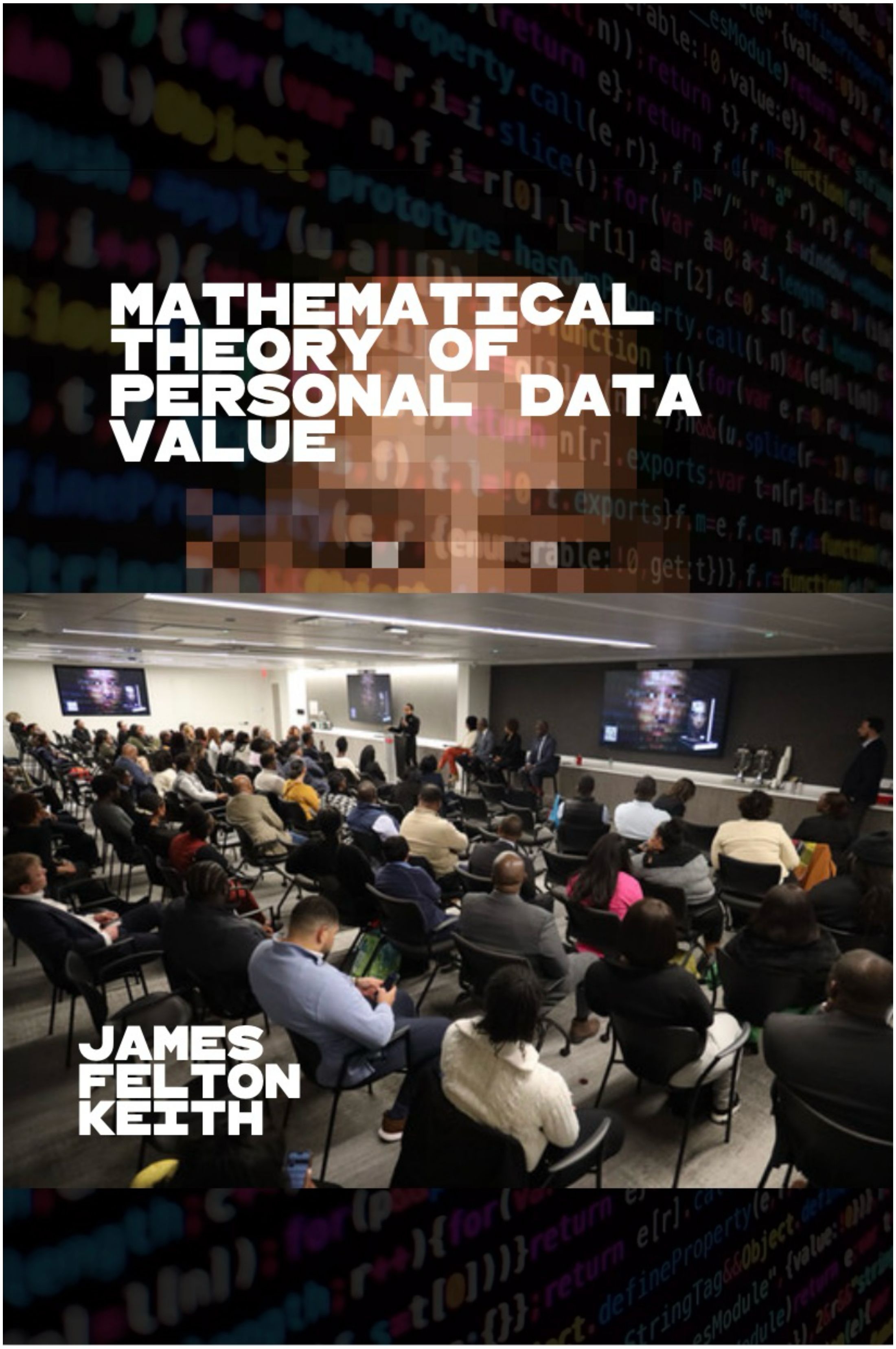
JFK KEITH INSTITUTE JAMES FELTON KEITH



MATHEMATICAL THEORY OF PERSONAL DATA VALUE

JAMES FELTON KEITH

Mathematical Theory of Personal Data Value: calculating the value of an individual's personal data input to the product development process

Crafting the data model

I’m mapping out a model to value personal data in product development, featuring Claude Shannon's 1948 Mathematical Theory of Communication. The focus is on information entropy and mutual information to gauge data's impact on product outcomes and consumer preferences.

Mapping the model

I'm integrating mutual information, entropy, data acquisition costs, and channel characteristics to evaluate how personal data influences product development outcomes, aiming to reduce uncertainty and boost product value.

Assessing data utility

I'm emphasizing the data's rare or unexpected quality, using *negative log probability* for information content, and considering relevance weighting to evaluate its impact on product development outcomes.

Synthesizing the data model

I'm mapping out the model by combining relevance weighting with mutual information, integrating over the data distribution, and assessing the channel's noise impact to determine the effective value of personal data.

Below is one way to develop a model that “prices” an individual’s personal data in the product development process using ideas from Claude Shannon’s Mathematical Theory of Communication. The approach is to treat personal data as a source of information that, when “transmitted” to the product development team, reduces uncertainty about customer needs or preferences. In other words, the “value” of the data is related to how much it improves decision-making—quantified as a reduction in uncertainty. One natural way to measure this is via **mutual information**.

In the following sections we describe the model step by step.

1. Setting Up the Framework

1.1. Define the Random Variables

- **XXX:** Represents the set of personal data collected from an individual. In practice, XXX might be multi-dimensional (e.g., demographics, behavior, preferences).
- **YYY:** Represents the product development outcomes that we care about. For instance, YYY might denote “customer satisfaction,” “purchase likelihood,” or any measure of how well the product aligns with customer needs.

1.2. The Communication Analogy

- **Source:** The individual’s personal data XXX acts like a source generating a message.
- **Channel:** The product development process is analogous to a communication channel that “receives” this information and uses it to reduce uncertainty about YYY.
- **Receiver/Decoder:** The decision-making process in product development that interprets the data and adapts the product accordingly.

2. Using Shannon’s Information Measures

2.1. Information Content

In Shannon’s theory, the information content (or “surprisal”) of a particular outcome xxx is given by

I(x)=−log 2P(x)I(x) = -\log_2 P(x)I(x)=−log2 P(x)

This means that rare (unexpected) pieces of data carry more “bits” of information.

2.2. Entropy of the Outcome

The uncertainty (entropy) about the product development outcome YYY before receiving any data is

H(Y)=−∑yP(y)log 2P(y)H(Y) = - \sum_{y} P(y) \log_2 P(y)H(Y)=−y∑ P(y)log2 P(y)

This represents our baseline uncertainty about YYY.

2.3. Conditional Entropy

After obtaining personal data XXX, the remaining uncertainty about YYY is

H(Y∣X)=−∑x,yP(x,y)log 2P(y|x)H(Y|X) = - \sum_{x,y} P(x,y) \log_2
P(y|x)H(Y∣X)=−x,y∑ P(x,y)log2 P(y∣x)

This quantifies the uncertainty in YYY after the personal data has been “transmitted” through the product development process.

2.4. Mutual Information

The key quantity is the **mutual information** between XXX and YYY:

I(X;Y)=H(Y)−H(Y∣X)I(X; Y) = H(Y) - H(Y|X)I(X;Y)=H(Y)−H(Y∣X)

This mutual information represents the **average reduction in uncertainty** about YYY when the personal data XXX is known. In our model, it is interpreted as the **value** of the personal data in reducing uncertainty in product development outcomes.

3. Modeling the Value of Personal Data

3.1. Basic Model

We define the value VVV of an individual’s personal data as:

V=I(X;Y)=H(Y)−H(Y∣X)V = I(X; Y) = H(Y) - H(Y|X)V=I(X;Y)=H(Y)−H(Y∣X)

- **H(Y)H(Y)H(Y):** Total uncertainty in outcomes without data.
- **H(Y∣X)H(Y|X)H(Y∣X):** Uncertainty remaining after the data is used.

A higher I(X;Y)I(X; Y)I(X;Y) means that the data is more informative (i.e., it reduces more uncertainty), which in turn implies higher value in shaping product development.

3.2. Incorporating Data Relevance

Not all personal data is equally useful. We can introduce a **relevance weighting factor** r(x)r(x)r(x) that accounts for how pertinent a particular piece of data xxx is to the product development process. One might modify the contribution of each data point as:

V(x)=r(x)×[−log 2P(x)]V(x) = r(x) \times [-\log_2 P(x)]V(x)=r(x)×[−log2 P(x)]

Then, the overall value could be aggregated (for example, by expectation):

V=∑xP(x) r(x) (−log 2P(x))V = \sum_{x} P(x) \, r(x) \, \bigl(-\log_2 P(x)\bigr)V=x∑ P(x)r(x)(−log2 P(x))

Alternatively, one could define a **weighted mutual information** measure, where the joint distribution P(x,y)P(x,y)P(x,y) is adjusted to reflect relevance.

3.3. Considering the "Channel"

In Shannon's theory, noise in the channel can degrade the message. Analogously, if the product development process (the "channel") misinterprets the personal data or if there is additional "noise" (e.g., data errors, misalignment with the product strategy), then the effective value of the data is lower. You could model this by introducing a fidelity factor η\etaη (with 0≤η≤10 \leq \eta \leq 10≤η≤1) so that:

V=η I(X;Y)V = \eta \, I(X; Y)V=ηI(X;Y)

Here, η\etaη represents the effectiveness with which the organization translates personal data into actionable insights. If the process is perfect (η=1 \eta = 1η=1), the full mutual information is realized; if not, only a fraction is used.

---

4. Example Application

Suppose a company is deciding between two product features, and the outcome YYY is binary (e.g., feature A leads to high customer satisfaction vs. low satisfaction). The baseline entropy is:

H(Y)=−[P(Y=1)log 2P(Y=1)+P(Y=0)log 2P(Y=0)]H(Y) = -\bigl[P(Y=1)\log_2 P(Y=1) + P(Y=0)\log_2 P(Y=0)\bigr]H(Y)=−[P(Y=1)log2 P(Y=1)+P(Y=0)log2 P(Y=0)]

If the personal data XXX (say, detailed customer behavior data) reduces the uncertainty to H(Y∣X)H(Y|X)H(Y∣X), then the value of the data is:

V=I(X;Y)=H(Y)−H(Y∣X)V = I(X; Y) = H(Y) - H(Y|X)V=I(X;Y)=H(Y)−H(Y∣X)

A higher value of VVV indicates that personal data significantly reduces uncertainty about customer satisfaction, guiding the company in making a better product decision.

---

5. Summary of the Model

1. **Identify Variables:**
    - XXX: Personal data.
    - YYY: Product development outcome.
2. **Quantify Uncertainty:**
    - H(Y)H(Y)H(Y): Entropy before receiving data.
    - H(Y∣X)H(Y|X)H(Y∣X): Conditional entropy after receiving data.
3. **Compute Value via Mutual Information:**
    - V=I(X;Y)=H(Y)−H(Y∣X)V = I(X; Y) = H(Y) - H(Y|X)V=I(X;Y)=H(Y)−H(Y∣X).
4. **Adjust for Relevance and Noise:**
    - Optionally include relevance r(x)r(x)r(x) and channel fidelity η\etaη.

This model frames the value of an individual's personal data in product development as the **amount of uncertainty reduced** by having that data—a perspective directly inspired by Shannon's theory of communication.

6. Final Thoughts

By grounding the model in Shannon's information measures, we can quantitatively assess how much an individual's personal data contributes to better decision-making in product development. This approach helps in both understanding the value of data and in guiding investments in data collection and processing infrastructure.