Information Theory and You

Putting the **fun** in fundamental limits

Overview

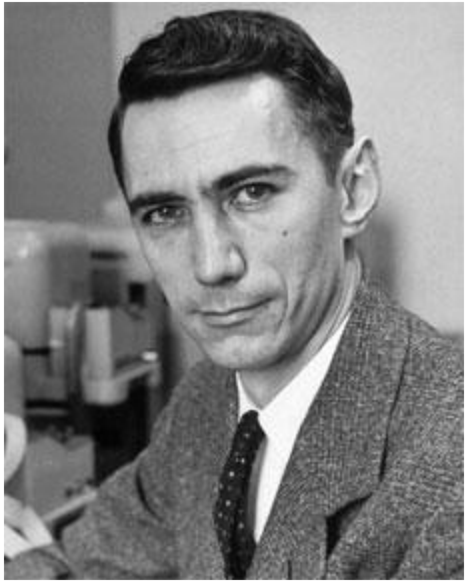
1. Takeaways
2. Claude Shannon
3. Entropy & Mutual Information
4. The Data Processing Inequality
5. Examples
6. Takeaways
7. Further Reading

Processing does not increase information, it makes existing information useful.

Good analysis can not make up for bad data.

Claude Shannon (1916 - 2001)

- UMichigan & MIT
- Key researcher at Bell Labs
- Made fundamental contributions to:
  - Digital circuit design
  - Telecommunications
  - Data compression
- Loved to juggle
- Could fly a plane
- "Father of Information Theory"



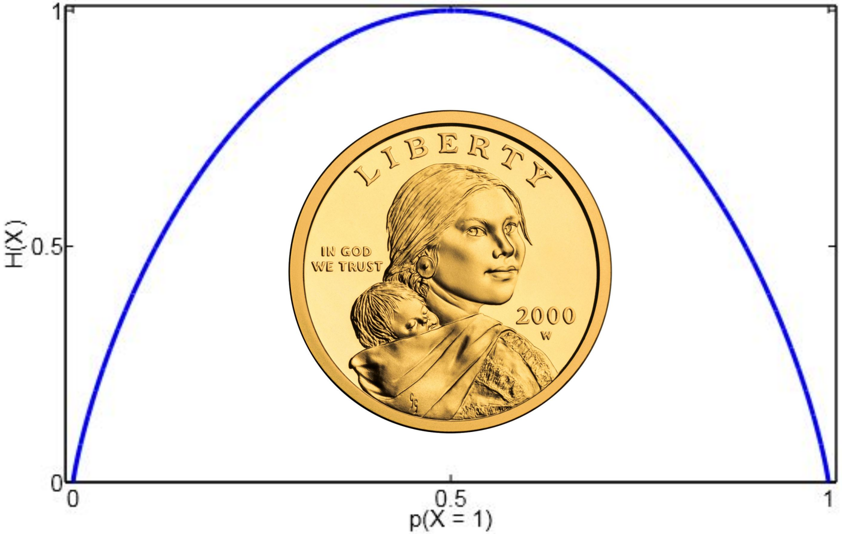
Entropy and Mutual Information

$$H(X) \triangleq -\sum_x p(x) log p(x)$$

"How uncertain am I about X?"

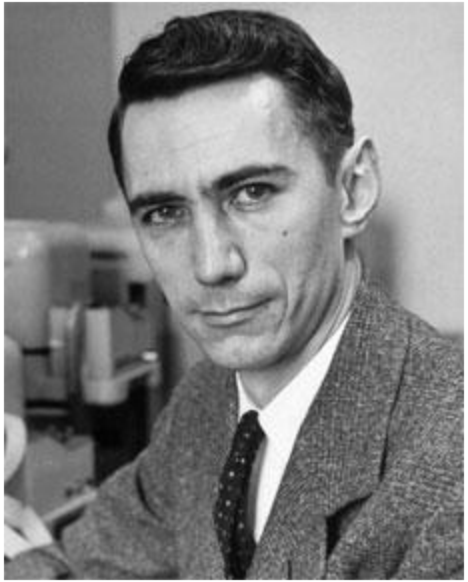
$$I(Y;X) \triangleq H(Y) - H(Y|X)$$

"How much does X tell me about Y?"



Claude Shannon (1916 - 2001)

My greatest concern was what to call it. I thought of calling it "information," but the word was overly used, so I decided to call it "uncertainty." When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it ***entropy****, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage."*



Data Processing Inequality

$$A \rightarrow B \rightarrow C$$

$$I(A;B) \geq I(A;C)$$

The Data Processing Inequality

world → *data* → *analysis*

↑	↑
measurement (basic science)	processing (statistics/ML)
challenges: representation noise	challenges: complexity expressiveness

Example: Averages

$$\Omega \rightarrow [x_1, x_2, x_3, x_4] \rightarrow \bar{x}$$

- O(1) space complexity
- O(n) time complexity
- Discards variance information

Example: Noise

Why care about noise?

$$x = \Omega + z$$

data = truth + noise

$$x - z = \Omega$$

data - noise = truth

Example: Noise

$$
\begin{tikzcd}
 & x \arrow[dr] & \\
\Omega \arrow[ur] \arrow[dr] & & f(x,z) \\
 & z \arrow[ur] &
\end{tikzcd}
$$

(research)

$$I(\Omega; f(x,z)) \geq I(\Omega; f(x))$$

Example: Foursquare

$$
\begin{tikzcd}
 & location \arrow[dr] & \\
\Omega \arrow[ur] \arrow[dr] & & pilgrim(location, checkins) \\
 & checkins \arrow[ur] &
\end{tikzcd}
$$

$$I(\Omega; pilgrim(location, checkins)) \geq I(\Omega; competitor(location))$$

Example: Google & TensorFlow

$$\Omega \begin{array}{l} \nearrow \quad DATA \rightarrow tf_{CPU}(DATA) \\ \searrow \quad data \rightarrow tf_{cpu}(data) \end{array}$$

$$I(\Omega; tf_{CPU}(DATA)) \geq I(\Omega; tf_{cpu}(data))$$

Example: Fitting Functions

$$\Omega \rightarrow x \begin{array}{l} \nearrow \; y = mx_* + b \\ \searrow \; y \sim N(\mu_*, \sigma_*^2) \end{array}$$

Linear regression - $O(d^3)$

Gaussian process - $O(n^3)$

Gaussian processes capture more information to create a better model, but you pay a computational cost when n >> d (i.e. almost always). When is this *useful?*

Example: Biotech

$$\Omega \rightarrow x_{bad} \rightarrow f_{good}(x_{bad})$$

$$I(\Omega; f_{ok}(x_{good})) \geq I(\Omega; f_{good}(x_{bad}))$$

Theranos???????

Processing does not increase information, it makes existing information useful.

Good analysis can not make up for bad data.

Claude Shannon
was a beast.

Further Reading