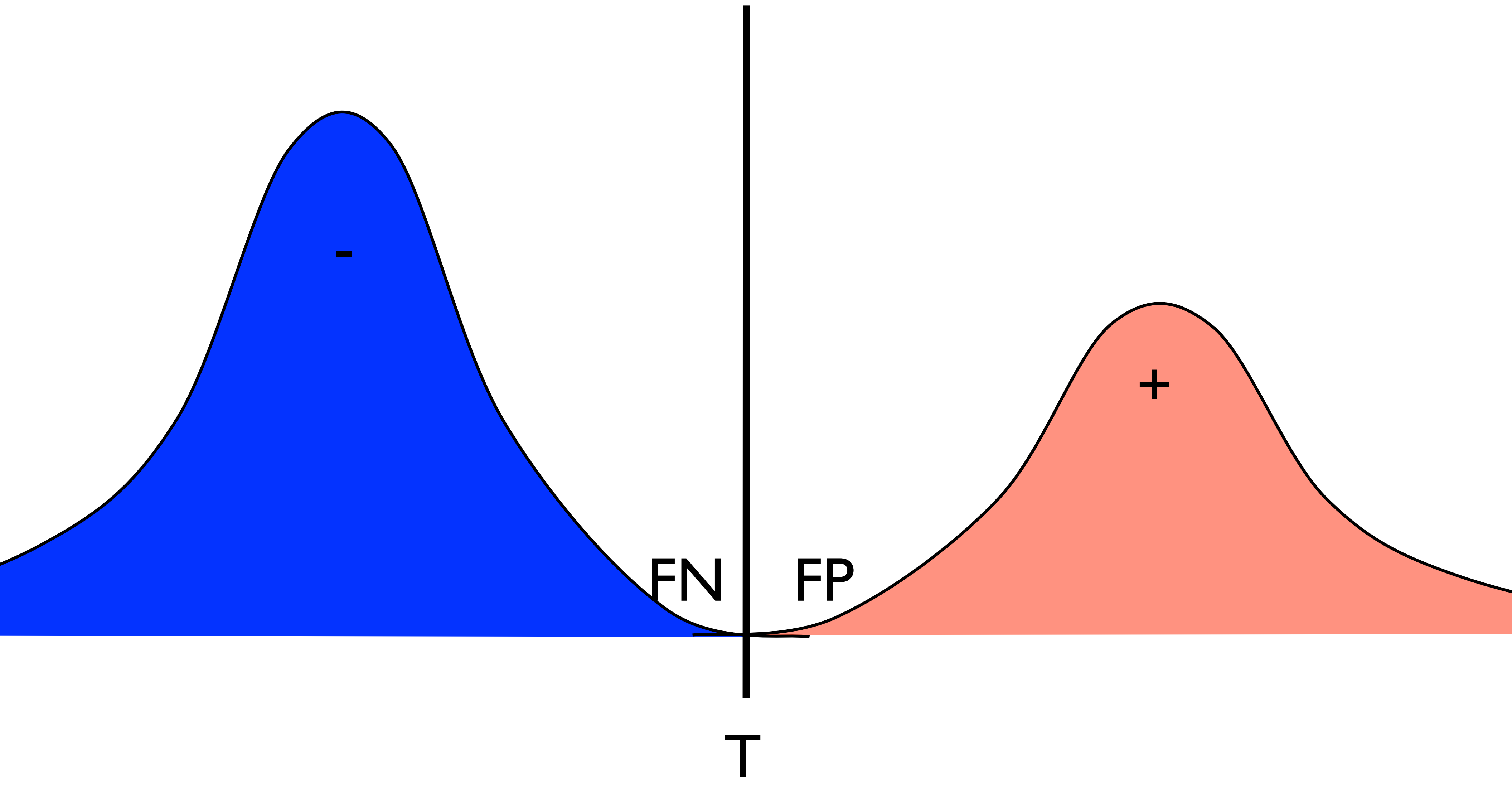
$O(D|S_1, \ldots, S_n) = O(D)L(S_1|D) \ldots L(S_n|D)$

$$\begin{aligned}\log O(D|S_1, \ldots, S_n) &= \log[O(D)L(S_1|D) \ldots L(S_n|D)] \\ &= \log[O(D)] + \log[O(S_1|D)] + \ldots + \log[O(S_n|D)] \\ &= W(D) + W(S_1|D) + \ldots + W(S_n|D)\end{aligned}$$

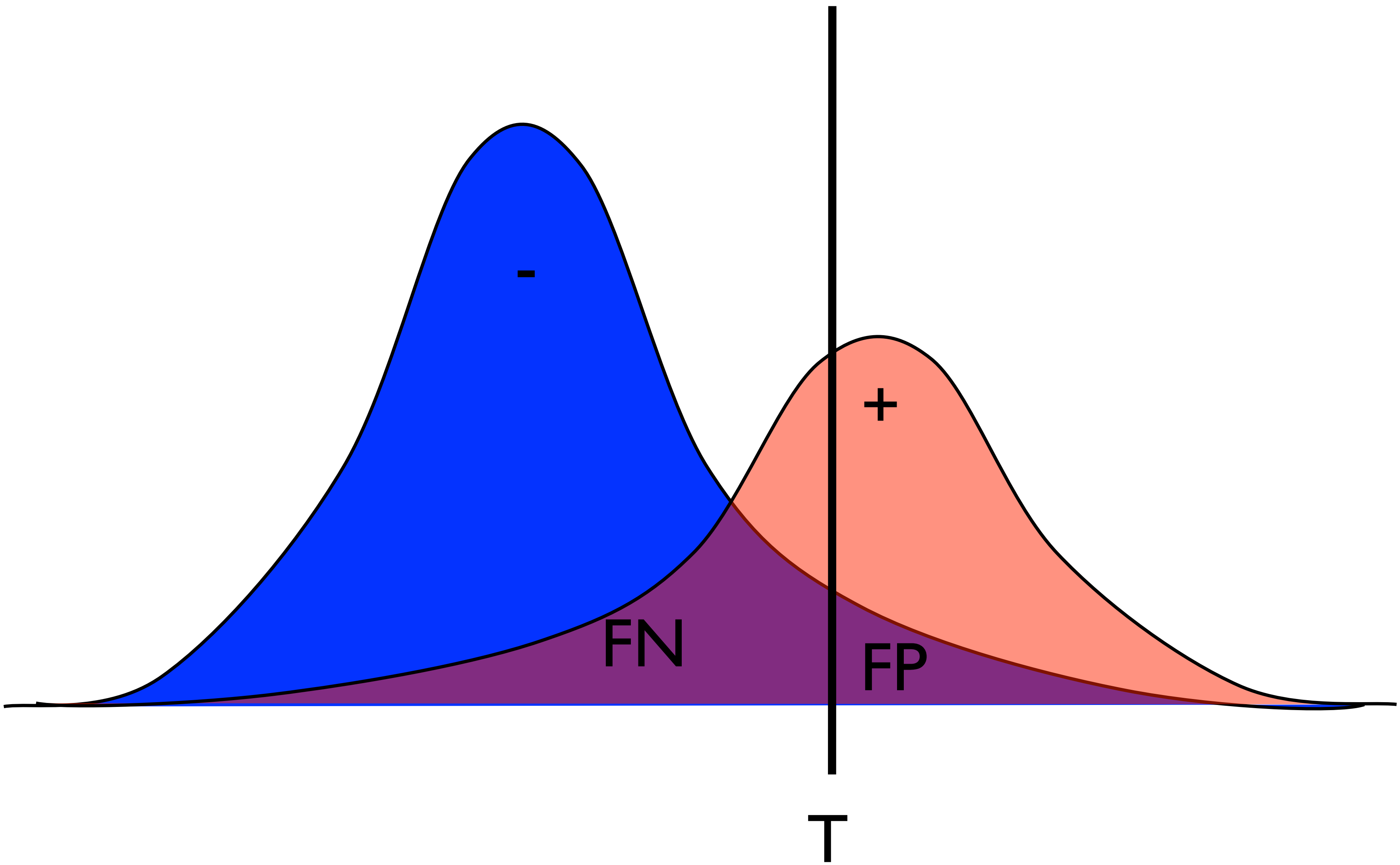
Test Thresholds



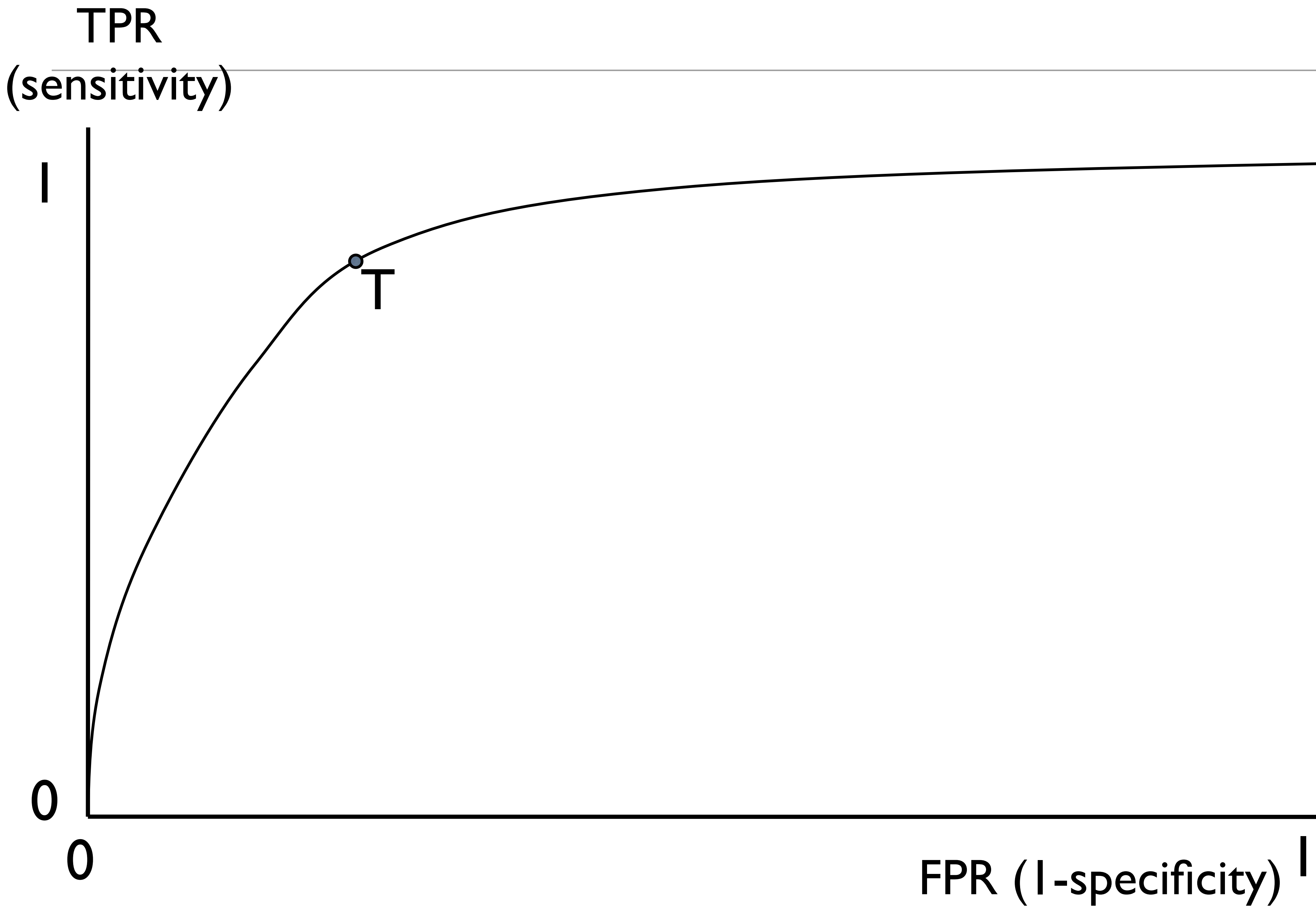
Wonderful Test



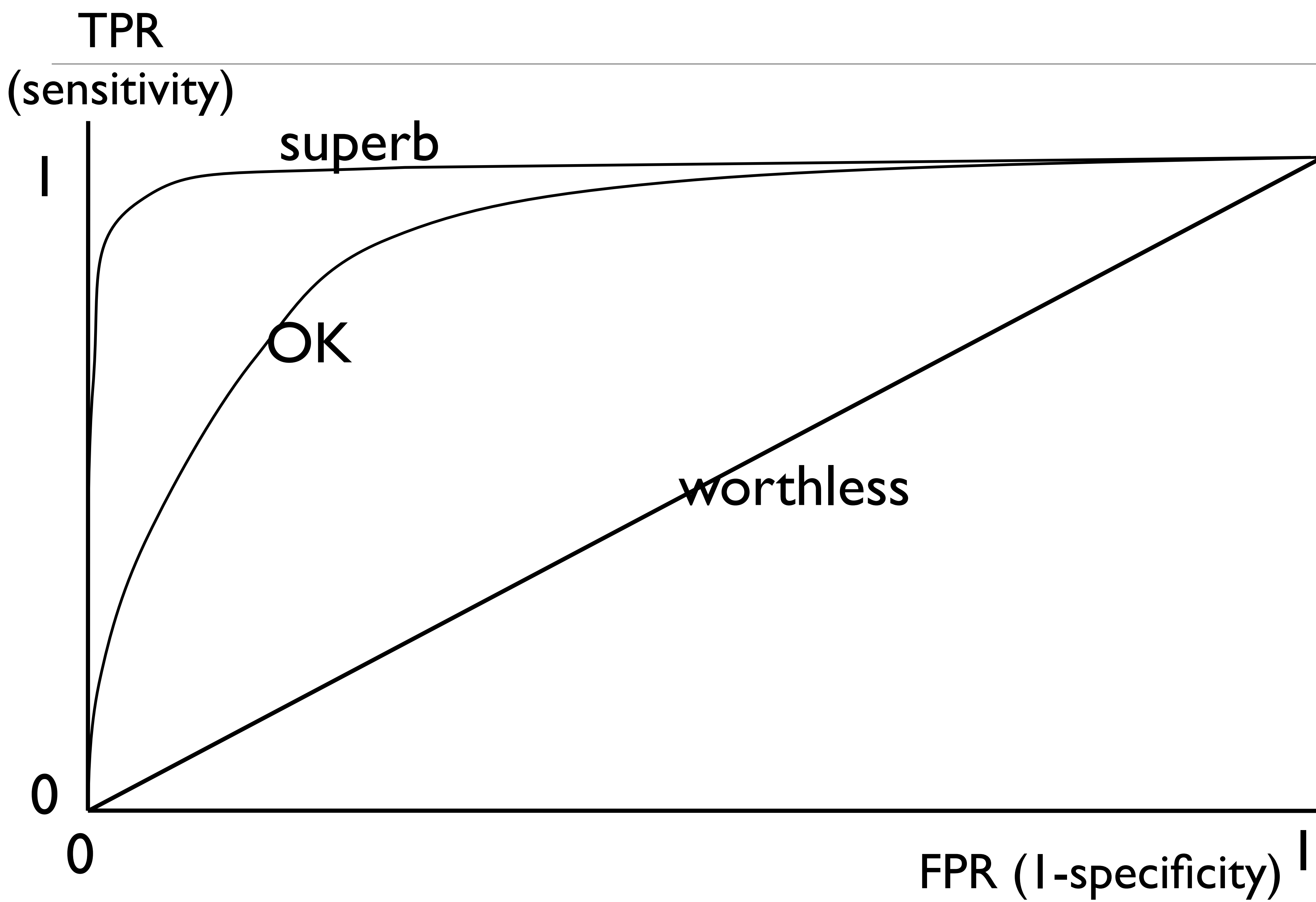
Test Thresholds Change Trade-off between Sensitivity and Specificity



Receiver Operator Characteristic (ROC) Curve



What makes a better test?



Rationality

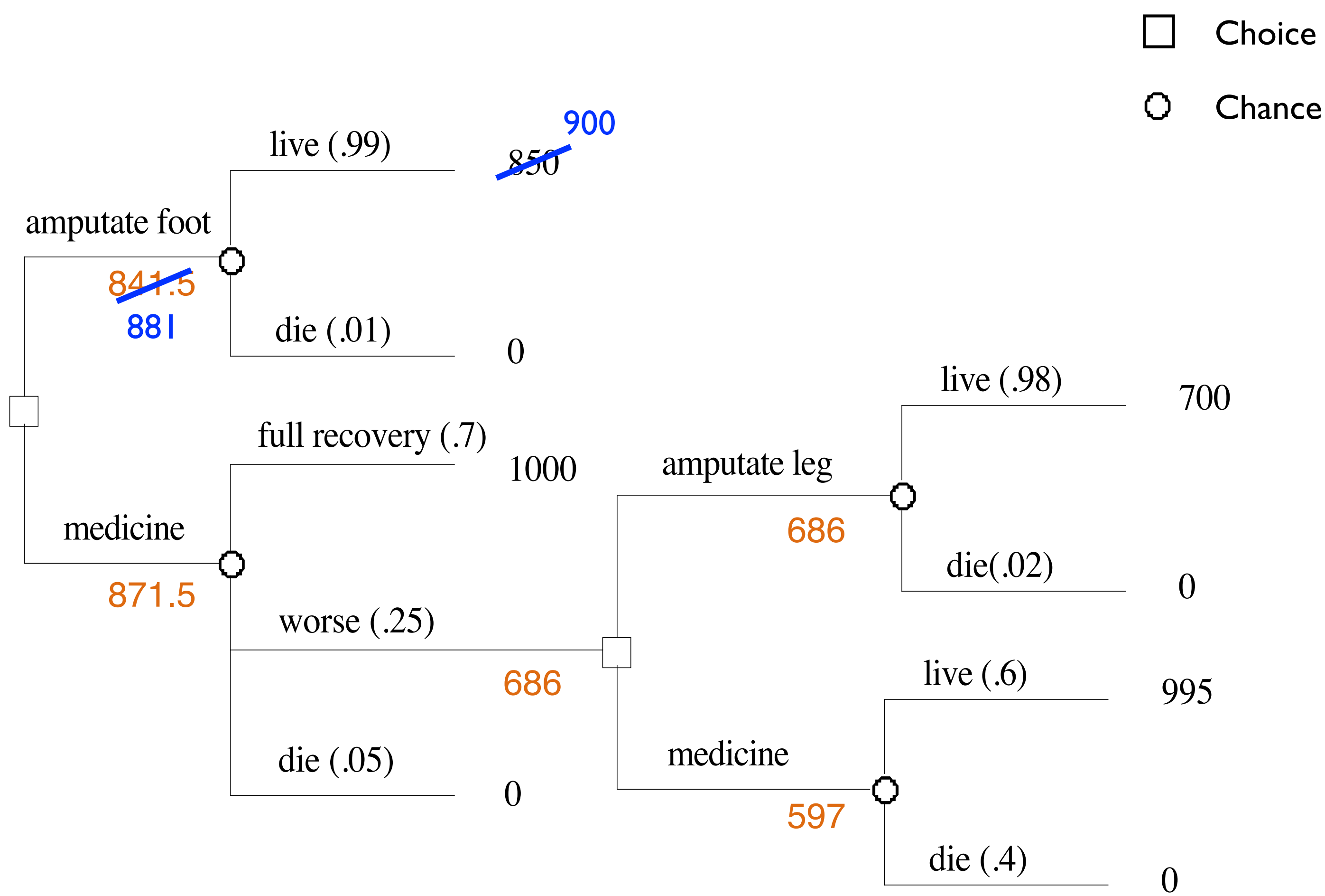
- Every action has a cost
- Principle of rationality
  - Act to maximize expected utility — *homo economicus*
  - Or minimize loss
- Utility measures the value (“goodness”) of an outcome, e.g.,
  - Life vs. death
  - Quality-adjusted life years (QALYs)

Case of a Man with Gangrene

- From Pauker's "Decision Analysis Service" at New England Medical Center Hospital, late 1970's.
- Man with gangrene of foot
- Choose to amputate foot or treat medically
- If medical treatment fails, patient may die or may have to amputate whole leg.
- What to do? How to reason about it?

Decision Tree for Gangrene Case

(Different sense of “Decision Tree” from ML/Classification!)



“Folding back” a Decision Tree

- The value of an outcome node is its utility
- The value of a chance node is the expected value of its alternative branches; i.e., their values weighted by their probabilities
- The value of a choice node is the maximum value of any of its branches

Where Do Utilities Come From?

- Standard gamble
  - Would you prefer (choose one of the following two):
    1. I chop off your foot
    2. We play a game in which a fair process produces a random number r between 0 and 1
    - If $r > 0.8$, I kill you; otherwise, you live on, healthy
  - If you're indifferent, that's the value of living without your foot!
  - I vary the 0.8 threshold until you are indifferent.
- Alas, difficult ascertainment problems!
  - Clearly depends on the individual
  - Not stable

Acute Renal Failure Program

- Differential Diagnosis of Acute Oliguric Renal Failure
  - “stop peeing”
- 14 potential causes, exhaustive and mutually exclusive
- 27 tests/questions/observations relevant to differential
  - “cheap”; therefore, ordering based on expected information gain
- 3 invasive tests (biopsy, retrograde pyelography, renal arteriography)
  - “expensive”; ordering based on (very naive) utility model
- 8 treatments (conservative, IV fluids, surgery for obstruction, steroids, antibiotics, surgery for clots, antihypertensive drugs, heparin)
  - expected outcomes are “better”, “unchanged”, “worse”

- Gorry, G. A., Kassirer, J. P., Essig, A., & Schwartz, W. B. (1973). Decision analysis as the basis for computer-aided management of acute renal failure. *The American Journal of Medicine*, 55(3), 473–484.

Question 5--What is the kidney size on plain film of the abdomen?

1. Small
2. Normal
3. Large
4. Very Large

Reply: 3

The current distribution is

Disease	Probability
OBSTR	0.80
FARF	0.12
PYE	0.04

Question 6—Was there a large fluid loss preceding the onset of oliguria?

Reply: No

The current distribution is

Disease	Probability
OBSTR	0.88
PYE	0.05
FARF	0.03

Question 7—What is the degree of Proteinuria?

1. 0
2. trace to 2+
3. 3+ to 4+

Reply: 1

The current distribution is

Disease	Probability
OBSTR	0.94
FARF	0.03
PYE	0.03

Question 8—Is there a history of prolonged hypotension preceding the onset of oliguria?

Reply: No

The current distribution is

Disease	Probability
OBSTR	0.96
PYE	0.03

Figure 1. *Typical interactive dialogue between the physician and the phase I computer program. The final diagnosis, which was arrived at after eight questions were asked, was urinary tract obstruction.*

Demo of Acute Renal Failure Program

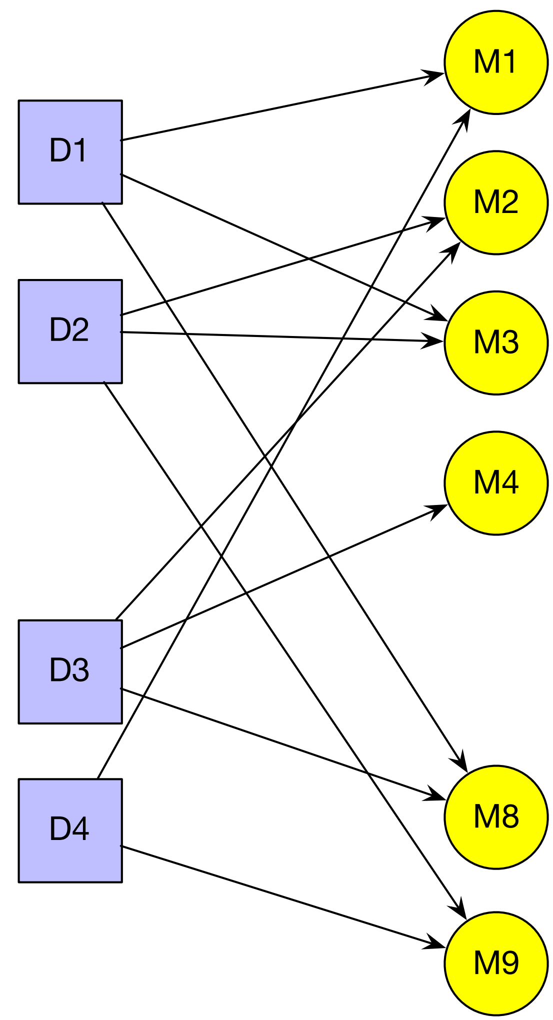
- Only the diagnostic portion
  - Original program also solved the decision analysis problem of what to do next
  - BADLY!
- 1990s GUI instead of 1970s terminal interface

“It thinks just the way I do!”

Bipartite Graph Model

- Multiple diseases
- Diseases are independent
- Manifestations depend only on which diseases are present
- Thus, they are conditionally independent

- This is a type of Bayes Network
- Computationally intractable
  - Complexity exponential in number of undirected cycles



Dialog/Internist/QMR ~1982

- ~500 diseases
  - (est. 70-75% of major diagnoses in internal medicine)
- ~3,500 manifestations
- (~15 man-years)

- By 1997, commercialized QMR had 766 Dx and 5498 Mx

Miller, R. A., Pople, H. E., & Myers, J. D. (1982). Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. The New England Journal of Medicine, 307(8), 468–476. http://doi.org/10.1056/NEJM198208193070803

Data in QMR

- For each Dx
  - List of associated Mx
    - with Evoking strength & Frequency
    - ~75 Mx per Dx
- For each Mx
  - Importance

Table 4. A Sample Manifestations List.*

DISPLAY WHICH MANIFESTATION LIST?
ALCOHOLIC HEPATITIS

AGE 16 TO 25 . . . 0 1
AGE 26 TO 55 . . . 0 3
AGE GTR THAN 55 . . . 0 2
ALCOHOL INGESTION RECENT HX . . . 2 4
ALCOHOLISM CHRONIC HX . . . 2 4
SEX FEMALE . . . 0 2
SEX MALE . . . 0 4
URINE DARK HX . . . 1 3
WEIGHT LOSS GTR THAN 10 PERCENT . . . 0 3
ABDOMEN PAIN ACUTE . . . 1 2
ABDOMEN PAIN COLICKY . . . 1 1
ABDOMEN PAIN EPIGASTRIUM . . . 1 2
ABDOMEN PAIN NON COLICKY . . . 1 2
ABDOMEN PAIN RIGHT UPPER QUADRANT . . . 1 3
ANOREXIA . . . 0 4
DIARRHEA ACUTE . . . 1 2
MYALGIA . . . 0 3
VOMITING RECENT . . . 0 4
ABDOMEN BRUIT CONTINUOUS RIGHT UPPER QUADRANT . . . 1 2
ABDOMEN BRUIT SYSTOLIC RIGHT UPPER QUADRANT . . . 1 2
ABDOMEN TENDERNESS RIGHT UPPER QUADRANT . . . 2 4
CONJUNCTIVA AND/OR MOUTH PALLOR . . . 1 2
FECES LIGHT COLORED . . . 1 2
FEVER . . . 0 4
HAND(S) DUPUYTRENS CONTRACTURE(S) . . . 1 2
JAUNDICE . . . 1 3
LEG(S) EDEMA BILATERAL SLIGHT OR MODERATE . . . 1 2
LIVER ENLARGED MASSIVE . . . 1 2
LIVER ENLARGED MODERATE . . . 1 3
LIVER ENLARGED SLIGHT . . . 1 2
PAROTID GLAND(S) ENLARGED . . . 1 2

Data in QMR

Frequency (Fr)	
1	Mx occurs rarely in Dx
2	Mx occurs in a substantial minority of cases of Dx
3	Mx occurs in roughly half of cases of Dx
4	Mx occurs in a substantial majority of cases of Dx
5	Mx occurs in essentially all cases of Dx

Evoking Strength (Ev)	
0	Nonspecific
1	Dx is a rare or unusual cause of Mx
2	Dx causes a substantial minority of instances of Mx
3	Dx is the most common but not overwhelming cause of Mx
4	Dx is the overwhelming cause of Mx
5	Mx is *pathognomonic* for Dx

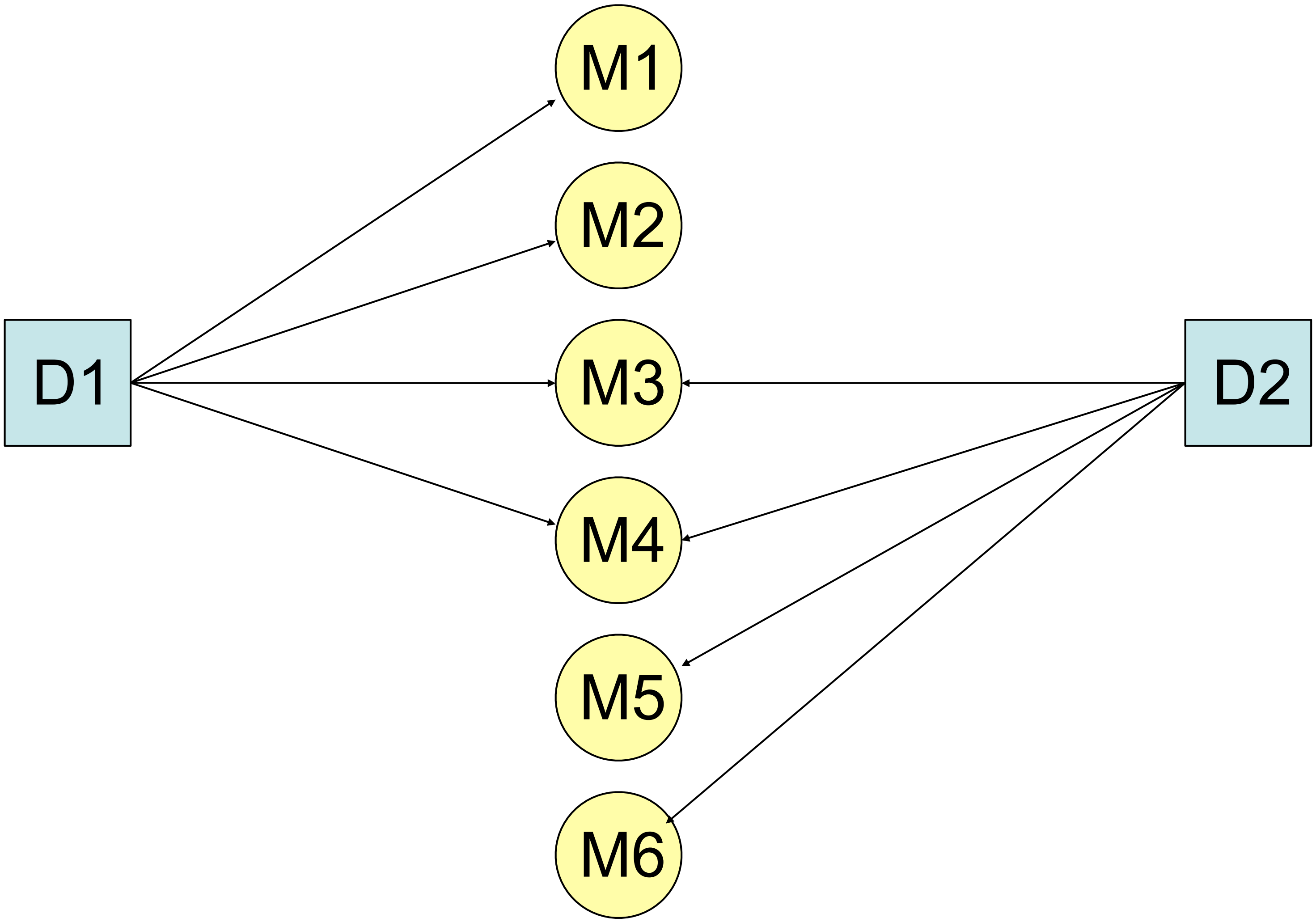
Importance (Im)	
1	Usually unimportant; occurs often in normal patients
2	May be important but can often be ignored
3	Medium importance, but unreliable indicator of disease
4	High importance, rarely disregarded
5	Absolutely must be explained by final diagnosis

Abductive Logic in QMR

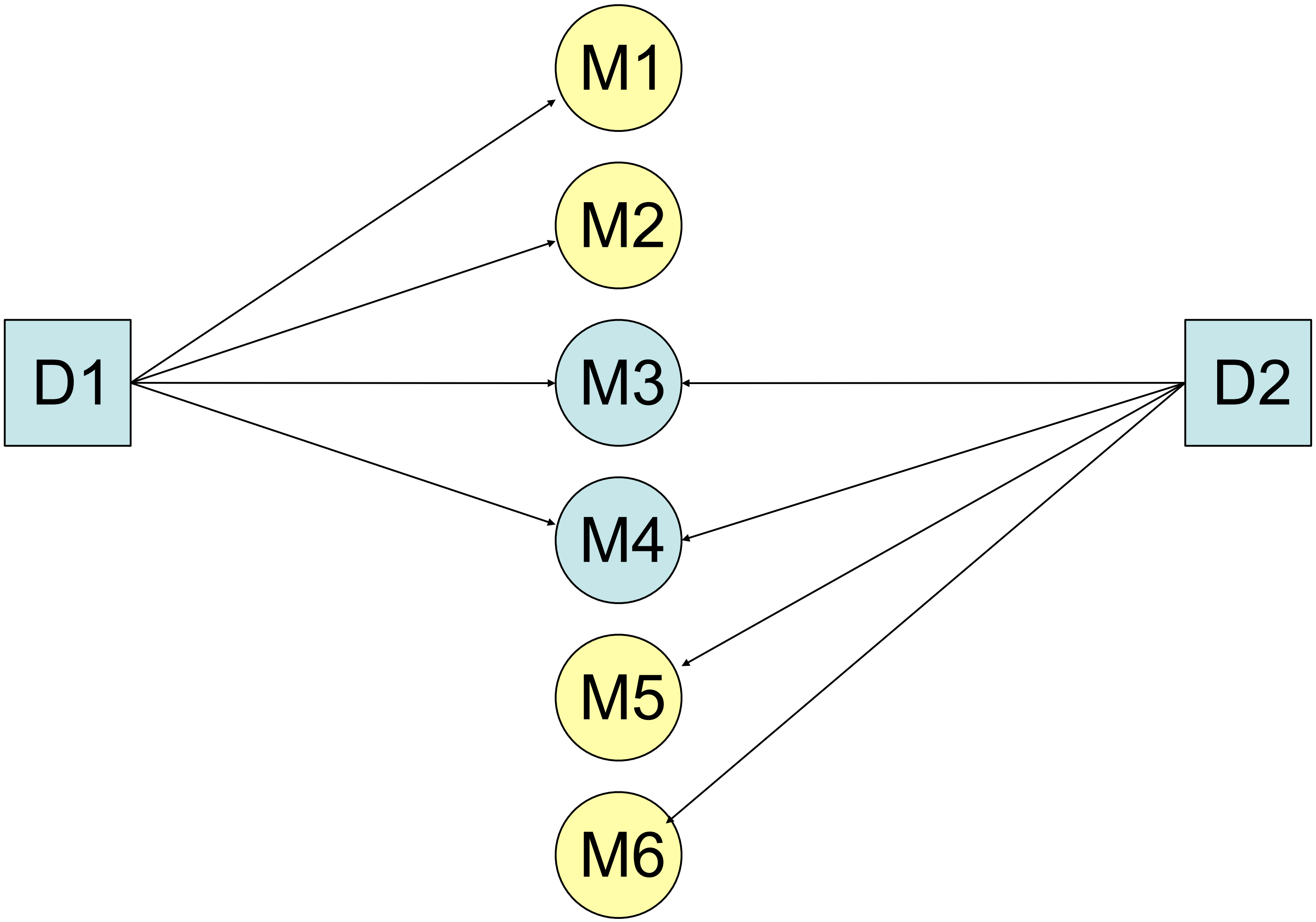
- List Mx of a case
  - Many demonstrated on NEJM Clinico-Pathological Conference cases
  - These are quite complex and challenging to doctors
- Evoke Dx's with high evoking strengths from Mx's
- Score Dx's
  - Positive:
    - Evoking strength of observed Manifestations
    - Scaled Frequency of causal links from confirmed Hypotheses
      - Scaling roughly exponential
  - Negative:
    - Frequency of predicted but absent Manifestations
    - Importance of unexplained Manifestations

- Form a differential around highest-scoring Dx

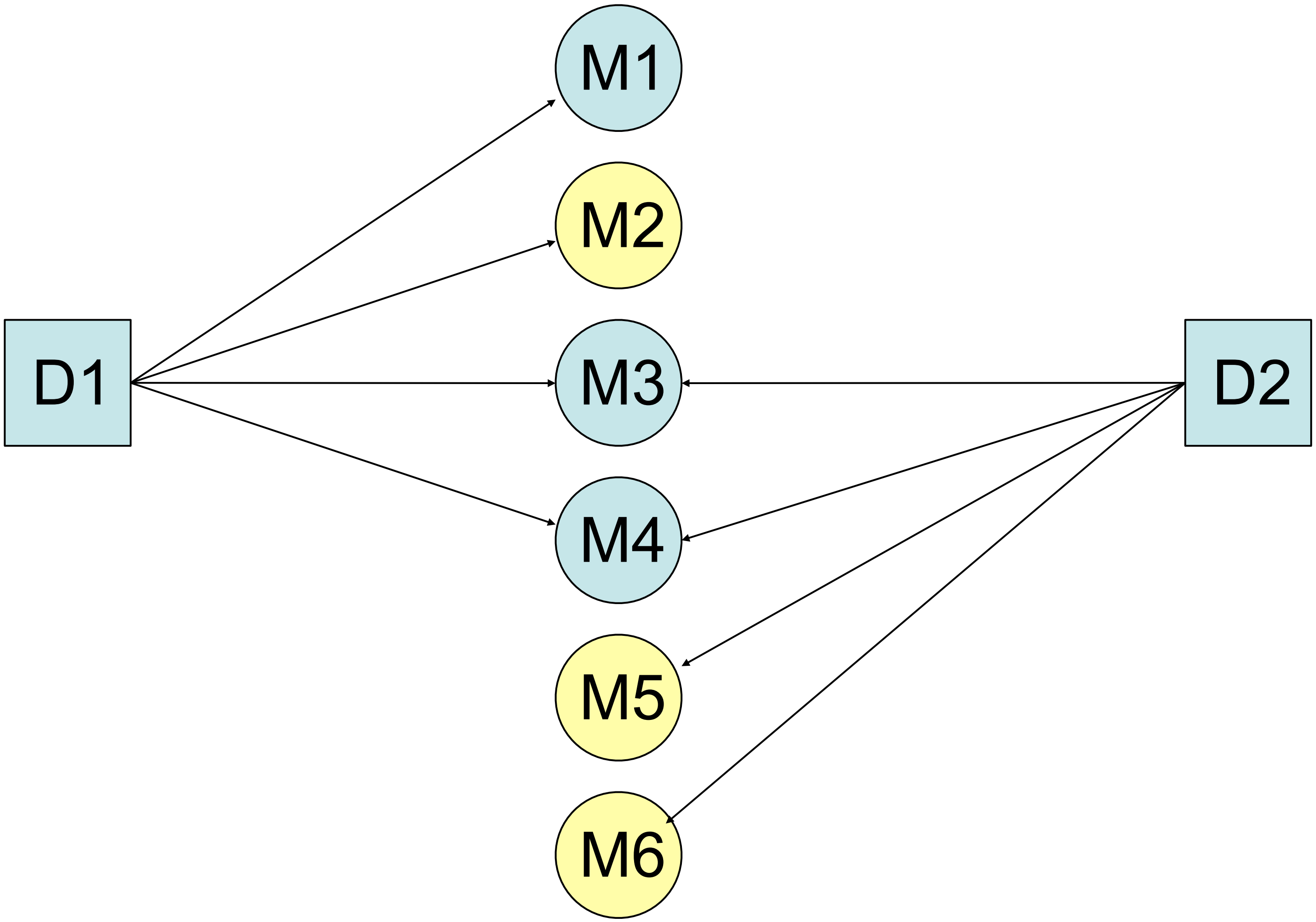
QMR Partitioning



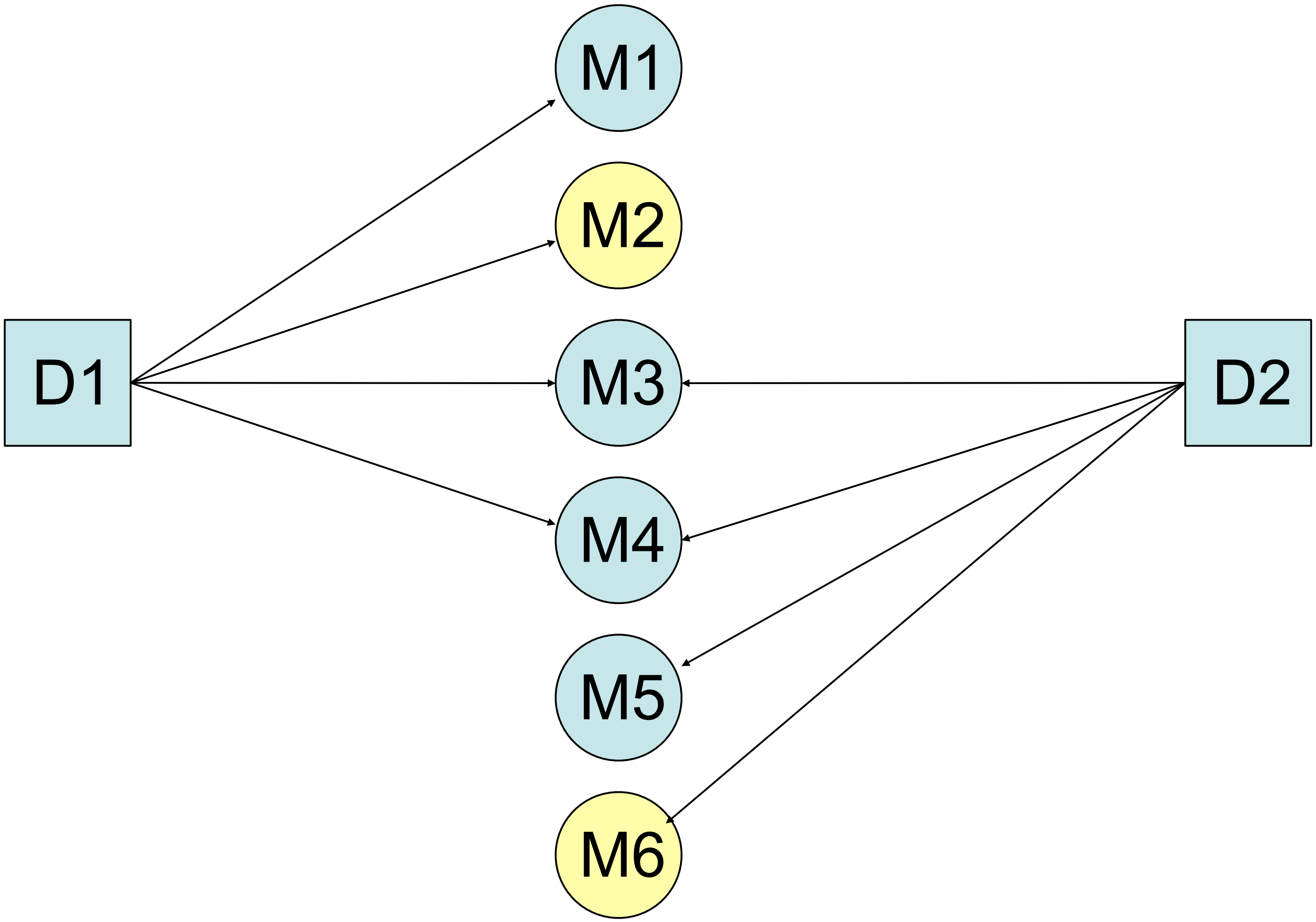
Competitors



Still Competitors



Probably Complementary



Multi-Hypothesis Diagnosis

- Set aside complementary hypotheses
  - … and manifestations predicted by them
- Solve diagnostic problem among competitors
  - differential determines questioning strategy: *pursue, rule-out, differentiate,* …
- Eliminate confirmed hypotheses and manifestations explained by them
- Repeat as long as there are coherent problems among the remaining data

Table 5. Summary of Results for Major Diagnoses in 19 Cases Used in the INTERNIST-I Evaluation.

CATEGORY	INTERNIST-I	CLINICIANS	DISCUSSANT
		no. of instances	
Total possible diagnoses	*43*	*43*	*43*
Definitive, correct	17	23	29
Tentative, correct	8	5	6
Failed to make correct diagnosis	18	15	8
Definitive, incorrect	5	8	11
Tentative, incorrect	6	5	2
Total no. of incorrect diagnoses	11	13	13
Total no. of errors in diagnosis	29	28	21

1990s Evaluation of Diagnostic Systems

- Evaluate: QMR, DXplain, Iliad, Meditel
- 105 cases (based on actual patients) created by 10 experts
- Results:
  - Coverage — fraction of real diagnoses included in program's KB
  - Correct — fraction of program's dx considered correct by experts
  - Rank — rank order of correct dx in program's list
  - Relevance — fraction of program's dx considered worthwhile by experts
  - Comprehensiveness — number of experts' dx included in program's top 20
  - Additional — "value added" dx by program

Berner, E. S., Webster, G. D., Shugerman, A. A., Jackson, J. R., Algina, J., Baker, A. L., et al. (1994). Performance of four computer-based diagnostic systems. The New England Journal of Medicine, 330(25), 1792–1796.

Table 1. Performance Scores of the Computer-Based Diagnostic Systems.

Variable and Sample Used*	Dxplain	Iliad	Meditel	QMR	Overall Analysis of Variance	P Value	Significant Pairwise Comparisons†
	mean (95 percent confidence interval)						
Diagnosis in Knowledge Base	0.91 (0.86–0.97)	0.76 (0.68–0.85)	0.85 (0.78–0.92)	0.73 (0.65–0.82)	$\chi^2 = 20.32$	<0.001	D vs. I, D vs. Q, M vs. Q
Correct Diagnosis							
105 cases	0.69 (0.60–0.78)	0.61 (0.52–0.70)	0.71 (0.62–0.79)	0.52 (0.43–0.62)	$\chi^2 = 11.58$	0.009	D vs. Q, M vs. Q
63 cases	0.79 (0.69–0.90)	0.76 (0.65–0.87)	0.89 (0.81–0.97)	0.71 (0.60–0.83)	$\chi^2 = 7.06$	0.070	—
Rank‡							
Diagnosis in program studied§	12.4 (9.5–15.3)	10.4 (8.0–12.8)	13.3 (10.5–16.1)	6.6 (3.0–10.3)	—	—	—
Diagnosis in all four programs¶	11.7 (8.3–15.1)	10.2 (7.5–12.9)	12.0 (8.8–15.3)	5.4 (3.7–7.1)	—	—	—
Relevance							
105 cases	0.24 (0.21–0.26)	0.19 (0.16–0.21)	0.22 (0.20–0.24)	0.37 (0.31–0.42)	F = 15.80	<0.001	Q vs. D, Q vs. M, Q vs. I, D vs. I, M vs. I
63 cases	0.26 (0.23–0.29)	0.21 (0.17–0.24)	0.23 (0.20–0.26)	0.46 (0.39–0.54)	F = 16.45	<0.001	Q vs. D, Q vs. M, Q vs. I, D vs. I
Comprehensiveness							
105 cases	0.38 (0.34–0.43)	0.25 (0.21–0.29)	0.38 (0.33–0.43)	0.28 (0.23–0.32)	F = 13.99	<0.001	D vs. I, D vs. Q, M vs. I, M vs. Q
63 cases	0.38 (0.33–0.44)	0.27 (0.22–0.32)	0.39 (0.32–0.46)	0.30 (0.25–0.35)	F = 5.05	0.004	D vs. I, D vs. Q, M vs. I, M vs. Q
Additional Diagnoses							
105 cases	2.3 (1.8–2.7)	2.0 (1.6–2.4)	2.1 (1.8–2.4)	1.8 (1.4–2.2)	F = 1.65	0.182	—
63 cases	2.6 (2.0–3.1)	2.2 (1.7–2.8)	2.2 (1.8–2.5)	2.0 (1.4–2.5)	F = 1.02	0.392	—

*The analyses of 105 cases were based on all cases included in the test, whereas the analyses of 63 cases were limited to the cases whose diagnoses were included in the knowledge base of all four programs.

†D denotes Dxplain, I Iliad, Q QMR, and M Meditel.

‡This variable could not be tested for significance because the sample varied in size according to the program used.

§This analysis included variable numbers of cases (72 for Dxplain, 64 for Iliad, 74 for Meditel, and 55 for QMR).

¶This analysis included variable numbers of cases (50 for Dxplain, 48 for Iliad, 56 for Meditel, and 45 for QMR).

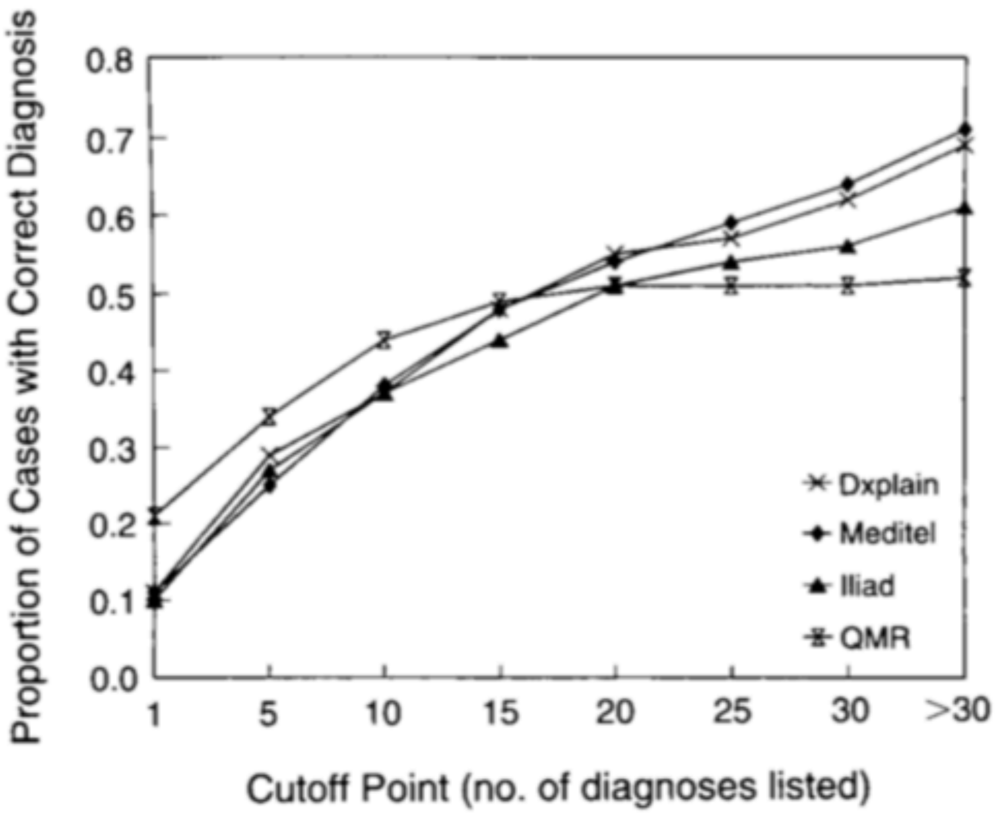
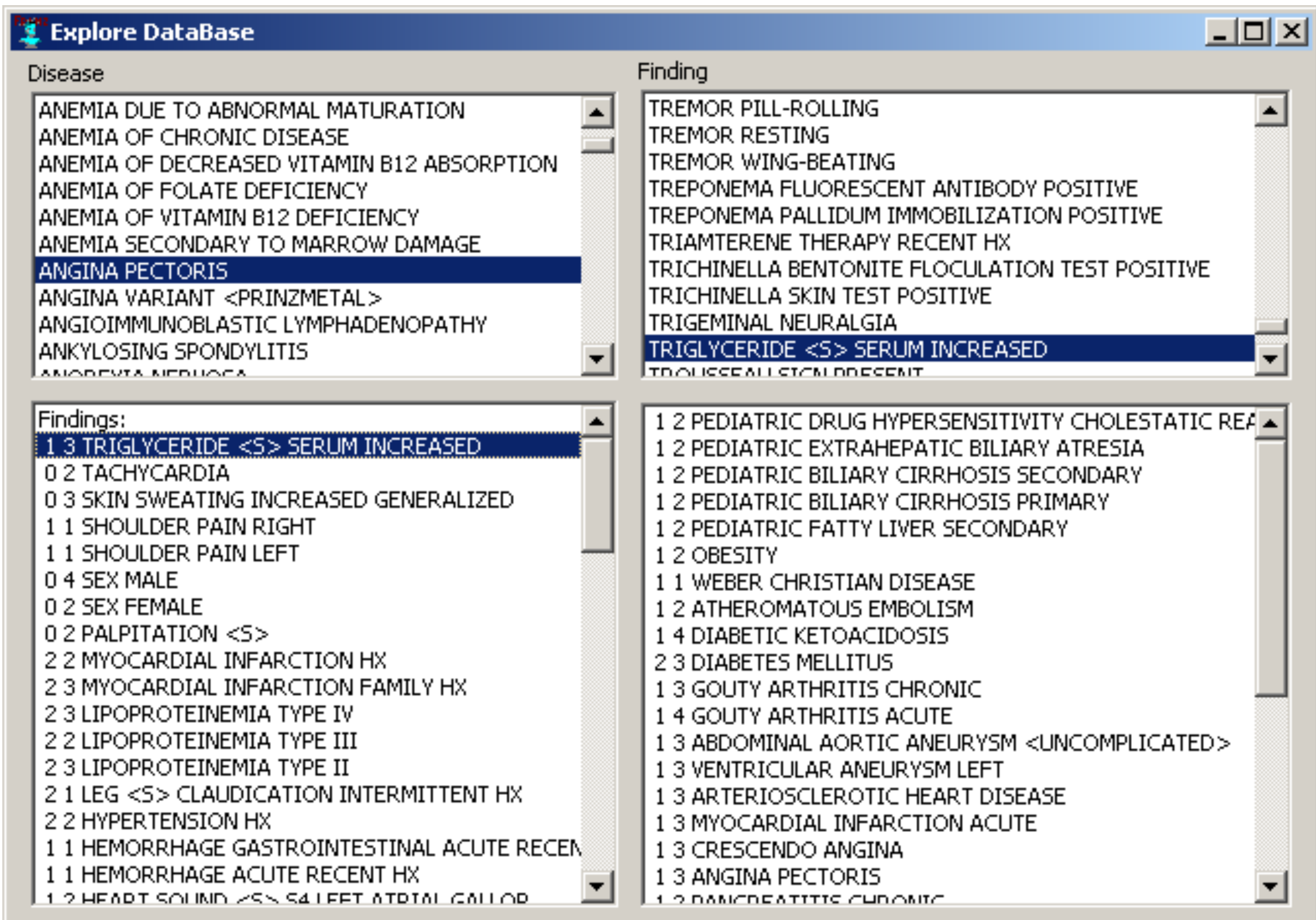


Figure 1. Proportion of Cases with a Correct Diagnosis in the Computer, According to the Cutoff Point Establishing the Numbers of Diagnoses Listed.

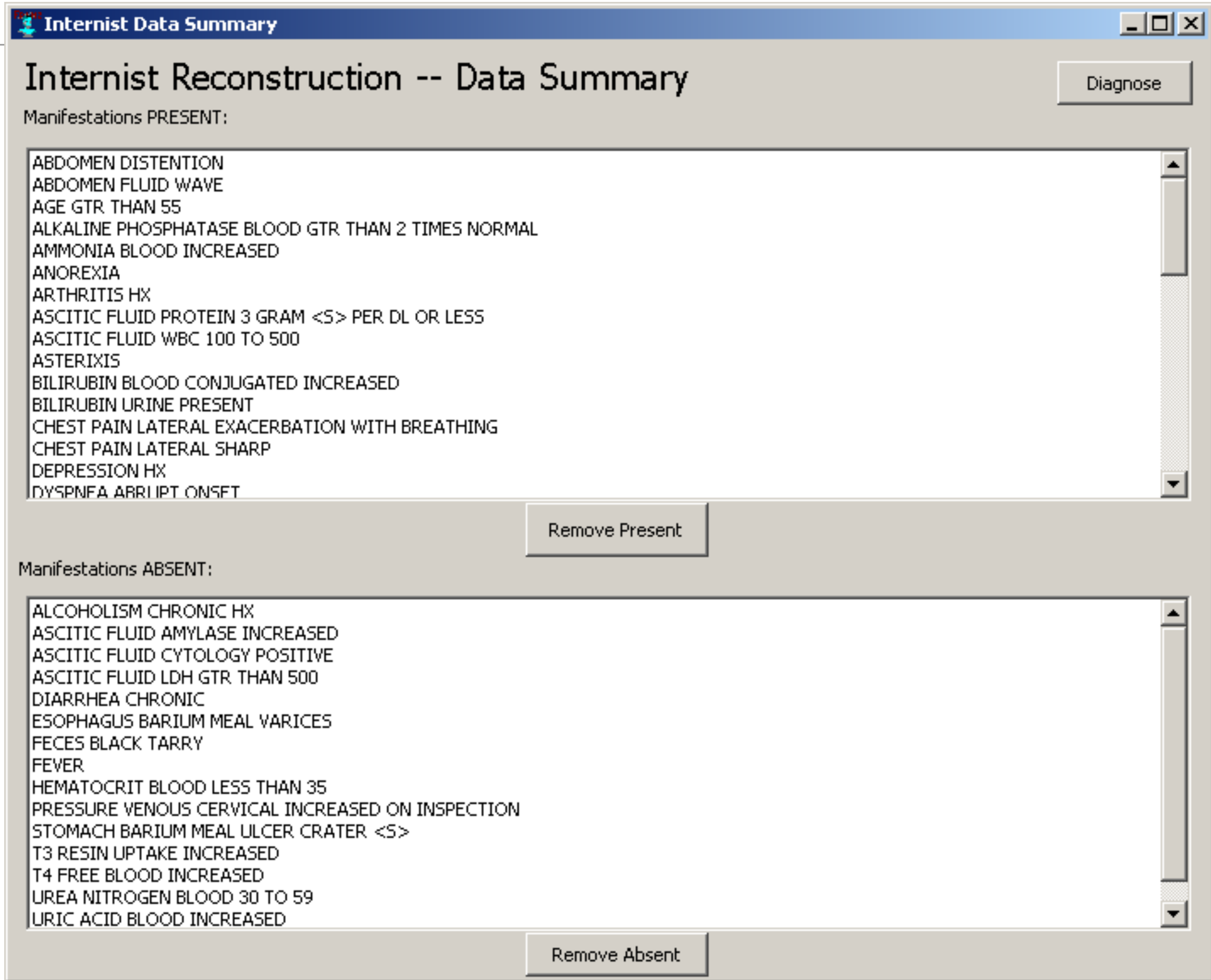
Evaluation Bottom Line

- … long lists of potential diagnoses. … many that a knowledgeable physician would regard as not being particularly helpful
- … each program suggested some diagnoses, though not highly likely ones, that the experts later agreed were worthy of inclusion in the differential diagnosis
- None performed consistently better or worse on all the measures
- Although the sensitivity and specificity … were not impressive, the programs have additional functions not evaluated
    - interactive display of signs and symptoms associated with diseases
    - relative likelihood of each dx (study only used ranking)
- Need to study effect of such programs on {physician, computer} team

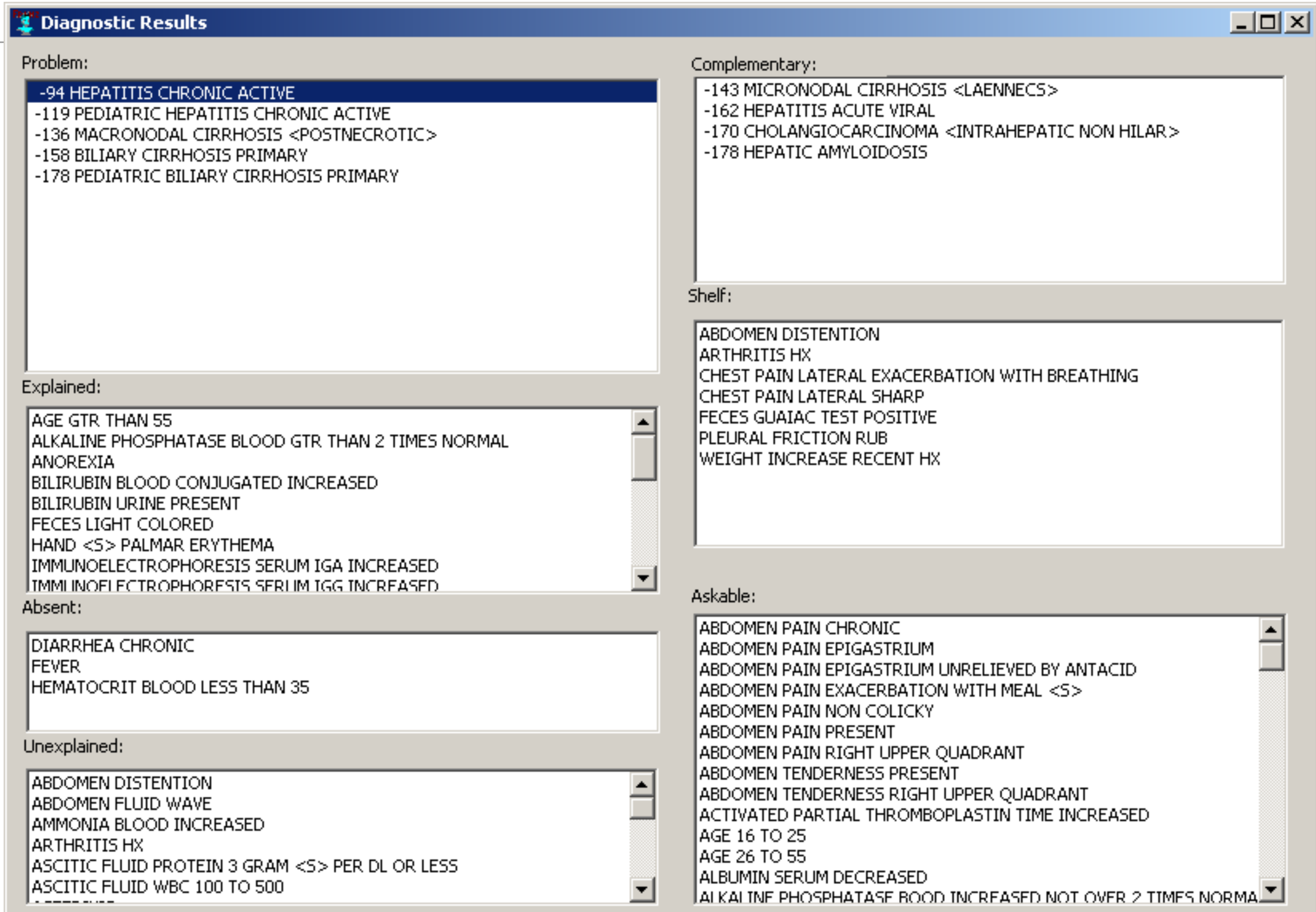
QMR Database



Example Case



Initial Solution



QMR-DT

- Interpret QMR data as a BN, with assumptions
  - Bipartite graph: marginal independence of Dx, conditional independence of Mx
  - Binary Dx and Mx
  - “Causal independence”—leaky *noisy-OR*
  - No distinction between Mx that predispose to a Dx and those that are a consequence of the Dx
  - Priors on Dx estimated from health statistics
    - problem of mapping QMR Dx names to ICD-9-CM
    - QMR treats age and gender as Mx, but QMR-DT conditions priors on them
    - No Evoking strengths are used
  - Estimate “leak” for each Mx from Importance values
- Use iterative diagnosis similar to QMR’s setting aside competitors, with Dx-Dx links altering priors on successive rounds
- Likelihood weighting to estimate posteriors

QMR-DT interpretation of Frequency and Importance

Table 1 A mapping between QMR frequencies and probabilities.

Frequency	$P(f^+ \mid \text{only } d_i^+)$
1	0.025
2	0.20
3	0.50
4	0.80
5	0.985

Table 2 A mapping between QMR imports and the probability that one or more significant diseases causes a finding f given that f is present.

Import	Fitted[a] $P(D_f \mid f)$	Std. Error $P(D_f \mid f)$
1	0.39	0.071
2	0.52	0.081
3	0.65	0.101
4	0.79	0.083
5	0.92	0.106

[a] The fitted $P(D_f \mid f)$ values were calculated by regressing the assessed values of $P(D_f \mid f)$ on the import values of the respective finding.

QMR-DT performance on Scientific American Medicine cases

Table 2 Ranks assigned to the reference diagnosis of the 23 SAM cases.

SAM case number	Algorithm QMR	TB	ITB	S	S/UD	S/UL
1	6	1	1	1	1	1
6	2	2	1	2	2	2
15	1	1	1	2	2	1
20	1	1	1	1	1	1
22	1	1	1	1	2	1
23†	–(1)	5(1)	20(1)	103(1)	4(1)	216(1)
25	3	1	2	1	2	6
27	1	1	3	1	1	1
28	1	2	1	1	1	1
29	3	4	11	9	6	106
30	5	2	3	7	17	36
31	12	9	11	24	166	255
33	2	2	17	2	1	1
34	1	6	12	4	4	445
35	1	1	3	1	2	2
37	2	17	2	2	7	8
40	1	1	1	1	1	352
42	4	1	3	2	2	1
46	1	1	1	1	1	1
47	1	1	1	1	1	1
50	1	1	2	1	1	1
51	2	2	5	57	22	30
53	3	1	1	1	1	1

Key:
– Reference diagnosis not ranked
† In case 23, we identified retrospectively an intermediate pathophysiologic state of malabsorption. The rank of malabsorption appears in parentheses for each algorithm.

Symptom Checkers

- Demo *K Health*
- BMJ article, 2015
  - 23 symptom checkers
  - 45 standardized patient vignettes
  - 3 levels of urgency:
    - emergent care needed: e.g., pulmonary embolism
    - non-emergent care reasonable: e.g., otitis media (ear ache)
    - self-care reasonable: e.g., viral infection
  - Goals
    - if diagnosis given, is right answer within top 20 (n=770)
    - if triage given, is it the right level of urgency (n=532)
  - Correct dx first in 34% of cases, within top 20 in 58%
  - Correct triage in 57% (80% in emergent, 55% non-emergent, 33% self-care)
    - different systems ranged from 33% to 78% average accuracy

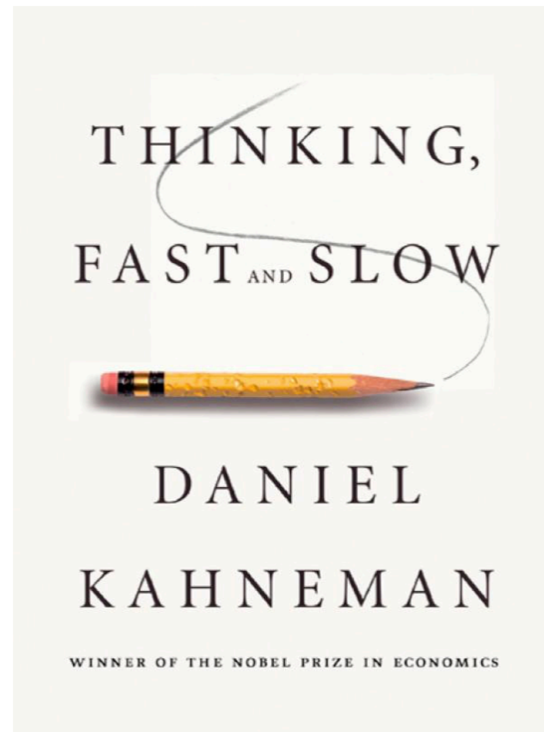
Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ (Clinical Research Ed), h3480–9. http://doi.org/10.1136/bmj.h3480

Symptom Checkers: BMJ conclusions

- The public is increasingly using the internet for self diagnosis and triage advice, and there has been a proliferation of computerized algorithms called symptom checkers that attempt to streamline this process
- Despite the growth in use of these tools, their clinical performance has not been thoroughly assessed
- Our study suggests that symptom checkers have deficits in both diagnosis and triage, and their triage advice is generally risk averse

Rationality under Resource Constraints

- Utility comes not only from the ultimate “patient” but from reasoning about the computational process
- McGyver’s utilities drop suddenly under deadline constraints
- Partial computation
  - Any-time algorithms
  - Simplify model
  - Approximate
- Kahneman
  - Fast: reflex, rules
  - Slow: deliberative



Horvitz, E. J. (1990). Rational metareasoning and compilation for optimizing decisions under bounded resources. Presented at Computational Intelligence ’89, Milan, Italy.

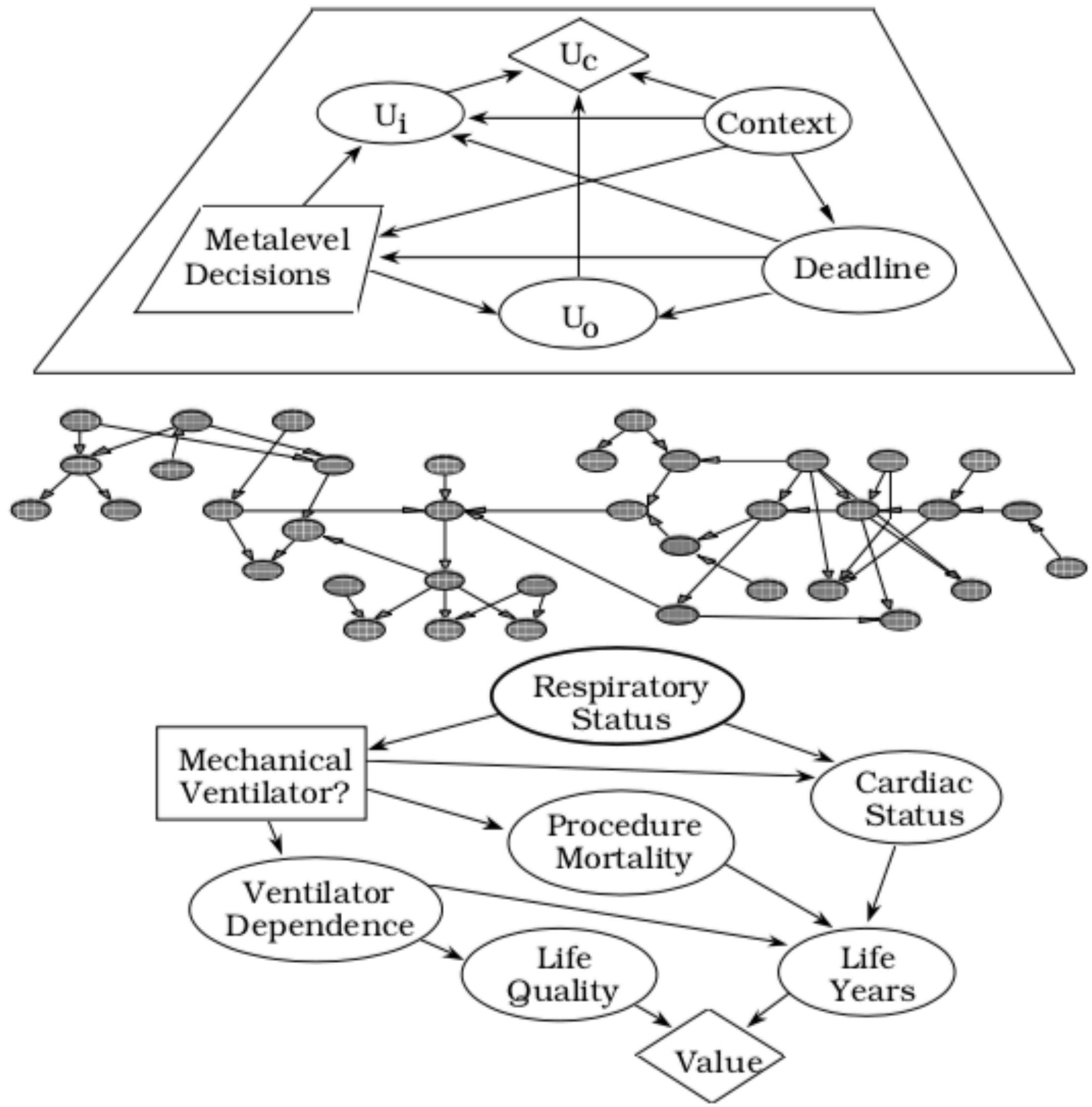
Meta-level Reasoning about How to Reason

- “the expected value of computation as a fundamental component of reflection about alternative inference strategies”
  - alternative methods (e.g., QMR’s question-asking strategies)
  - degree of refinement (e.g., incremental algorithms can stop early)
- Value of information, value of computation, value of experimentation

Horvitz, E., Cooper, G. F., & Heckerman, D. (1989). Reflection and Action Under Scarce Resources - Theoretical Principles and Empirical Study. Presented at the IJCAI.

A Time-Pressured Decision Problem

- decision-theoretic metareasoning
- belief network representing propositions and dependencies in intensive care physiology
- close-up on “Respiratory Status” node and its relationship to current decision problem
    - “A 75yo woman in ICU has sudden breathing difficulties”
    - Should we start mechanical ventilation?



Horvitz, E., Cooper, G. F., & Heckerman, D. (1989). Reflection and Action Under Scarce Resources - Theoretical Principles and Empirical Study. Presented at the IJCAI.

Reinforcement Learning for Speeding up Diagnosis

- Rather than heuristics, use MDP formulation and RL
- State space: set of positive and negative findings
- Action space: ask about a finding, or conclude a diagnosis
- Reward: correct or incorrect (**single**) diagnosis
- Finite horizon imposed by limit on number of questions
- Discount factor encourages short question sequences
- Standard q-learning framework, using double-deep NN strategy
- Magic sauce:
    - Encourage asking questions likely to have positive answers because of sparsity, by *reward shaping*: add extra reward; policy still optimal
    - Identify reduced finding space by *feature rebuilding.*

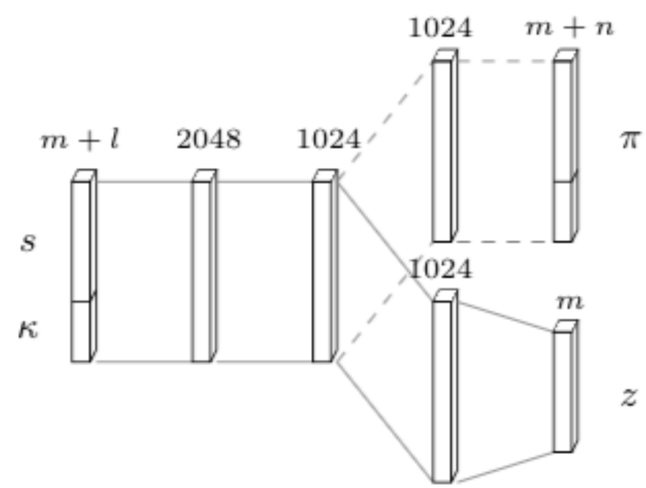
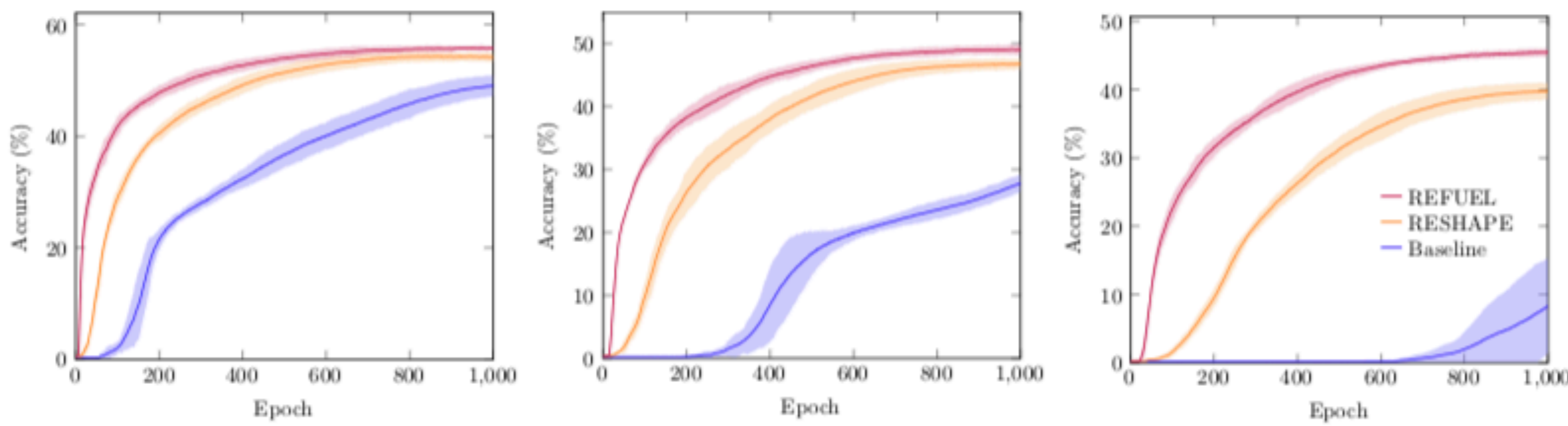


Figure 1: Dual neural network architecture. The upper branch is the policy π of an agent. The lower branch is the feature rebuilding part of sparse features.

Peng, Y.-S., Tang, K.-F., Lin, H.-T., & Chang, E. Y. (2018). REFUEL - Exploring Sparse Features in Deep Reinforcement Learning for Fast Disease Diagnosis. Presented at NeurIPS.

REFUEL Performance

- Simulated data: 650 diseases and 376 symptoms
-



(a) 200 diseases (b) 300 diseases (c) 400 diseases

Figure 2: Experiments on 3 datasets of different disease numbers. The curves show the training accuracy of three methods. REFUEL (red line) uses reward shaping and feature rebuilding; RESHAPE (yellow line) only uses reward shaping; Baseline (blue line) adopts none of them. The solid line is the averaged result of 5 different random seeds. The shaded area represents two standard deviations.