Announcement

- HW4 on BNs due today

Find a data sample that justifies the following interchange with Dr. Bayes

Is the patient male or female?
Male
then administer treatment A

Is the patient male or female?
Female
then administer treatment A

Is the patient male or female?
Unknown
then administer treatment B

How to proceed?

- Build an empirical model (well, hybrid – in fact mostly analytic)
- Dr. Bayes interactions are constraints on model parameters
- Then, either
  - Choose parameters to satisfy constraints
    and
    make up data to yield these parameters
  - OR Convince ourselves of their inconsistency

Building a Model
(for Dr. Bayes)

- Build! Don't just wait for an inspiration
- What are the random variables?
  - (what relevant features change across individuals?)
  - G: Gender (male / female)
  - T: Treatment (a / b)
  - I: Improvement (yes / no)

  (All Boolean)
- Joint has …?

  8 numbers, 7 parameters

Joint Distribution

- Three Boolean random variables: Gender m/f, Treatment a/b, Improvement y/n
- N patients

(y / n)	a	b
m	may / man	mby / mbn
f	fay / fan	fby / fbn

"may" is the number of males who improved after treatment "a"

divide each count by N to get estimated probabilities
(sample averages) which will then sum to 1

Constraints on Parameters

- Is the patient male or female?
  Male [Female]
  then administer treatment A
- Meaning in the model?
- $P(I{=}y \mid G{=}m, T{=}a) > P(I{=}y \mid G{=}m, T{=}b)$
- $P(I{=}y \mid G{=}f, T{=}a) > P(I{=}y \mid G{=}f, T{=}b)$
- Is the patient male or female?
  Unknown
  then administer treatment B
- $P(I{=}y \mid T{=}a) < P(I{=}y \mid T{=}b)$

Constraints on Parameters

- P(I=y | G=m, T=a) > P(I=y | G=m, T=b)
- may / ma > mby / mb
- may / (may + man) > mby / (mby + mbn)

- P(I=y | G=f, T=a) > P(I=y | G=f, T=b)
- fay / (fay + fan) > fby / (fby + fbn)

- P(I=y | T=a) < P(I=y | T=b)
- ay / a < by / b

- Some search and arithmetic…

Dr. Bayes

- Gender m/f, Treatment a/b, Improvement y/n
- 100 patients:
  - 50 m 50 f
  - 50 a 50 b

(y / n)	a	b
male	25/15	5/5
female	9/1	32/8

- P(y|m,a) >? P(y|m,b)
- P(y|m,a) = 25/40 = 0.625 P(y|m,b) = 5/10 = 0.5

- P(y|f,a) >? P(y|f,b)
- P(y|f,a) = 9/10 = 0.9 P(y|f,b) = 32/40 = 0.8

- P(y|a) = 34/50 = 0.68 P(y|b) = 37/50 = 0.74

Simpson's "Paradox"
real-world examples

- Two hospitals in a city
  - Hospital A is better for minor injuries
  - Hospital A is better for serious injuries
  - Hospital B is better if unknown
- Berkeley engineering college
  - Overall the college discriminates against females and for males in admission
  - But each department discriminates against males and for females in admission

Machine Learning

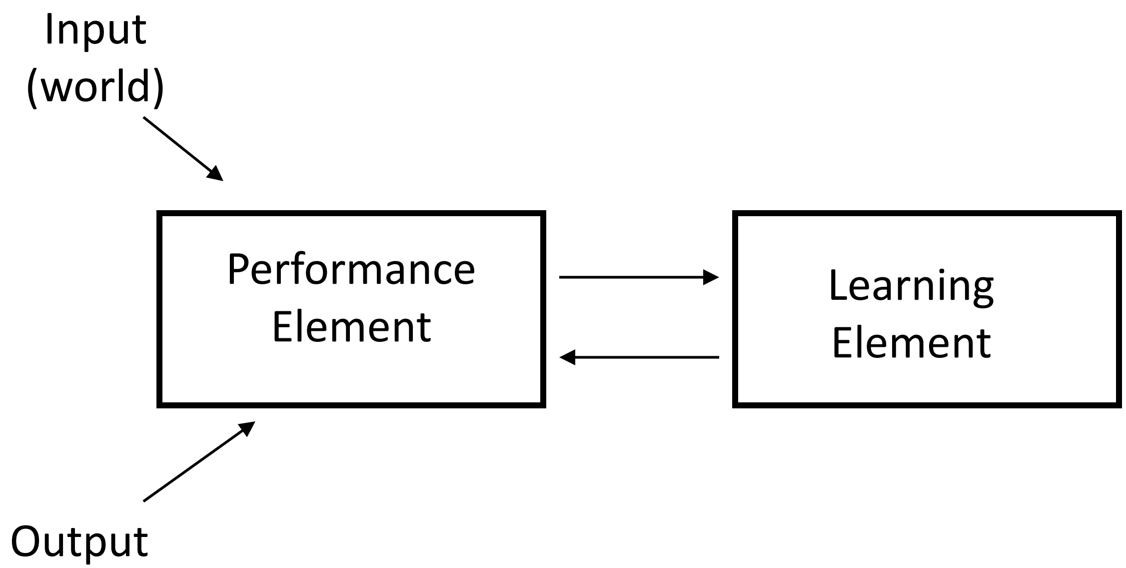
More Human Involvement

Less Human Involvement

Mostly This One

- Programming (?)
- Rote Learning
- Supervised Learning from Examples
- Semi-supervised Learning
- Learning by Analogy
- Learning from Observation (active learning)
- Data Mining / Unsupervised Learning
- Discovery

General Learning Paradigm



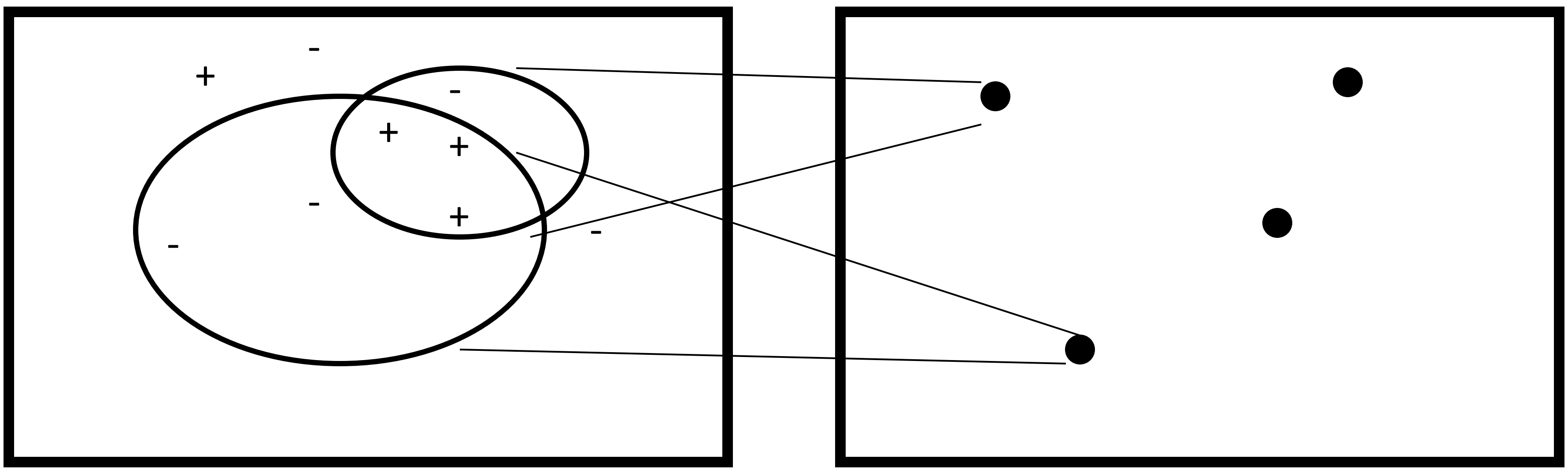
Peformance Element task is usually either

Classification	or	Problem Solving
Chest X-ray diag.		Planning
Insurance risk eval.		Network configuration
Handwritten recog.		Adaptive user interface
…		…

Important Distinctions and Ideas

- On line (incremental / streaming) vs. Batch
- Supervised / Unsupervised / Semi-supervised
  (compare w/ reinforcement learning)
- Generative vs. Discriminative
- Two Spaces for a Learner
  - Example Space – all possible inputs
  - Hypothesis (concept) Space
    - All possible outputs (concepts)
    - Each partitions the input space
- Training Set for Supervised Learning

Machine Learning as an Empirically Guided Search through the Hypothesis Space



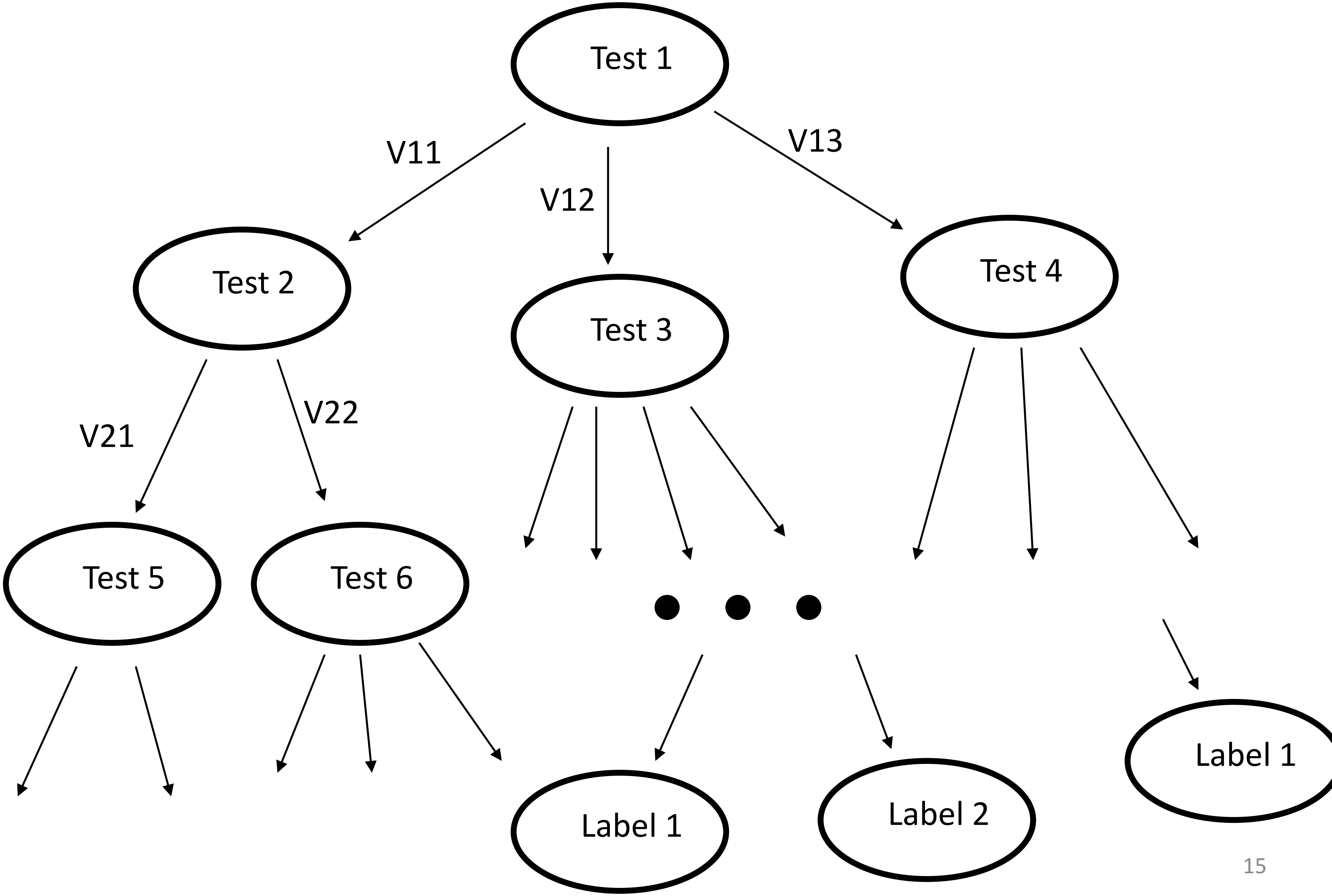
Examples

Hypotheses

Learning Decision Trees for Classification

- Ross Quinlan
  - ID3
  - C4.5
  - C5.0 (commercial product)
  - AI / ML
- Breiman, Friedman, Olshen, & Stone
  - CART
  - Statistics

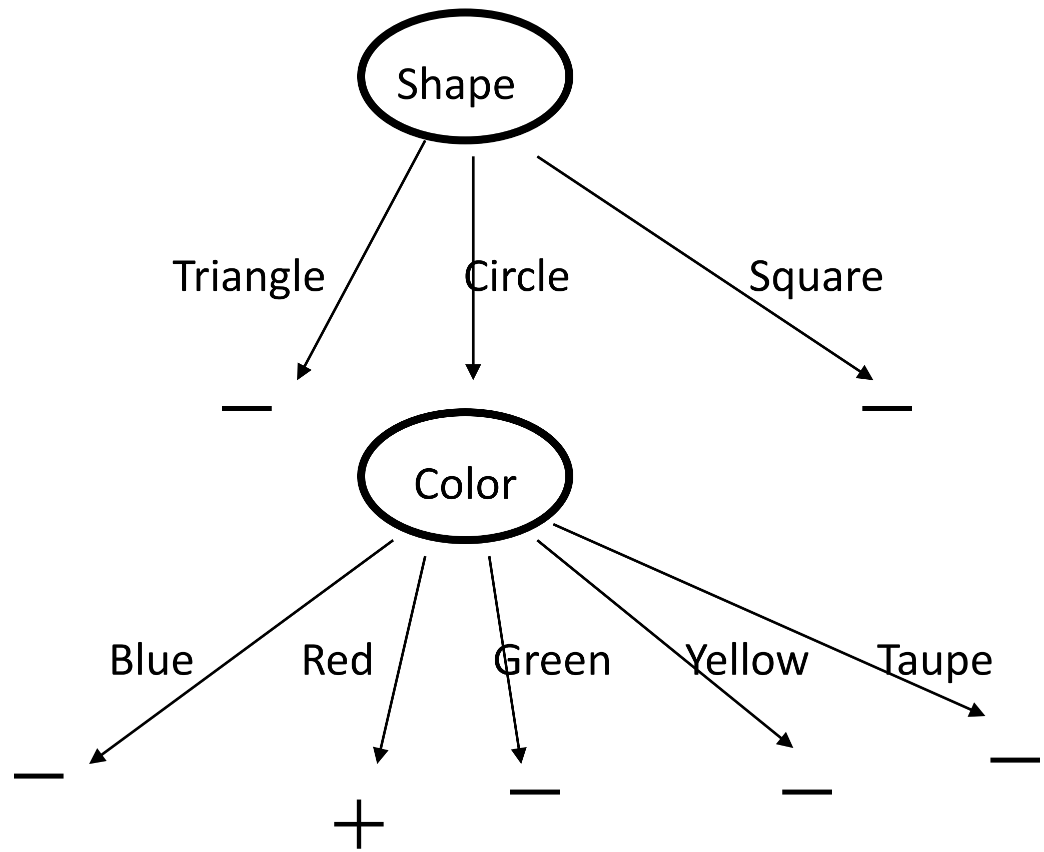
What is a decision tree?



Suppose I like circles that are red

(I might not be aware of the rule)

- Features:
  - Owner
    - John, Mary, Sam
  - Size
    - Large, Small
  - Shape
    - Triangle, Circle, Square
  - Texture
    - Rough, Smooth
  - Color
    - Blue, Red, Green, Yellow, Taupe



$$\forall x\ [\text{Like}(x) \Leftrightarrow (\text{Circle}(x) \wedge \text{Red}(x))]$$

Decision Tree Learning by Hill Climbing

- If node is homogeneous (or good enough) then STOP
- Choose most useful test on which to split
- Recur on Children

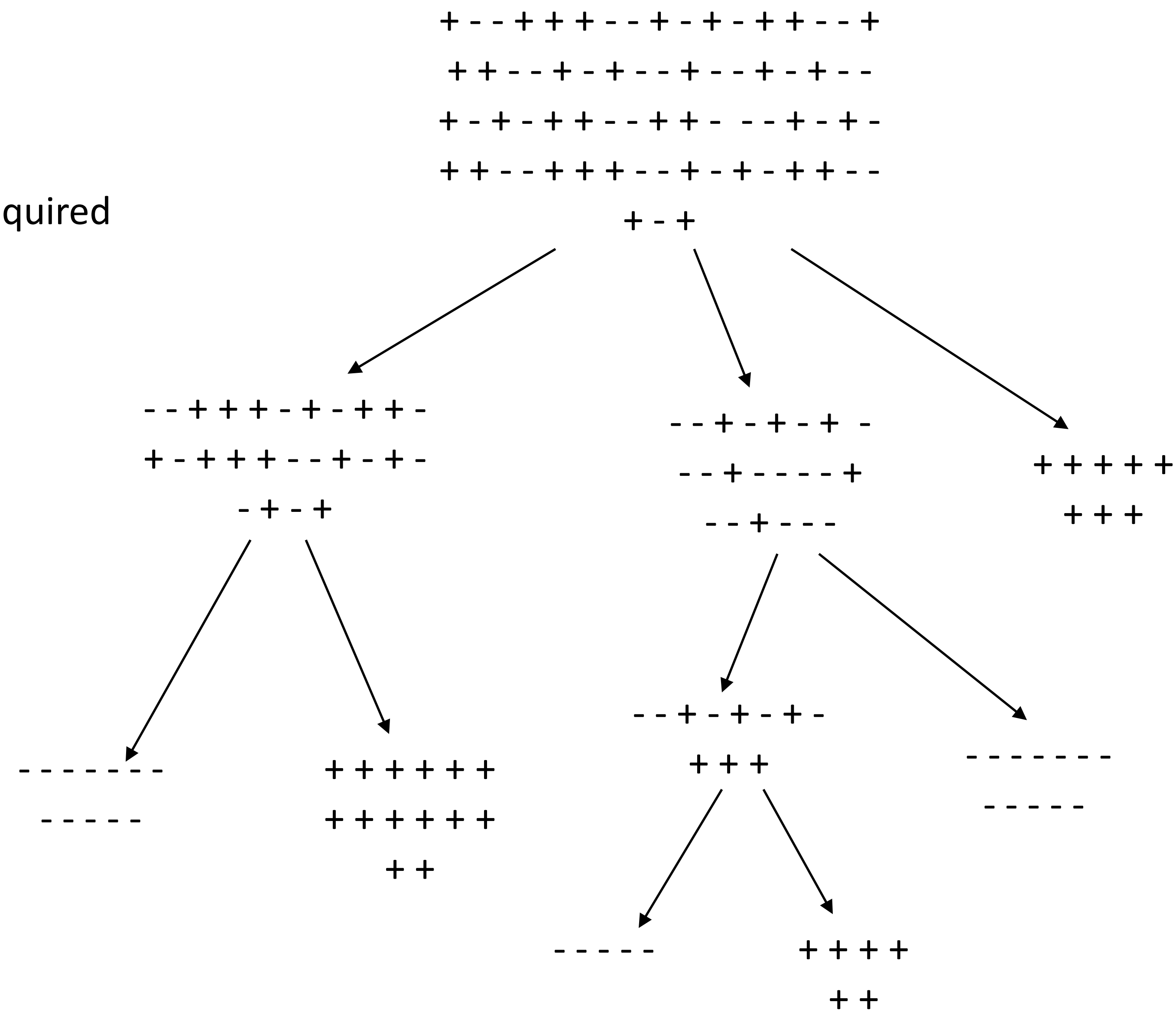
Hypothesis space is the set of all decision trees

Training Data

Highly Disorganized

High Entropy

Much Information Required



Highly Organized

Low Entropy

Little Information Required

What makes a (test / split / feature) useful?

- Improved homogeneity
  - Entropy reduction
  - Information gain
- To evaluate a split utility
  - Measure entropy / information required before
  - Measure entropy / information required after
  - Subtract
- Expected number of bits to communicate the label of an item chosen randomly from a set

Measuring Information

H denotes *Information Need* or *Entropy*

- H(S) = bits required to label some $x \in S$
- What is the upper bound if label $\in$ {+,-}
- What is $H(S_1)$?

S_1 = + + +

Measuring Information

- H(S) = bits required to label some $x \in S$
- What is the upper bound if label $\in \{+,-\}$
- What is $H(S_1)$?
- What is $H(S_2)$? $S_2 =$ – –
 – –

Measuring Information

- $H(S)$ = bits required to label some $x \in S$
- What is the upper bound if label $\in \{+,-\}$
- What is $H(S_1)$?
- What is $H(S_2)$?
- What is $H(S_3)$?

S_3 =

\+ + + + + + + + + + + +
\+ + + + + + + + + + + +
\+ + + + + + + + + + + +
\+ + + + + + + + + + + +

Measuring Information

- H(S) = bits required to label some $x \in S$
- What is the upper bound if label $\in \{+,-\}$
- What is $H(S_1)$?
- What is $H(S_2)$? $S_4 = \; + -$
- What is $H(S_3)$?
- What is $H(S_4)$?

Measuring Information

- H(S) = bits required to label some $x \in S$
- What is the upper bound if label $\in \{+,-\}$
- What is $H(S_1)$?
- What is $H(S_2)$?
- What is $H(S_3)$?
- What is $H(S_4)$?
- What is $H(S_5)$?

$$S_5 = \begin{matrix} + & + & + & + & + & + & + & + & + & + & + & + \\ + & + & + & + & + & + & + & + & + & + & + & + \\ - & - & - & - & - & - & - & - & - & - & - & - \\ - & - & - & - & - & - & - & - & - & - & - & - \end{matrix}$$

Measuring Information

- H(S) = bits required to label some $x \in S$
- What is the upper bound if label $\in \{+,-\}$
- What is $H(S_1)$?
- What is $H(S_2)$?
- What is $H(S_3)$?
- What is $H(S_4)$?
- What is $H(S_5)$?
- What is $H(S_6)$?

$S_6 =$

+ + + + + + + + + + + + +
+ + + + + + + + + + + + +
+ + + + + + + + + + + + +
+ + + + + + + + + + + + −

Think of *expected* number of bits

$H(S_6)$ should be closer to 0 than to 1
Information theory / coding theory is relevant

Measuring Information

- H(S) = bits required to label some $x \in S$
- Label $\in$ {A,B,C,D,E,F}, Upper bound now?
- What is $H(S_7)$?

FOR	SAY
A	1
B	01
C	0000
D	0001
E	0010
F	0011

S_7 =

F A B B A A B A D
A A A D A B E A F
A A B B A C A E B
A A A B C

=

A A A A A A A A
A A A A A A A A   16
B B B B B B B B   8
C C D D E E F F   2 2 2 2

Sometimes needs 4 bits / label (worse than 3)

Measuring Information

What is the expected number of bits?

- 16/32 use 1 bit
- 8/32 use 2 bits
- 4 x 2/32 use 4 bits

S_7 =

A A A A A A A A	16
A A A A A A A A	
B B B B B B B B	8
C C D D E E F F	2 2 2 2

FOR	SAY
A	1
B	01
C	0000
D	0001
E	0010
F	0011

0.5(1) + 0.25(2) + 0.0625(4) + 0.0625(4) + 0.0625(4) + 0.0625(4)

= 0.5 + 0.5 + 0.25 + 0.25 + 0.25 + 0.25

= 2

H(S) = $$\sum_{v \in Labels} -\Pr(v) \cdot \log_2(\Pr(v))$$

Information Gain

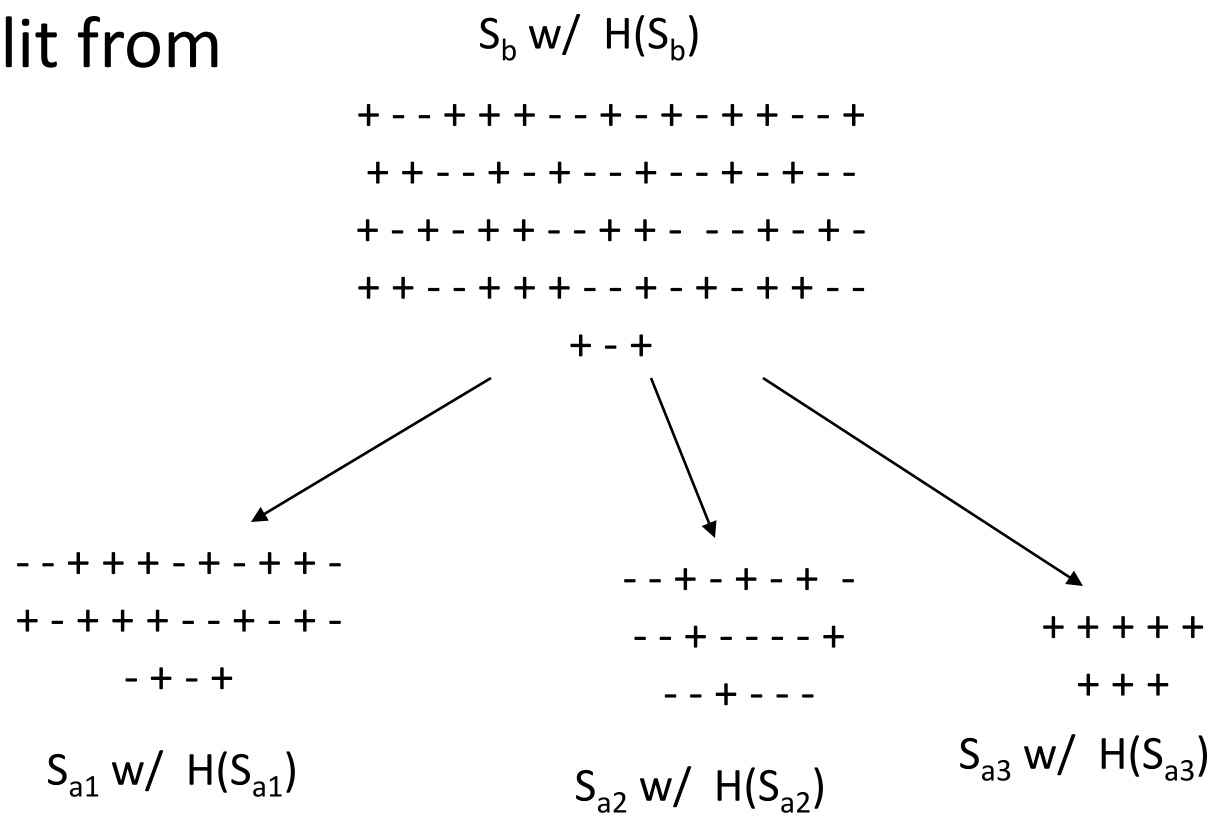
Subtract Information required after split from before

Information required:

Before $H(S_b)$

After $Pr(S_{a1}) \cdot H(S_{a1}) +$
$Pr(S_{a2}) \cdot H(S_{a2}) +$
$Pr(S_{a3}) \cdot H(S_{a3})$

Estimate probabilities using sample counts



Information Gain = $$\mathbf{H}(S_b) - \sum_i \mathbf{H}(S_{ai}) \frac{|S_{ai}|}{|S_b|}$$

Choosing the Most Useful Test

- Estimate information gain for each test

Information Gain = $$\mathbf{H}(S_b) - \sum_i \mathbf{H}(S_{ai}) \frac{|S_{ai}|}{|S_b|}$$

- Choose the highest