# Automatic Discovery and Generation of Visual Design Characteristics

## Application to Visual Conjoint Analysis

Ankit Sisodia, Alex Burnap and Vineet Kumar
*Yale School of Management*

University of Illinois at Urbana Champaign

May 2023

# To Do

- Replace Tables that do not fit by figures (pictures of the Table)
- Summarize key point method in 2 slides
- Elie Feit method in a slide
- Play up the downsides of ground truth.
- Human Faces disentanglement
- Evidence on characteristics thinking versus Gestalt thinking - studies, quotes etc. any evidence.
- VAEs versus GANs for disentanglement – add a couple of slides
- (Optional) Can we show entanglement versus disentanglement from ML literature using DSprites etc.?
- (V) What are the boundary conditions? When would this method not work?
- Why would using product characteristics work?
- Why does supervision help?

# VAE v GAN

| # | Topic | VAE | GAN | Source |
|---|---|---|---|---|
| 1 | Disentanglement Performance | High | Low | [Lee et al., 2020] |
| 2 | Quality of generated image | Low | High | [Lee et al., 2020] |
| 3 | Training instability | Low | High | [Lee et al., 2020] |
| 4 | Local v Global Concepts | Global | Local | [Gabbay, Cohen, and Hoshen, 2021] |
| 5 | Data requirement | Low | High | [Karras et al., 2020] |
| 6 | Ability to work on small or detailed objects | No | Yes | [Locatello et al., 2020] |

- **[1,2,3]** According to Lee et al. [2020]: "VAE-based approaches are effective in learning useful disentangled representations in various tasks, but their generation quality is generally worse than the state-of-the-arts, which limits its applicability to the task of realistic synthesis. On the other hand, GAN based approaches can achieve the high-quality synthesis with a more expressive decoder and without explicit likelihood estimation. However, they tend to learn comparably more entangled representations than the VAE counterparts and are notoriously difficult to train, even with recent techniques to stabilize the training."
- **[4]:** According to Gabbay, Cohen, and Hoshen [2021]: "Such methods that rely on a pretrained unconditional StyleGAN generator are mostly successful in manipulating highly-localized visual concepts (e.g. hair color), while the control of global concepts (e.g. age) seems to be coupled with the face identity."
- **[5]:** According to Karras et al. [2020]: "Acquiring, processing, and distributing the $10^5 - 10^6$ images required to train a modern high-quality, high-resolution GAN is a costly undertaking. The key problem with small datasets is that the discriminator overfits to the training examples; its feedback to the generator becomes meaningless and training starts to diverge."
- **[6]** According to Locatello et al. [2020]: "It is however interesting to notice how the GAN based methods perform especially well on the data sets SmallNORB and MPI3D where VAE based approaches struggle with reconstruction as the objects are either too detailed or too small."

# Disentanglement on Faces


(a) Baldness (-6, 6) (b) Face width (0, 6) (c) Gender (-6, 6) (d) Mustache (-6, 0)

Figure 1: Qualitative comparisons on CelebA. Traversal ranges are shown in parentheses. Some attributes are only manifested in one direction of a latent variable, so we show a one-sided traversal. Most semantically similar variables from a $\beta$-VAE are shown for comparison.

Source: From Chen et al. [2018]

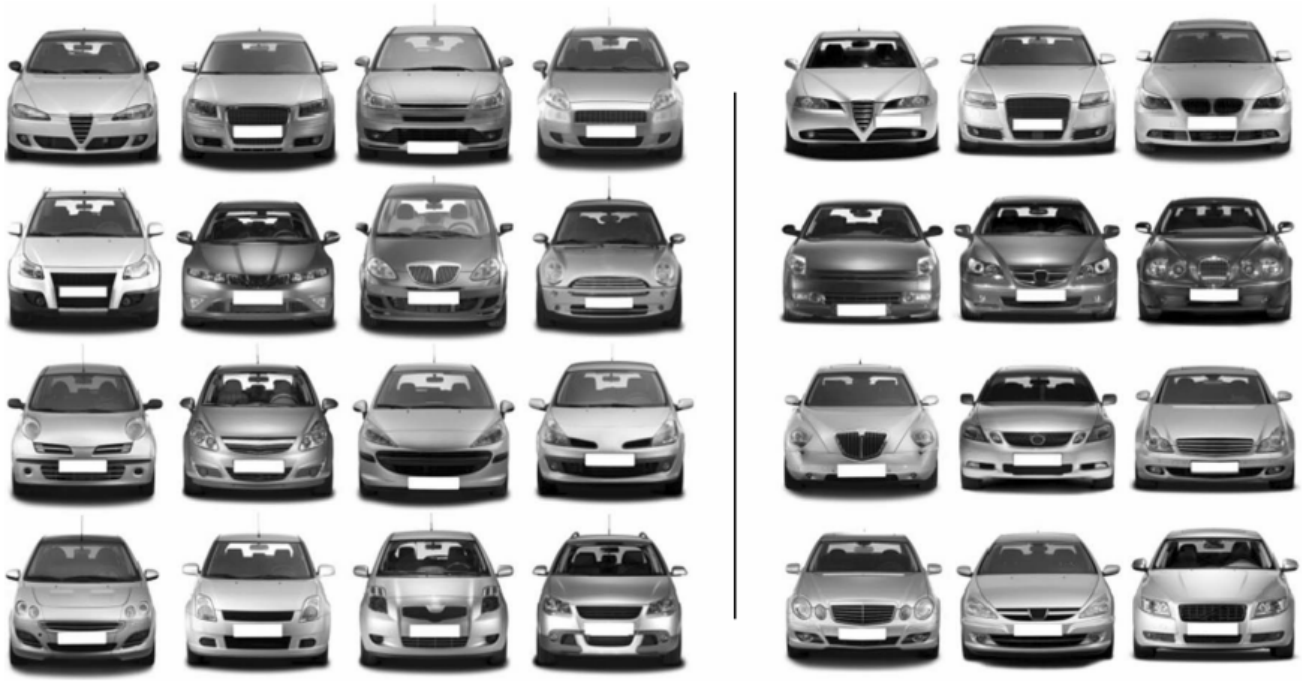
# Feature Point Method (Text)

1. We procured a metallic-grey specimen car of each of the 28 car models from various dealerships.
2. We then had the frontal design of each car photographed in a professional studio under standardized conditions.
3. Using morphing software (Perrett et al. 1994), we created two separate morphs, one with the photographs of the 16 compact cars and another with the 12 premium cars. We first defined 50 characteristic feature points of each frontal design (e.g., vertex of headlights, grill, windshield).
4. Next, the morphing software computed the mean position of each feature point across all models within a segment.
5. It then warped the images of all individual cars to the prototypical proportions (see the appendix; feature points also indicated) and averaged the color values of the corresponding pixels to create the morph (Benson and Perrett 1993).
6. We calculated a prototypicality score for each design by summing the Euclidian distances of each of the 50 feature points of the car from the corresponding feature points in the morphed (prototypical) car and inverting the overall score.
7. To calculate an objective measure of design complexity, we ran several computer algorithms to compress each image file. Perception research and algorithmic information theory (AIT; Donderi 2006) posit that a compressed image file can accurately measure picture complexity.

Source: Landwehr, Labroo, and Herrmann [2011]

# Feature Point Method (Images)

The 16 compact cars (left panel) and the 12 premium cars (right panel)



Source: Landwehr, Labroo, and Herrmann [2011]

# Feature Point Method (Images pt 2)

The morph of the 16 compact cars (left panel) and the 12 premium cars (right panel) with the positions of the feature points indicated

Sample dot patterns employed to validate fluency measures: left panel, feature points of the VW Polo; middle panel, feature points of morph; right panel, random feature points

Source: Landwehr, Labroo, and Herrmann [2011]

# Elea Feit Method

| Scenario 2 of 3 | Vehicle 1 | Vehicle 2 | Vehicle 3 | Vehicle 4 |
|---|---|---|---|---|
| Styling |  |  |  |  |
| AWD/FWD | All Wheel Drive (AWD) | Front Wheel Drive (FWD) | All Wheel Drive (AWD) | All Wheel Drive (AWD) |
| Fuel Economy | 20 mpg city | 16 mpg city | 26 mpg city | 26 mpg city |
| Engine | 4 cylinder hybrid | 6 cylinder | 4 cylinder | 4 cylinder hybrid |
| Seating | 8 passengers | 8 passengers | 5 passengers | 7 passengers |
| Cargo Capacity | 35 Cu. Ft. (about 7 large suitcases) | 35 Cu. Ft. (about 7 large suitcases) | 35 Cu. Ft. (about 7 large suitcases) | 35 Cu. Ft. (about 7 large suitcases) |
| Max Cargo Capacity (seats folded down) | small (60 Cu. Ft.) | small (60 Cu. Ft.) | small (60 Cu. Ft.) | small (60 Cu. Ft.) |
| Price (MSRP) | $24,999 | $24,999 | $27,999 | $24,999 |
| **Which of these vehicles would you be most likely to buy?** | ○ | ○ | ○ | ○ |

Figure 1: A choice task that includes images in addition to other verbal attributes.

Dotson, Jeffrey P. and Beltramo, Mark A. and Feit, Elea McDonnell and Smith, Randall C., Modeling the Effect of Images on Product Choices (April 12, 2019). Available at SSRN: https://ssrn.com/abstract=2282570

# Visual design matters across many product categories . . .

**Cars**

**Fashion**

**Furniture**

# . . . even for mundane categories like yogurt


*"We worked hard to get the packaging right . . . American yogurt has always been sold in containers with relatively narrow openings. In Europe yogurt containers are wider and squatter, and that's what I wanted for Chobani."*

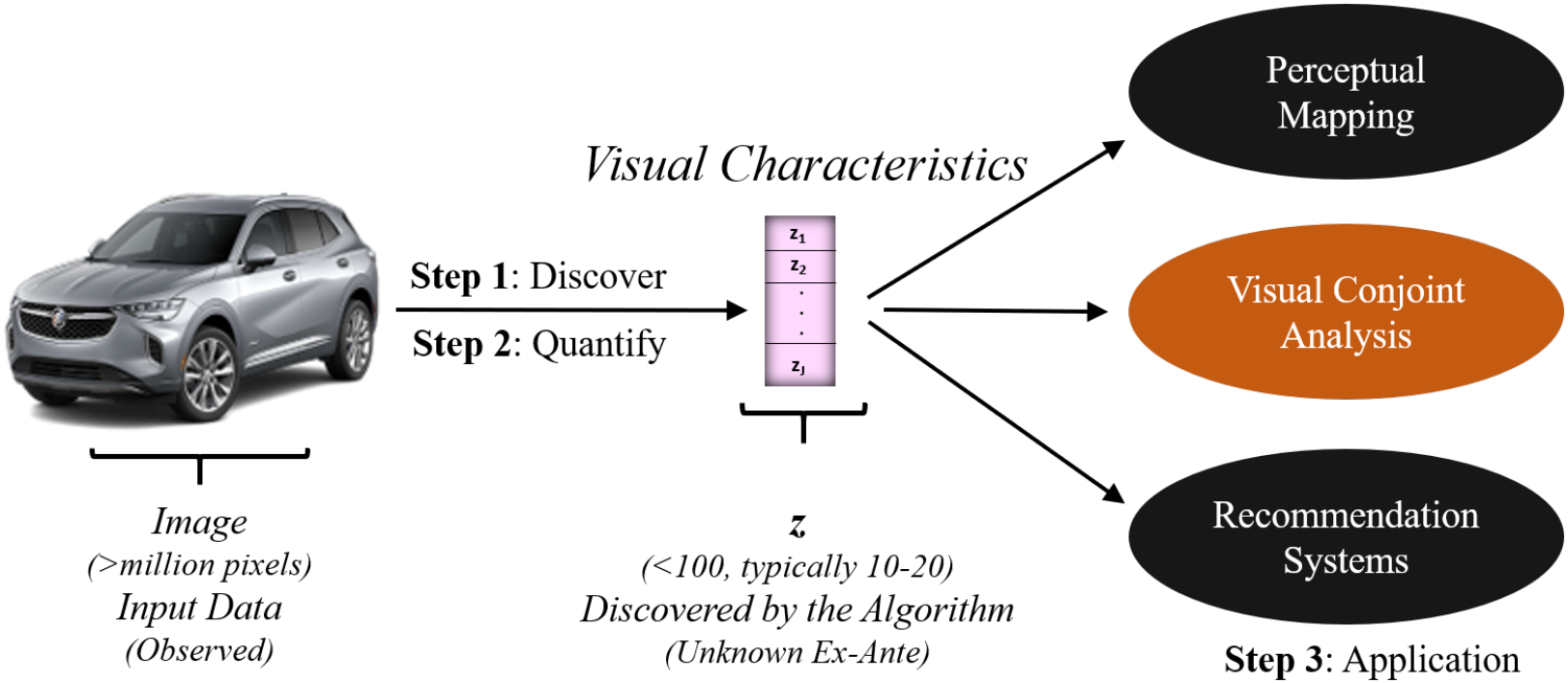
*—Hamdi Ulukaya, Founder & CEO, Chobani*

# Visual design matters

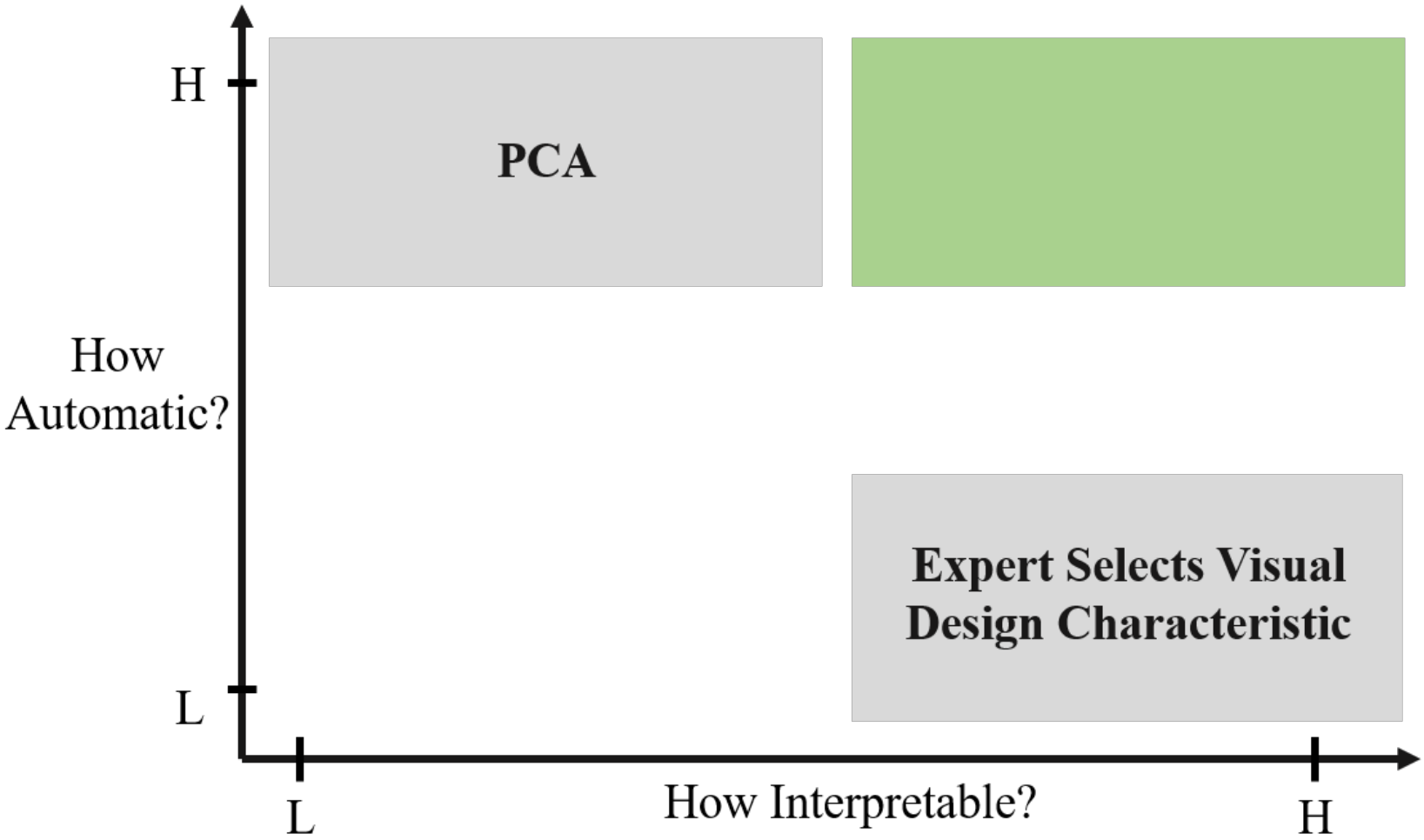
*"Exterior look/design is the top reason shoppers avoid a particular vehicle (30%), followed by cost (17%)."*

*—JD Power Avoider Study 2015*

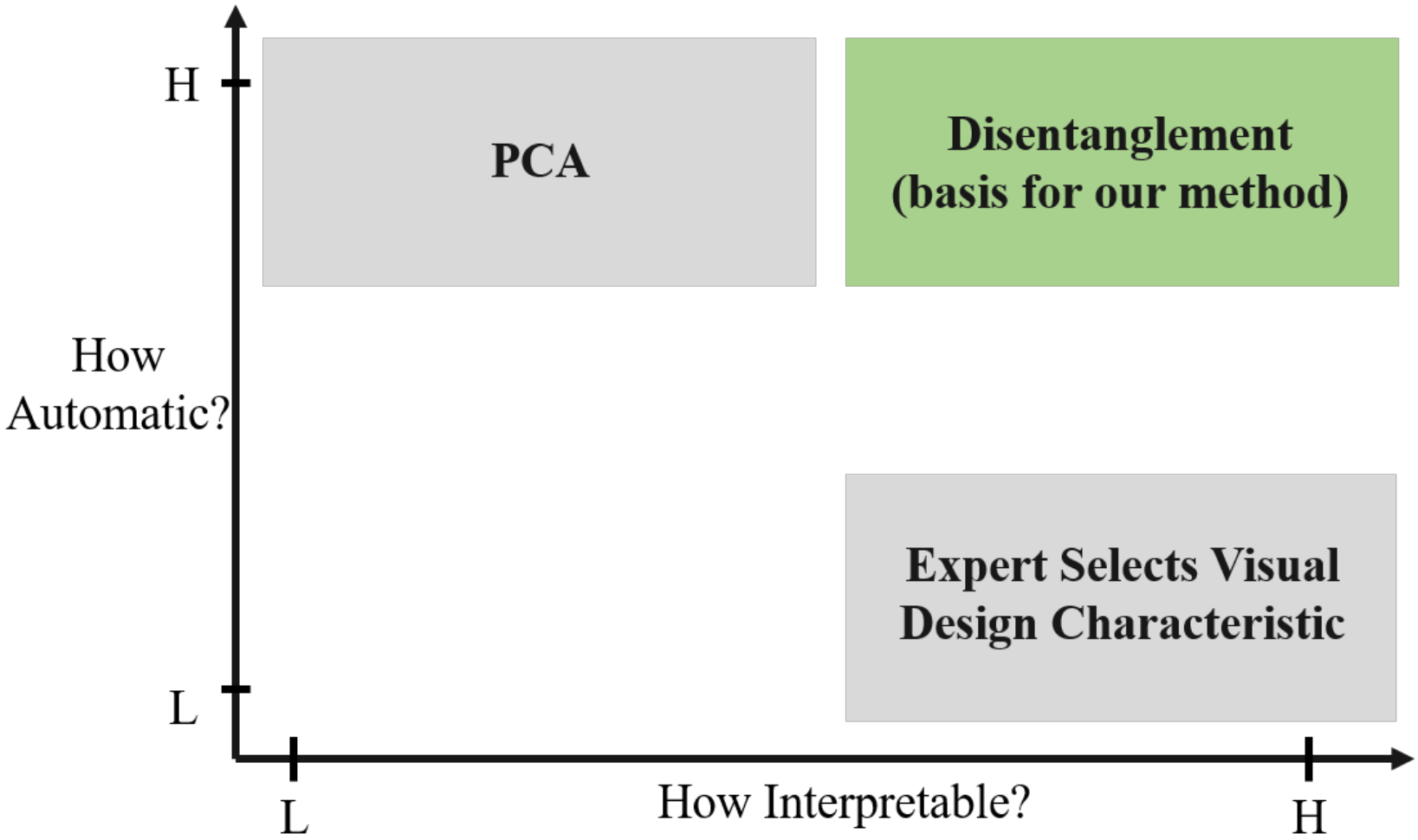
# Research Goals: Discover & quantify visual design characteristics using visual data



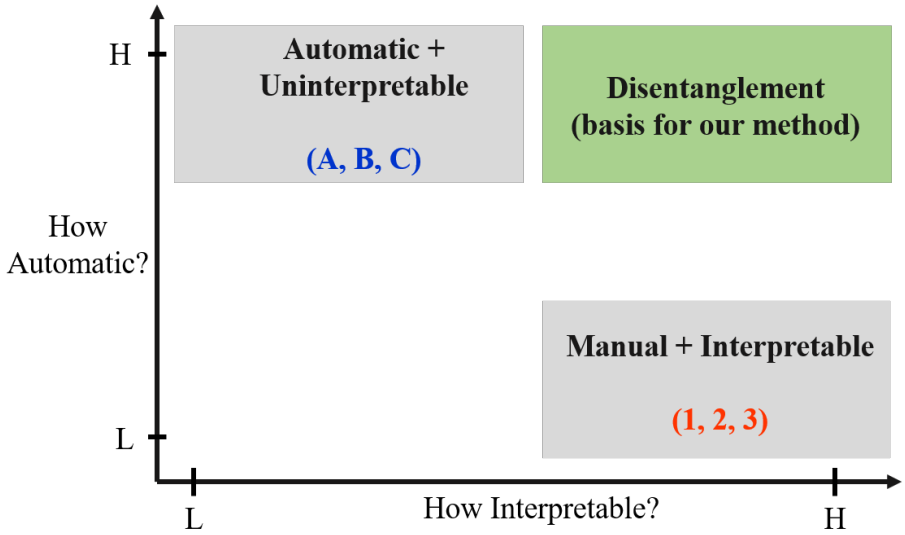
# Modeling Visual Characteristics: A comparison of methods



# Modeling Visual Characteristics: A comparison of methods



# Modeling Visual Characteristics: A comparison of methods



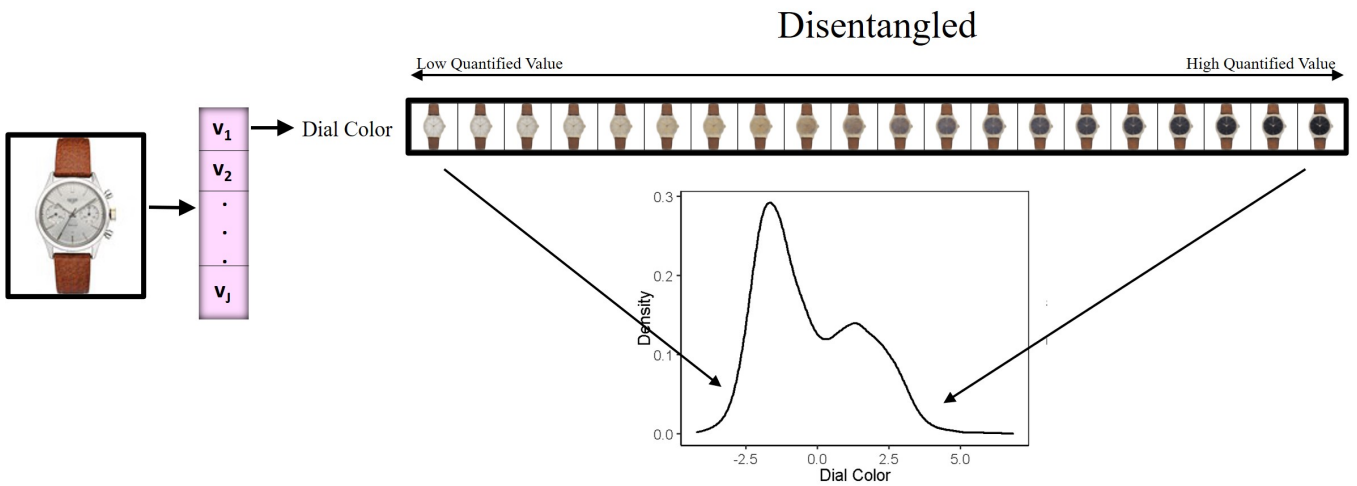

**Automatic + Uninterpretable**

- A - Bajari, P. L. et al. (2021) : Hedonic prices and quality adjusted price indices powered by AI, *CENMAP working paper*
- B - Law, S., et al. (2019) : Take a look around: using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology (TIST)*
- C - Aubry, S., et al. (2019) : Machine learning, human experts, and the valuation of real assets. *CFS Working Paper Series*

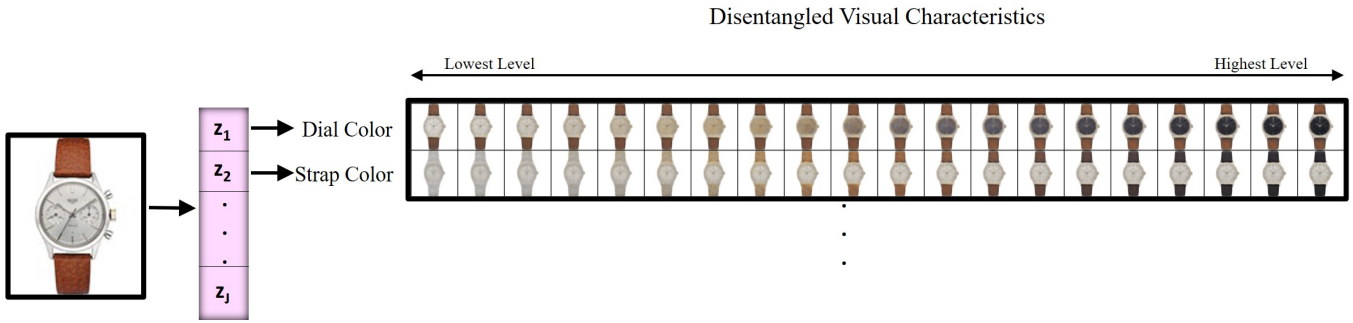
**Manual + Interpretable**

- 1 - Zhang, M. et al. (2022) : Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from yelp. *Management Science*
- 2 - Liu, Y., et al. (2017) : The effects of products' aesthetic design on demand and marketing-mix effectiveness: The role of segment prototypicality and brand consistency. *Journal of Marketing*
- 3 - Zhang, S., et al. (2021) : What makes a good image? Airbnb demand analytics leveraging interpretable image features. *Management Science*

# Disentangled v Entangled Representation



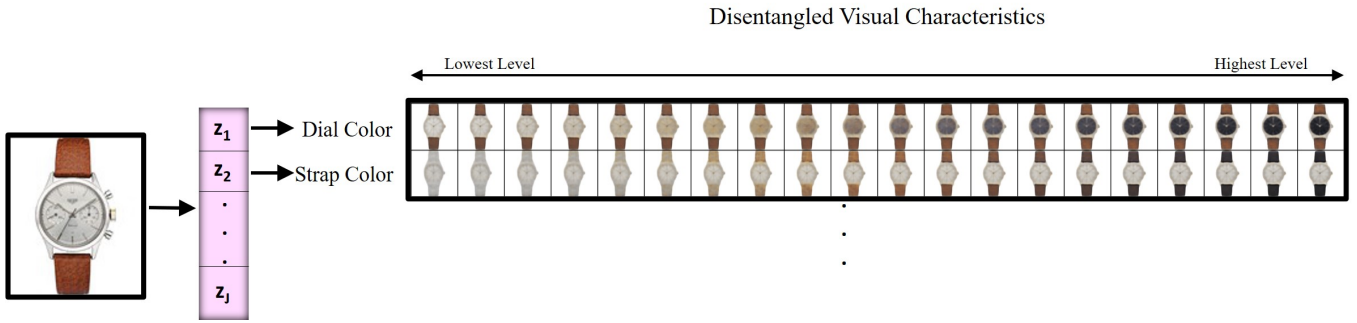
# Disentangled v Entangled Representation



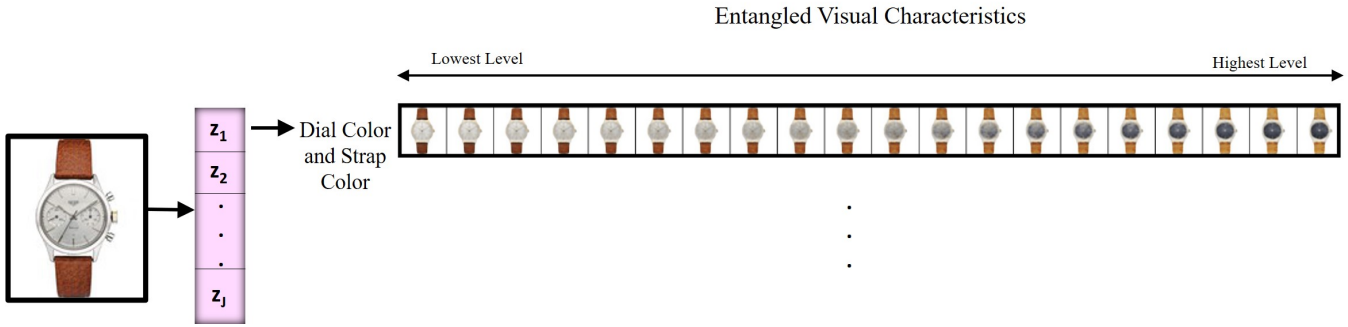

$v_1$ (disentangled): only dial color changes

$v_2$ (disentangled): only strap color changes

# Disentangled v Entangled Representation




$v_1$ (disentangled): only dial color changes $v_2$ (disentangled): only strap color changes




$v_1$ (entangled): both dial color and strap color changes

# What is disentanglement?

## Bengio et al (2013)

*"A disentangled representation can be defined as one where **single latent units** are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors"*

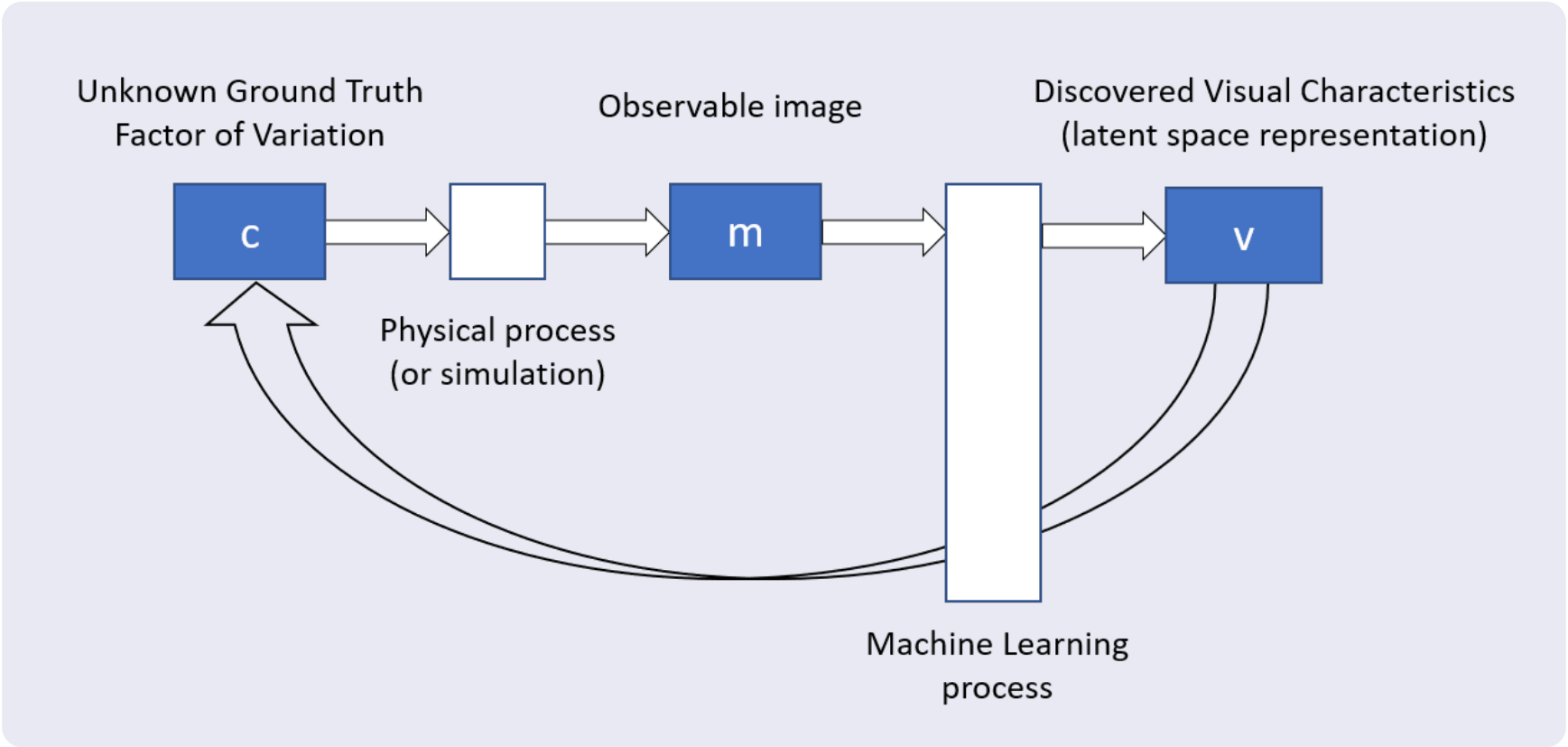
# What is disentanglement?

## Bengio et al (2013)

*"A disentangled representation can be defined as one where **single latent units** are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors"*

- Latent Units (**v**): Dimensions in the model's latent space
- Generative factors (**c**): Human-interpretable true characteristics

# What is disentanglement?

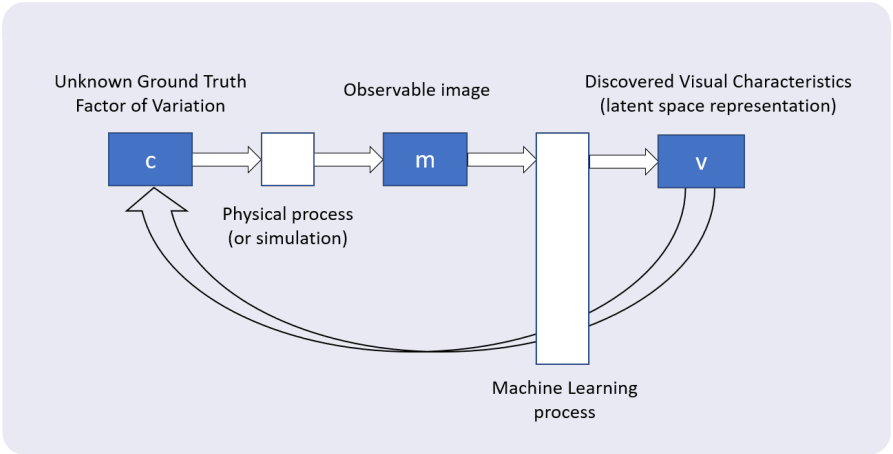



Goal of machine learning process: Recover latent space $\mathbf{v}(\mathbf{m}(\mathbf{c}))$ and make correspondence $\boxed{\mathbf{c} \longleftrightarrow \mathbf{v}}$

# What is disentanglement?

## Bengio et al (2013)

*"A disentangled representation can be defined as one where **single latent units** are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors"*



# Representation Learning

Burgess et al. [2017] describes this in more detail:

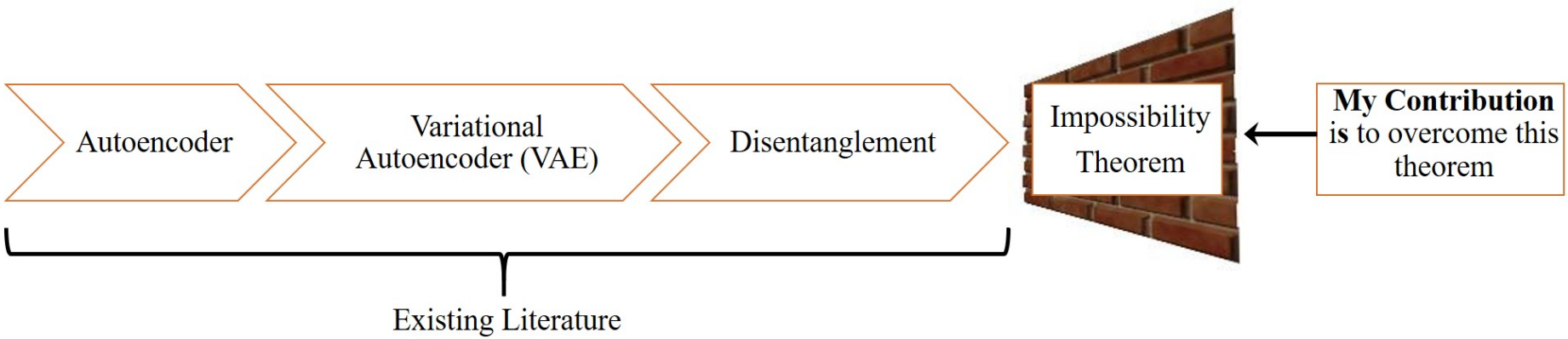
> *"A disentangled representation can be defined as one where single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors [Bengio, Courville, and Vincent, 2013]. For example, a model trained on a dataset of 3D objects might learn independent latent units sensitive to single independent data generative factors, such as object identity, position, scale, lighting or colour, similar to an inverse graphics model [Kulkarni et al., 2015]. A disentangled representation is therefore factorised and often interpretable, whereby different independent latent units learn to encode different independent ground-truth generative factors of variation in the data."*

# Disentanglement Representation Learning

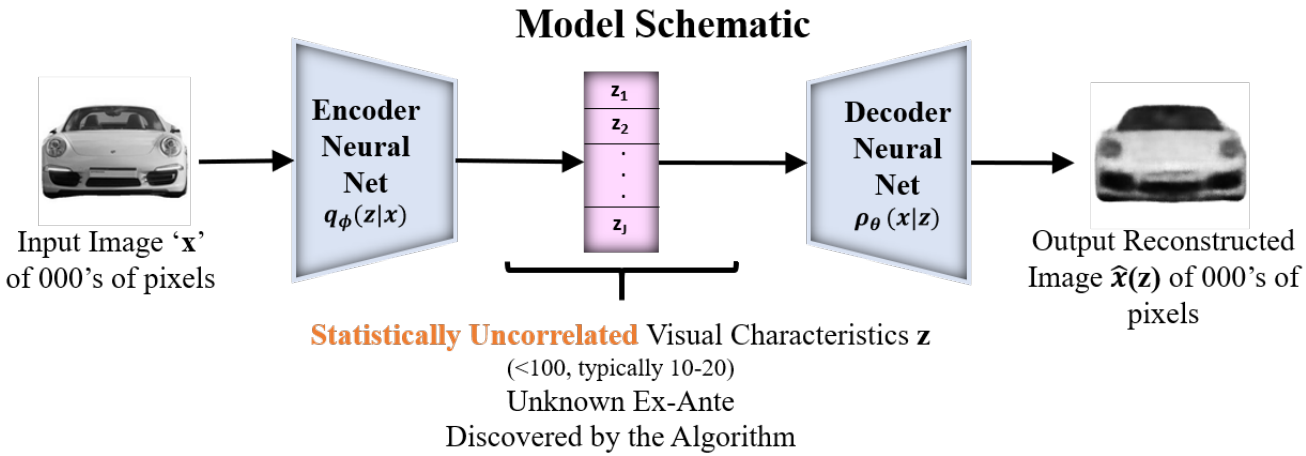
According to Bengio, Courville, and Vincent [2013]:

> *"learning representations of the data that make it easier to extract useful information when building classifiers or other predictors."*

# Roadmap of Methodological Approach

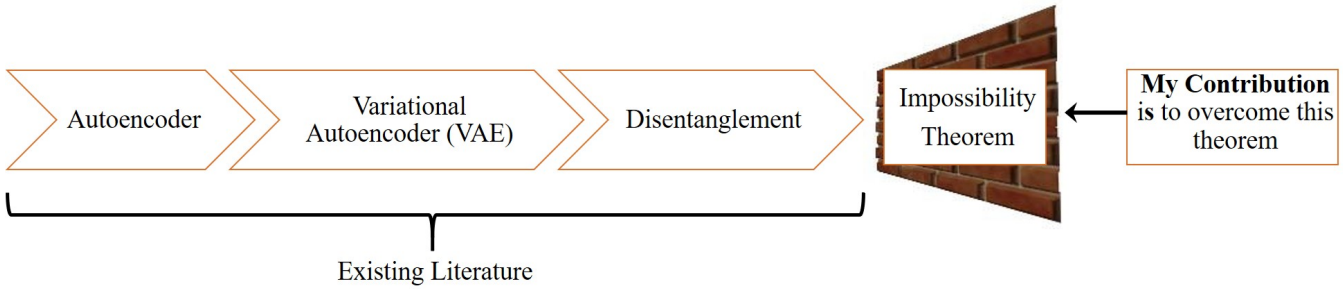


# Models in Existing Literature




| Model | Goal |
|---|---|
| Autoencoder (AE) | Reconstruction accuracy |
| Variational Autoencoder (VAE) | . . . + structured latent space |
| Disentanglement | . . . + . . . + statistically independent latent space |

# Impossibility Theorem

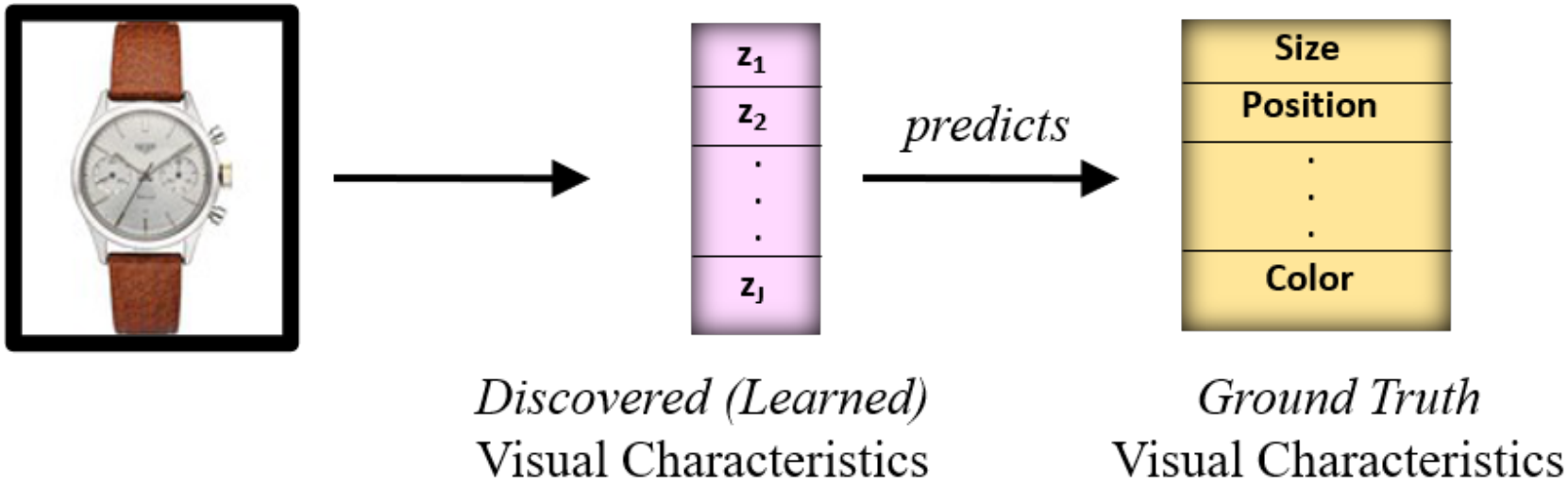



## Impossibility Theorem

Unsupervised (*i.e. only images*) learning of disentangled representations is *fundamentally impossible* except under certain restrictive conditions.

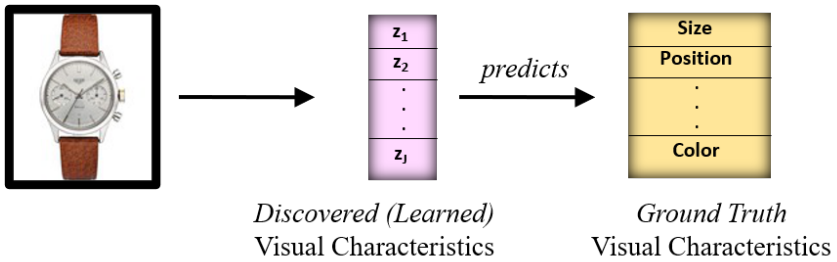
**Implication**: Every disentangled representation can have other equivalent entangled representations.

# Overcoming Impossibility Theorem



# Overcoming Impossibility Theorem

Common approach to ground truth in ML is to get humans to label[1]



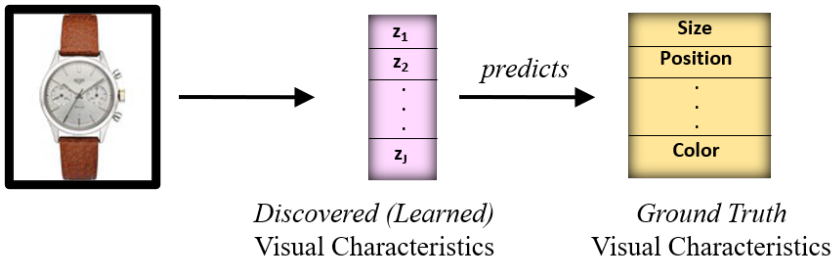

## What's the Problem?

- Ground truth on visual characteristics is unknown. In fact, these are precisely what we want to find.

[1] Locatello, Francesco, et al. "Disentangling factors of variation using few labels." ICLR. 2020.
Gyawali, Prashnna K. et al. "Learning to disentangle inter-subject anatomical variations in electrocardiographic data." IEEE Transactions on Biomedical Engineering. 2021.

# Overcoming Impossibility Theorem

Common approach to ground truth in ML is to get humans to label[1]



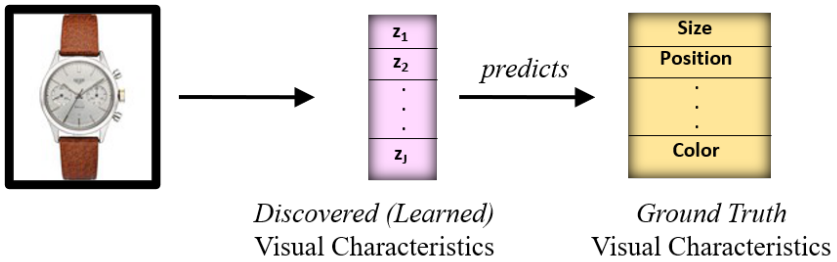

## What's the Problem?

- Ground truth on visual characteristics is unknown. In fact, these are precisely what we want to find.
- Researcher needs to determine what are the true characteristics to focus on

[1] Locatello, Francesco, et al. "Disentangling factors of variation using few labels." ICLR. 2020.
Gyawali, Prashnna K. et al. "Learning to disentangle inter-subject anatomical variations in electrocardiographic data." IEEE Transactions on Biomedical Engineering. 2021.

# Overcoming Impossibility Theorem

Common approach to ground truth in ML is to get humans to label[1]



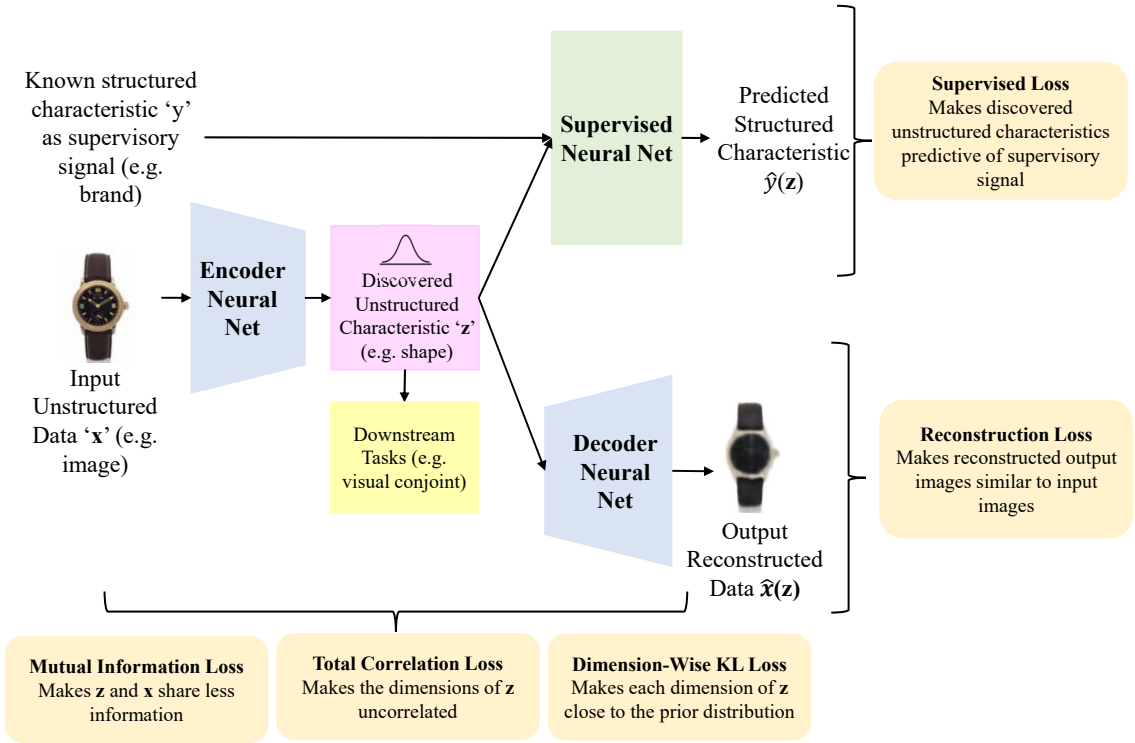

## What's the Problem?

- Ground truth on visual characteristics is unknown. In fact, these are precisely what we want to find.
- Researcher needs to determine what are the true characteristics to focus on
- Need to ensure humans understand what these labels are and how to quantify them for each image

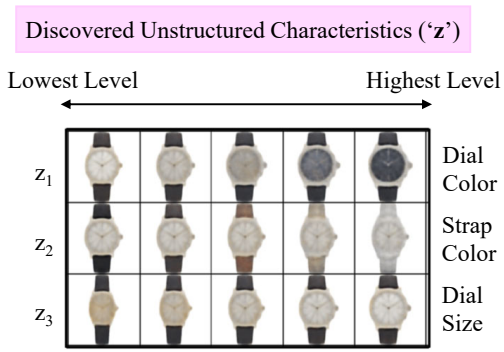
[1] Locatello, Francesco, et al. "Disentangling factors of variation using few labels." ICLR. 2020.
Gyawali, Prashnna K. et al. "Learning to disentangle inter-subject anatomical variations in electrocardiographic data." IEEE Transactions on Biomedical Engineering. 2021.

# Schematic of Proposed Approach

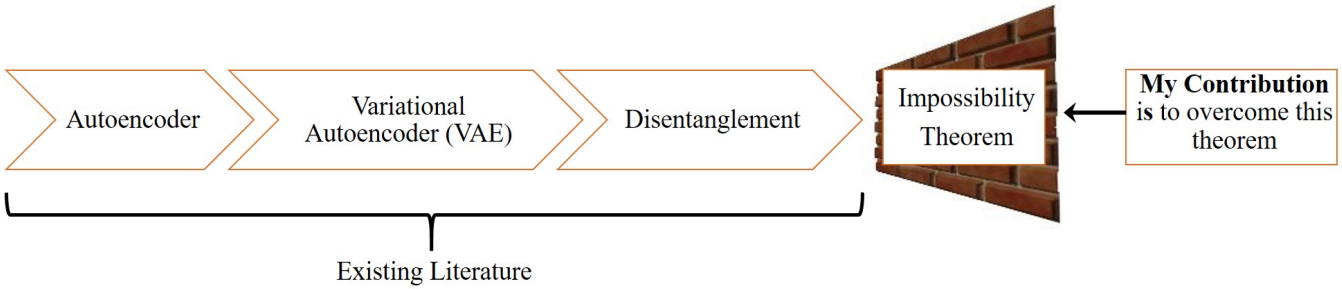
a




b



# Contribution




**Solution** without ground truth on visual characteristics:
Leverage structured product characteristics to provide a supervisory signal for disentanglement

**Why?** See Next Slide

# Temporary Slide

Our approach is motivated by the result from Locatello et al. [2020] that even weak supervision with noisy metrics of ground truth is able to achieve good disentanglement. *We posit and empirically find that product characteristics act as such noisy measures that are correlated with ground truth corresponding to visual characteristics.* The ML literature has always relied on some version of the ground truth.

1. First, consider a characteristic like material, e.g. silver that provides a certain visual look to a product. Material more broadly is known to significantly affect visual appearance and consumer perceptions [Fleming, 2014].
2. Second, a product characteristic like brand is likely to strongly impact visual look of a product. The signature of the brand design is often visibly present and apparent from the product's appearance to consumers, especially for product categories with visible consumption [Ferraro, Kirmani, and Matherly, 2013, Liu, Dzyabura, and Mizik, 2020, Simonson and Schmitt, 1997] or luxury brands [Lee, Hur, and Watkins, 2018, Megehee and Spake, 2012].
3. Third, consider the role of price, which is strictly speaking not a product characteristic, since it can be set by the retailer. However, many brands, especially luxury brands, maintain carefully curated pricing tiers with strong consumer associations.

# Table of Notation

| Symbol | Category | Meaning |
|---|---|---|
| $\mathbf{x}$ | Input Data | Product image |
| $\mathbf{y}$ | Input Data | Supervisory signal(s) |
| $\hat{\mathbf{x}}$ | Output Data | Reconstructed image |
| $\hat{\mathbf{y}}$ | Output Data | Predicted Supervisory Signal(s) |
| $\mathbf{z}$ | Latent Space | Visual characteristic vector |
| $\mathbf{z}_{\text{inf}}$ | Subset of Latent Space | Informative visual characteristic |
| $p(\mathbf{z})$ | Model | Prior distribution |
| $p_\theta(\mathbf{x}\|\mathbf{z})$ | Decoder Neural Net | Conditional Probability of Generating Image Data given Latent Space |
| $q_\phi(\mathbf{z}\|\mathbf{x})$ | Encoder Neural Net | Conditional Probability of Latent Space given Image Data |
| $p_w(\mathbf{y}\|\mathbf{z})$ | Supervisory Neural Net | Conditional Probability of Supervisory Signal given Latent Space |
| $\theta$ | Weights of Neural Net | Decoder's parameters |
| $\phi$ | Weights of Neural Net | Encoder's parameters |
| $w$ | Weights of Neural Net | Supervisory Net's parameters |
| $\mathbf{E}_{q_\phi(\mathbf{z}\|\mathbf{x})}\left[\log p_\theta(\mathbf{x}\|\mathbf{z})\right]$ | Loss Function | Reconstruction Loss |
| $I_q(\mathbf{z},\mathbf{x})$ | Loss Function | Mutual Information Loss |
| $KL\left[q(\mathbf{z})\|\|\prod_{j=1}^{J} q(z_j)\right]$ | Loss Function | Total Correlation Loss |
| $\sum_{j=1}^{J} KL\left[q(z_j)\|\|p(z_j)\right]$ | Loss Function | Dimension KL Divergence Loss |
| $P(\hat{y}(\mathbf{z}), y)$ | Loss Function | Supervised Loss |
| $\mathcal{L}(\theta,\phi,\beta;\mathbf{x},\mathbf{z})$ | Loss Function | Total Loss |
| $J$ | Hyperparameter | Dimensionality of latent space |
| $\alpha$ | Hyperparameter | Weight on Mutual Information Loss |
| $\beta$ | Hyperparameter | Weight on Total Correlation Loss |
| $\gamma$ | Hyperparameter | Weight on Dimension KL Divergence Loss |
| $\delta$ | Hyperparameter | Weight on Supervised Loss |

# Model Estimation

- Learn model parameters by minimizing loss $L(\theta, \phi; \mathbf{x}, \mathbf{z})$ of integrated model
- $\theta$ and $\phi$ are encoder and decoder parameters; $\mathbf{x}$ are images and $\mathbf{z}$ are visual characteristics

$$\underbrace{L(\theta, \phi, \mathbf{w}); \mathbf{x}, \mathbf{z})}_{\text{Total Loss}} = \underbrace{\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\substack{\text{Reconstruction} \\ \text{Loss}}} + \alpha \underbrace{I_q(\mathbf{z}, \mathbf{x})}_{\substack{\text{Mutual} \\ \text{Information} \\ \text{Loss}}} + \beta \underbrace{KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]}_{\substack{\text{Total Correlation} \\ \text{Loss}}}$$

$$+\gamma \underbrace{\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]}_{\substack{\text{Dimension-Wise} \\ \text{KL Divergence Loss}}} + \delta \underbrace{P(\hat{\mathbf{y}}(\mathbf{z}), \mathbf{y})}_{\substack{\text{Supervised} \\ \text{Loss}}}$$

| **Loss Term** | **Why is this term included?** |
|---|---|
| Reconstruction | Promotes accurate reconstruction of images |

# Model Estimation

- Learn model parameters by minimizing loss $L(\theta, \phi; \mathbf{x}, \mathbf{z})$ of integrated model
- $\theta$ and $\phi$ are encoder and decoder parameters; $\mathbf{x}$ are images and $\mathbf{z}$ are visual characteristics

$$\underbrace{L(\theta, \phi, \mathbf{w}); \mathbf{x}, \mathbf{z})}_{\text{Total Loss}} = \underbrace{\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\substack{\text{Reconstruction} \\ \text{Loss}}} + \alpha \underbrace{I_q(\mathbf{z}, \mathbf{x})}_{\substack{\text{Mutual} \\ \text{Information} \\ \text{Loss}}} + \beta \underbrace{KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]}_{\substack{\text{Total Correlation} \\ \text{Loss}}}$$

$$+\gamma \underbrace{\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]}_{\substack{\text{Dimension-Wise} \\ \text{KL Divergence Loss}}} + \delta \underbrace{P(\hat{\mathbf{y}}(\mathbf{z}), \mathbf{y})}_{\substack{\text{Supervised} \\ \text{Loss}}}$$

| **Loss Term** | **Why is this term included?** |
|---|---|
| Reconstruction | Promotes accurate reconstruction of images |
| Mutual Information | Minimizes redundant information |

# Model Estimation

- Learn model parameters by minimizing loss $L(\theta, \phi; \mathbf{x}, \mathbf{z})$ of integrated model
- $\theta$ and $\phi$ are encoder and decoder parameters; $\mathbf{x}$ are images and $\mathbf{z}$ are visual characteristics

$$\underbrace{L(\theta, \phi, \mathbf{w}); \mathbf{x}, \mathbf{z})}_{\text{Total Loss}} = \underbrace{\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\substack{\text{Reconstruction} \\ \text{Loss}}} + \alpha \underbrace{I_q(\mathbf{z}, \mathbf{x})}_{\substack{\text{Mutual} \\ \text{Information} \\ \text{Loss}}} + \beta \underbrace{KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]}_{\substack{\text{Total Correlation} \\ \text{Loss}}}$$

$$+\gamma \underbrace{\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]}_{\substack{\text{Dimension-Wise} \\ \text{KL Divergence Loss}}} + \delta \underbrace{P(\hat{\mathbf{y}}(\mathbf{z}), \mathbf{y})}_{\substack{\text{Supervised} \\ \text{Loss}}}$$

| **Loss Term** | **Why is this term included?** |
|---|---|
| Reconstruction | Promotes accurate reconstruction of images |
| Mutual Information | Minimizes redundant information |
| **Total Correlation** | Promotes statistical independence between visual characteristics |

# Model Estimation

- Learn model parameters by minimizing loss $L(\theta, \phi; \mathbf{x}, \mathbf{z})$ of integrated model
- $\theta$ and $\phi$ are encoder and decoder parameters; $\mathbf{x}$ are images and $\mathbf{z}$ are visual characteristics

$$\underbrace{L(\theta, \phi, \mathbf{w}); \mathbf{x}, \mathbf{z})}_{\text{Total Loss}} = \underbrace{\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\substack{\text{Reconstruction} \\ \text{Loss}}} + \alpha \underbrace{I_q(\mathbf{z}, \mathbf{x})}_{\substack{\text{Mutual} \\ \text{Information} \\ \text{Loss}}} + \beta \underbrace{KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]}_{\substack{\text{Total Correlation} \\ \text{Loss}}}$$

$$+\gamma \underbrace{\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]}_{\substack{\text{Dimension-Wise} \\ \text{KL Divergence Loss}}} + \delta \underbrace{P(\hat{\mathbf{y}}(\mathbf{z}), \mathbf{y})}_{\substack{\text{Supervised} \\ \text{Loss}}}$$

| Loss Term | Why is this term included? |
|---|---|
| Reconstruction | Promotes accurate reconstruction of images |
| Mutual Information | Minimizes redundant information |
| **Total Correlation** | Promotes statistical independence between visual characteristics |
| Dimension-Wise KL | Penalizes deviations from a prior |

# Model Estimation

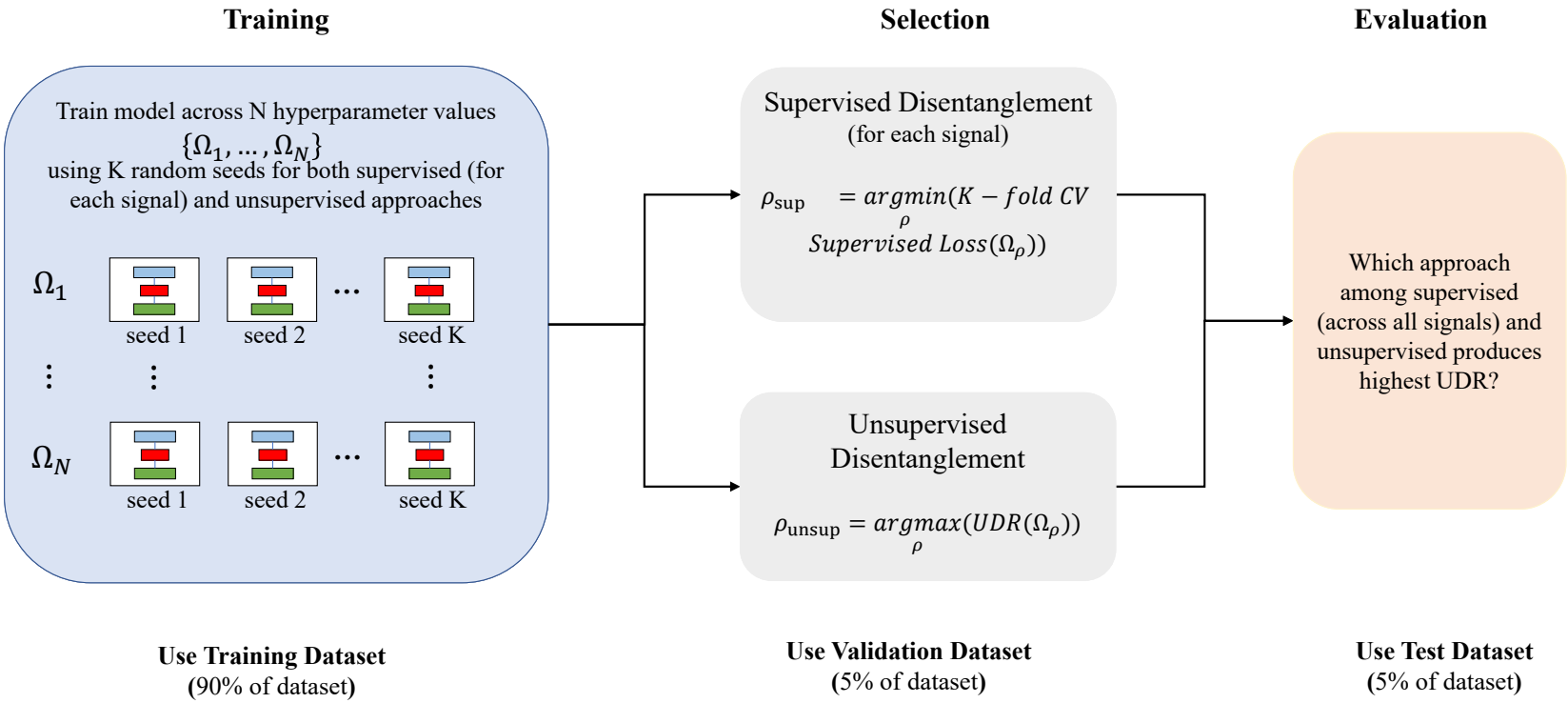
- Learn model parameters by minimizing loss $L(\theta, \phi; \mathbf{x}, \mathbf{z})$ of integrated model
- $\theta$ and $\phi$ are encoder and decoder parameters; $\mathbf{x}$ are images and $\mathbf{z}$ are visual characteristics

$$\underbrace{L(\theta, \phi, \mathbf{w}); \mathbf{x}, \mathbf{z})}_{\text{Total Loss}} = \underbrace{\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\substack{\text{Reconstruction} \\ \text{Loss}}} + \alpha \underbrace{I_q(\mathbf{z}, \mathbf{x})}_{\substack{\text{Mutual} \\ \text{Information} \\ \text{Loss}}} + \beta \underbrace{KL\left[q(\mathbf{z})||\prod_{j=1}^{J} q(z_j)\right]}_{\substack{\text{Total Correlation} \\ \text{Loss}}}$$

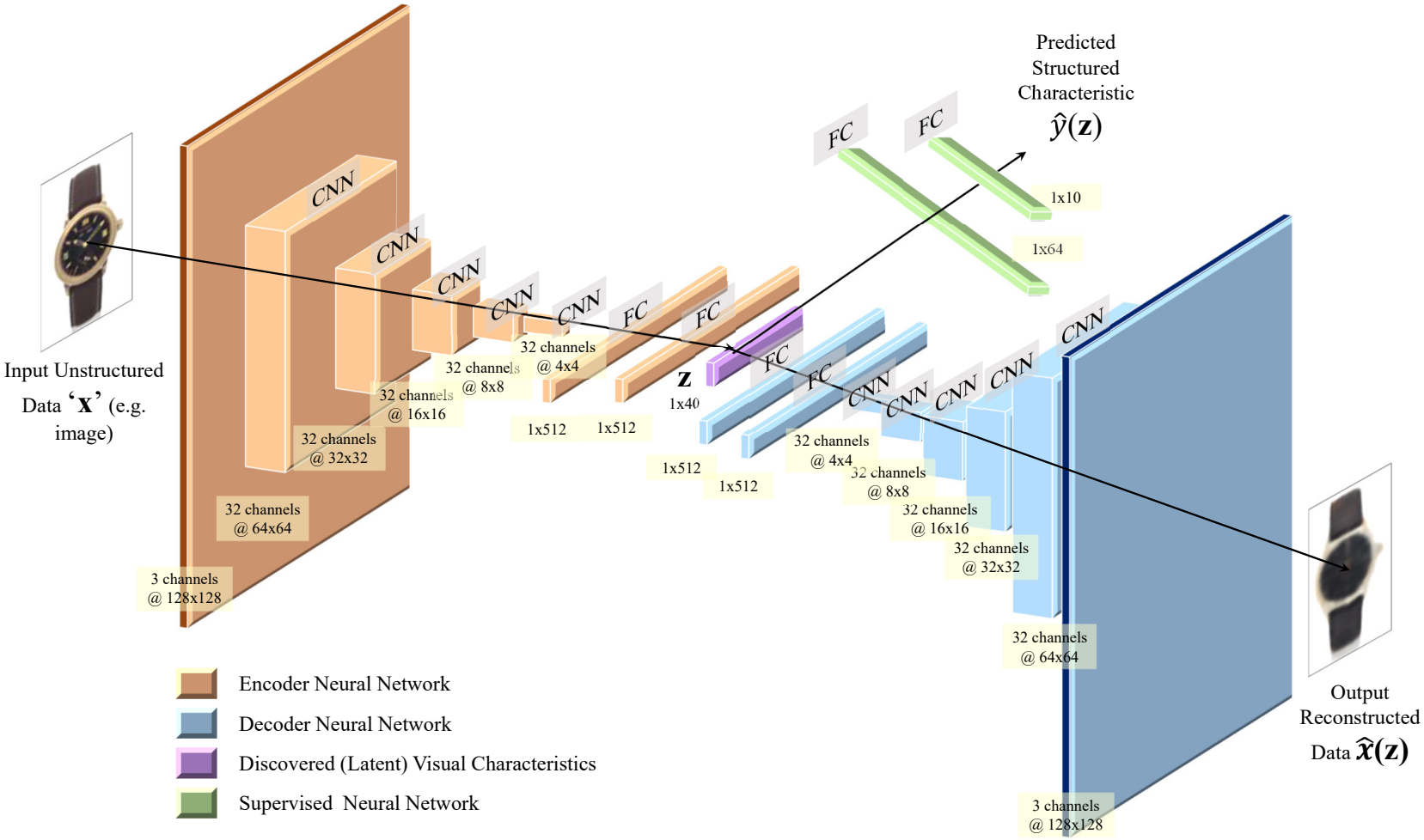
$$+\gamma \underbrace{\sum_{j=1}^{J} KL\left[q(z_j)||p(z_j)\right]}_{\substack{\text{Dimension-Wise} \\ \text{KL Divergence Loss}}} + \delta \underbrace{P(\hat{\mathbf{y}}(\mathbf{z}), \mathbf{y})}_{\substack{\text{Supervised} \\ \text{Loss}}}$$

| **Loss Term** | **Why is this term included?** |
|---|---|
| Reconstruction | Promotes accurate reconstruction of images |
| Mutual Information | Minimizes redundant information |
| **Total Correlation** | Promotes statistical independence between visual characteristics |
| Dimension-Wise KL | Penalizes deviations from a prior |
| **Supervised** | Provides a signal to address the impossibility theorem |

# Model Training, Selection, & Evaluation



# Model Architecture



# Disentanglement Evaluation Metric

Estermann, Marks, and Yanik [2020] details the value of UDR, which we quote below:

> *"There are no labels available for many real-life applications and for some data, generative factors of interest are hard or impossible for humans to annotate.*

# Disentanglement Evaluation Metric

Estermann, Marks, and Yanik [2020] details the value of UDR, which we quote below:

> *"Recently, Duan et al. [8] defined a new, unsupervised heuristic for evaluating the disentanglement performance of models, based on the assumption that* ***models that disentangle well are more likely to be similar to each other than the ones that do not disentangle*** *[16, 17, 18, 19].*

# Disentanglement Evaluation Metric

Estermann, Marks, and Yanik [2020] details the value of UDR, which we quote below:

> *"Recently, Duan et al. [8] defined a new, unsupervised heuristic for evaluating the disentanglement performance of models, based on the assumption that* ***models that disentangle well are more likely to be similar to each other than the ones that do not disentangle*** *[16, 17, 18, 19]. They demonstrate that this Unsupervised Disentanglement Ranking (UDR) correlates well with metrics that rely on previously annotated labels across various models and datasets [8]."*

# Disentanglement Evaluation Metric

Unsupervised Disentanglement Ranking [2]: Metric to evaluate the quality of disentanglement

$$\beta, \delta = \underset{\beta \in \mathbb{R}, \delta \in \mathbb{R}}{\arg\max} \; UDR$$

- **Intuition**: Discover representations that are invariant to model perturbations
- **Range** of UDR: 0 to 1 (Higher is better)
- Does not require access to ground truth labels

[2] Duan et. al. "Unsupervised model selection for variational disentangled representation learning." ICLR 2020

# Data: Luxury watches auctioned at Christie's.

| | |
|---|---|
| Number of Watches | 6857 |
| Brand | Audemar's Piguet, Cartier, Patek Philippe, Rolex, Others |
| Movement | Automatic, Mechanical, Quartz |
| Material | Ceramic, Diamond, Gold, Platinum, Steel, Titanium |
| Year of Manufacture | Pre-1950s, 1950s, 1960s, 1970s, 1980s, 1990s, 2000s, 2010s |
| Case Diameter (in mm) | Min (9 mm); Median (37 mm); Max (62 mm) |
| Hammer Price (in 2000 dollars) | Min ($364); Median ($6,971); Max ($950,196) |
| Auction Year | 2001-2020 |
| Auction Location | Amsterdam, Dubai, Hong Kong, London, New York, Online |

# Example of Entanglement and Disentanglement in Visual Characteristics

a) Focal Watch

b) Entangled Model

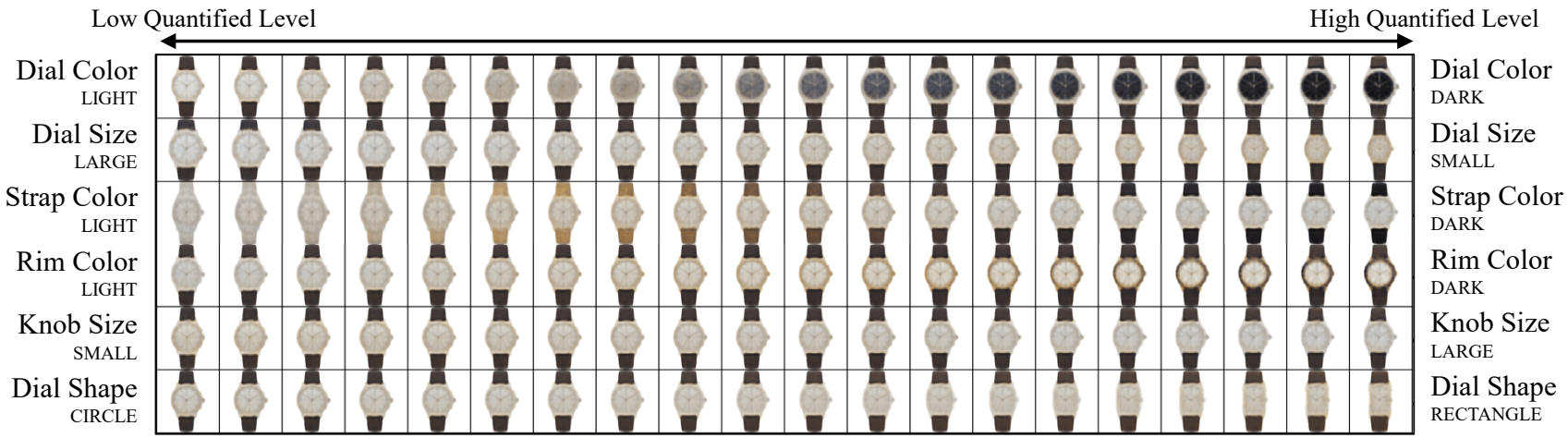
c) Disentangled Model

*Notes:* **a**: Focal watch **b**: Entangled model outputs a characteristic that changes both the dial color and strap color as its the level is changed. **c**: Disentangled model outputs two independent characteristics for dial color and strap color.

# Comparison of Different Supervisory Approaches

| Number of Signals | Supervisory Signals | UDR |
|---|---|---|
| 2 | Brand & Material | 0.363 |
| 2 | Circa & Movement | 0.357 |
| 2 | Brand & Circa | 0.309 |
| 3 | Brand, Material & Movement | 0.242 |
| 2 | Circa & Material | 0.184 |
| 1 | Brand | 0.135 |
| 0 | Unsupervised | 0.131 |
| 1 | Material | 0.128 |
| 2 | Material & Movement | 0.122 |
| 2 | Brand & Movement | 0.121 |
| 1 | Movement | 0.116 |
| 1 | Circa | 0.112 |
| 1 | Price | 0.076 |

# Discovered Visual characteristics



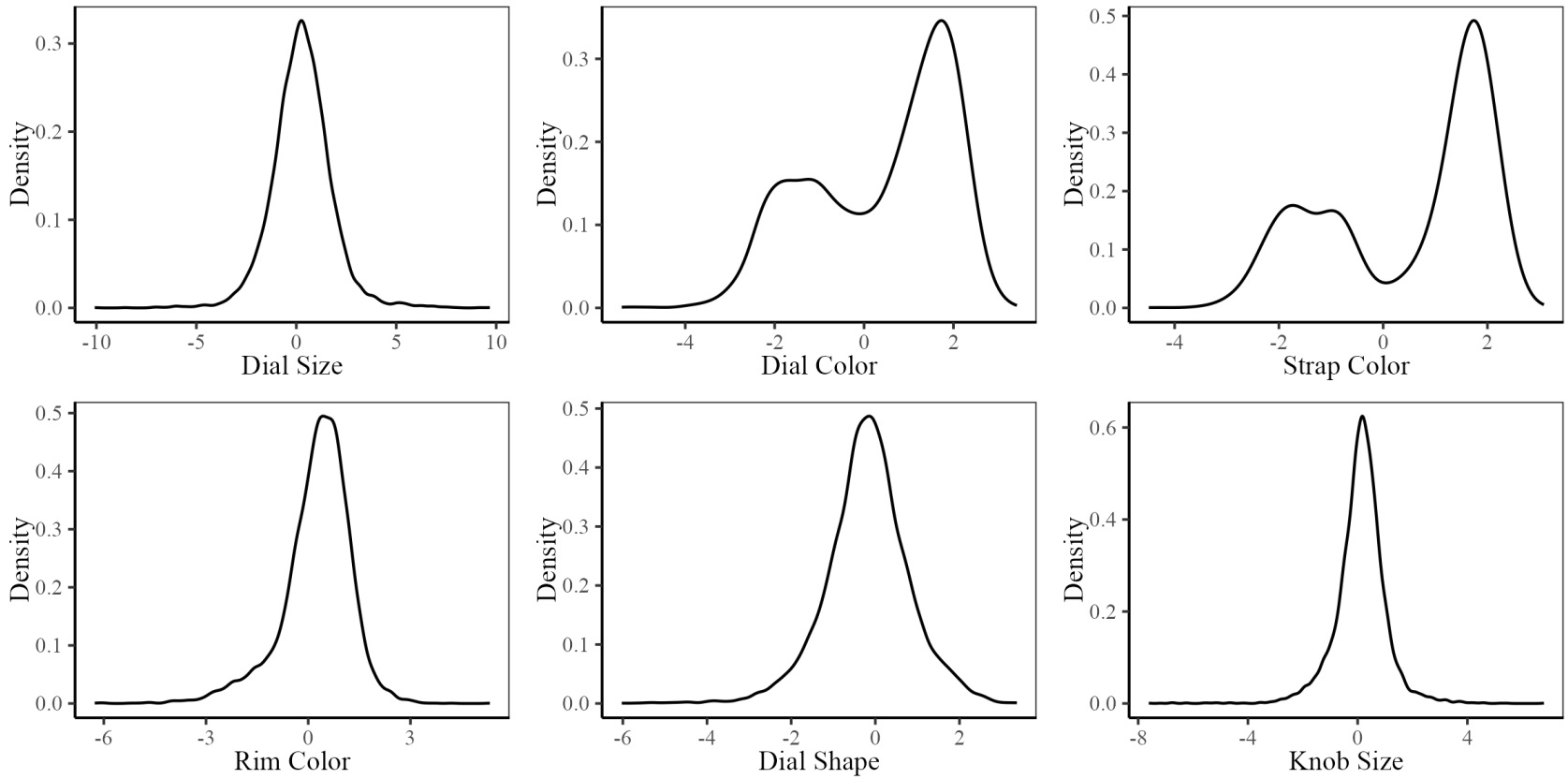
# Summary Statistics of Discovered Visual characteristics (from 'Brand+Material' Signal)

| Visual characteristic | Mean | SD | Min | Max |
|---|---|---|---|---|
| Dial Size | 0.28 | 1.49 | −11.08 | 9.68 |
| Dial Color | 0.38 | 1.59 | −5.42 | 3.42 |
| Strap Color | 0.50 | 1.58 | −4.50 | 3.08 |
| Rim Color | 0.25 | 1.01 | −6.27 | 5.33 |
| Dial Shape | −0.19 | 0.99 | −6.03 | 3.36 |
| Knob Size | 0.11 | 0.93 | −7.61 | 6.79 |

# Correlations Between Visual Characteristics

| | Dial Size | Dial Color | Strap Color | Rim Color | Dial Shape | Knob Size |
|---|---|---|---|---|---|---|
| Dial Size | **1.00** | 0.17 | -0.08 | -0.03 | -0.02 | 0.00 |
| Dial Color | 0.17 | **1.00** | 0.03 | -0.00 | 0.09 | -0.02 |
| Strap Color | -0.08 | 0.03 | **1.00** | -0.11 | -0.03 | -0.04 |
| Rim Color | -0.03 | -0.00 | -0.11 | **1.00** | 0.09 | -0.01 |
| Dial Shape | -0.02 | 0.09 | -0.03 | 0.09 | **1.00** | 0.05 |
| Knob Size | 0.00 | -0.02 | -0.04 | -0.01 | 0.05 | **1.00** |

# Density of Discovered Visual characteristics (from 'Brand+Material' Signal)



# Surveys to Validate Interpretability


**Bezel**: Ring around the watch dial or face

**Crown**: little knob on the side of the watch used to set time

**Date Window**: Indicates the date

**Dial**: Main face of the watch (over which hands move)

**Hands**: Indicate time

**Hour Marker**: Indicators where the hands point to tell the time

**Lug**: Connects the dial to the strap

**Strap**: Secures the watch to the wrist

# Validating Interpretability of Each Characteristic's Meaning

Starting from the image on the left, **what part of the watch changes the most** as you go from left to right? Carefully check both large and small visual aspects. Go through each part of the watch one by one before selecting any option. Refer to the above image to see parts of the watch.

Note: Images are low-quality on purpose

- ○ Bezel
- ○ Crown
- ○ Date Window
- ○ Dial
- ○ Hands
- ○ Hour Marker
- ○ Lug
- ○ Strap

How is that part of the watch changing?

# Validating Interpretability of Quantification of Each Characteristic

Which pair of watches in your judgment are more similar in terms of dial color than the other pair? (ignore all the other features of the watches)

Left

○

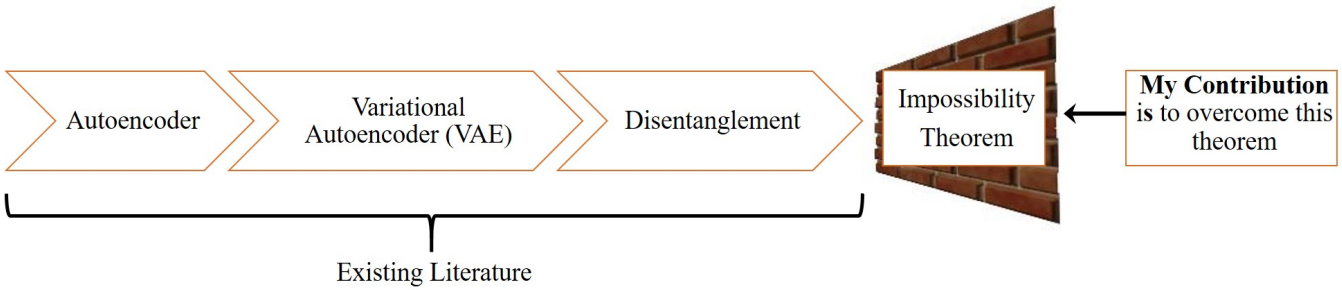
Right

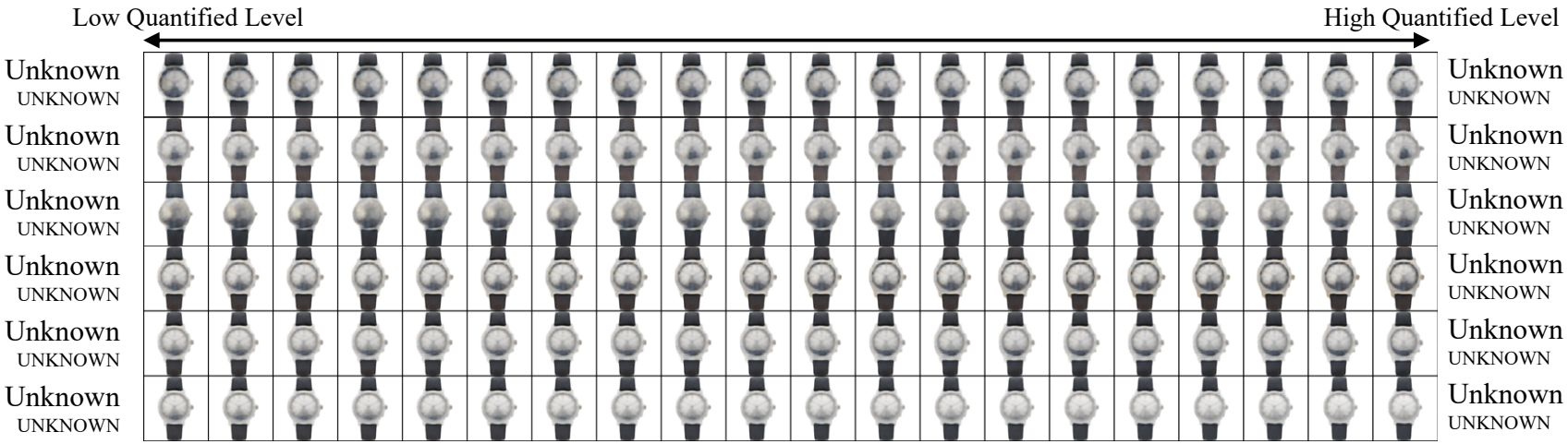
○

# Human Interpretation of Visual Characteristics and Quantification

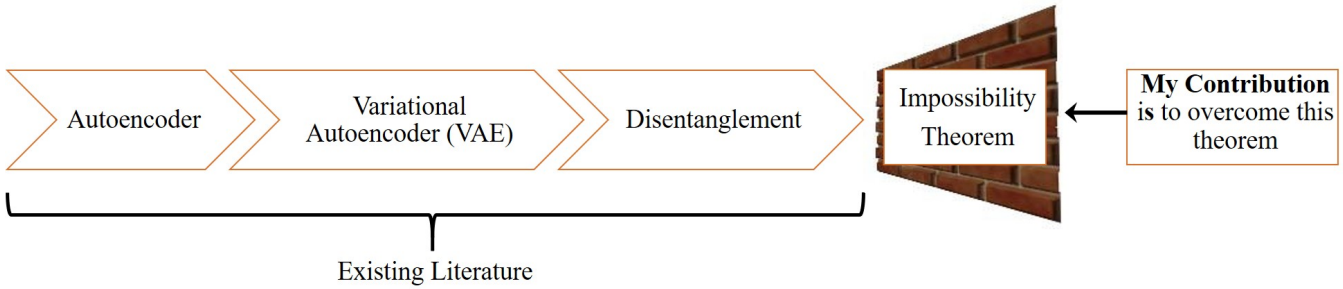
| Visual characteristic | Semantic Meaning | Quantification |
|---|---|---|
| Dial Size | 81% | 83% |
| Dial Color | 84% | 92% |
| Strap Color | 96% | 92% |
| Rim Color | 90% | 88% |
| Dial Shape | 91% | 68% |
| Knob Size | 73% | 85% |

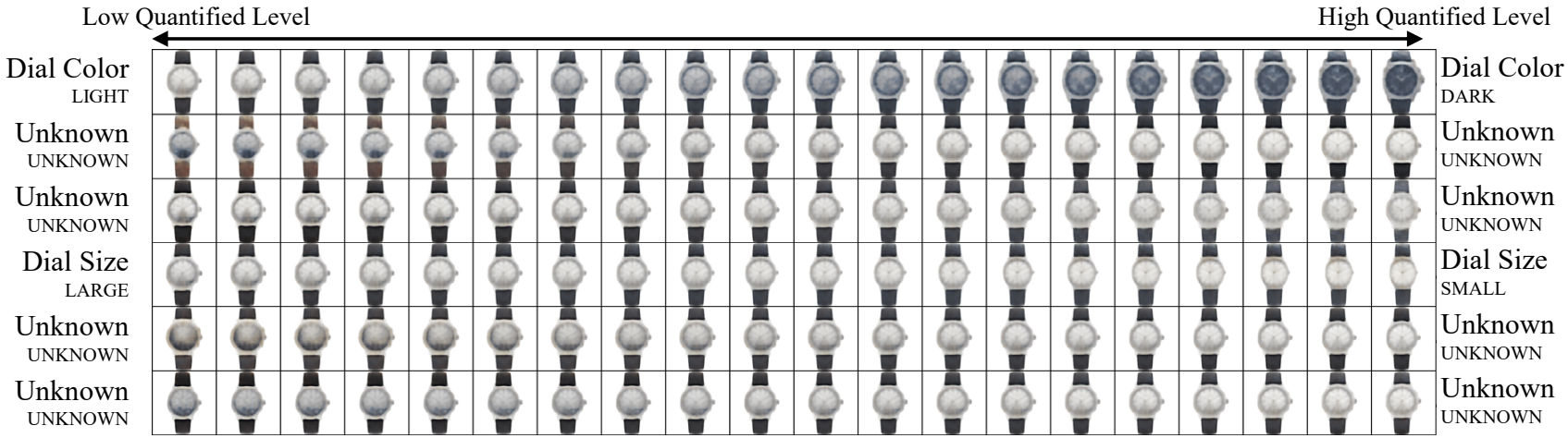
# Discovered Visual characteristics from Autoencoders



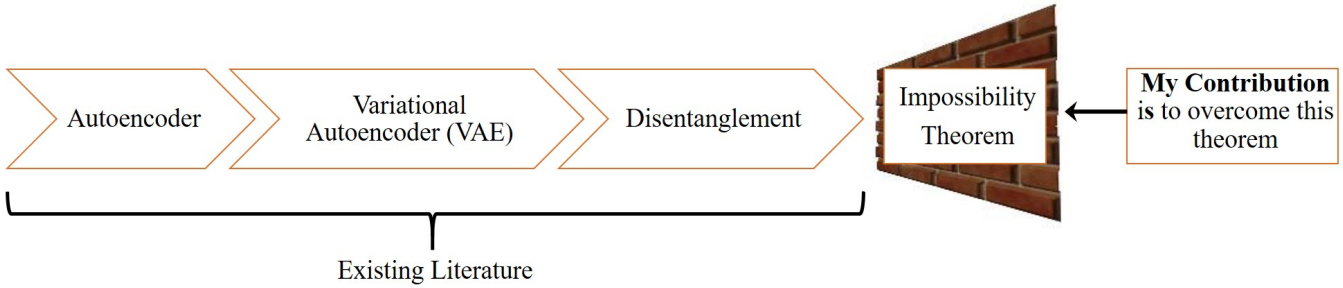



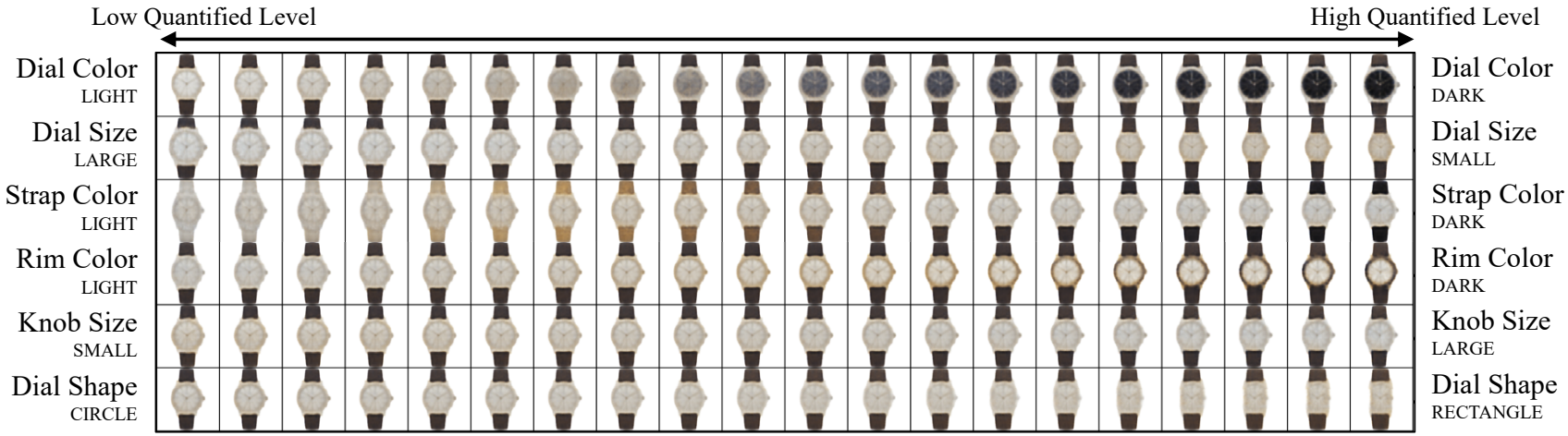
# Discovered Visual characteristics from Variational Autoencoders



# Discovered Visual characteristics from Disentanglement Model



# Example choice-based conjoint (CBC) question in conjoint survey.

Consider the two watches below that vary **only on visual style**. Of these two, which watch would you prefer more (for yourself)?

Select

Select

Next

# Conjoint Survey Design Elements

| Stage | Name | Purpose |
|---|---|---|
| 1 | Introduction | Explain purpose of study and obtain consent. |

# Conjoint Survey Design Elements

| Stage | Name | Purpose |
|---|---|---|
| 1 | Introduction | Explain purpose of study and obtain consent. |
| 2 | Category Identification | Open-ended questions to determine whether respondents were able to identify what category (e.g. shoes) a blurry image belonged to. |

# Conjoint Survey Design Elements

| Stage | Name | Purpose |
|---|---|---|
| 1 | Introduction | Explain purpose of study and obtain consent. |
| 2 | Category Identification | Open-ended questions to determine whether respondents were able to identify what category (e.g. shoes) a blurry image belonged to. |
| 3 | Instructional Manipulation Check (IMC) | Attention check “trap question” for post-hoc respondent filtering. |

# Conjoint Survey Design Elements

| Stage | Name | Purpose |
|---|---|---|
| 1 | Introduction | Explain purpose of study and obtain consent. |
| 2 | Category Identification | Open-ended questions to determine whether respondents were able to identify what category (e.g. shoes) a blurry image belonged to. |
| 3 | Instructional Manipulation Check (IMC) | Attention check "trap question" for post-hoc respondent filtering. |
| 4 | Choice-Based Conjoint (CBC) Instructions | Explain upcoming conjoint choice question tasks with instructions to choose based only on visual style. |

# Conjoint Survey Design Elements

| Stage | Name | Purpose |
|---|---|---|
| 1 | Introduction | Explain purpose of study and obtain consent. |
| 2 | Category Identification | Open-ended questions to determine whether respondents were able to identify what category (e.g. shoes) a blurry image belonged to. |
| 3 | Instructional Manipulation Check (IMC) | Attention check “trap question” for post-hoc respondent filtering. |
| 4 | Choice-Based Conjoint (CBC) Instructions | Explain upcoming conjoint choice question tasks with instructions to choose based only on visual style. |
| 5 | “Warm Up” CBC Practice | Help respondents understand the range of watch designs before making real choices. |

# Conjoint Survey Design Elements

| Stage | Name | Purpose |
|---|---|---|
| 1 | Introduction | Explain purpose of study and obtain consent. |
| 2 | Category Identification | Open-ended questions to determine whether respondents were able to identify what category (e.g. shoes) a blurry image belonged to. |
| 3 | Instructional Manipulation Check (IMC) | Attention check “trap question” for post-hoc respondent filtering. |
| 4 | Choice-Based Conjoint (CBC) Instructions | Explain upcoming conjoint choice question tasks with instructions to choose based only on visual style. |
| 5 | “Warm Up” CBC Practice | Help respondents understand the range of watch designs before making real choices. |
| 6 | 15 CBC questions | Elicit respondent choice of preferred watch design |

# Conjoint Survey Design Elements

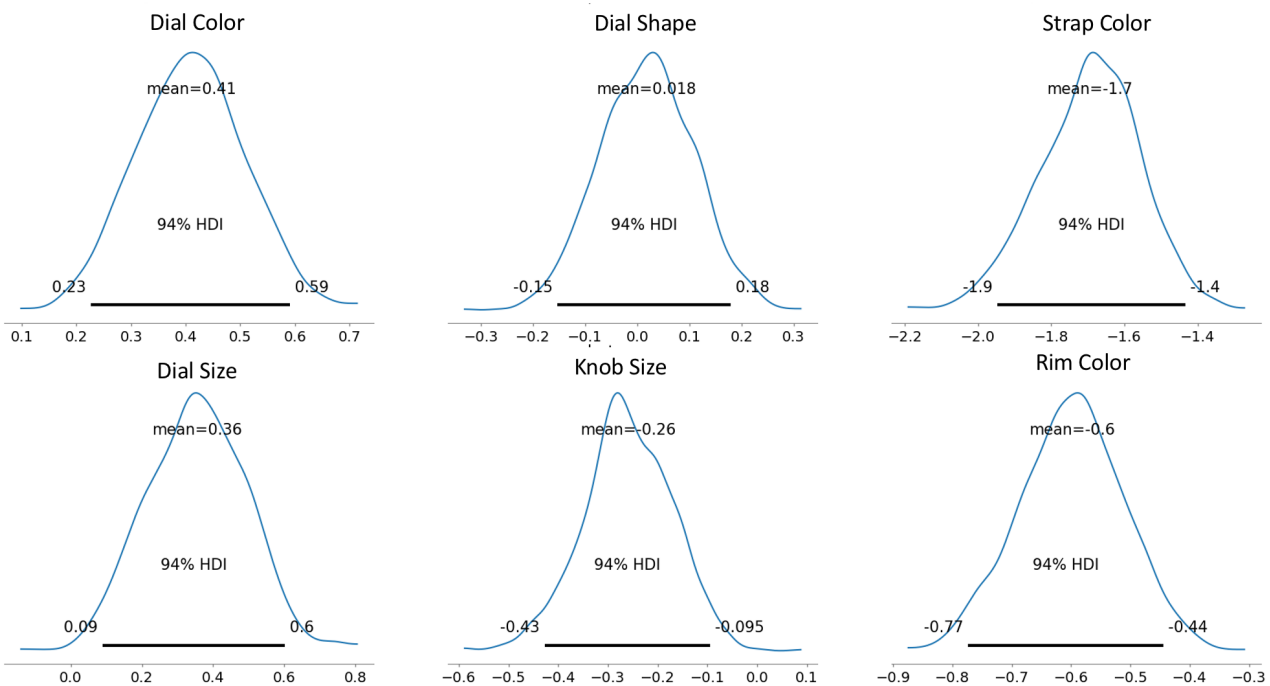
| Stage | Name | Purpose |
|---|---|---|
| 1 | Introduction | Explain purpose of study and obtain consent. |
| 2 | Category Identification | Open-ended questions to determine whether respondents were able to identify what category (e.g. shoes) a blurry image belonged to. |
| 3 | Instructional Manipulation Check (IMC) | Attention check “trap question” for post-hoc respondent filtering. |
| 4 | Choice-Based Conjoint (CBC) Instructions | Explain upcoming conjoint choice question tasks with instructions to choose based only on visual style. |
| 5 | “Warm Up” CBC Practice | Help respondents understand the range of watch designs before making real choices. |
| 6 | 15 CBC questions | Elicit respondent choice of preferred watch design |
| 7 | Respondent Information | Obtain demographic and psychographic variables |

# Utility Model

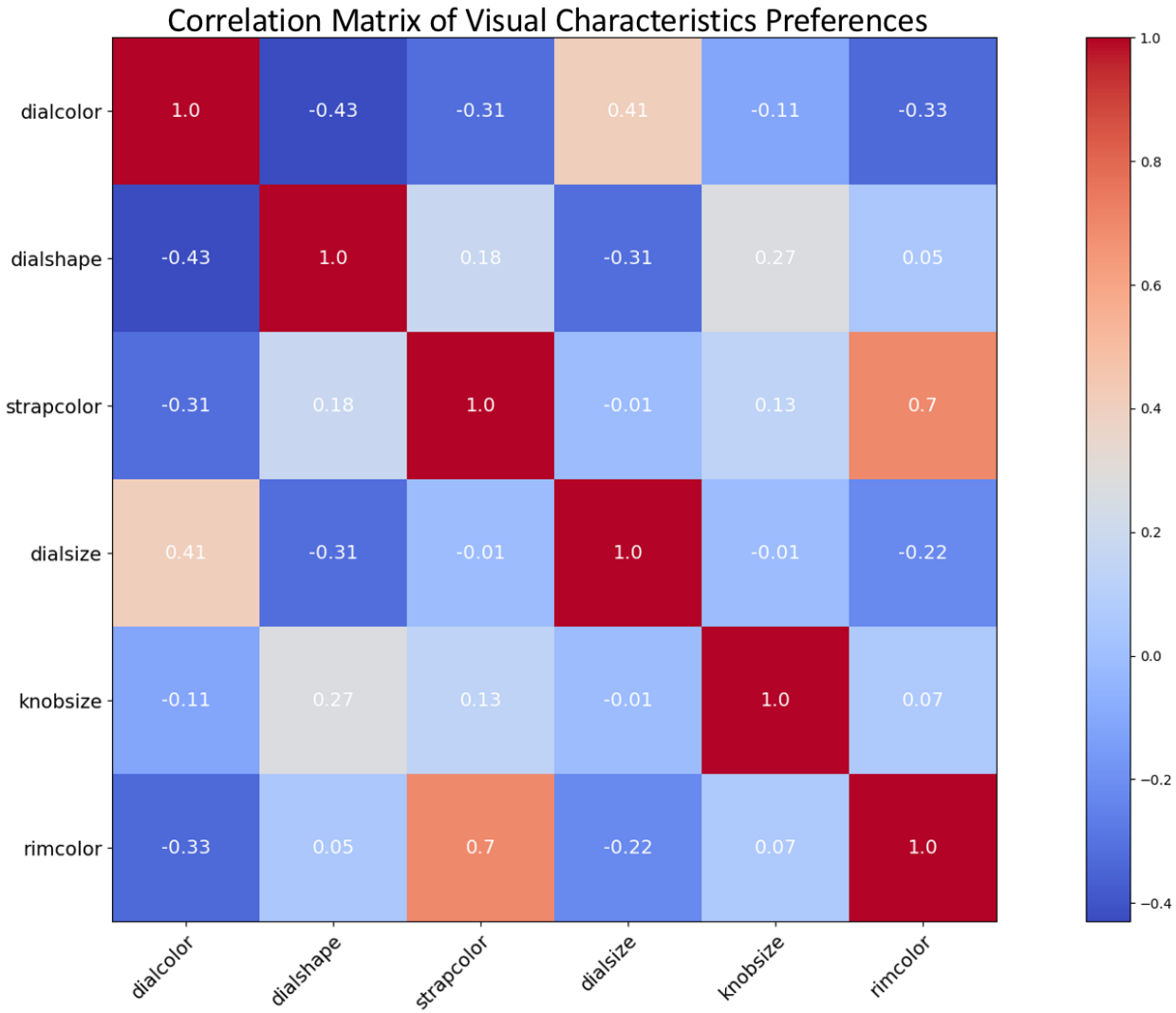
$$
\begin{aligned}
\bar{\beta}_\Theta &\sim \mathcal{N}(\mathbf{0}, \sigma_\Theta^2) \\
\Sigma &\sim \mathcal{N}(\bar{\beta}_\Theta, \Sigma_\Theta) \\
\Sigma_{\mathbf{fi}} &\sim \mathrm{LKJ}(\eta) \\
\Sigma_\beta &= \mathbf{D}(\sigma_\beta)\Sigma_\beta\mathbf{D}(\sigma_\beta) \\
\mathbf{fi}_i &\sim \mathcal{N}(\Sigma^T \mathbf{r}_i, \Sigma_\beta) \\
u_i^j &= z_j\beta_i + \epsilon_{ij} \\
y_i^{j,j'} &\sim \mathrm{Bernoulli}(\omega_i(j,j')) \\
\text{where} \quad \omega_i(j,j') &= \frac{\exp(u_i^j)}{\exp(u_i^j) + \exp(u_i^{j'})}
\end{aligned}
$$

where $\mathrm{LKJ}(\eta)$ is a Cholesky factorization of the correlation matrix $\Sigma_{\mathbf{fi}}$ of the individual "part-worth" preference vector over visual characteristics [Lewandowski, Kurowicka, and Joe, 2009]. $\mathbf{D}(\cdot)$ denotes a diagonal matrix, $\mathbf{r}_i$ are consumer covariates, $u_i^j$ is the utility customer $i$ gets from watch design $j$, and $\epsilon_{ij}$ is a Gumbel random variable. The Bernoulli probability parameter $\omega_i(j,j')$ is specified by the logit function, and $\{j,j'\}_i$ denotes the set of all pairwise choice comparisons for watches $j, j' \in J$ that customer $i$ chose over in the conjoint survey. Note that $\sigma_\Theta^2$, $\Sigma_\Theta$, $\eta$ are researcher-defined hyperparameters chosen via model selection using prediction accuracy on the validation data split as the evaluation metric.

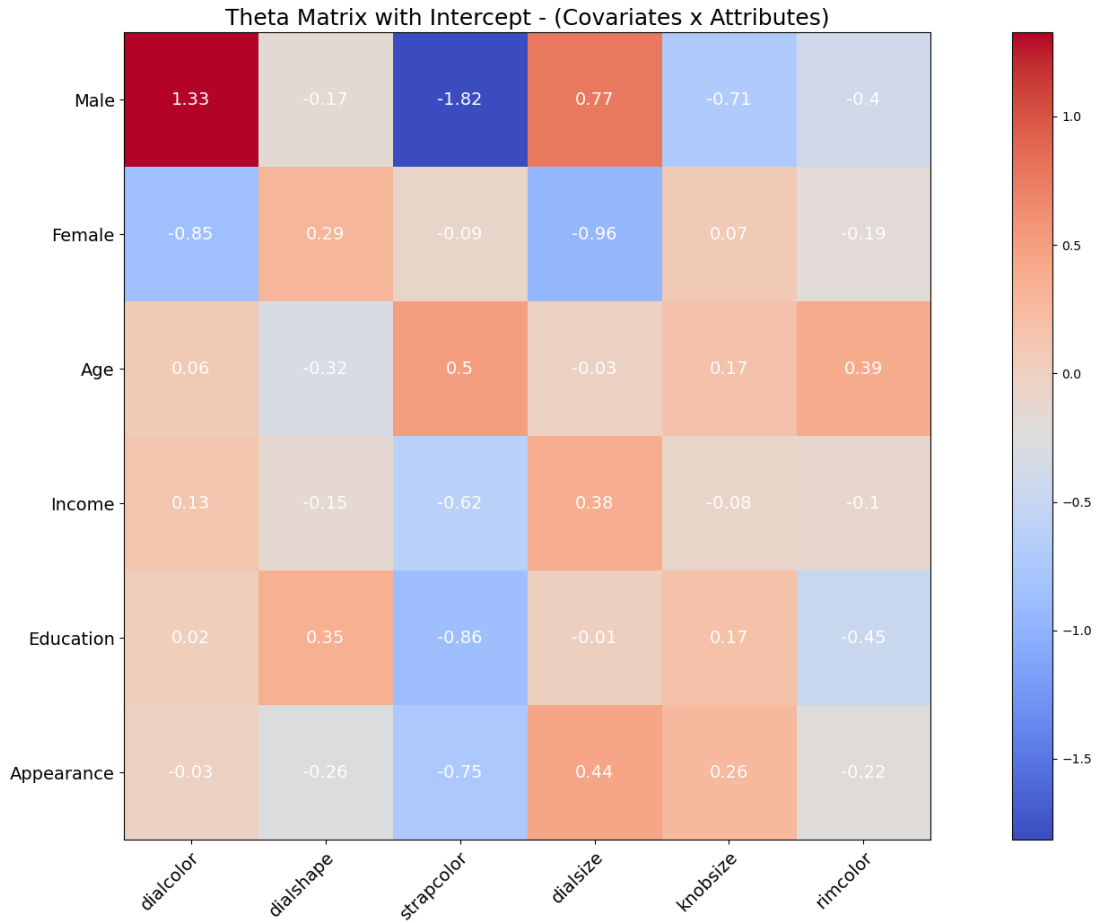
# Posterior Distributions of Population-Level Preference Coefficients $\beta$



# Correlation of population-level preference parameters $\beta$



# Heatmap of expectation of theta matrix relating consumer covariates with preferences over visual characteristics.



# Conjoint Model Accuracy (Generated Watches)

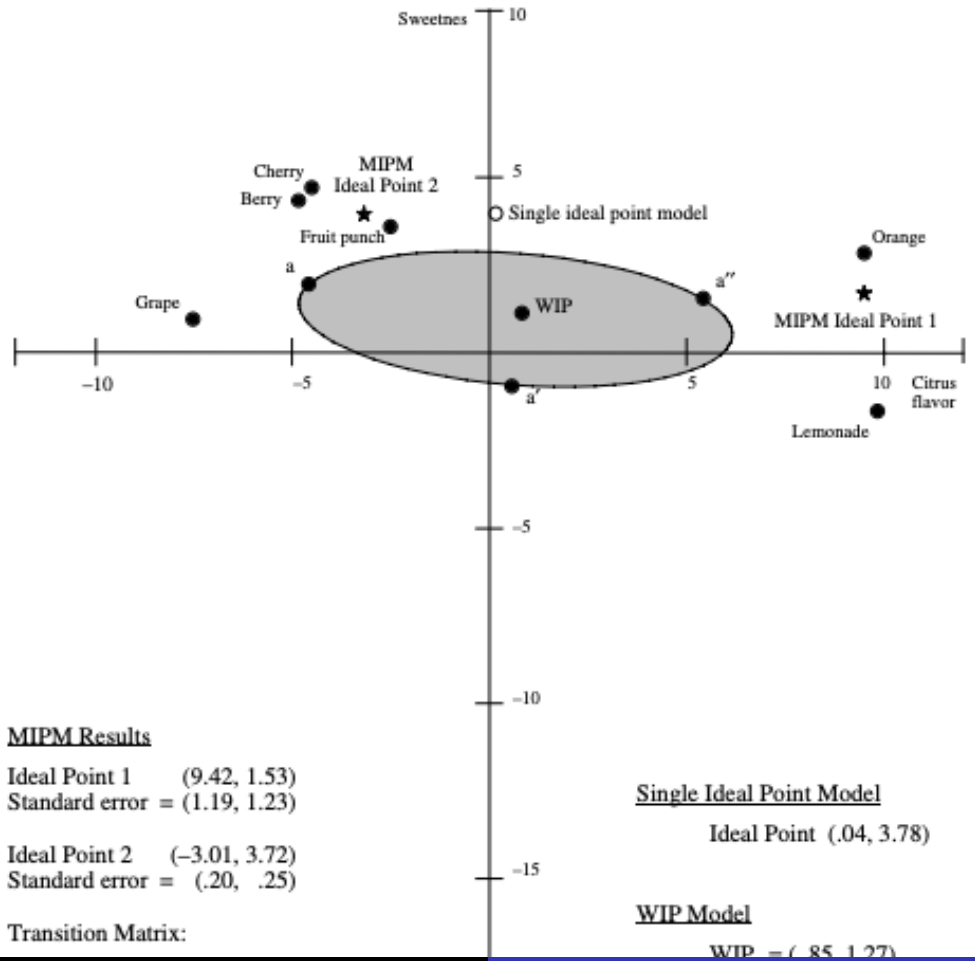
| Model | Hit Rate | (Std. Dev.) |
|---|---|---|
| Logit Model (Homogeneous) | 63.16% | (2.34%) |
| Pretrained Deep Learning Model (Heterogeneous) | 68.31% | (1.54%) |
| HB Model (Heterogeneous) | 72.33% | (0.85%) |

- Pretrained Deep learning model is trained on millions of images, and has millions of parameters
- Our model also has lots of parameters, but all predictions are based on only 6 visual characteristics

# Ideal Points

- What do consumers in a segment prefer most?

Figure 2
MIPM RESULTS FOR HOUSEHOLD 057



# Generated “Ideal Point” Watches for Two Segments

Segment 1:
“Ideal Point” Watch Design

Segment 2:
“Ideal Point” Watch Design

# The End

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013), "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, 35 (8), 1798–1828.

Burgess, C., I. Higgins, A. Pal, Loic Matthey, Nick Watters, G. Desjardins, and Alexander Lerchner "Understanding disentangling in $\beta$-VAE," "Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems," (2017).

Chen, Ricky T. Q., Xuechen Li, Roger B Grosse, and David K Duvenaud "Isolating Sources of Disentanglement in Variational Autoencoders," "Advances in Neural Information Processing Systems," pages 2615–2625 (2018).

Estermann, Benjamin, Markus Marks, and Mehmet Fatih Yanik (2020), "Robust Disentanglement of a Few Factors at a Time using rPU-VAE," *Advances in Neural Information Processing Systems*, 33, 13387–13398.

Ferraro, Rosellina, Amna Kirmani, and Ted Matherly (2013), "Look at me! Look at me! Conspicuous brand usage, self-brand connection, and dilution," *Journal of Marketing Research*, 50 (4), 477–488.

Fleming, Roland W (2014), "Visual perception of materials and their properties," *Vision research*, 94, 62–75.

Gabbay, Aviv, Niv Cohen, and Yedid Hoshen (2021), "An image is worth more than a thousand words: Towards disentanglement in the wild," *Advances in Neural Information Processing Systems*, 34.

Karras, Tero, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila "Training Generative Adversarial Networks with Limited Data," "Advances in Neural Information Processing Systems," Vol. 33., pages 12104–12114 (2020).

Kulkarni, Tejas D., William F. Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum "Deep convolutional inverse graphics

network," "Advances in Neural Information Processing Systems," pages 2539–2547 (2015).

Landwehr, Jan R, Aparna A Labroo, and Andreas Herrmann (2011), "Gut liking for the ordinary: Incorporating design fluency improves automobile sales forecasts," *Marketing Science*, 30 (3), 416–429.

Lee, Jung Eun, Songyee Hur, and Brandi Watkins (2018), "Visual communication of luxury fashion brands on social media: effects of visual complexity and brand familiarity," *Journal of Brand Management*, 25, 449–462.

Lee, Wonkwang, Donggyun Kim, Seunghoon Hong, and Honglak Lee "High-fidelity synthesis with disentangled representation," "European Conference on Computer Vision," pages 157–174, Springer (2020).

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe (2009), "Generating random correlation matrices based on vines and extended onion method," *Journal of multivariate analysis*, 100 (9), 1989–2001.

Liu, Liu, Daria Dzyabura, and Natalie Mizik (2020), "Visual listening in: Extracting brand image portrayed on social media," *Marketing Science*, 39 (4), 669–686.

Locatello, Francesco, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen "Weakly-supervised disentanglement without compromises," "International Conference on Machine Learning," pages 6348–6359, PMLR (2020).

Locatello, Francesco, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem "Disentangling Factors of Variations Using Few Labels," "International Conference on Learning Representations," (2020).

Megehee, Carol M and Deborah F Spake (2012), "Consumer enactments of archetypes using luxury brands," *Journal of business research*, 65 (10), 1434–1442.

Simonson, Alex and Bernd H Schmitt (1997), *Marketing*

*aesthetics: The strategic management of brands, identity, and image* Simon and Schuster.