# The Event Space Isomorphism: Arch-Native and Human-Native Partitions of the Byte Alphabet

Claude and MJC

February 11, 2026

**Abstract**

We show that the byte partition discovered by SVD of skip-bigram tables (the *arch-native* event space) and the partition defined by human knowledge of byte semantics (the *human-native* event space) are related by a permutation that approximately preserves the inner product structure of the reduced SVD subspace. On the output side (right singular vectors), the optimal permutation achieves up to 85.7% frequency-weighted accuracy and 0.661 NMI. On the input side, the arch-native partition is a *refinement* of the human-native partition: it discovers sub-category structure (vowel vs. consonant, mid-word vs. boundary) within the dominant "lowercase" class that the human labeling does not distinguish. This asymmetry is the E→N map made concrete: the arch-native events *are* the numbers, and the human labels are a coarsening.

## 1 Two Partitions of $\{0, \ldots, 255\}$

### 1.1 Arch-native: SVD sign bits

Given a skip-bigram count table $C_{xy}^{(g)}$ at offset $g$, we form the centered conditional matrix $A_{xo} = P(o|x) - P(o)$ and compute its SVD. The top 3 left singular vectors $u_0, u_1, u_2 \in \mathbb{R}^{256}$ define a partition of input bytes into $2^3 = 8$ groups by the sign pattern $(\operatorname{sgn} u_0(b), \operatorname{sgn} u_1(b), \operatorname{sgn} u_2(b))$. Similarly, the right singular vectors $v_0, v_1, v_2$ partition *output* bytes into 8 groups. These are the **arch-native** partitions: they are what the data's statistical structure discovers without human labels.

### 1.2 Human-native: semantic byte classes

We define 8 groups from common knowledge of the ASCII/byte encoding:

| Group | Name | #Bytes | Data fraction |
|---|---|---|---|
| H0 | lowercase (a–z) | 26 | 67.3% |
| H1 | uppercase (A–Z) | 26 | 3.5% |
| H2 | digits (0–9) | 10 | 2.3% |
| H3 | whitespace (space, \n, tab) | 3 | 15.5% |
| H4 | XML markup (<, >, /, &) | 4 | 2.7% |
| H5 | punctuation (., ; : ! ? ’ ” - etc.) | 13 | 6.9% |
| H6 | other printable (= + * — # etc.) | 15 | 1.3% |
| H7 | non-printable (control, high bytes) | 159 | 0.5% |

These labels require zero computation—they come from shared human background knowledge of how text works.

## 2 The Permutation

For each offset $g \in \{0, \ldots, 15\}$, we seek a permutation $\pi_g : \{0, \ldots, 7\} \to \{0, \ldots, 7\}$ that maximizes the frequency-weighted accuracy

$$\mathrm{Acc}(\pi) = \sum_{i=0}^{7} \sum_{b:\, \mathrm{SVD}(b)=i,\, \mathrm{Human}(b)=\pi(i)} \frac{f_b}{N}$$

where $f_b$ is the count of byte $b$ in the data. Since $8! = 40{,}320$, we find the global optimum by brute force.

We also compute the **Normalized Mutual Information** (NMI) between the two partitions (frequency-weighted), which is permutation-invariant and measures how much knowing one partition tells you about the other, and the **centroid cosine similarity**: the mean cosine between the frequency-weighted centroid of each SVD group and its matched human group, projected into the 3-dimensional SVD subspace.

## 3 Results

### 3.1 Output side (V): strong isomorphism

| Offset | Acc% | NMI | Cos | Notable mapping |
|---|---|---|---|---|
| 0 | 60.3 | 0.604 | 0.225 | lower↔SVD3, ws↔SVD5 |
| 1 | 73.9 | 0.639 | 0.553 | lower↔SVD7, ws↔SVD1 |
| 3 | 75.2 | 0.547 | 0.339 | lower↔SVD1, ws↔SVD7 |
| 4 | 68.9 | 0.532 | 0.395 | lower↔SVD5, ws↔SVD7 |
| 9 | **82.3** | 0.607 | 0.716 | lower↔SVD7, ws↔SVD3 |
| 10 | 70.9 | 0.590 | 0.407 | lower↔SVD1, punct↔SVD0 |
| 11 | **85.7** | **0.661** | 0.542 | lower↔SVD7, ws↔SVD1 |
| 13 | 68.1 | 0.600 | 0.801 | lower↔SVD7, ws↔SVD1 |
| **Mean (all 16)** | **62.2** | **0.524** | **0.504** | |

At offset 11, 85.7% of all byte occurrences are correctly classified under the optimal permutation. The NMI of 0.661 means the two partitions share about two-thirds of their information content.

### 3.2 Input side (U): refinement

| Offset | Acc% | NMI | Cos | Key structure |
|---|---|---|---|---|
| 0 | 40.7 | 0.359 | 0.371 | SV1 splits lower into word-interior vs. boundary |
| 1 | 47.9 | 0.375 | 0.395 | 5 SVD groups contain lowercase subsets |
| 4 | **55.1** | 0.387 | **0.884** | Best centroid alignment; ws perfectly isolated |
| 8 | 46.6 | 0.323 | 0.350 | ws perfectly isolated (SVD7 = 14.2%) |
| 11 | 46.4 | 0.428 | 0.838 | digits isolated; ws isolated |
| 12 | 50.4 | 0.387 | 0.659 | lowercase split: 30% + 17% in two groups |
| **Mean (all 16)** | **44.3** | **0.344** | **0.514** | |

Input-side accuracy is systematically lower (44% vs. 62%). The reason is clear from the confusion matrices: **lowercase letters (67% of data) are split across 3–5 SVD groups**, each capturing a different predictive feature. For instance, at offset 0:

- SVD group 7: 21.0% lowercase (the "text continuation" group)
- SVD group 2: 13.8% lowercase (the "XML-adjacent text" group)
- SVD group 5: 13.8% lowercase (the "word-internal" group)

Meanwhile, the smaller human classes map cleanly:

- **Whitespace** is perfectly isolated into a single SVD group at 11 of 16 offsets (the group contains $\geq$14.2% ws and $<$1% non-ws).
- **Digits** are isolated at offsets 0, 1, 9, 11, 14, 15.
- **Non-printable** bytes always map to a single group (0.5% of data).

## 4 The Asymmetry Is the E→N Map

The key observation: the input-side arch-native partition is a *refinement* of the human-native partition. Every SVD group is approximately a subset of one human class. The human partition is obtained by *merging* SVD groups:

$$\underbrace{\mathrm{SVD}_2 \cup \mathrm{SVD}_5 \cup \mathrm{SVD}_7}_{\text{“lowercase”}} \quad \underbrace{\mathrm{SVD}_6}_{\text{“uppercase”}} \quad \underbrace{\mathrm{SVD}_4 \cap \mathrm{digits}}_{\text{“digits”}} \quad \underbrace{\mathrm{SVD}_4 \cap \mathrm{ws}}_{\text{“whitespace”}}$$

This is exactly the **E→N** map from our framework. The arch-native events (E) are the fine-grained statistical reality. The human-native labels (N) are a numbering system—a coarsening that assigns names to clusters of events. The permutation $\pi$ is not a bijection between equal partitions; it is the best *alignment* of a fine partition with a coarse one.

On the output side, the asymmetry reverses: the human-native partition is closer to the arch-native one because output bytes are classified by "what predicts me" (the distributional context),

which closely tracks human categories. Space is predicted after word-ending letters; digits after other digits or colons; uppercase after periods and newlines. The *distributional definition* of a category matches the *semantic definition*.

## 5 Inner Product Preservation

In the 3-dimensional SVD subspace, each byte $b$ has coordinates $(u_0(b), u_1(b), u_2(b))$. The frequency-weighted centroid of each group defines a point in this space. If the isomorphism preserves inner product structure, then the centroid of SVD group $i$ should be close to the centroid of human group $\pi(i)$.

We measure this by the cosine similarity between paired centroids. On the input side, the mean cosine across all offsets is **0.514** (range: 0.195–0.884). On the output side, the mean is **0.504** (range: 0.113–0.863).

The best cases (cosine $> 0.8$) occur at offsets 4, 11, 12 (input) and offsets 7, 9, 13 (output). These are the offsets where the SVD's statistical partition most closely aligns with human semantic intuition.

The moderate mean cosine ($\approx 0.5$) reflects the refinement structure: SVD centroids for lowercase subgroups point in *different* directions within the lowercase region, while the human "lowercase" centroid averages them. The alignment is partial because the SVD resolves structure that the human labeling collapses.

## 6 Pair Agreement

For all pairs of active bytes, we ask: do the two partitions agree on whether they belong to the same group?

| | Input (U) | Output (V) |
|---|---|---|
| Same-group agreement | 44.3% | — |
| Different-group agreement | 76.4% | — |

The asymmetry is diagnostic: the partitions rarely *disagree* about bytes being in different groups (76% agreement), but the arch-native partition splits groups that the human partition merges (only 44% same-group agreement). This confirms the refinement interpretation.

## 7 Conclusion

The arch-native (SVD) and human-native (semantic) event spaces are not identical partitions, but they are related by a permutation that preserves the essential structure. The V-side (output) isomorphism is strong: 85.7% accuracy at best, 62% mean. The U-side (input) is a refinement: the SVD discovers sub-category structure within the human "lowercase" class that reflects genuine predictive distinctions (vowel/consonant, word position, XML context).

This is the E→N map made explicit. The events are the arch-native statistical reality; the numbers are the human labels we assign to make them intelligible. The permutation $\pi$ is not arbitrary—it consistently maps whitespace to whitespace, digits to digits, XML to XML. The fact that it also discovers that "lowercase" is not one event but several is not a failure of the isomorphism; it is the isomorphism *doing its job*, revealing structure that human labeling merely approximates.

**Data:** 262,144 bytes of enwik9. 16 offsets. SVD: 300 power iterations, top 8 components. Permutation: brute-force 8!. Tool: `es_iso.c`.