Representations
Representations

What?
• Given data (in certain representation), produce a representation which provides a better understanding of the data

Why?
• Several ML models require data in a specific representation to work well
  Usually $\mathbb{R}^d$, sometimes as a similarity function,
  occasionally graphs, rarely as curved spaces

• Enhance the signal in data
  Discover underlying structure, suppress noise

• Improve computational efficiency and decrease space usage
  Dimensionality reduction, can use simpler models
Dimension Reduction: A Successful Example
Any kind of data processing results in information loss.

**Theorem (Data Processing Inequality):** Suppose $X \rightarrow Y \rightarrow Z$, then $I(X;Y) \geq I(X;Z)$

No clever manipulation of the data (deterministic or randomized) can improve inference or provide more information about the underlying process $X$ than $Y$ itself.
Data Processing Inequality: Suppose $X \rightarrow Y \rightarrow Z$, then $I(X;Y) \geq I(X;Z)$

Proof: Consider $I(X;(Y,Z)) = H(X) - H(X|YZ)$

$$= H(X) - H(X|Z) + H(X|Z) - H(X|YZ)$$

$$= I(X;Z) + I(X;Y|Z) \geq 0$$

$$= I(X;Y) + I(X;Z|Y) = 0$$

[b/c of the Markovian property $X \perp Z | Y$]

The theorem follows.
Data Processing Inequality: If $X \rightarrow Y \rightarrow Z$, then $I(X;Y) \geq I(X;Z)$

This seems like bad news:
Any processing/re-representation of data can only result in information loss about the underlying process.

Catch:
If we are smart about our processing, we can ensure that we retain important aspects of data that are useful for our understanding of the underlying process, and lose all the frivolous/uninteresting information.

Example:
Suppose we want a representation for effective nearest neighbors, then we only need to retain ordinal information ($a$ is closer to $b$ than $c$)
Metric Embeddings
A Motivating Example

Given a data in a ‘dissimilarity between objects’ form

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$x_2$</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$x_3$</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>$x_4$</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

How can we come up a vectoral representation, which respects the relations?
- To gain better understanding of the relationships between data
- If we embed the data in $(\mathbb{R}^d, L_2)$ we can apply off-the-shelf models.
Metric Embeddings

**Goal:** Given a metric space \((X, \rho)\) want to embed it in a normed space \((\mathbb{R}^d, L_p)\).

Measuring the quality of an embedding:

Given two metric spaces \((X, \rho)\) and \((Y, \sigma)\). A mapping \(f : X \rightarrow Y\) is called a **D-embedding** of \(X\) into \(Y\) (for \(D \geq 1\)) if there exists an \(r > 0\) s.t. for all \(x, x' \in X\),

\[
 r \cdot \rho(x, x') \leq \sigma(f(x), f(x')) \leq D \cdot r \cdot \rho(x, x')
\]

- \(D\) is called the **distortion** of the embedding \(f\)
- If \(D = 1\), then \(f\) is **distance-preserving** and thus called **isometric** (typically \(r=1\))
- If \(D \geq 1\), and \(r \leq (1/D)\), then \(f\) is a **contraction**

Why normed spaces?
- easier to deal with
- we have a better understanding
Embeddings into $L_\infty$

**Theorem (Fréchet):** An $n$-point metric space $(X, \rho)$ can be isometrically embedded into $L^n_\infty$

**Proof:**
Consider the mapping 

$$f(x) = \begin{bmatrix} \rho(x, x_1) \\ \rho(x, x_2) \\ \vdots \\ \rho(x, x_n) \end{bmatrix}$$

**Observation:**
- $f$ is a contraction, ie $\forall u, v \in X$, $\|f(u) - f(v)\|_{L^n_\infty} \leq \rho(u, v)$
  - Why? By triangle inequality $\forall u, v, x_i \in X$, $\rho(u, x_i) - \rho(v, x_i) \leq \rho(u, v)$ in particular, $
  \max_i |\rho(u, x_i) - \rho(v, x_i)| \leq \rho(u, v)$
  - $\|f(u) - f(v)\|_{L^n_\infty} \leq \rho(u, v)$

- $\forall u, v \in X$, $\exists i$ s.t. $\rho(u, v) = (f(u) - f(v))_i$
  - Why? 
  For row $i$ corresponding to $v$
  $$(f(u) - f(v))_i = \rho(u, v)$$
Fréchet Embedding

**Theorem (Fréchet):** An $n$-point metric space $(X, \rho)$ can be **isometrically** embedded into $L^n_{\infty}$

**Proof:**
Consider the mapping

$$f(x) = \begin{bmatrix} \rho(x, x_1) \\ \rho(x, x_2) \\ \vdots \\ \rho(x, x_n) \end{bmatrix}$$

**Observation:**

- $\forall u, v \in X, \quad \| f(u) - f(v) \|_{L^n_{\infty}} \leq \rho(u, v)$
- $\forall u, v \in X, \exists i$ s.t. $\rho(u, v) = (f(u) - f(v))_i$

$$\rho(u, v) = (f(u) - f(v))_i \leq |f(u) - f(v)|_i \leq \max_i |f(u) - f(v)| = \| f(u) - f(v) \|_{L^n_{\infty}} \leq \rho(u, v)$$

$L_{\infty}$ is a universal space!
Good news: $L_\infty$ is a universal space... for finite metric spaces

Some issues:

• The target dimension is huge ($d = n$). Can it be reduced? well... we can drop it down to $n – 1$
  (second observation can be refined by one coordinate)

How a significant improvement?
Incompressibility result

**Theorem (Incompressibility of general metric spaces):**
If $Z$ is a normed space that $D$-embeds all $n$ points metric space, then

- $\dim(Z) = \Omega(n)$ for $D < 3$
- $\dim(Z) = \Omega(n^{1/2})$ for $D < 5$
- $\dim(Z) = \Omega(n^{1/3})$ for $D < 7$

More compression requires bigger distortion