Differential Diagnosis

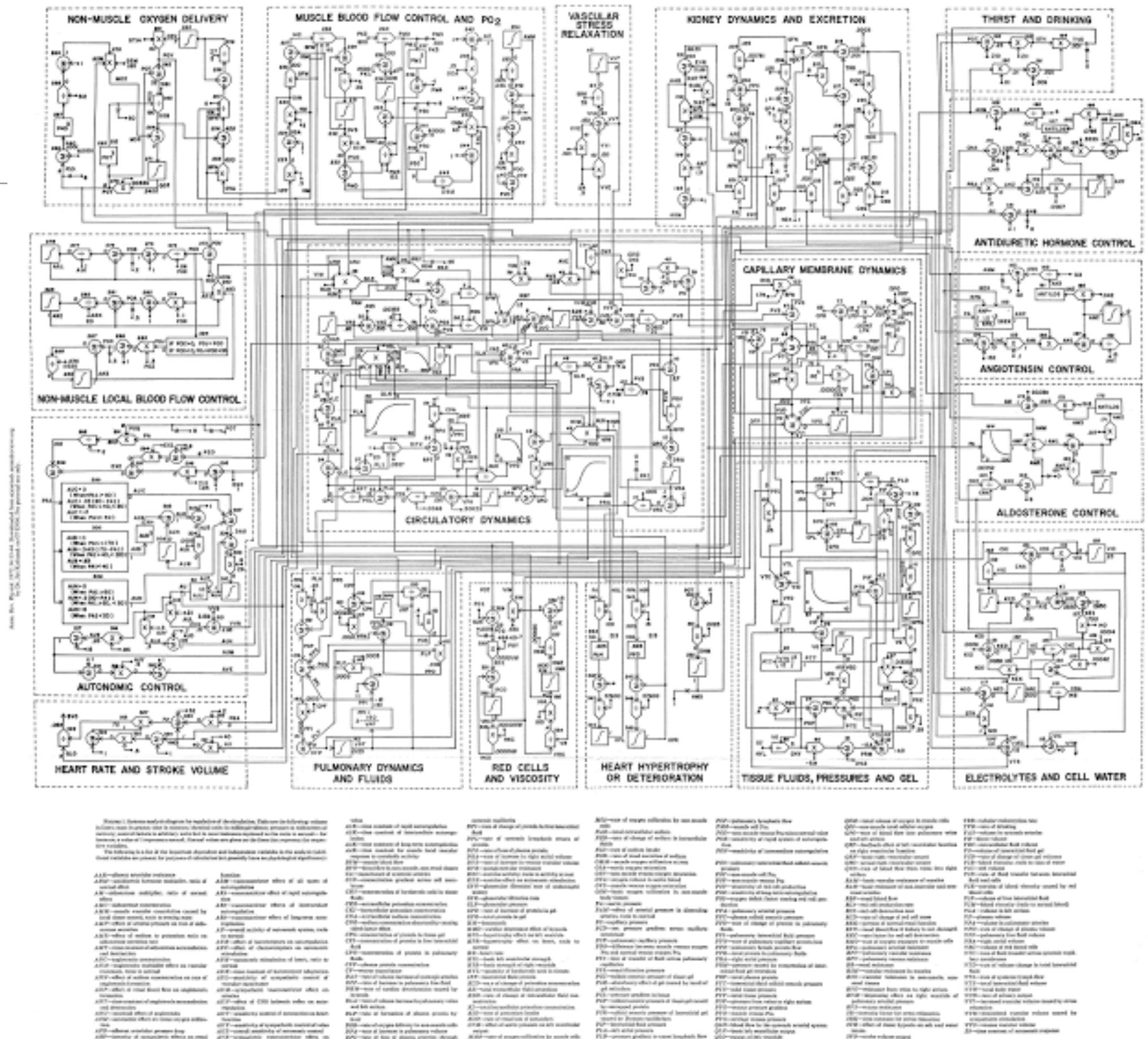"**Diagnosis** is the identification of the nature and cause of a certain phenomenon"
"**differential diagnosis** is the distinguishing of a particular disease or condition from others that present similar clinical features"
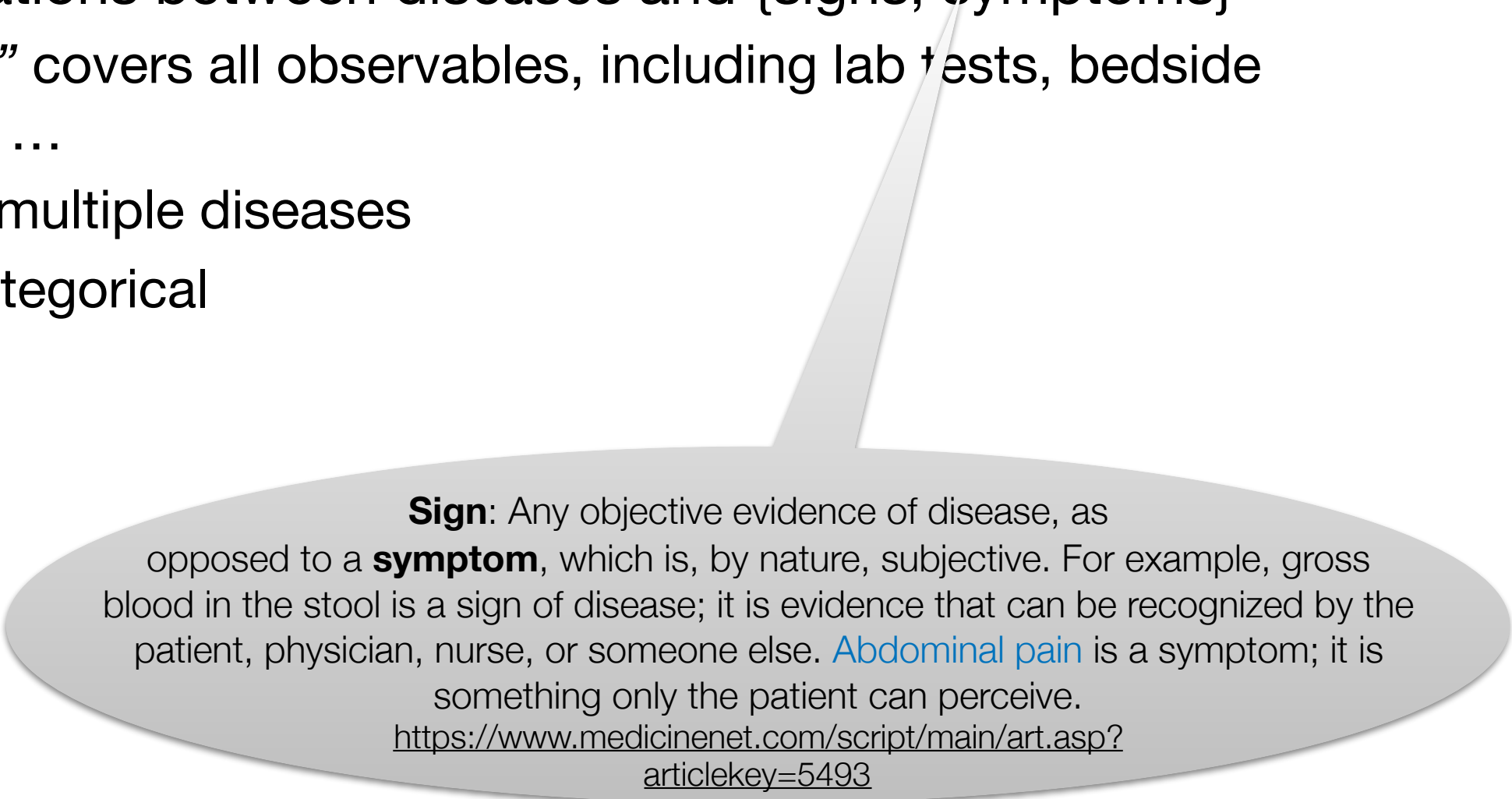
—Wikipedia

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

A Stop Any blue boxes checked
        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}
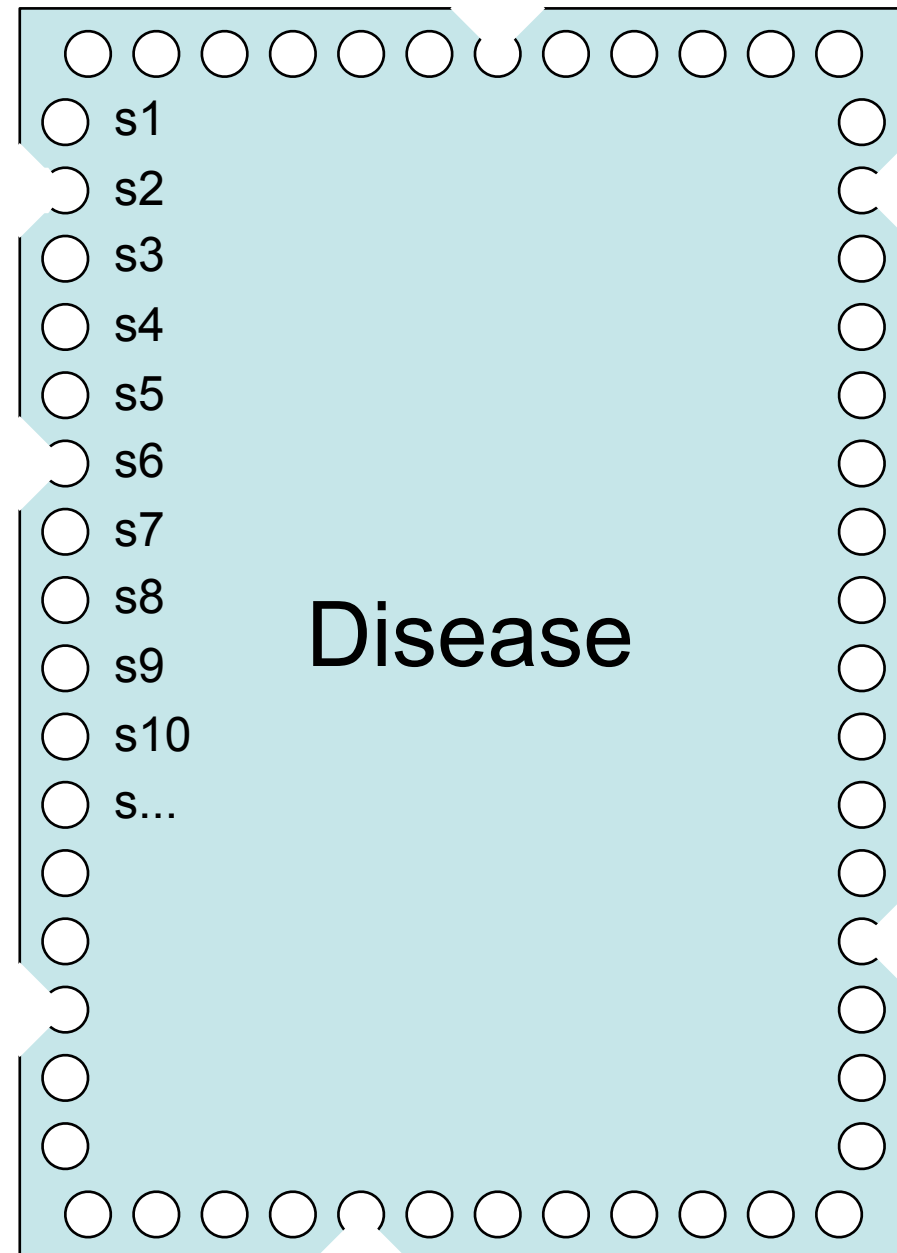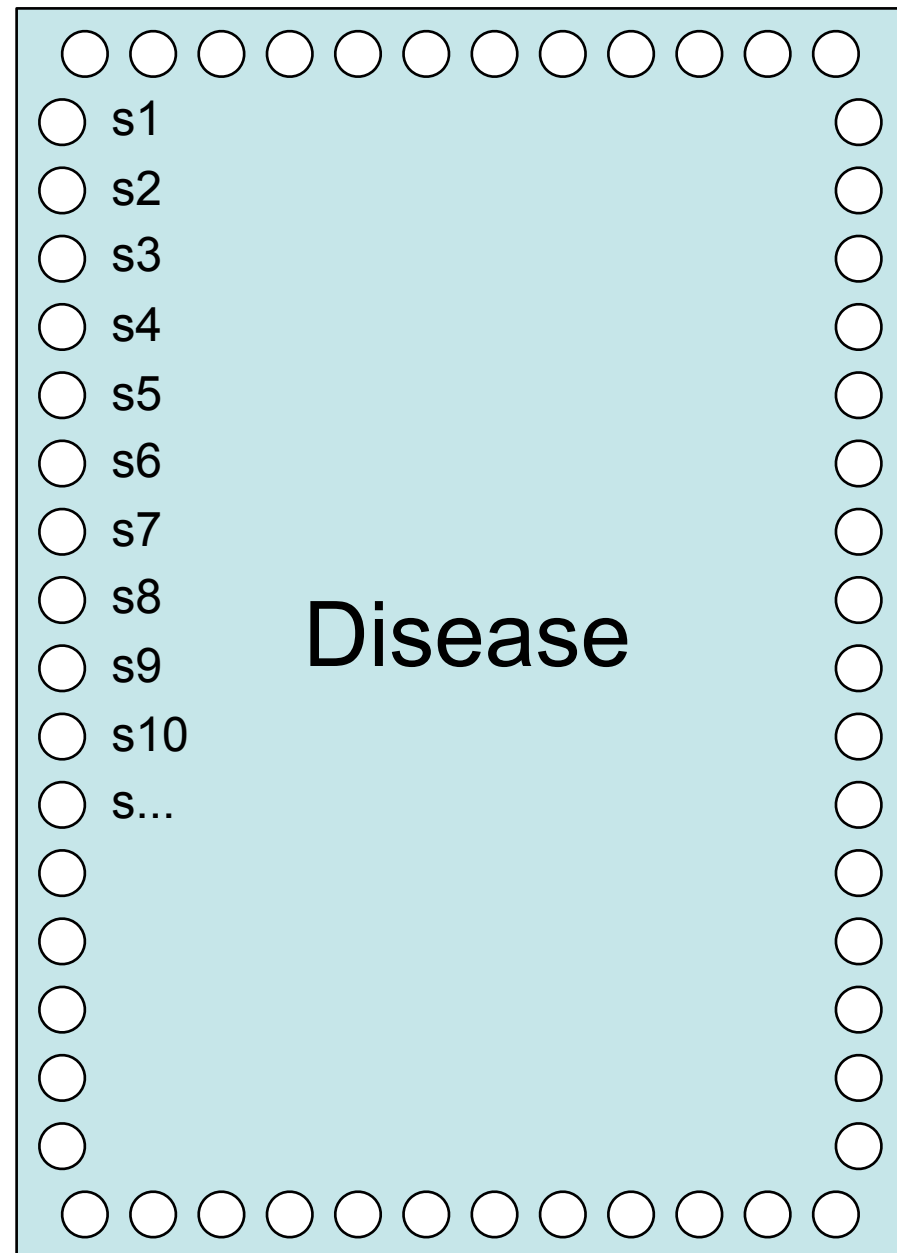
Diagnosis by Card Selection
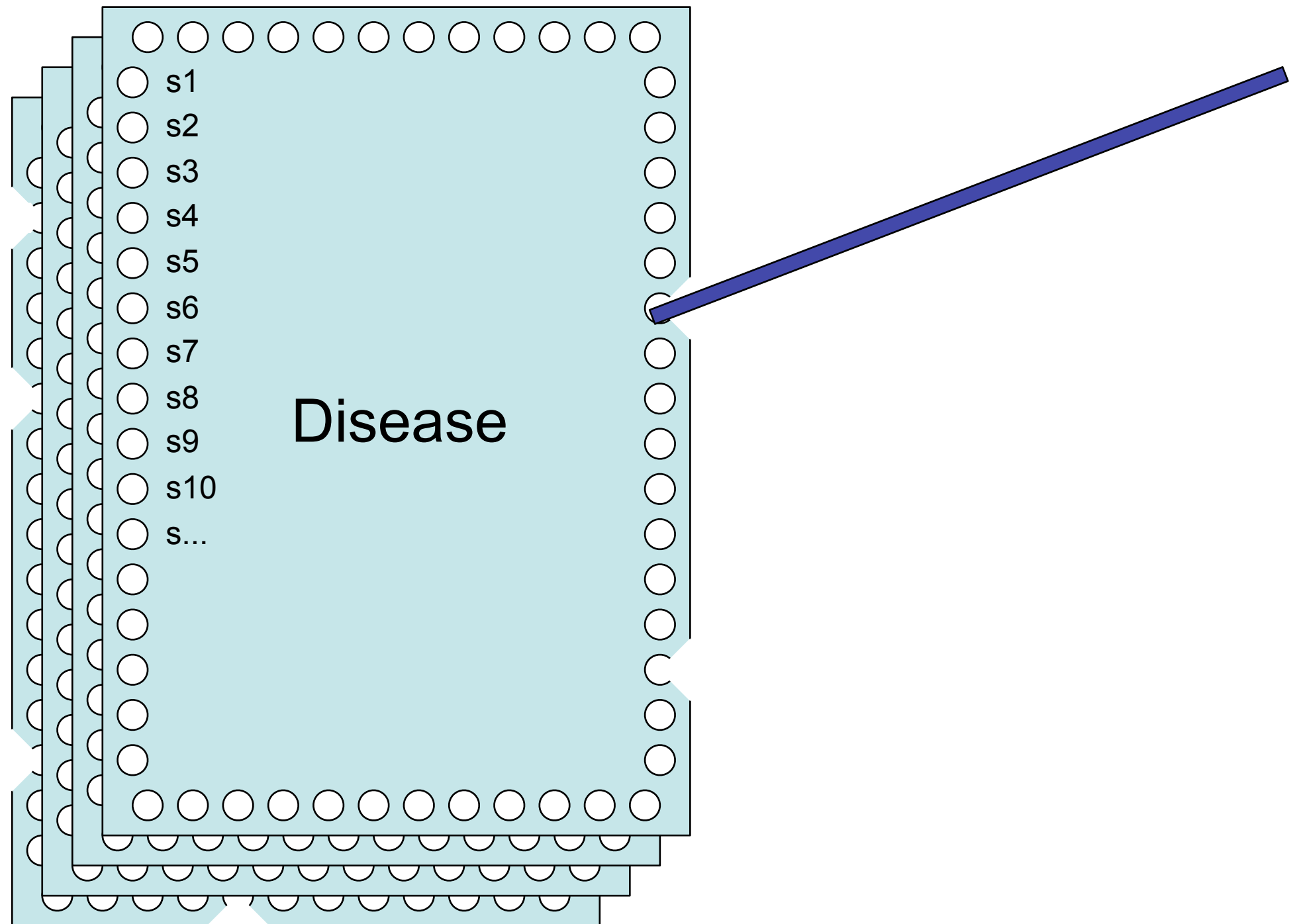
Naïve Bayes

- **Exhaustive** and **Mutually Exclusive** disease hypotheses (1 and only 1)
- **Conditionally independent** observables (manifestations)
- $P(D_i)$, $P(M_{ij}|D_i)$
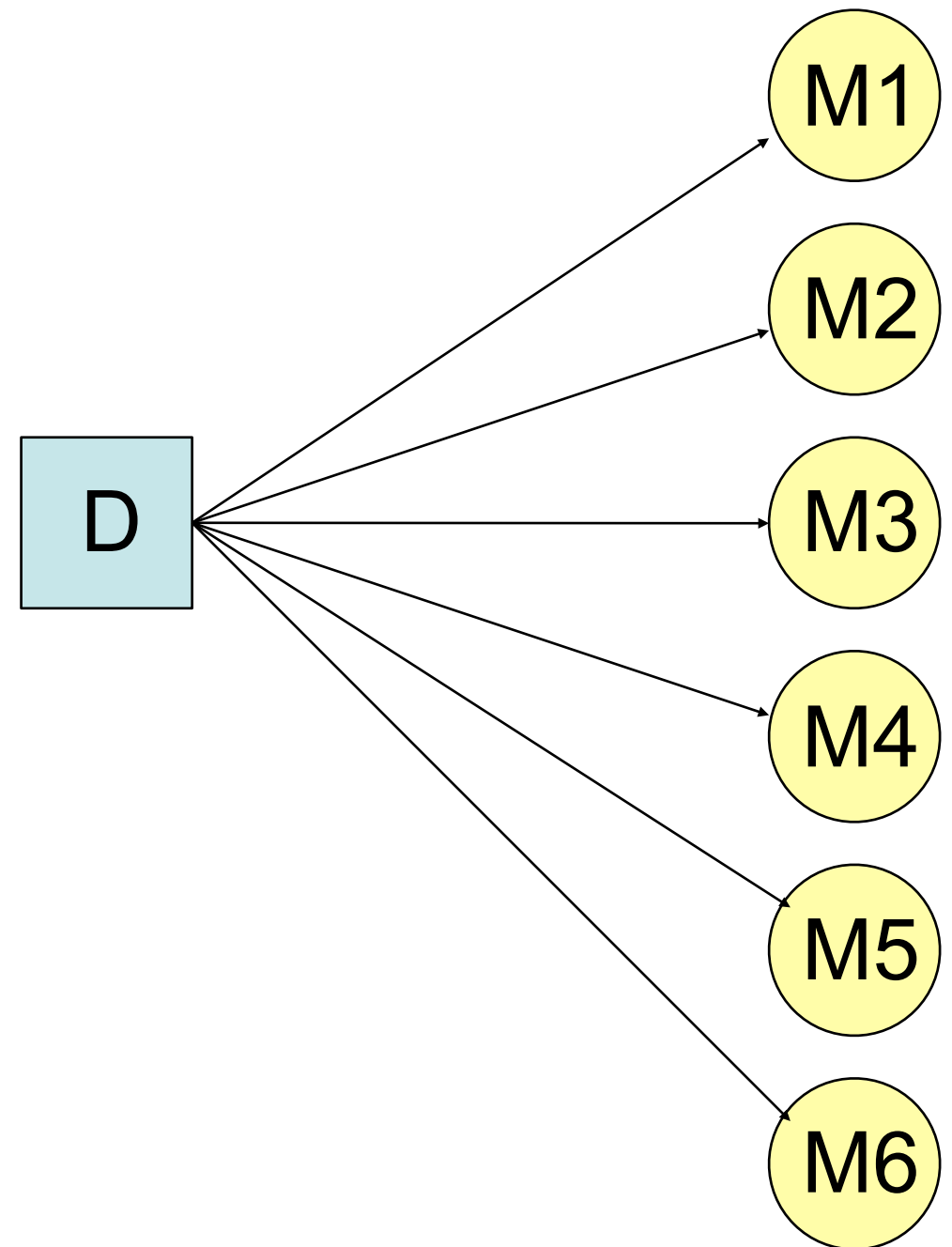
How certain are we after a test?
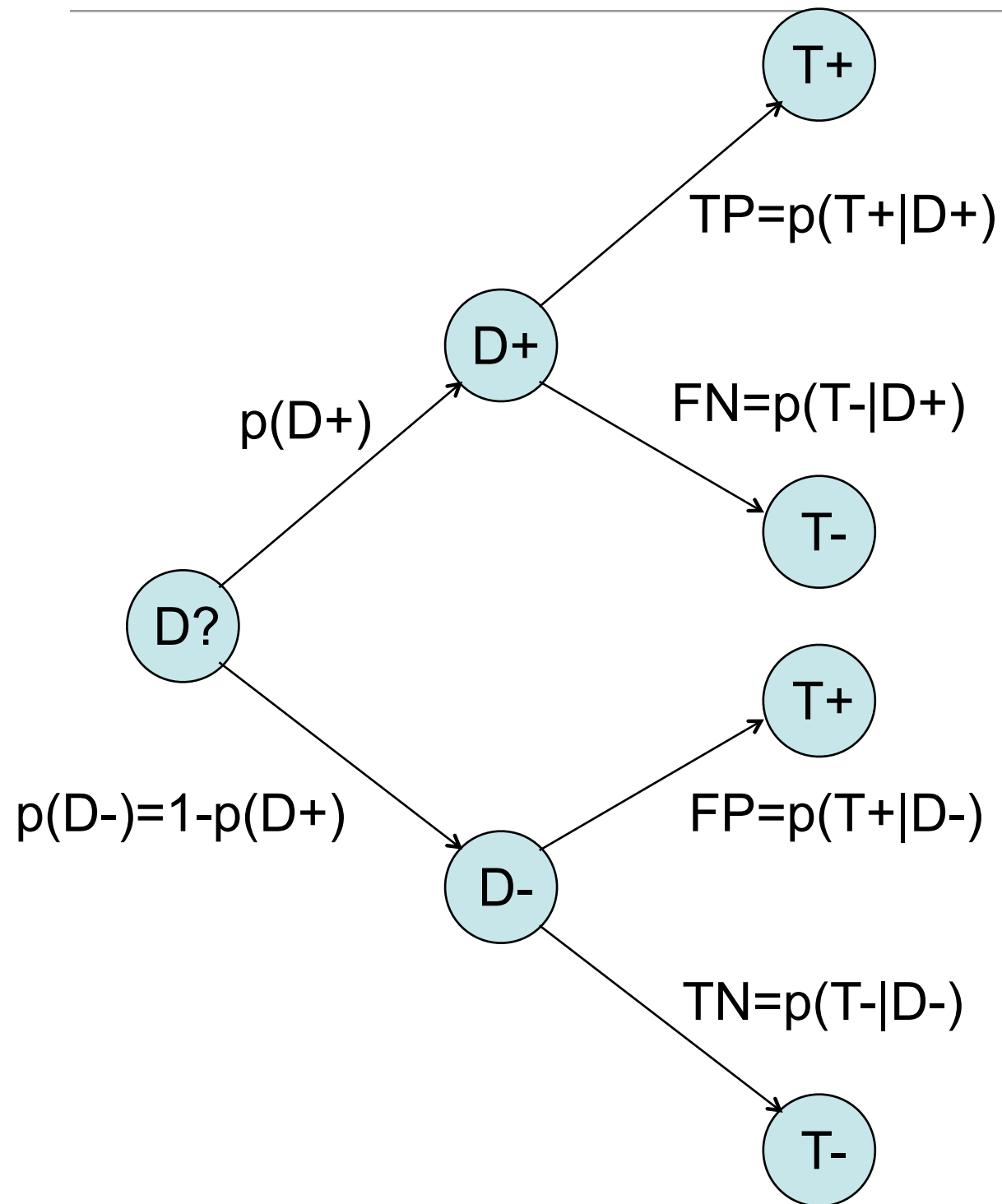
Imagine P(D+) = .001 (it's a rare disease)
Accuracy of test = P(T+|D+) = P(T-|D-) = .95

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

| D_1: 0.12 D_2: 0.37 ... D_n: 0.03 | obs S_i → | D_1: 0.19 D_2: 0.30 ... D_n: 0.01 | obs S_j → | D_1: 0.08 D_2: 0.59 ... D_n: 0.05 | obs S_k → | D_1: 0.01 D_2: 0.96 ... D_n: 0.00 |
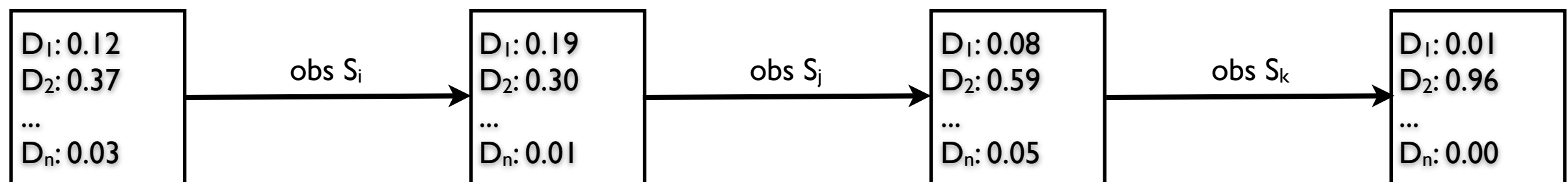
- After the j-th observation,

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

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

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

$$\log O(D|S_1, \ldots, S_n) = \log[O(D)L(S_1|D) \ldots L(S_n|D)]$$
$$= \log[O(D)] + \log[O(S_1|D)] + \ldots + \log[O(S_n|D)]$$
$$= W(D) + W(S_1|D) + \ldots + W(S_n|D)$$
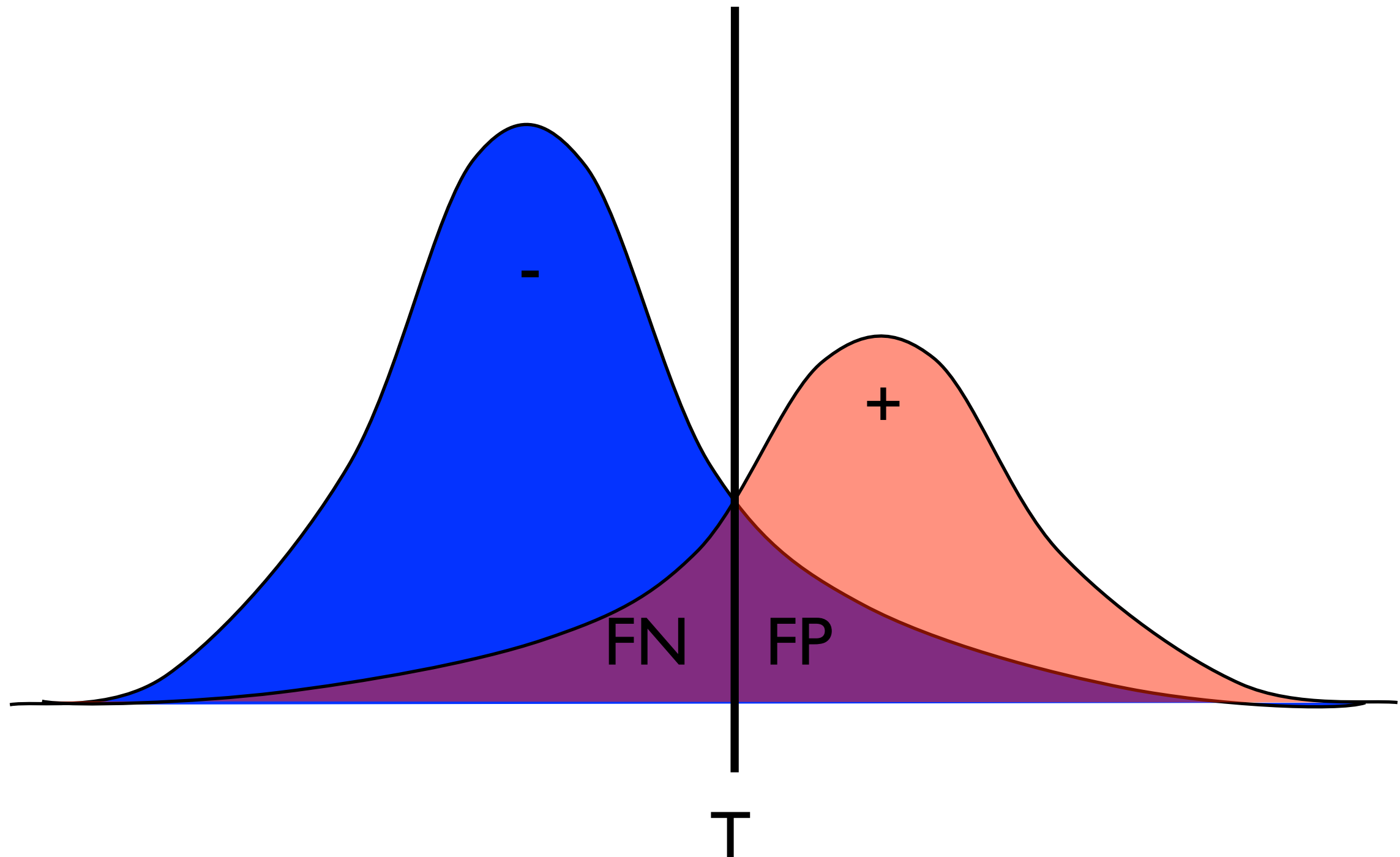
Test Thresholds

Wonderful Test
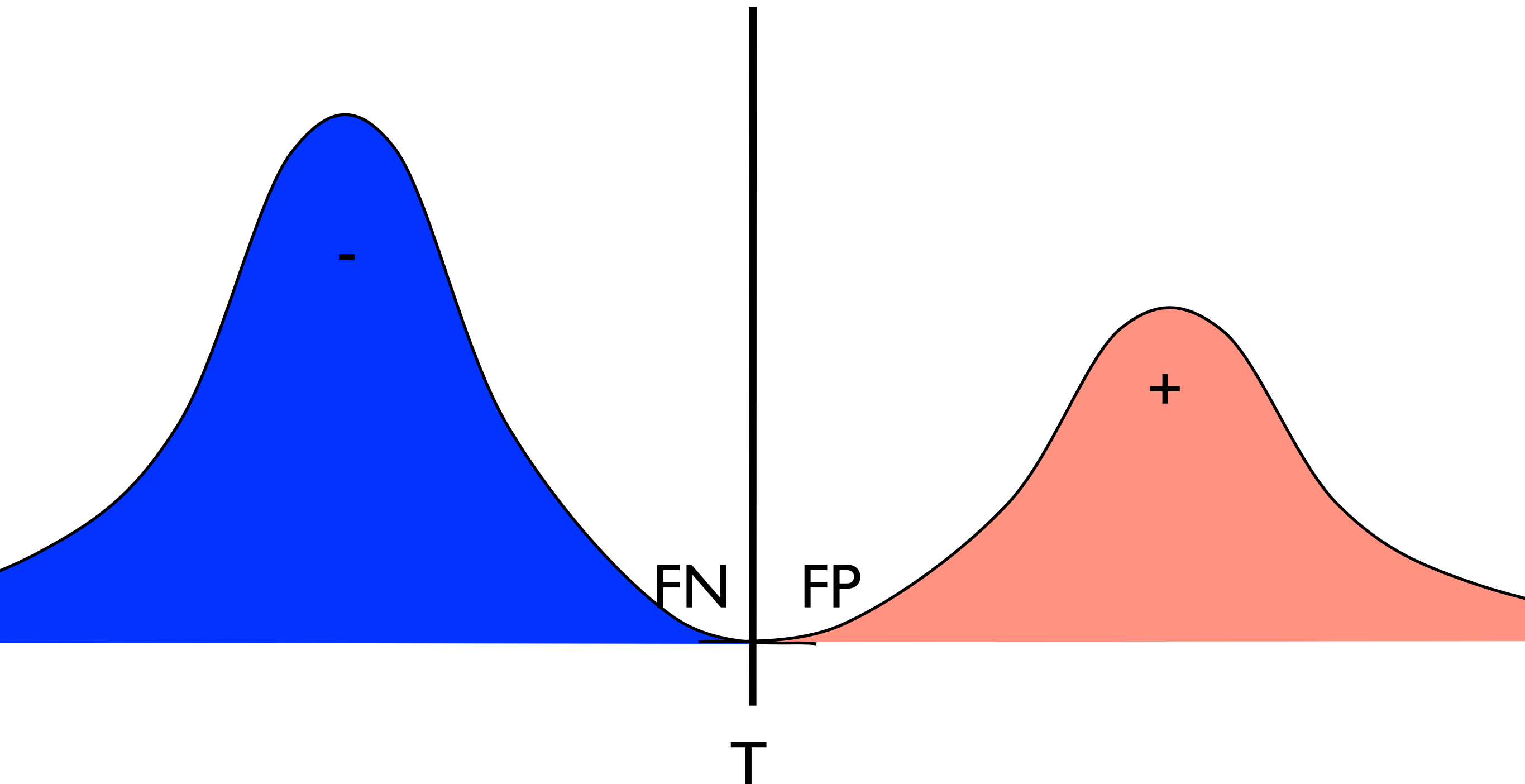
Test Thresholds Change Trade-off between Sensitivity and Specificity
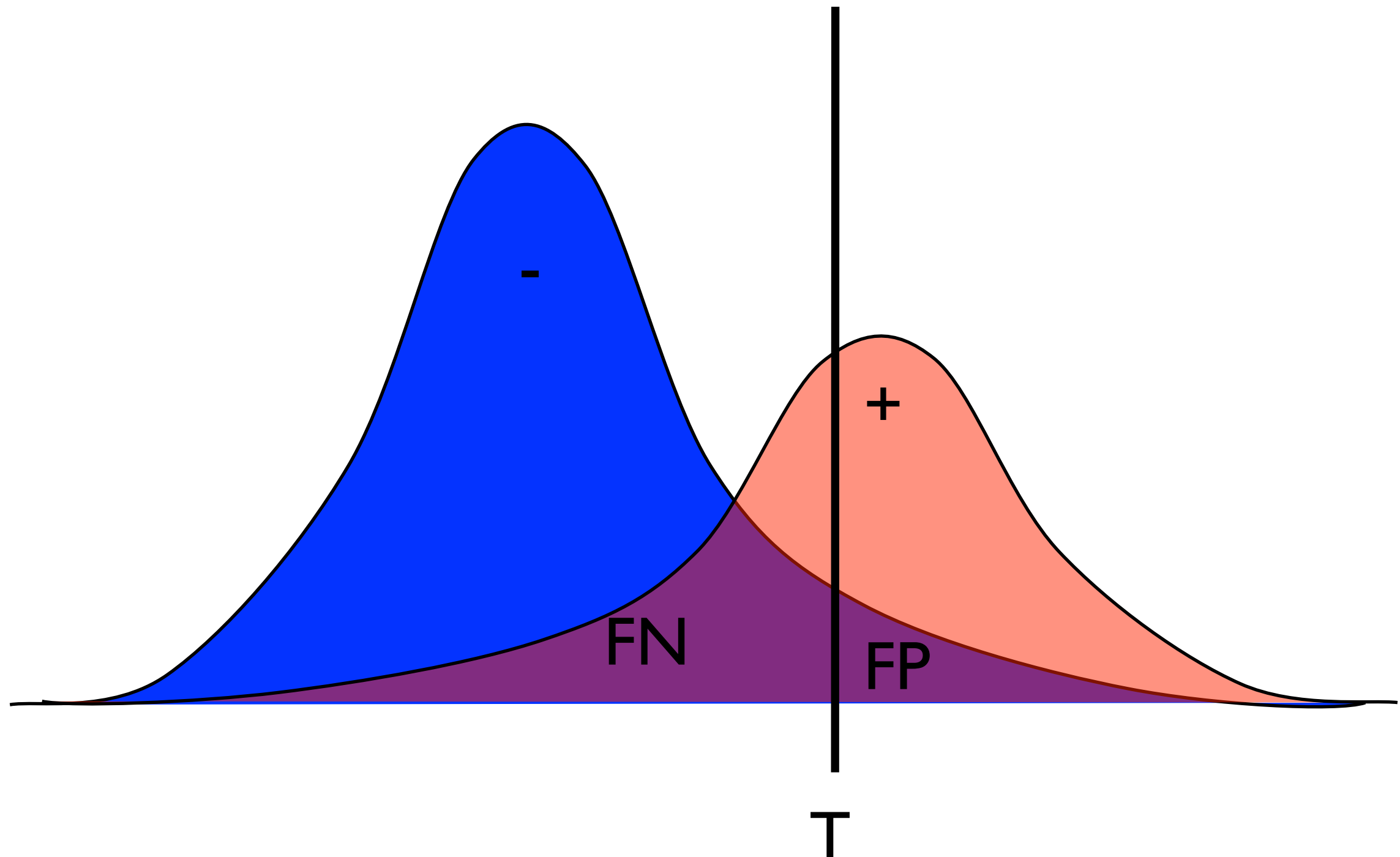
Receiver Operator Characteristic (ROC) Curve

TPR
(sensitivity)

1

T

0

0   FPR (1-specificity)   1

What makes a better test?

TPR (sensitivity)

superb

OK

worthless

FPR (1-specificity)
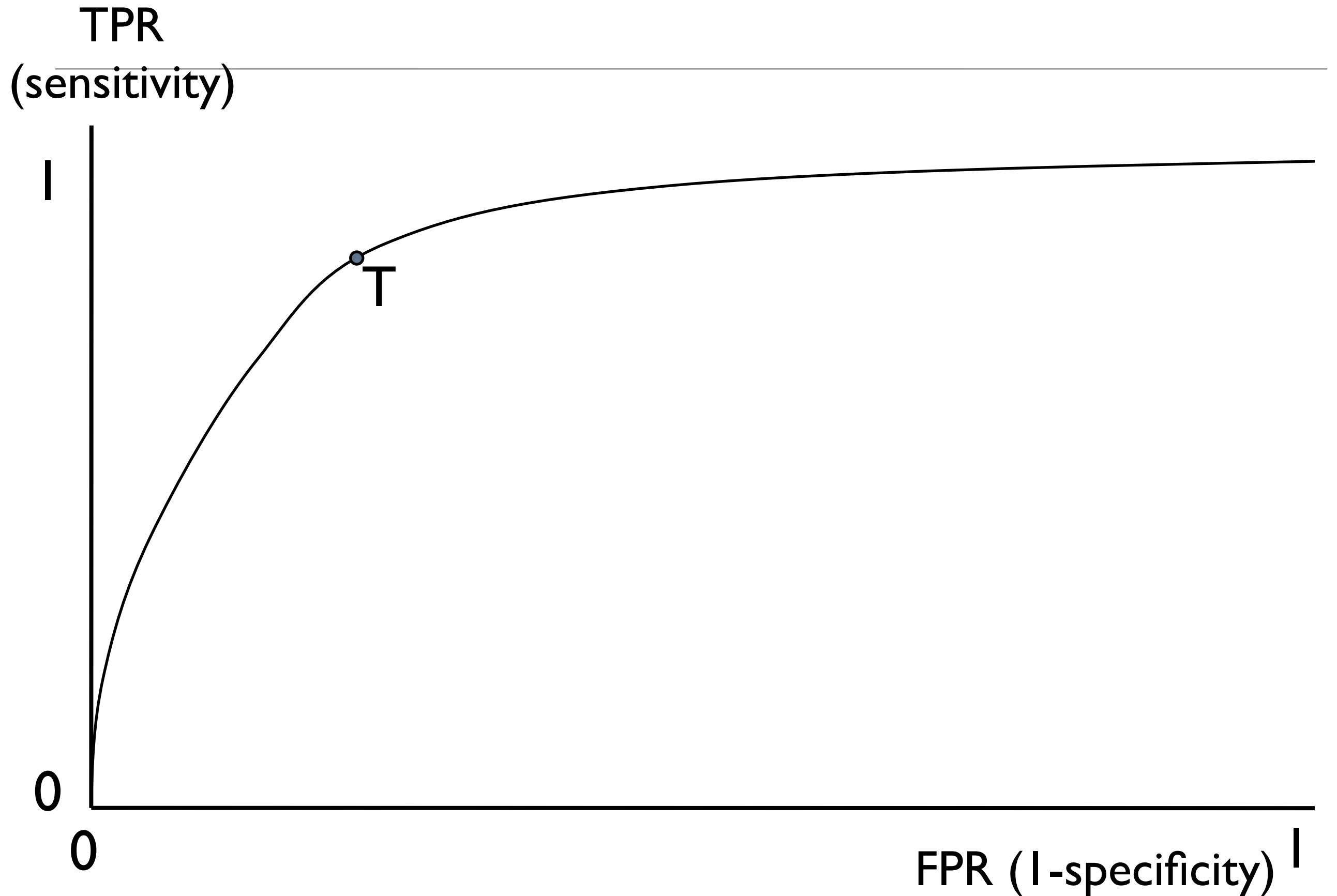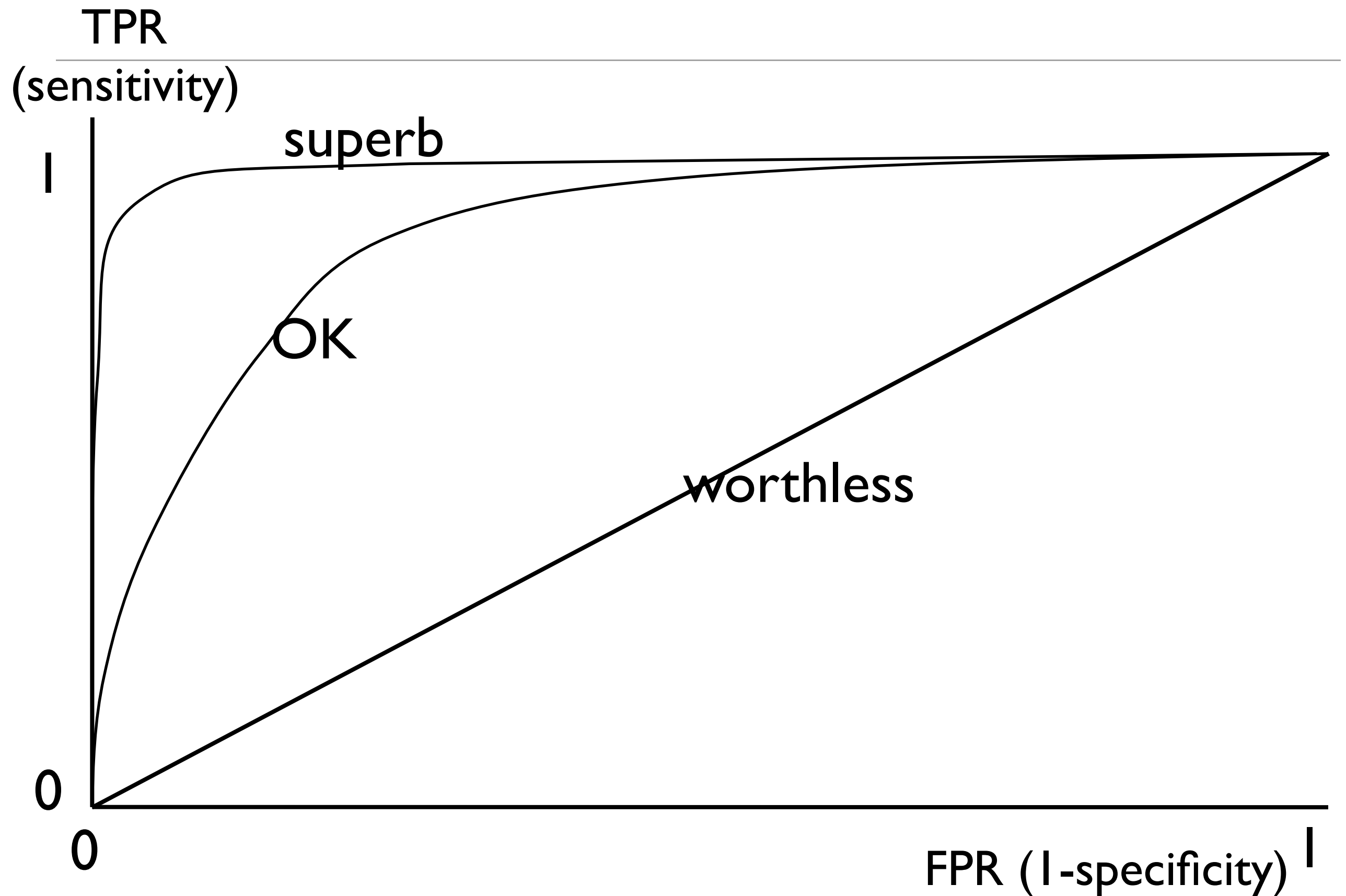
Rationality

- 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)

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
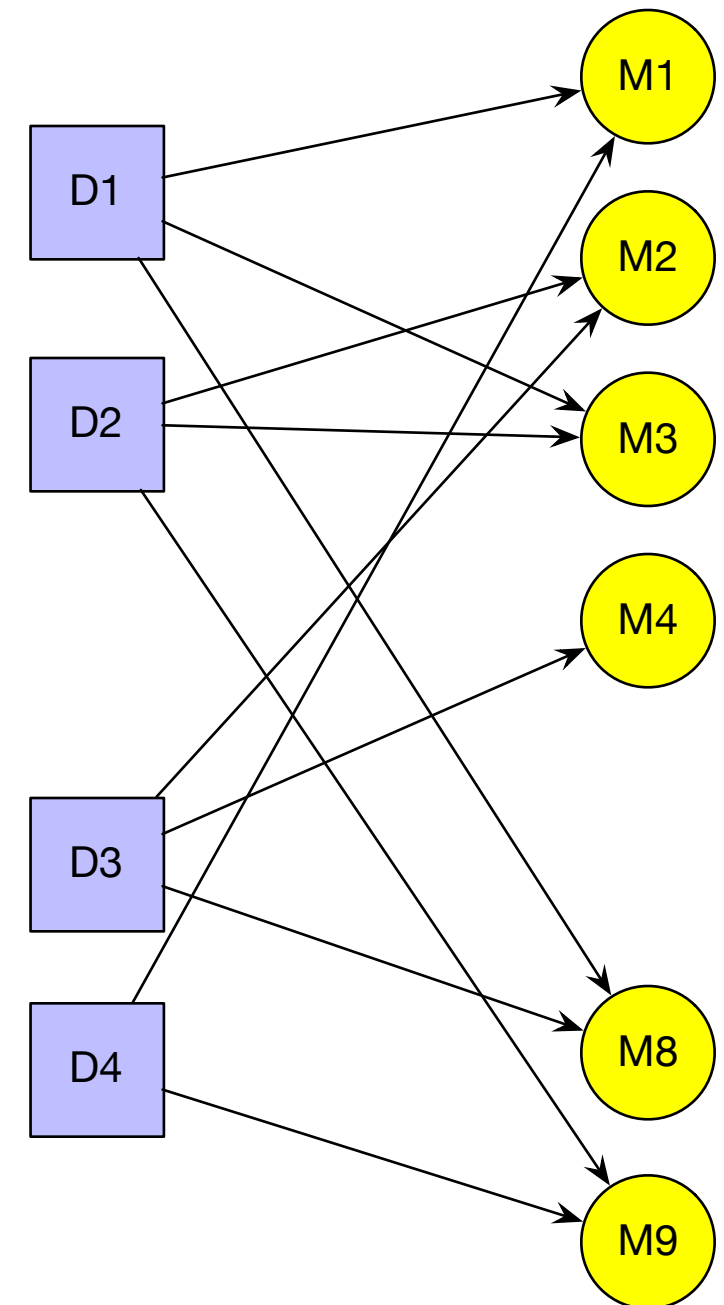
Dialog/Internist/QMR ~1982

- ~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

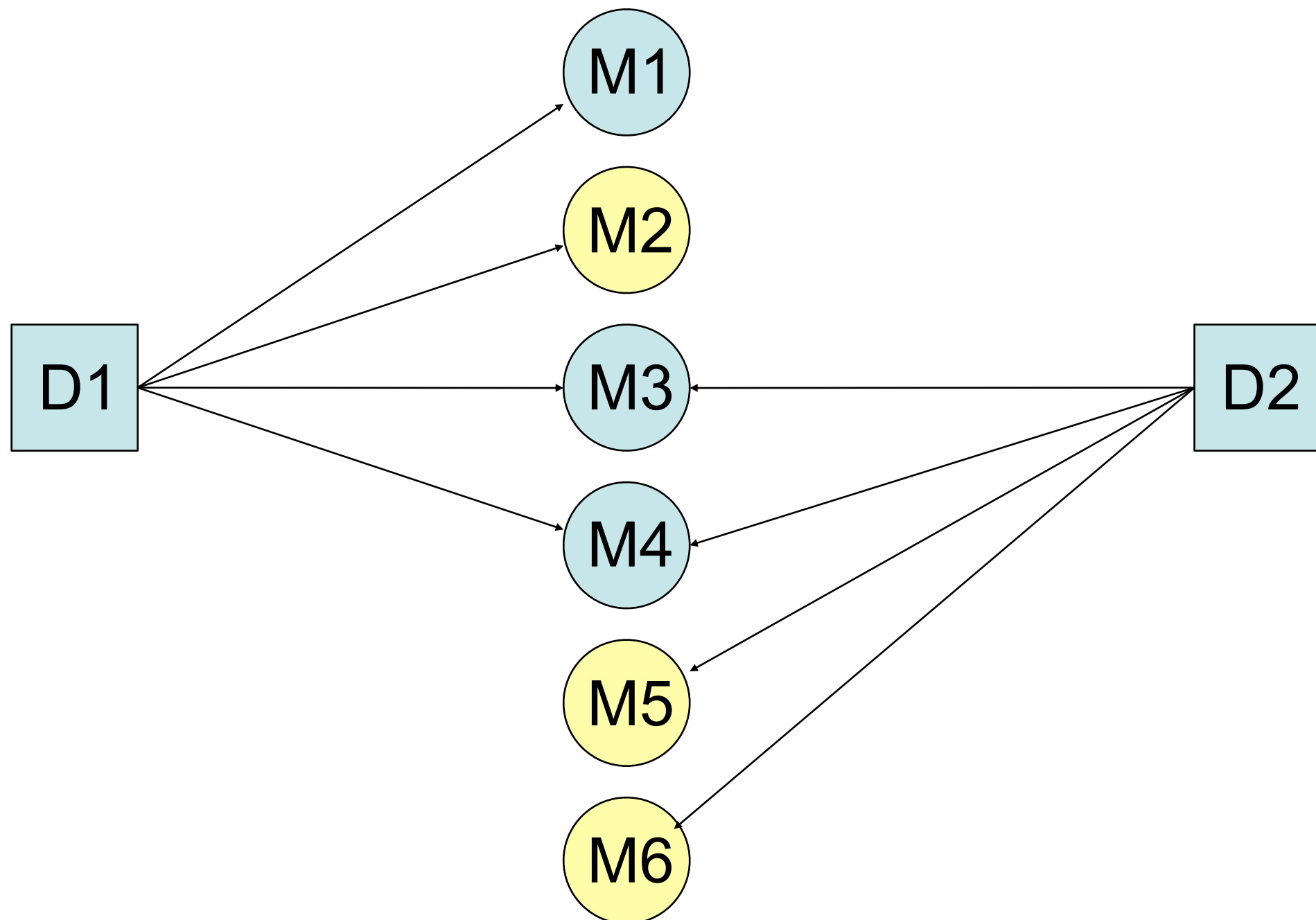- Form a differential around highest-scoring Dx
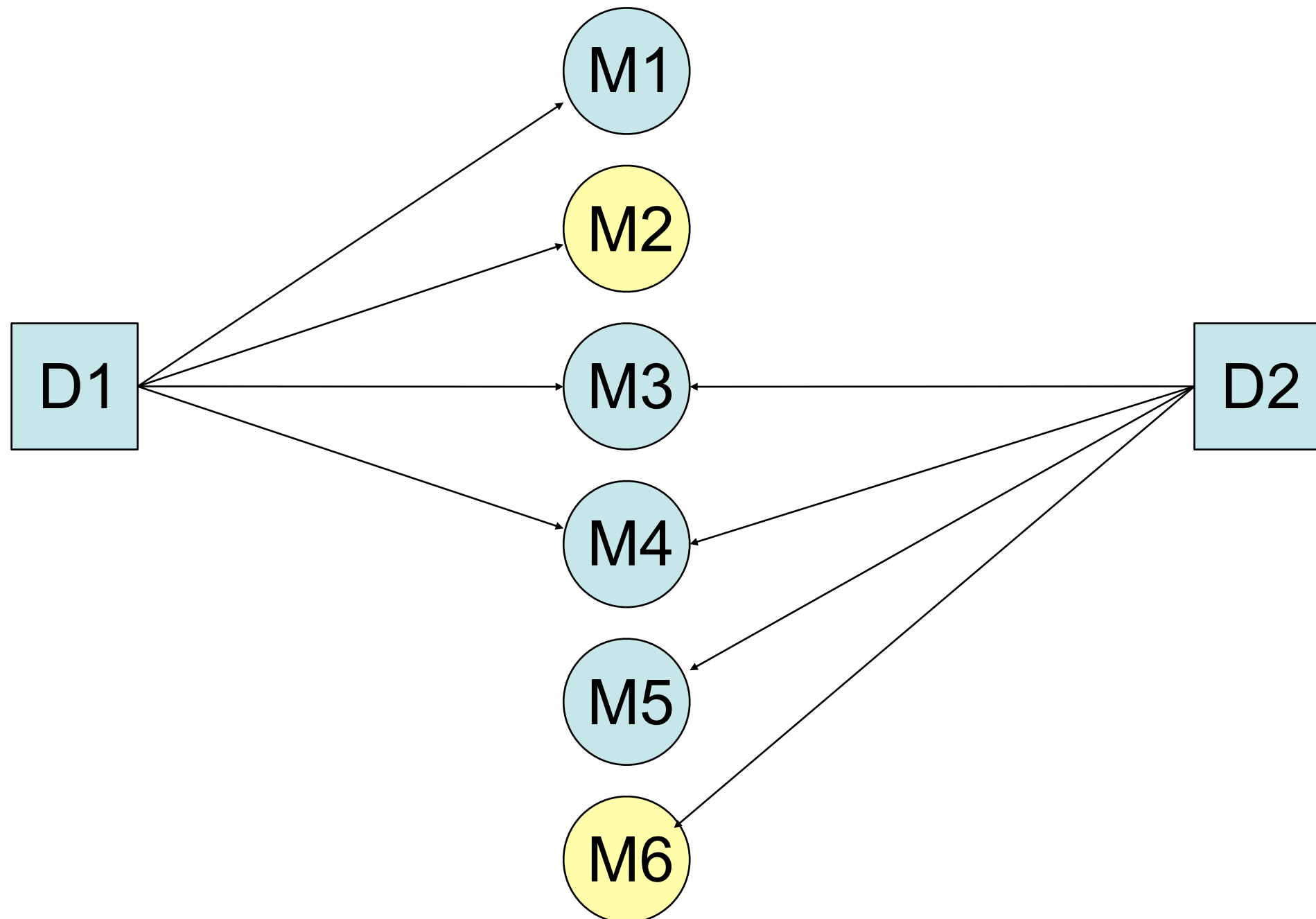
QMR Partitioning
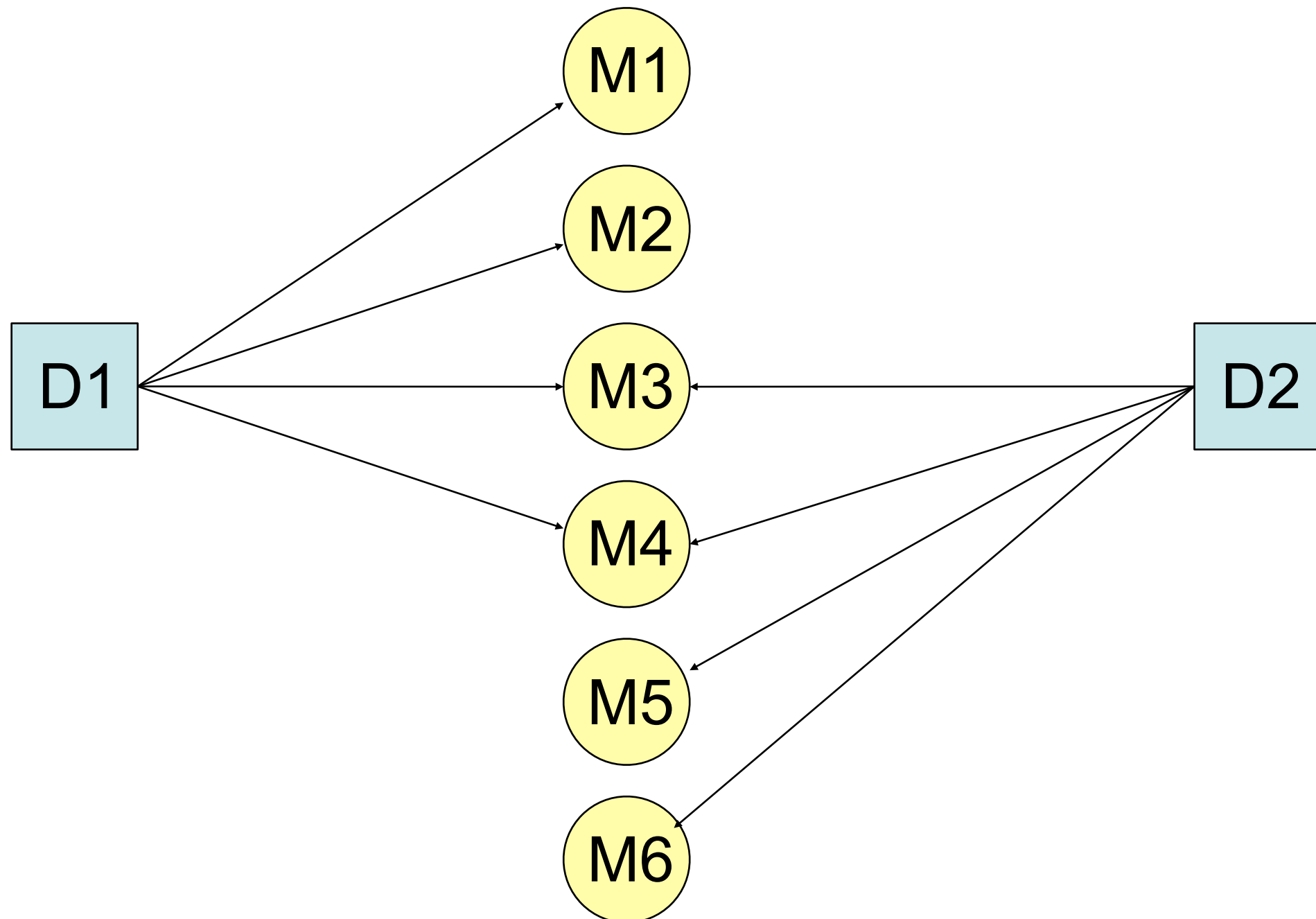
Competitors
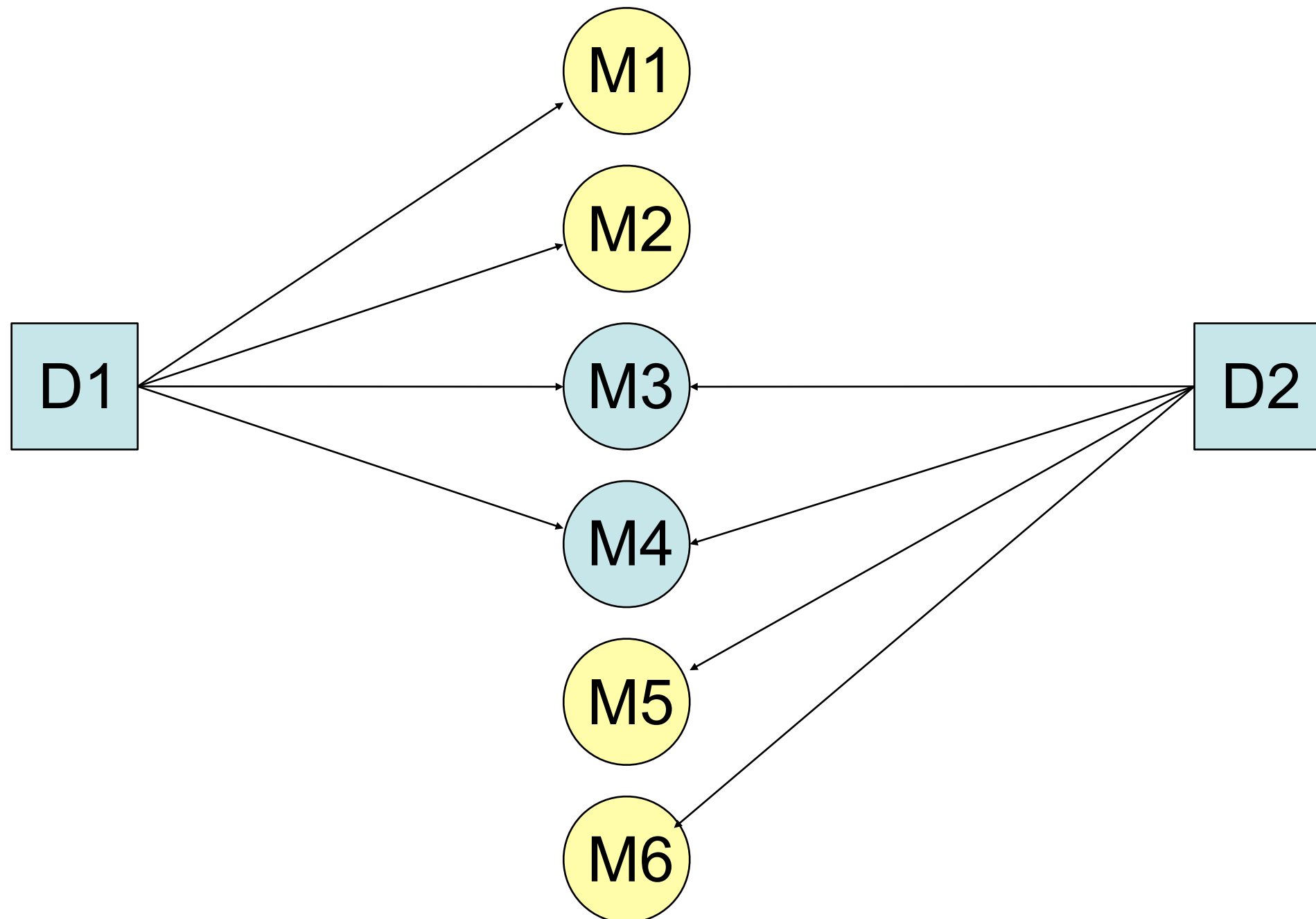
Still Competitors

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.

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

Internist Data Summary

Internist Reconstruction -- Data Summary

[Diagnose]

Manifestations PRESENT:

```
ABDOMEN DISTENTION
ABDOMEN FLUID WAVE
AGE GTR THAN 55
ALKALINE PHOSPHATASE BLOOD GTR THAN 2 TIMES NORMAL
AMMONIA BLOOD INCREASED
ANOREXIA
ARTHRITIS HX
ASCITIC FLUID PROTEIN 3 GRAM <S> PER DL OR LESS
ASCITIC FLUID WBC 100 TO 500
ASTERIXIS
BILIRUBIN BLOOD CONJUGATED INCREASED
BILIRUBIN URINE PRESENT
CHEST PAIN LATERAL EXACERBATION WITH BREATHING
CHEST PAIN LATERAL SHARP
DEPRESSION HX
DYSPNEA ABRUPT ONSET
```

[Remove Present]

Manifestations ABSENT:

```
ALCOHOLISM CHRONIC HX
ASCITIC FLUID AMYLASE INCREASED
ASCITIC FLUID CYTOLOGY POSITIVE
ASCITIC FLUID LDH GTR THAN 500
DIARRHEA CHRONIC
ESOPHAGUS BARIUM MEAL VARICES
FECES BLACK TARRY
FEVER
HEMATOCRIT BLOOD LESS THAN 35
PRESSURE VENOUS CERVICAL INCREASED ON INSPECTION
STOMACH BARIUM MEAL ULCER CRATER <S>
T3 RESIN UPTAKE INCREASED
T4 FREE BLOOD INCREASED
UREA NITROGEN BLOOD 30 TO 59
URIC ACID BLOOD INCREASED
```

[Remove Absent]

Initial Solution

Diagnostic Results

Problem:
- -94 HEPATITIS CHRONIC ACTIVE
- -119 PEDIATRIC HEPATITIS CHRONIC ACTIVE
- -136 MACRONODAL CIRRHOSIS <POSTNECROTIC>
- -158 BILIARY CIRRHOSIS PRIMARY
- -178 PEDIATRIC BILIARY CIRRHOSIS PRIMARY

Complementary:
- -143 MICRONODAL CIRRHOSIS <LAENNECS>
- -162 HEPATITIS ACUTE VIRAL
- -170 CHOLANGIOCARCINOMA <INTRAHEPATIC NON HILAR>
- -178 HEPATIC AMYLOIDOSIS

Explained:
- AGE GTR THAN 55
- ALKALINE PHOSPHATASE BLOOD GTR THAN 2 TIMES NORMAL
- ANOREXIA
- BILIRUBIN BLOOD CONJUGATED INCREASED
- BILIRUBIN URINE PRESENT
- FECES LIGHT COLORED
- HAND <S> PALMAR ERYTHEMA
- IMMUNOELECTROPHORESIS SERUM IGA INCREASED
- IMMUNOELECTROPHORESIS SERUM IGG INCREASED

Shelf:
- ABDOMEN DISTENTION
- ARTHRITIS HX
- CHEST PAIN LATERAL EXACERBATION WITH BREATHING
- CHEST PAIN LATERAL SHARP
- FECES GUAIAC TEST POSITIVE
- PLEURAL FRICTION RUB
- WEIGHT INCREASE RECENT HX

Absent:
- DIARRHEA CHRONIC
- FEVER
- HEMATOCRIT BLOOD LESS THAN 35

Unexplained:
- ABDOMEN DISTENTION
- ABDOMEN FLUID WAVE
- AMMONIA BLOOD INCREASED
- ARTHRITIS HX
- ASCITIC FLUID PROTEIN 3 GRAM <S> PER DL OR LESS
- ASCITIC FLUID WBC 100 TO 500

Askable:
- ABDOMEN PAIN CHRONIC
- ABDOMEN PAIN EPIGASTRIUM
- ABDOMEN PAIN EPIGASTRIUM UNRELIEVED BY ANTACID
- ABDOMEN PAIN EXACERBATION WITH MEAL <S>
- ABDOMEN PAIN NON COLICKY
- ABDOMEN PAIN PRESENT
- ABDOMEN PAIN RIGHT UPPER QUADRANT
- ABDOMEN TENDERNESS PRESENT
- ABDOMEN TENDERNESS RIGHT UPPER QUADRANT
- ACTIVATED PARTIAL THROMBOPLASTIN TIME INCREASED
- AGE 16 TO 25
- AGE 26 TO 55
- ALBUMIN SERUM DECREASED
- ALKALINE PHOSPHATASE BOOD INCREASED NOT OVER 2 TIMES NORMA

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

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
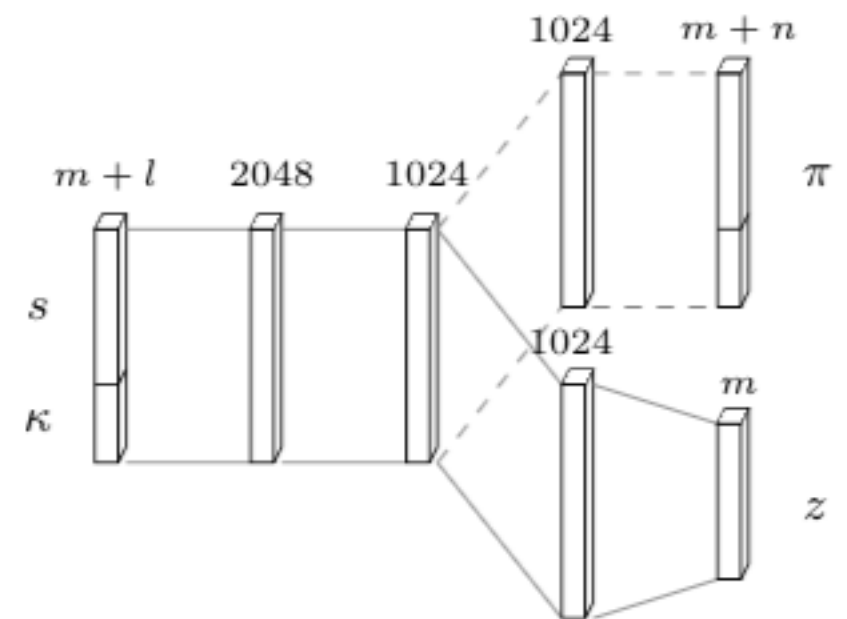 - Identify reduced finding space by *feature rebuilding.*

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
.

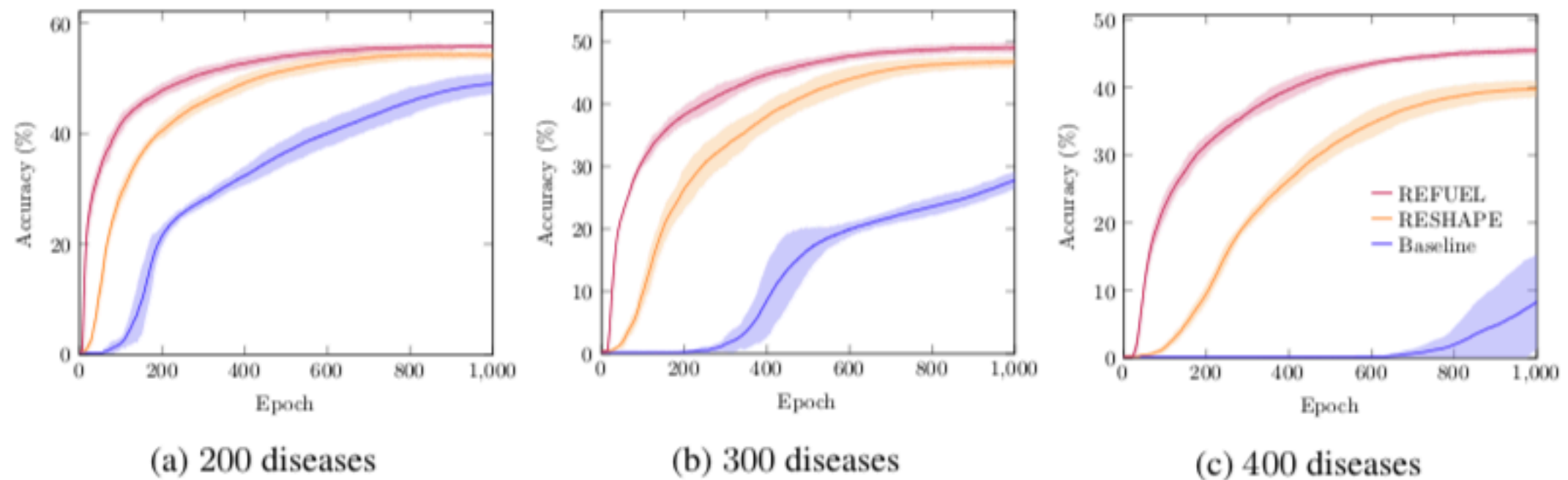(a) 200 diseases   (b) 300 diseases   (c) 400 diseases

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.