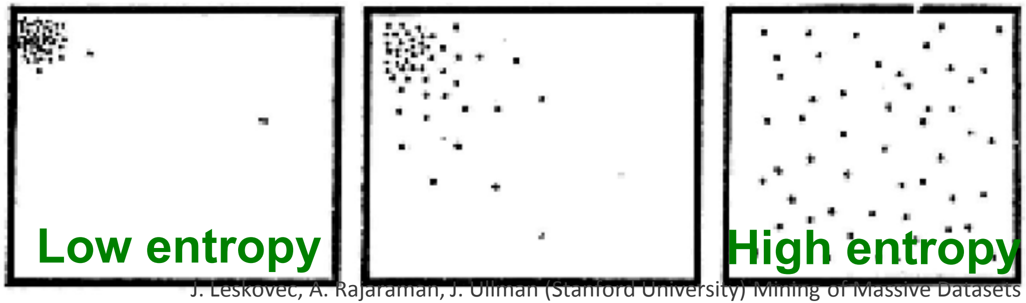
Why Information Gain? Entropy

- **Entropy:** What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from **X**'s distribution?
- **The entropy of X:** $H(X) = -\sum_{j=1}^{m} p_j \log p_j$
  - "High Entropy": **X** is from a uniform (boring) distribution
    - A histogram of the frequency distribution of values of **X** is **flat**
  - "Low Entropy": **X** is from varied (peaks and valleys) distribution
    - A histogram of the frequency distribution of values of **X** would have many lows and one or two highs



Why Information Gain? Entropy

- Suppose I want to predict **Y** and I have input **X**
  - **X** = College Major
  - **Y** = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Math	Yes
History	No

- **From this data we estimate**
  - $P(Y = Yes) = 0.5$
  - $P(X = Math \,\&\, Y = No) = 0.25$
  - $P(X = Math) = 0.5$
  - $P(Y = Yes \mid X = History) = 0$
- **Note:**
  - $H(Y) = -½*\log_2(½)-½*\log_2(½)$ **= 1**
  - $H(X) = 1.5$

Why Information Gain? Entropy

- Suppose I want to predict **Y** and I have input **X**
  - **X** = College Major
  - **Y** = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Math	Yes
History	No

- **Def: Specific Conditional Entropy**
  - **H(Y | X=v)** = The entropy of **Y** among only those records in which **X** has value **v**
  - **Example:**
    - H(Y|X=Math) = 1
    - H(Y|X=History) = 0
    - H(Y|X=CS) = 0

Why Information Gain?

- Suppose I want to predict **Y** and I have input **X**
  - **X** = College Major
  - **Y** = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Math	Yes
History	No

- **Def: Conditional Entropy**
  - **H(Y | X) =** The average specific conditional entropy of **Y**
    - **=** if you choose a record at random what will be the conditional entropy of **Y**, conditioned on that row's value of **X**
    - **=** Expected number of bits to transmit **Y** if both sides will know the value of **X**
  - **=** $\sum_j P(X = v) H(Y|X = v)$

Why Information Gain?

- Suppose I want to predict **Y** and I have input **X**

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Math	Yes
History	No

- **H(Y | X) =** The average specific conditional entropy of **Y**

$$= \sum_j P(X = v)H(Y|X = v)$$

- **Example:**

V_j	$P(X=v_j)$	$H(Y\|X=v_j)$
Math	0.5	1
History	0.25	0
CS	0.25	0

- **So:** H(Y|X)=0.5*1+0.25*0+0.25*0 = **0.5**

Why Information Gain?

- Suppose I want to predict **Y** and I have input **X**

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
Math	Yes
History	No

- **Def: Information Gain**
  - **IG(Y|X)** = I must transmit **Y**. **How many bits on average would it save me if both ends of the line knew X?**

$$IG(Y|X) = H(Y) - H(Y \mid X)$$

- **Example:**
  - H(Y) = 1
  - H(Y|X) = 0.5
  - Thus **IG(Y|X) = 1 – 0.5 = 0.5**

What is Information Gain used for?

- Suppose you are trying to predict whether someone is going live past 80 years
- From historical data you might find:
  - IG(LongLife | HairColor) = 0.01
  - IG(LongLife | Smoker) = 0.2
  - IG(LongLife | Gender) = 0.25
  - IG(LongLife | LastDigitOfSSN) = 0.00001
- **IG tells us how much information about Y is contained in X**
  - **So attribute X that has high IG(Y|X) is a good split!**