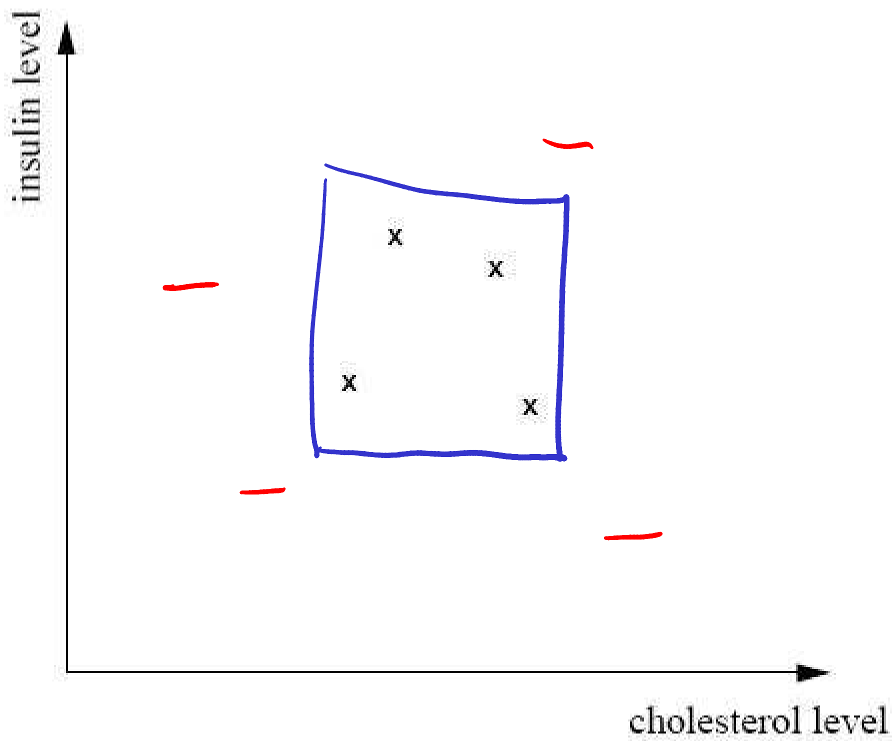
CS340: Bayesian concept learning

Kevin Murphy

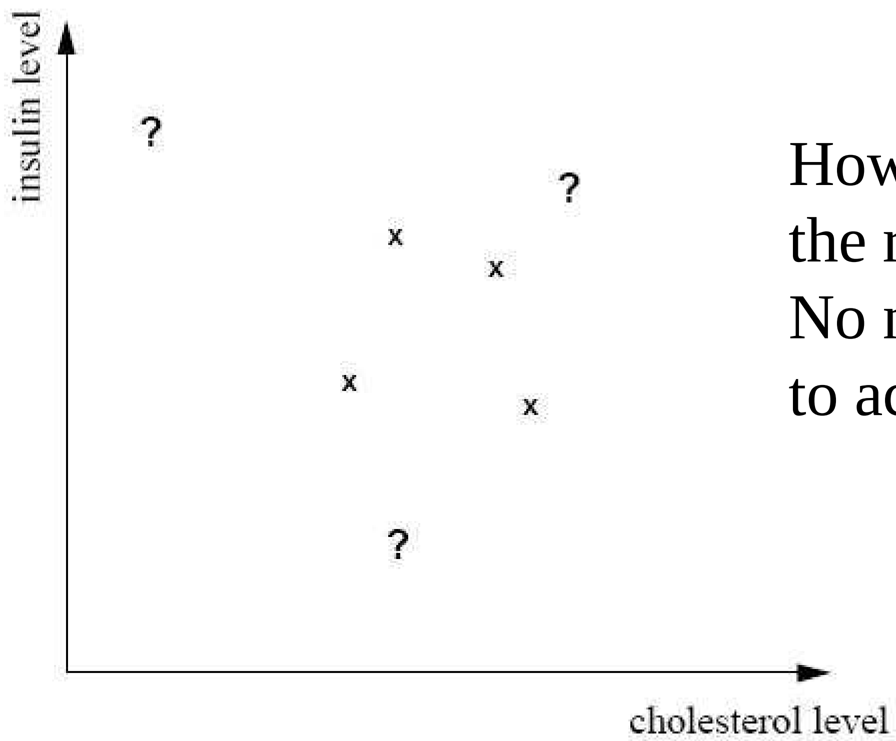
Based on Josh Tenenbaum's PhD thesis (MIT BCS 1999)

“Concept learning” (binary classification) from positive and negative examples



"healthy levels"

Concept learning from positive only examples

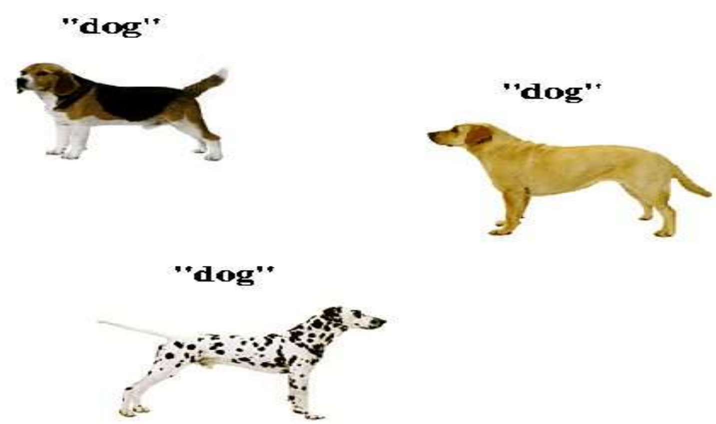


How far out should
the rectangle go?
No negative examples
to act as an upper bound.

"healthy levels"

Human learning vs machine learning/ statistics

- Most ML methods for learning "concepts" such as "dog" require a large number of positive and negative examples
- But people can learn from small numbers of positive only examples (look at the doggy!)
- This is called "one shot learning"



Everyday inductive leaps

How can we learn so much about . . .

- Meanings of words
- Properties of natural kinds
- Future outcomes of a dynamic process
- Hidden causal properties of an object
- Causes of a person’s action (beliefs, goals)
- Causal laws governing a domain

. . . from such limited data?

The Challenge

- How do we generalize successfully from very limited data?
  - Just one or a few examples
  - Often only positive examples
- Philosophy:
  - Induction called a “problem”, a “riddle”, a “paradox”, a “scandal”, or a “myth”.
- Machine learning and statistics:
  - Focus on generalization from many examples, both positive and negative.

The solution: Bayesian inference

- Bayes' rule: $P(H \mid D) = \dfrac{P(H)P(D \mid H)}{P(D)}$

- Various compelling (theoretical and experimental) arguments that one should represent one's beliefs using probability and update them using Bayes rule

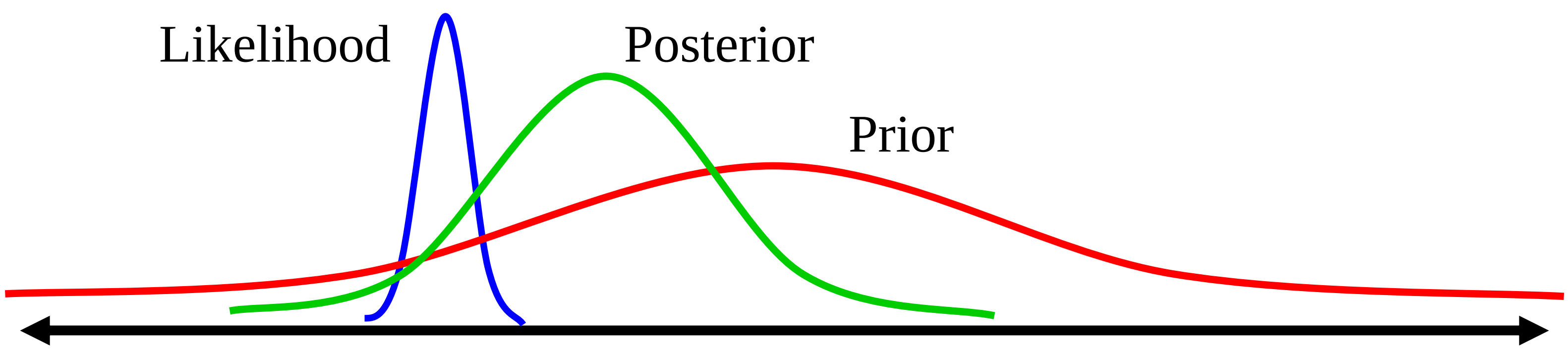
Bayesian belief updating

Posterior probability

Likelihood

Prior probability

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$



Bayesian inference = Inverse probability theory

Derivation of Bayes rule

- By the defn of conditional prob

$$p(A = a|B = b) = \frac{p(A = a, B = b)}{p(B = b)} \quad \text{if } p(B = b) > 0$$

- By chain rule

$$p(A = a, B = b) = p(B = b|A = a)p(A = a)$$

- By rule of total probability

$$p(B = b) = \sum_{a'} p(B = b, A = a')$$

- Hence we get Bayes' rule

$$p(A = a|B = b) = \frac{p(B = b|A = a)p(A = a)}{\sum_{a'} p(B = b|A = a)p(A = a)}$$

Bayesian inference: key ingredients

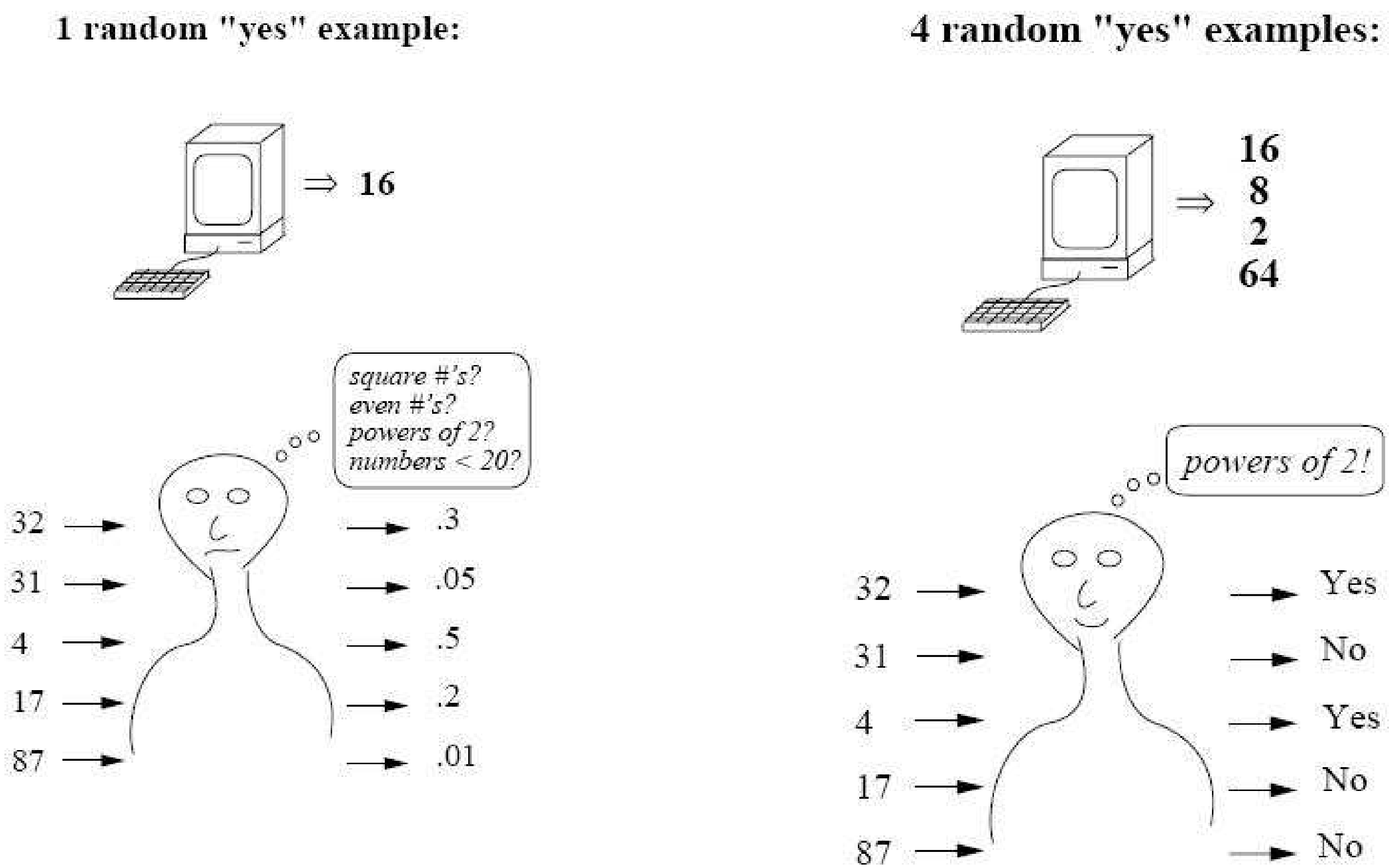
- Hypothesis space H
- Prior p(h)
- Likelihood p(D|h)
- Algorithm for computing posterior p(h|D)

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$

Two examples

- The “number game” – inferring abstract patterns from sequences of integers
- The “healthy levels game” – inferring rectangles from points in R^2

The number game



- Learning task:
  - Observe one or more examples (numbers)
  - Judge whether other numbers are “yes” or “no”.

The number game

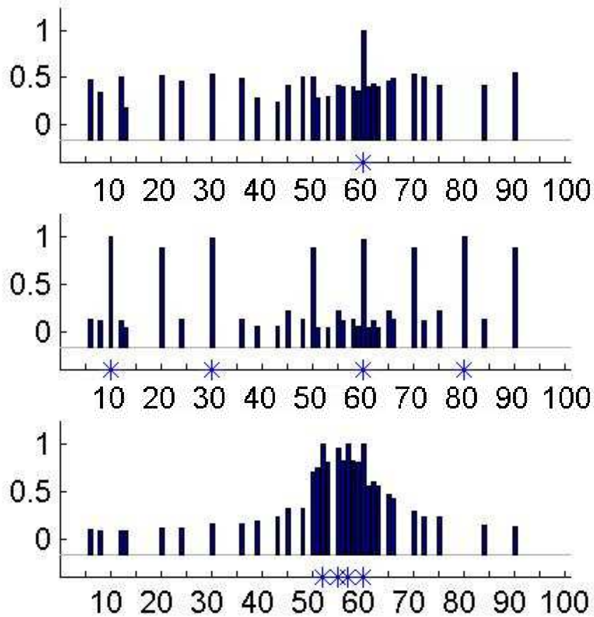
Examples of "yes" numbers	Hypotheses
60	multiples of 10 even numbers ? ? ?
60 80 10 30	multiples of 10 even numbers
60 63 56 59	numbers "near" 60

Human performance

60

60 80 10 30

60 52 57 55



Diffuse similarity

Rule:
"multiples of 10"

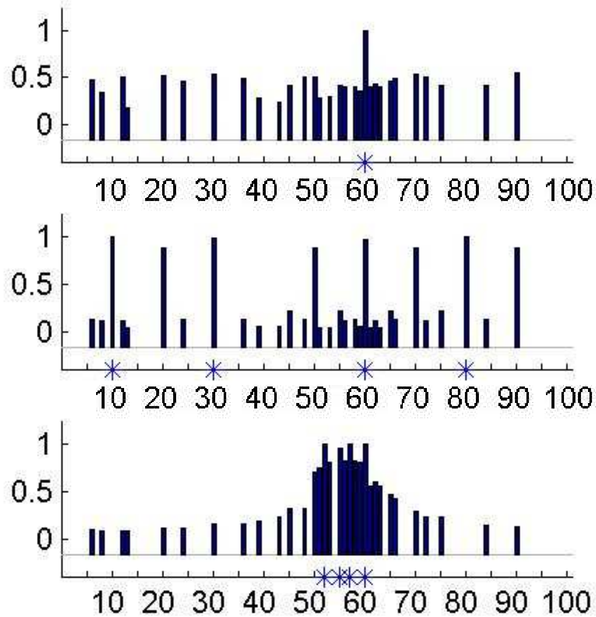
Focused similarity:
numbers near 50-60

Human performance

60

60 80 10 30

60 52 57 55



Diffuse similarity

Rule:
"multiples of 10"

Focused similarity:
numbers near 50-60

Some phenomena to explain:

- People can generalize from just positive examples.
- Generalization can appear either graded (uncertain) or all-or-none (confident).

Bayesian model

- H: Hypothesis space of possible concepts:
- $X = \{x_1, \ldots, x_n\}$: n examples of a concept C.
- Evaluate hypotheses given data using Bayes' rule:

$$p(h \mid X) = \frac{p(X \mid h)\, p(h)}{\sum_{h' \in H} p(X \mid h')\, p(h')}$$

  - $p(h)$ ["prior"]: domain knowledge, pre-existing biases
  - $p(X|h)$ ["likelihood"]: statistical information in examples.
  - $p(h|X)$ ["posterior"]: degree of belief that h is the true extension of C.

Hypothesis space

- Mathematical properties (~50):
  - odd, even, square, cube, prime, …
  - multiples of small integers
  - powers of small integers
  - same first (or last) digit
- Magnitude intervals (~5000):
  - all intervals of integers with endpoints between 1 and 100
- Hypothesis can be defined by its **extension**

$$h = \{x : h(x) = 1,\ x = 1, 2, \ldots, 100\}$$

Likelihood p(X|h)

- **Size principle**: Smaller hypotheses receive greater likelihood, and exponentially more so as *n* increases.

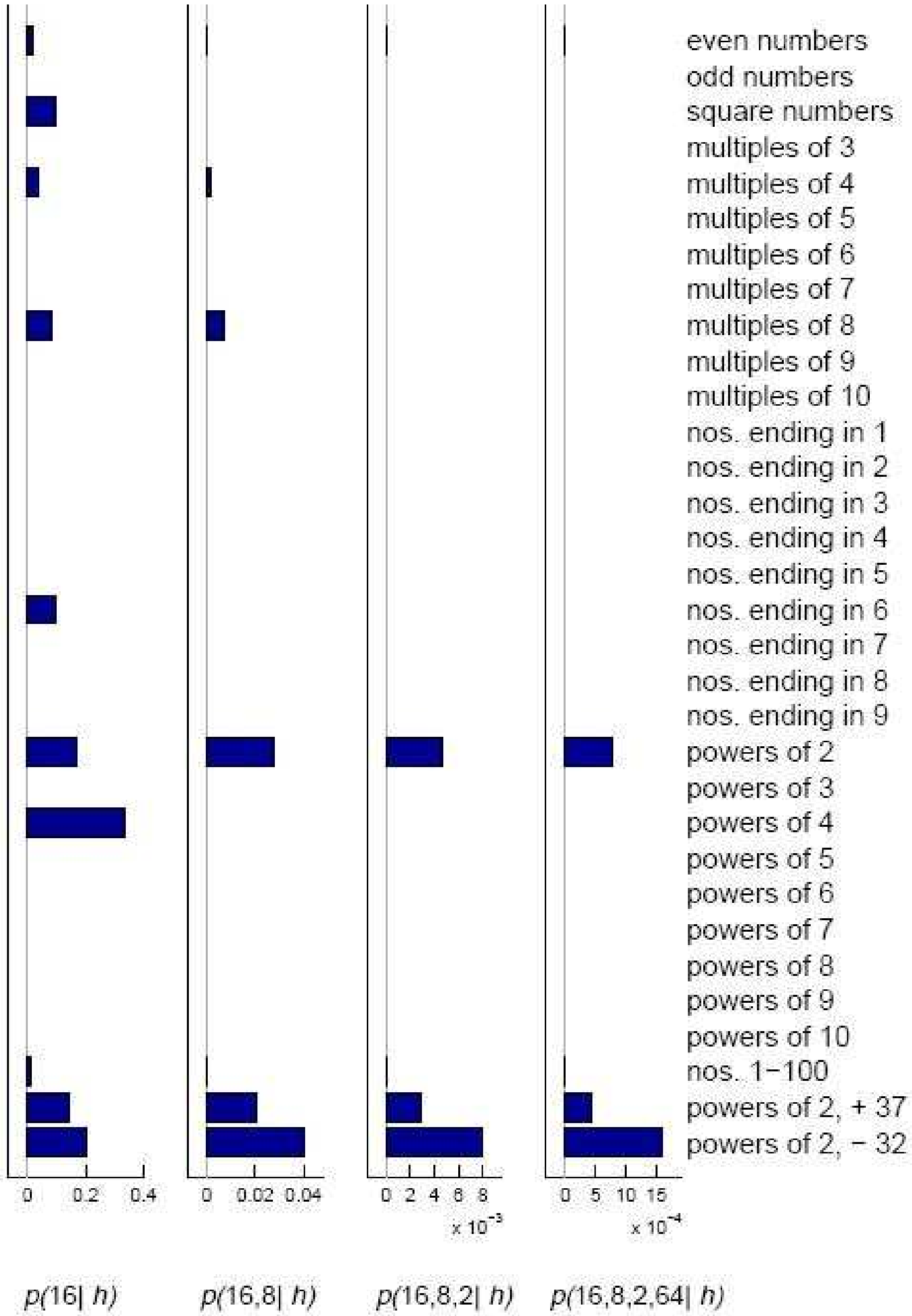
$$p(X \mid h) = \left[\frac{1}{\text{size}(h)} \right]^n \text{ if } x_1, \ldots, x_n \in h$$

$$= 0 \text{ if any } x_i \notin h$$

- Follows from assumption of randomly sampled examples (**strong sampling**).
- Captures the intuition of a representative sample.

Example of likelihood

- X={20,40,60}
- H1 = multiples of 10 = {10,20,…,100}
- H2 = even numbers = {2,4,…,100}
- H3 = odd numbers = {1,3,…,99}
- P(X|H1) = 1/10 * 1/10 * 1/10
- p(X|H2) = 1/50 * 1/50 * 1/50
- P(X|H3)   = 0


even numbers
odd numbers
square numbers
multiples of 3
multiples of 4
multiples of 5
multiples of 6
multiples of 7
multiples of 8
multiples of 9
multiples of 10
nos. ending in 1
nos. ending in 2
nos. ending in 3
nos. ending in 4
nos. ending in 5
nos. ending in 6
nos. ending in 7
nos. ending in 8
nos. ending in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
nos. 1−100
powers of 2, + 37
powers of 2, − 32
0 0.2 0.4
0 0.02 0.04
0 2 4 6 8
x 10^-3
0 5 10 15
x 10^-4
p(16| h)
p(16,8| h)
p(16,8,2| h)
p(16,8,2,64| h)

Likelihood function

- Since $p(\vec{x}|h)$ is a distribution over vectors of length n, we require that, for all h, $\sum_{\vec{x}} p(x|h) = 1$
- It is easy to see this is true, e.g., for h=even numbers, n=2

$$\sum_{x_1=1}^{100} \sum_{x_2=1}^{100} p(x_1, x_2|h) = \sum_{x_1=1}^{100} \sum_{x_2=1}^{100} p(x_1|h)p(x_2|h) = \sum_{x_1 \in even} \sum_{x_2 \in even} \frac{1}{50}\frac{1}{50} = 1$$

- If x is fixed, we do not require $\sum_h p(X|h) = 1$
- Hence we are free to multiply the likelihood by any constant independent of h

Illustrating the size principle

h_1

2	4	6	8	10
12	14	16	18	20
22	24	26	28	30
32	34	36	38	40
42	44	46	48	50
52	54	56	58	60
62	64	66	68	70
72	74	76	78	80
82	84	86	88	90
92	94	96	98	100

h_2

Illustrating the size principle

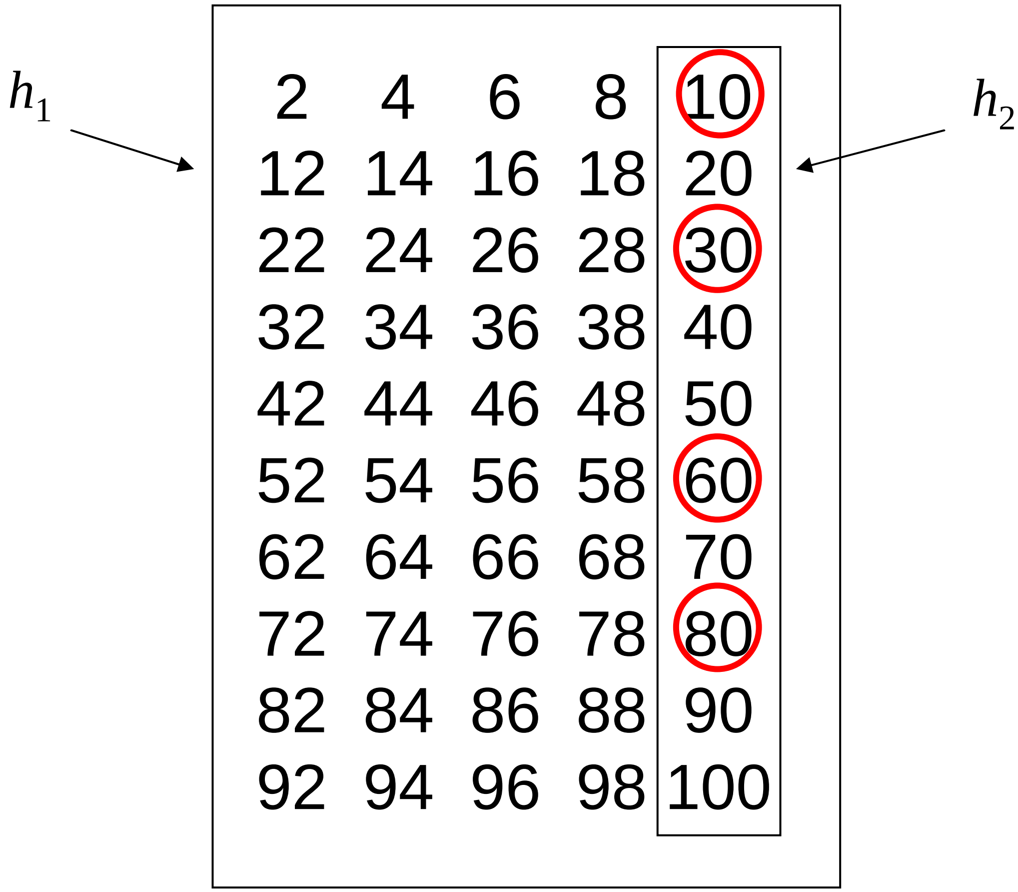
h_1

2	4	6	8	10
12	14	16	18	20
22	24	26	28	30
32	34	36	38	40
42	44	46	48	50
52	54	56	58	60
62	64	66	68	70
72	74	76	78	80
82	84	86	88	90
92	94	96	98	100

h_2

Data slightly more of a coincidence under h_1

Illustrating the size principle



Data *much* more of a coincidence under h_1