Comments: Seeing is Believing

- Piqueras et al., "An image is worth K topics"
- Sunderland and Morucci, "Counterfactual inpainting"
- Arnold et al., "Measuring media slant"

10 Sep 2025

Dean Knox **Penn**

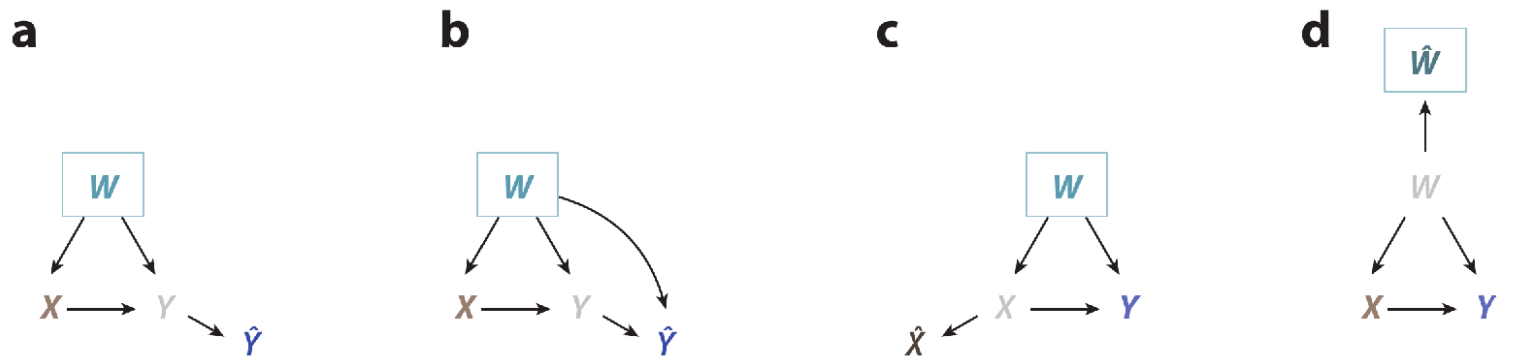


Figure 3

Causal structures of theory and measurement. Data environments correspond to the theory of **Figure 1**. Panels *a* and *b* illustrate environments in which analysts are unable to directly observe the outcome Y and thus must resort to a learned measure $\hat{Y}$ that is either (*a*) uncontaminated or (*b*) contaminated by confounders. The other two panels depict cases in which (*c*) the treatment X or (*d*) the confounders W cannot be observed, so that analysts can only adjust for learned proxies ($\hat{X}$ or $\hat{W}$).

Source: Dean Knox, Christopher Lucas, and Wendy Tam Cho. 2022. "Testing Causal Theories with Learned Proxies, "*Annual Review of Political Science*.

Piqueras et al.,

"An image is worth K topics"

Piqueras et al.,
"An image is worth K topics"

- Each image $i \in \{1, \ldots, N\}$
  - Has treatment variables $\vec{x}_i$...

Piqueras et al.,
"An image is worth K topics"

- Each image $i \in \{1, \ldots, N\}$
  - Has treatment variables $\vec{x}_i$...
  - ... causing it to belong to topic k with proportion θ_{ik}

Piqueras et al.,
"An image is worth K topics"

- Each image $i \in \{1, \ldots, N\}$
  - Has treatment variables $\vec{x}_i$...
  - ... causing it to belong to topic k with proportion θ_{ik}
- Topic $k \in \{1, \ldots, K\}$ has embedding $\vec{\beta}_k \in \mathbb{R}^D$
  - Represents position of image that is "pure" instance of k

Piqueras et al., "An image is worth K topics"

- Each image $i \in \{1, \ldots, N\}$
  - Has treatment variables $\vec{x}_i$...
  - ... causing it to belong to topic k with proportion θ_{ik}
- Topic $k \in \{1, \ldots, K\}$ has embedding $\vec{\beta}_k \in \mathbb{R}^D$
  - Represents position of image that is "pure" instance of k
- Then, image i has embedding position $\vec{z}_i \in \mathbb{R}^D$ that is convex combination of $\vec{\beta}_k$ with weight θ_{ik} + noise

Piqueras et al., "An image is worth K topics"

- Each image $i \in \{1, \ldots, N\}$
  - Has treatment variables $\vec{x}_i$...
  - ... causing it to belong to topic k with proportion θ_{ik}
- Topic $k \in \{1, \ldots, K\}$ has embedding $\vec{\beta}_k \in \mathbb{R}^D$
  - Represents position of image that is "pure" instance of k
- Then, image i has embedding position $\vec{z}_i \in \mathbb{R}^D$ that is convex combination of $\vec{\beta}_k$ with weight θ_{ik} + noise
- **QOI:** coefs. of multinomial logit regression $\vec{\theta}_i \sim x_i$

Some brief comments

- There are *two* relevant notions of embeddings!
  - Unobserved "true" embeddings $\vec{z}_i$ which create $\mathbf{Image}_i$
  - Estimated $\hat{\vec{z}}_i$, a projection of $\mathbf{Image}_i$ from pretrained model
- "True" $\vec{z}_i$ theorized to be causally downstream from "true" unit-level outcome, the topic prop. $\vec{\theta}_i \in \Delta_K$

Some brief comments

- There are *two* relevant notions of embeddings!
  - Unobserved "true" embeddings $\vec{z}_i$ which create $\mathbf{Image}_i$
  - Estimated $\hat{\vec{z}}_i$, a projection of $\mathbf{Image}_i$ from pretrained model
- "True" $\vec{z}_i$ theorized to be causally downstream from "true" unit-level outcome, the topic prop. $\vec{\theta}_i \in \Delta_K$
- Outcome subgraph is thus $\vec{\theta} \rightarrow z \rightarrow \mathbf{Image} \rightarrow \hat{z}$
  - If there is direct treatment contamination $x \rightarrow \mathbf{Image}$, we cannot infer that $\mathbf{Corr}(x, \hat{z}) \neq 0$ implies $x \rightarrow \vec{\theta}$ causation

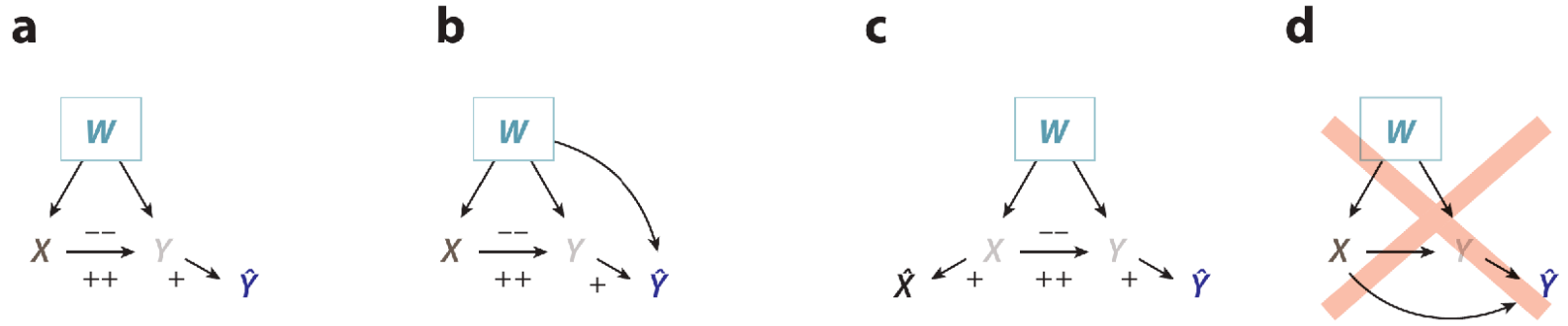


Figure 5

Learned outcomes. In the settings shown, the true outcome Y is unknown, but the proxy $\hat{Y}$ can be estimated from auxiliary information. In all cases, both confounders W and treatment X are known, and the analyst adjusts for W. (*a*) In a simple case, Y has an on-average monotonic effect on an uncontaminated proxy $\hat{Y}$, and the $X \rightarrow Y$ effect is known to exhibit distributional monotonicity of an unknown direction (indicated by the presence of both positive and negative signs). In this case, positive (negative) distributional monotonicity in $X \rightarrow Y$ is guaranteed to produce weakly positive (negative) $\text{Cov}(\hat{X}, \hat{Y}|W)$, and the sign of the $X \rightarrow Y$ effect can be identified. (*b*) This result holds even when $\hat{Y}$ is contaminated by W, so long as these contextual factors are adjusted for in a subsequent regression. (*c*) The same result still holds when X is also imperfectly, but on-average monotonically, learned. (*d*) However, if the proxy is contaminated by the treatment itself, association between X and $\hat{Y}$ cannot be interpreted as evidence for the theorized $X \rightarrow Y$ effect.

Source: Dean Knox, Christopher Lucas, and Wendy Tam Cho. 2022. "Testing Causal Theories with Learned Proxies, "*Annual Review of Political Science*.

Some brief comments

- There are *two* relevant notions of embeddings!
  - Unobserved "true" embeddings $\vec{z}_i$ which create $\mathbf{Image}_i$
  - Estimated $\hat{\vec{z}}_i$, a projection of $\mathbf{Image}_i$ from pretrained model
- "True" $\vec{z}_i$ theorized to be causally downstream from "true" unit-level outcome, the topic prop. $\vec{\theta}_i \in \Delta_K$
- Outcome subgraph is thus $\vec{\theta} \rightarrow z \rightarrow \mathbf{Image} \rightarrow \hat{z}$
  - If there is direct treatment contamination $x \rightarrow \mathbf{Image}$, we cannot infer that $\mathbf{Corr}(x, \hat{z}) \neq 0$ implies $x \rightarrow \vec{\theta}$ causation

Some brief comments

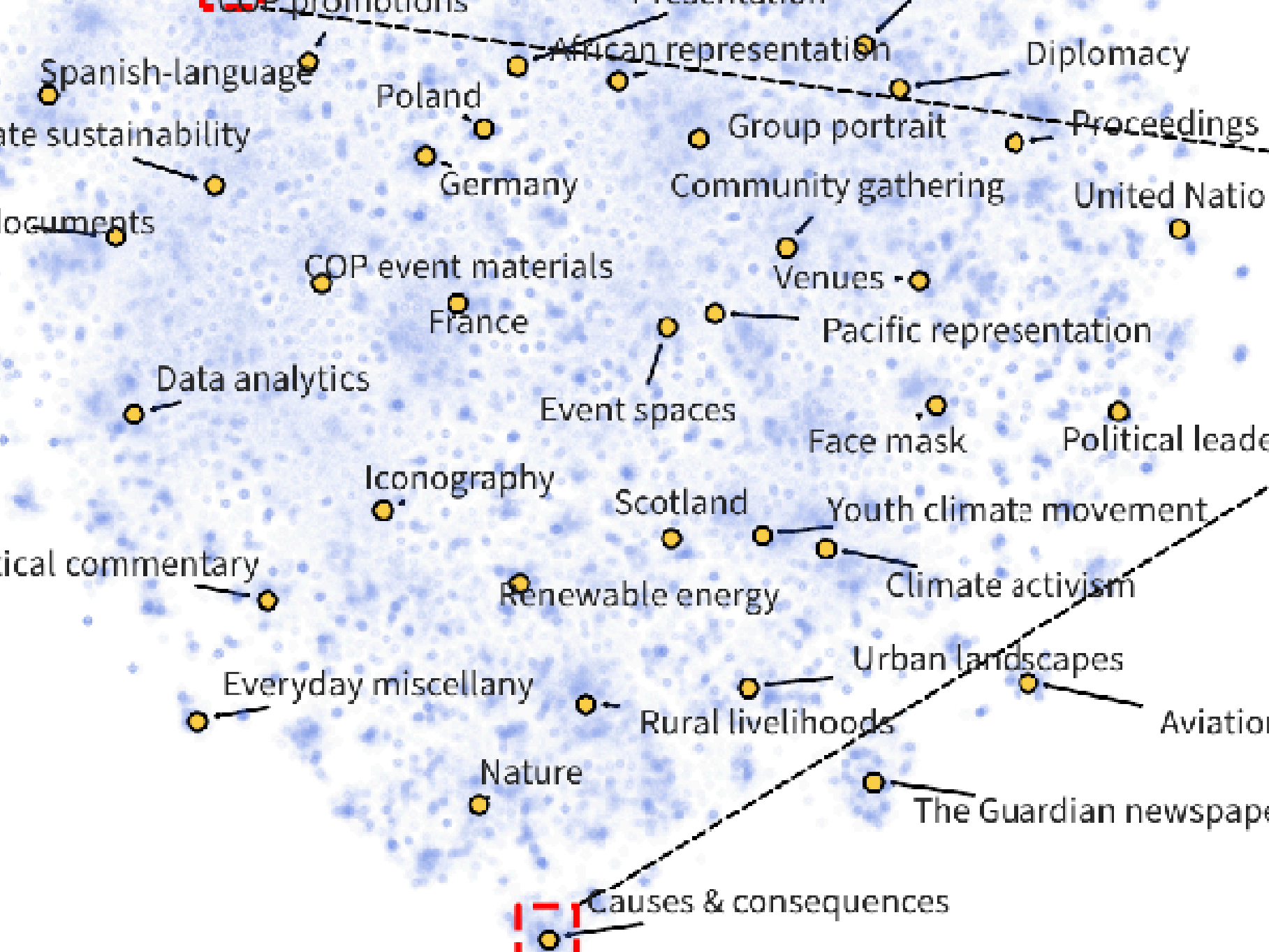
- May be worth thinking about statistical identification
  - CLIP space has $D = 768$ for 400M web images
  - Effective dimensionality of COP social media images is far lower (paper refers to $\tilde{D} \approx 6$ "principal visual dimensions")

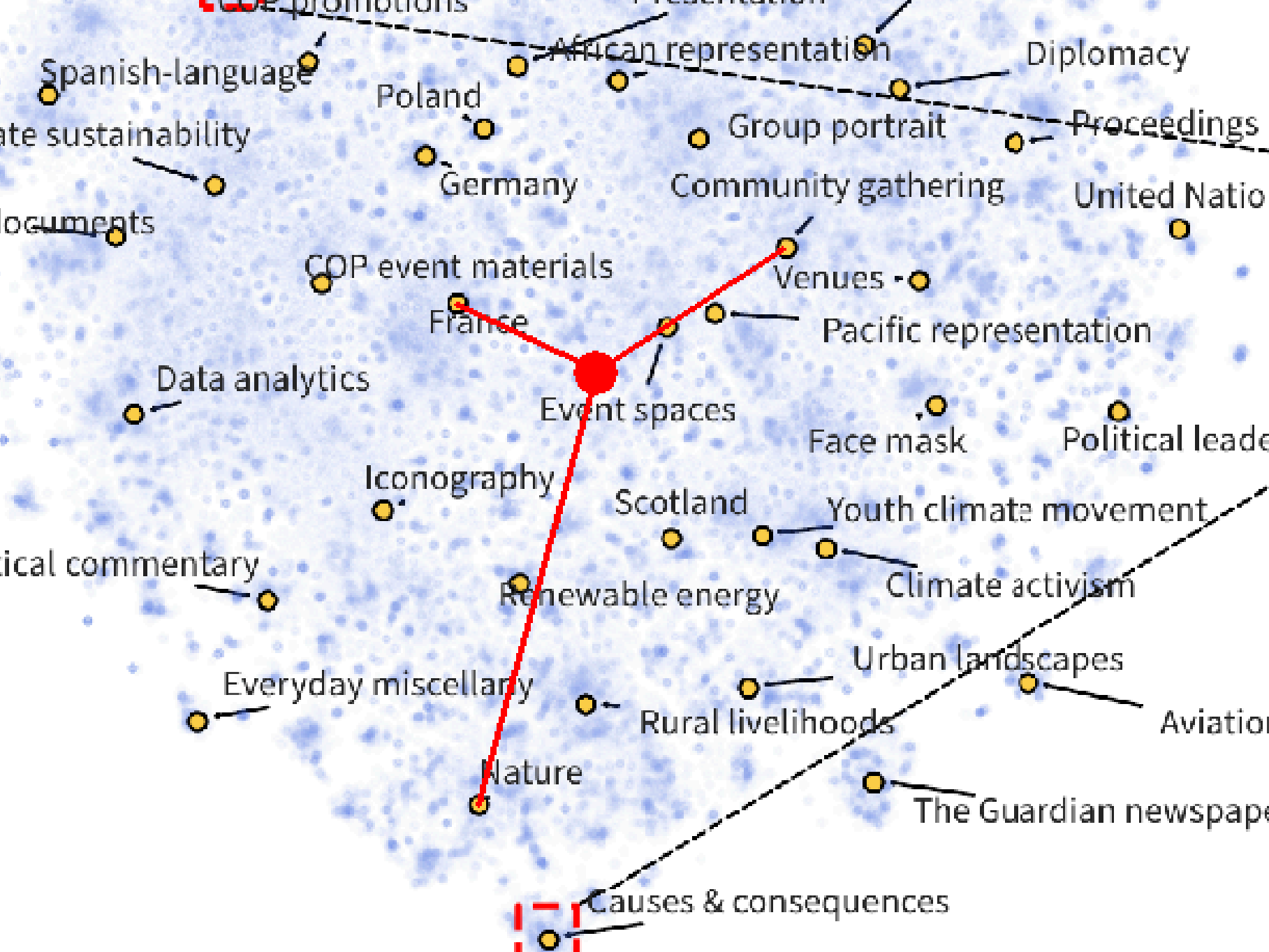
Some brief comments

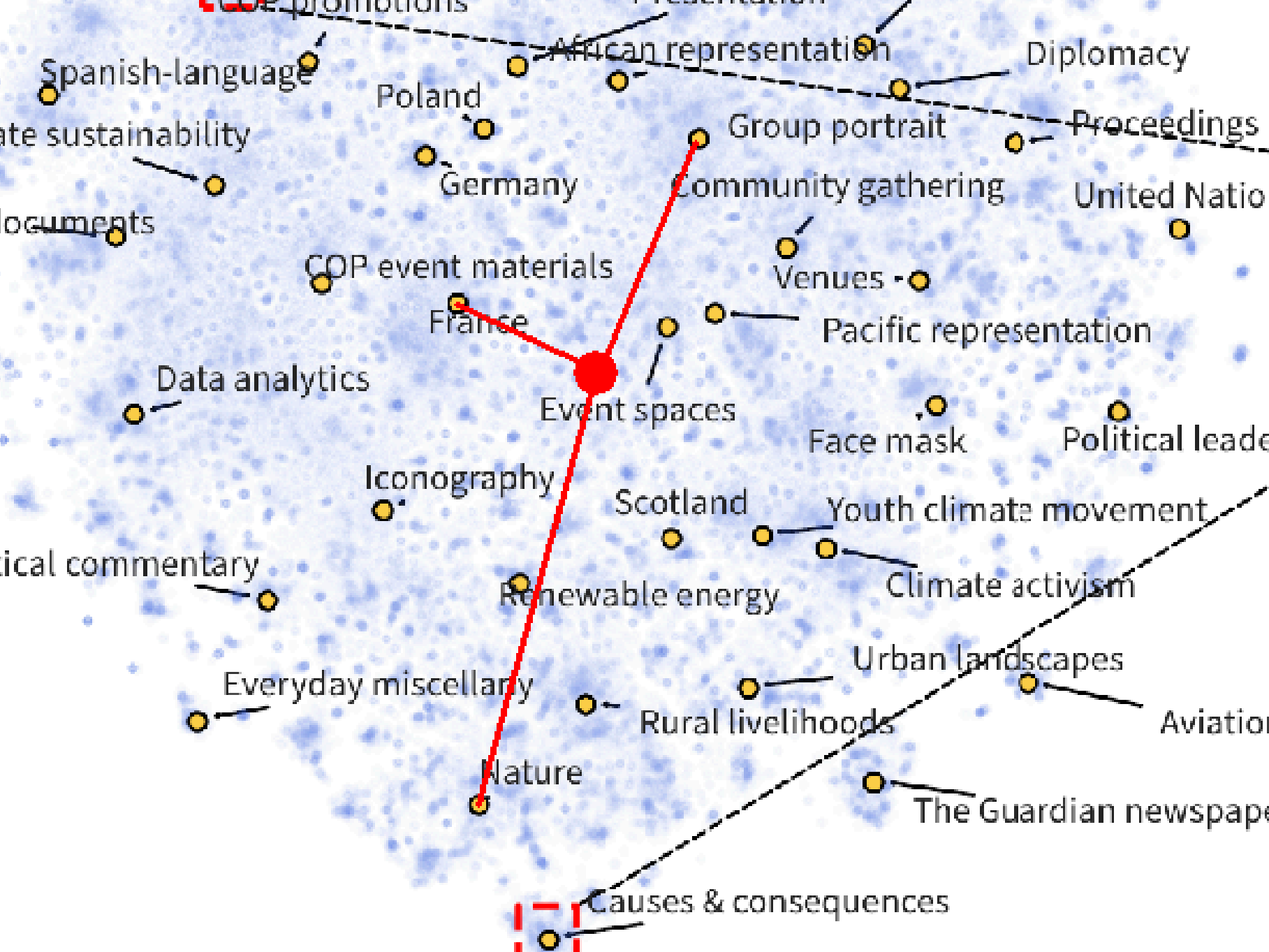
- May be worth thinking about statistical identification
  - CLIP space has $D = 768$ for 400M web images
  - Effective dimensionality of COP social media images is far lower (paper refers to $\tilde{D} \approx 6$ "principal visual dimensions")
- Model expresses $\vec{z}_i$ as convex combination of $K = 45$ latent topics, so $K \gg \tilde{D}$
  - But by Carathéodory's theorem, any $\vec{\theta}_i$ can be equally well expressed by combination of $\tilde{D} + 1$ points on hull of $\vec{\theta}$s

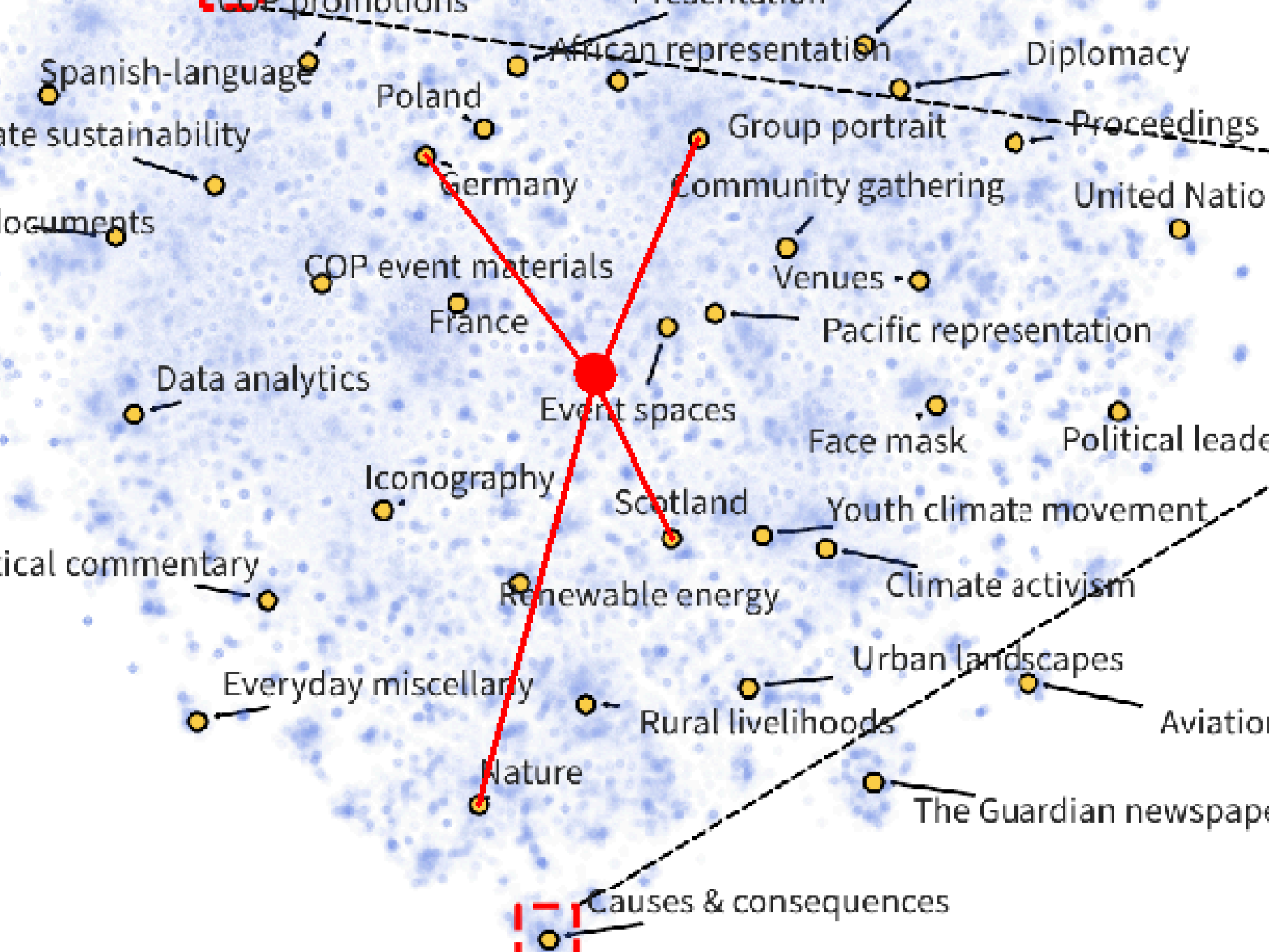
Some brief comments

- May be worth thinking about statistical identification
  - CLIP space has $D = 768$ for 400M web images
  - Effective dimensionality of COP social media images is far lower (paper refers to $\tilde{D} \approx 6$ "principal visual dimensions")
- Model expresses $\vec{z}_i$ as convex combination of $K = 45$ latent topics, so $K \gg \tilde{D}$
  - But by Carathéodory's theorem, any $\vec{\theta}_i$ can be equally well expressed by combination of $\tilde{D} + 1$ points on hull of $\vec{\theta}$s
- Key estimates seem heavily shaped by regularization
  - Without it, $\vec{\beta}$ slide around freely; not constrained to lie on hull
  - Many choices for which $\tilde{D} + 1$ elements of $\vec{\theta}_i$ are non-zero


Spanish-language
Poland
African representation
Diplomacy
Proceedings
Group portrait
Germany
Community gathering
United Natio
COP event materials
Venues
France
Pacific representation
Data analytics
Event spaces
Face mask
Political leade
Iconography
Scotland
Youth climate movement
Climate activism
Renewable energy
Urban landscapes
Everyday miscellany
Rural livelihoods
Aviation
Nature
The Guardian newspape
Causes & consequences


Spanish-language
African representation
Diplomacy
Poland
Group portrait
Proceedings
Germany
Community gathering
United Natio
COP event materials
Venues
France
Pacific representation
Data analytics
Event spaces
Face mask
Political leade
Iconography
Scotland
Youth climate movement
Renewable energy
Climate activism
Urban landscapes
Everyday miscellany
Rural livelihoods
Aviation
Nature
The Guardian newspape
Causes & consequences


Spanish-language
Poland
African representation
Diplomacy
ate sustainability
Group portrait
Proceedings
Germany
Community gathering
United Natio
documents
COP event materials
Venues
France
Pacific representation
Data analytics
Event spaces
Face mask
Political leade
Iconography
Scotland
Youth climate movement
ical commentary
Renewable energy
Climate activism
Urban landscapes
Everyday miscellany
Rural livelihoods
Aviation
Nature
The Guardian newspape
Causes & consequences


African representation
Diplomacy
Spanish-language
Poland
Proceedings
Group portrait
Germany
Community gathering
United Natio
COP event materials
Venues
France
Pacific representation
Data analytics
Event spaces
Face mask
Political leade
Iconography
Scotland
Youth climate movement
Climate activism
Renewable energy
Urban landscapes
Everyday miscellany
Rural livelihoods
Aviation
Nature
The Guardian newspape
Causes & consequences

Sunderland and Morucci,
"Counterfactual inpainting"

Sunderland and Morucci, "Counterfactual inpainting"

- Image I composed of *spatially separable*:
  - Background confounders X
  - High-level treatment concept T (indirectly)
- T can manifest via many graphical renderings A
- Viewers consume image I and produce outcome Y

Figure 3: Causal Graph for Inference with Graphical Treatments

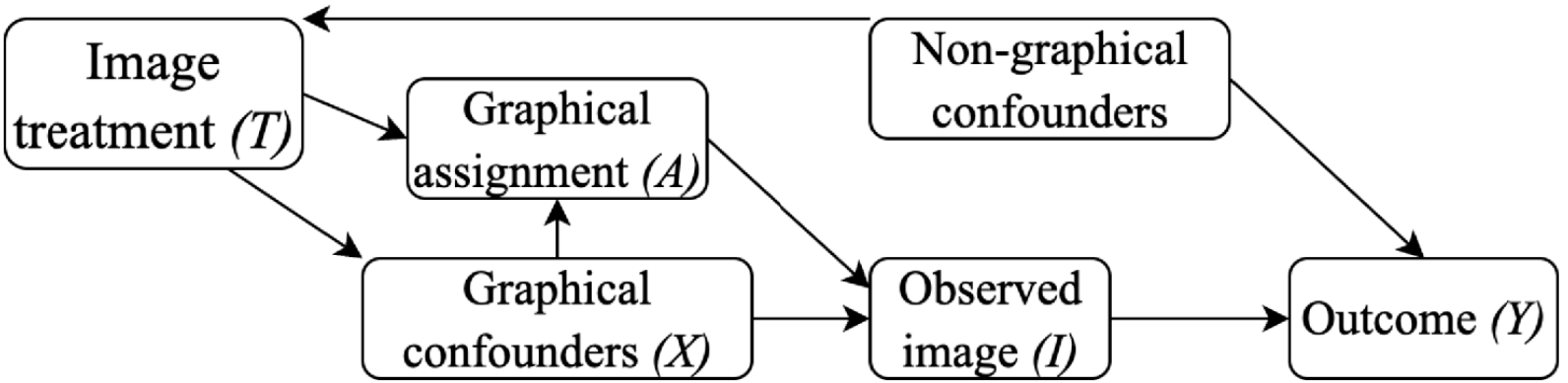
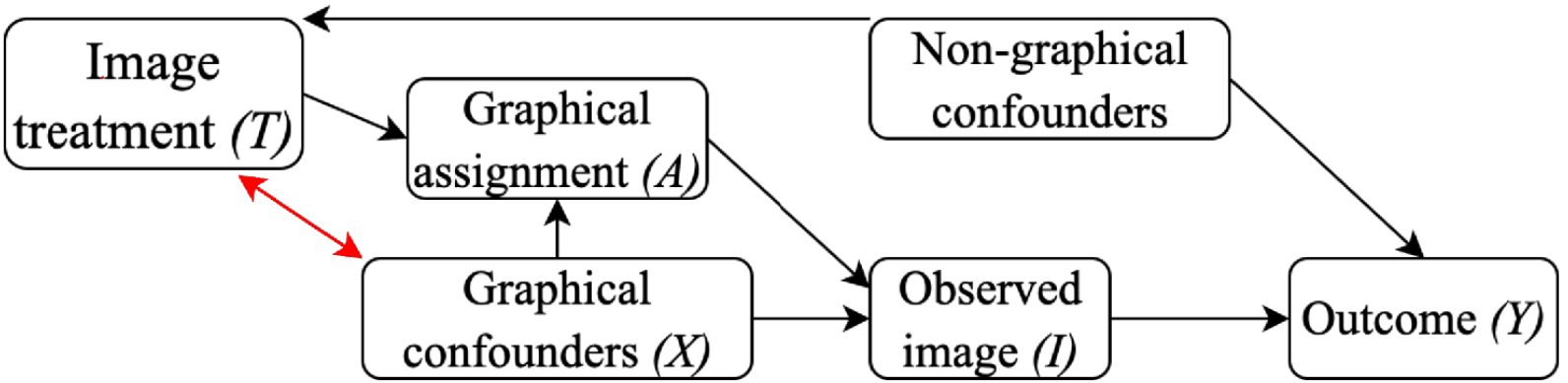


Figure 3: Causal Graph for Inference with Graphical Treatments



Some brief comments

- One reasonable "ATE" is of T on Y, i.e. $\mathbb{E}[Y(t=1) - Y(t=0)]$
- Quantity defined in paper is currently not an "ATE" $\int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_t}(a) - \int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_c}(a)$

Some brief comments

- One reasonable "ATE" is of $\boldsymbol{T}$ on $\boldsymbol{Y}$, i.e. $\mathbb{E}[\boldsymbol{Y}(\boldsymbol{t}=\mathbf{1}) - \boldsymbol{Y}(\boldsymbol{t}=\mathbf{0})]$
- Quantity defined in paper is currently not an "ATE" $\int \mathbb{E}[\boldsymbol{Y}(\boldsymbol{a}, \boldsymbol{X})] \mathrm{d}\boldsymbol{p}_{\boldsymbol{A}_t}(\boldsymbol{a}) - \int \mathbb{E}[\boldsymbol{Y}(\boldsymbol{a}, \boldsymbol{X})] \mathrm{d}\boldsymbol{p}_{\boldsymbol{A}_c}(\boldsymbol{a})$
- Exact nature of proposed intervention is unclear

Some brief comments

- One reasonable "ATE" is of T on Y, i.e. $\mathbb{E}[Y(t=1) - Y(t=0)]$
- Quantity defined in paper is currently not an "ATE" $\int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_t}(a) - \int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_c}(a)$
- Exact nature of proposed intervention is unclear
  - $Y(a, X)$ fixes confounder to its observed value

Some brief comments

- One reasonable "ATE" is of T on Y, i.e. $\mathbb{E}[Y(t=1) - Y(t=0)]$
- Quantity defined in paper is currently not an "ATE" $\int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_t}(a) - \int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_c}(a)$
- Exact nature of proposed intervention is unclear
  - $Y(a, X)$ fixes confounder to its observed value
  - Original graph includes the $T \to X \to A$ path; ATE would involve $Y(a, X(a))$ which is presumably not the intent

Some brief comments

- One reasonable "ATE" is of T on Y, i.e. $\mathbb{E}[Y(t=1) - Y(t=0)]$
- Quantity defined in paper is currently not an "ATE" $\int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_t}(a) - \int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_c}(a)$
- Exact nature of proposed intervention is unclear
  - $Y(a, X)$ fixes confounder to its observed value
  - Original graph includes the $T \to X \to A$ path; ATE would involve $Y(a, X(a))$ which is presumably not the intent
- Exact nature of $p_{A_t}(a)$ and $p_{A_c}(a)$ not defined

Some brief comments

- One reasonable "ATE" is of T on Y, i.e. $\mathbb{E}[Y(t=1) - Y(t=0)]$
- Quantity defined in paper is currently not an "ATE" $\int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_t}(a) - \int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_c}(a)$
- Exact nature of proposed intervention is unclear
  - $Y(a, X)$ fixes confounder to its observed value
  - Original graph includes the $T \to X \to A$ path; ATE would involve $Y(a, X(a))$ which is presumably not the intent
- Exact nature of $p_{A_t}(a)$ and $p_{A_c}(a)$ not defined
  - Could be misinterpreted as PDFs, perhaps $\mathrm{d}F_{A(t)}(a)$ instead

Some brief comments

- One reasonable "ATE" is of T on Y, i.e. $\mathbb{E}[Y(t=1) - Y(t=0)]$
- Quantity defined in paper is currently not an "ATE" $\int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_t}(a) - \int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_c}(a)$
- Exact nature of proposed intervention is unclear
  - $Y(a, X)$ fixes confounder to its observed value
  - Original graph includes the $T \to X \to A$ path; ATE would involve $Y(a, X(a))$ which is presumably not the intent
- Exact nature of $p_{A_t}(a)$ and $p_{A_c}(a)$ not defined
  - Could be misinterpreted as PDFs, perhaps $\mathrm{d}F_{A(t)}(a)$ instead
  - Unconditional $p_{A(t)}(a)$, not $p_{A(t)}(a|X)$; ignores $X \to A$

Some brief comments

- One reasonable "ATE" is of T on Y, i.e. $\mathbb{E}[Y(t=1) - Y(t=0)]$
- Quantity defined in paper is currently not an "ATE" $\int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_t}(a) - \int \mathbb{E}[Y(a, X)] \mathrm{d}p_{A_c}(a)$
- Exact nature of proposed intervention is unclear
  - $Y(a, X)$ fixes confounder to its observed value
  - Original graph includes the $T \to X \to A$ path; ATE would involve $Y(a, X(a))$ which is presumably not the intent
- Exact nature of $p_{A_t}(a)$ and $p_{A_c}(a)$ not defined
  - Could be misinterpreted as PDFs, perhaps $\mathrm{d}F_{A(t)}(a)$ instead
  - Unconditional $p_{A(t)}(a)$, not $p_{A(t)}(a|X)$; ignores $X \to A$
  - Also seems inconsistent with the proposed implementation

Arnold et al.,
"Measuring media slant"

Arnold et al., "Measuring media slant"

- Somewhat tweaked notation

Arnold et al., "Measuring media slant"

- Somewhat tweaked notation
- Properties of images and image pairs:
  - $\vec{E}_i \in \mathbb{R}^{768}$: embedding of image i in CLIP space
  - $\mathbf{sim}(\vec{E}_i, \vec{E}_j)$: cosine similarity of images i, j

Arnold et al., "Measuring media slant"

- Somewhat tweaked notation
- Properties of images and image pairs:
  - $\vec{E}_i \in \mathbb{R}^{768}$: embedding of image i in CLIP space
  - $\mathbf{sim}(\vec{E}_i, \vec{E}_j)$: cosine similarity of images i, j
- Properties of media outlets and outlet pairs:
  - $O_i \in \{1, \ldots, 10\}$: identity of media outlet posting image i
  - $S_p \in \mathbb{R}^K$: slant of outlet p in 1-dim. latent space
  - $\mathbf{sim}(S_p, S_q)$: multiplicative similarity of outlets p, q

Arnold et al., "Measuring media slant"

- Somewhat tweaked notation
- Properties of images and image pairs:
  - $\vec{E}_i \in \mathbb{R}^{768}$: embedding of image i in CLIP space
  - $\mathbf{sim}(\vec{E}_i, \vec{E}_j)$: cosine similarity of images i, j
- Properties of media outlets and outlet pairs:
  - $O_i \in \{1, \ldots, 10\}$: identity of media outlet posting image i
  - $S_p \in \mathbb{R}^K$: slant of outlet p in 1-dim. latent space
  - $\mathbf{sim}(S_p, S_q)$: multiplicative similarity of outlets p, q
- $\mathbf{sim}(S_{O_i}, S_{O_j})$: sim. of outlets posting images i, j

Some brief comments

- What is real and what is not?

Some brief comments

- What is real and what is not?
  - Which assumptions define the theorized "true" DGP?
  - Which are modeling conveniences that you are agnostic to?

Some brief comments

- What is real and what is not?
  - Which assumptions define the theorized "true" DGP?
  - Which are modeling conveniences that you are agnostic to?
- **Choice 1:** the existence of a "true" embedding space
  - $\boldsymbol{d_i} \in \mathbb{R}^{\mathbf{768}}$ treated as perfectly measured attribute of $\boldsymbol{i}$

Some brief comments

- What is real and what is not?
  - Which assumptions define the theorized "true" DGP?
  - Which are modeling conveniences that you are agnostic to?
- **Choice 1:** the existence of a "true" embedding space
  - $d_i \in \mathbb{R}^{768}$ treated as perfectly measured attribute of i
- **Choice 2:** the additive & multiplicative effects model
  - $\mathbf{sim}(d_i, d_j) = \mathbf{tanh}(\mu + \alpha_{O_i} + \alpha_{O_j} + \mathbf{sim}(O_i, O_j) + \varepsilon_{ij})$

Some brief comments

- What is real and what is not?
  - Which assumptions define the theorized "true" DGP?
  - Which are modeling conveniences that you are agnostic to?
- **Choice 1:** the existence of a "true" embedding space
  - $d_i \in \mathbb{R}^{768}$ treated as perfectly measured attribute of i
- **Choice 2:** the additive & multiplicative effects model
  - $\mathbf{sim}(d_i, d_j) = \mathbf{tanh}(\mu + \alpha_{O_i} + \alpha_{O_j} + \mathbf{sim}(O_i, O_j) + \varepsilon_{ij})$
- $\uparrow \alpha_{O_i}$ means $\uparrow$ image i's similarity to every j

Some brief comments

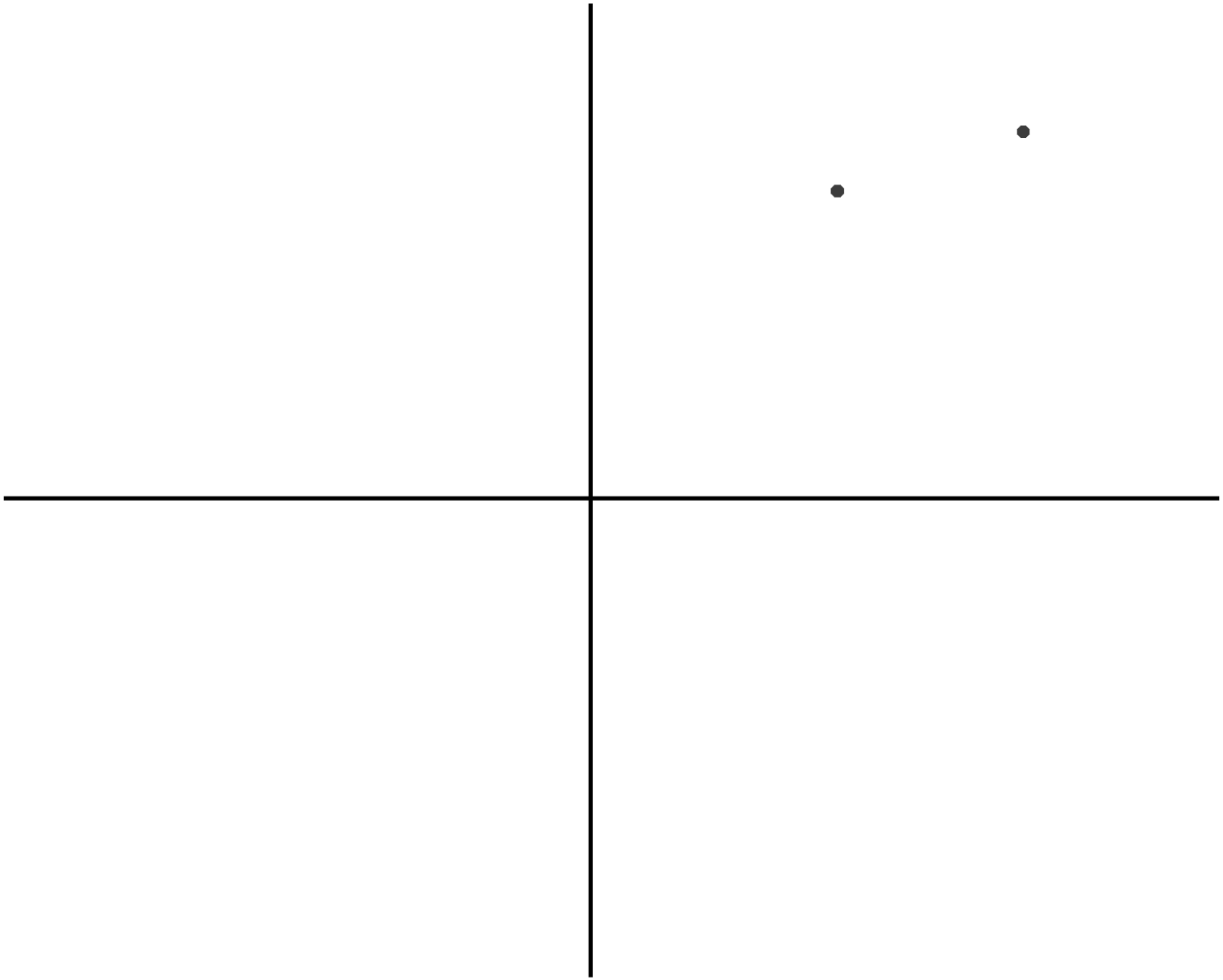
- What is real and what is not?
  - Which assumptions define the theorized "true" DGP?
  - Which are modeling conveniences that you are agnostic to?
- **Choice 1:** the existence of a "true" embedding space
  - $d_i \in \mathbb{R}^{768}$ treated as perfectly measured attribute of i
- **Choice 2:** the additive & multiplicative effects model
  - $\mathbf{sim}(d_i, d_j) = \mathbf{tanh}(\mu + \alpha_{O_i} + \alpha_{O_j} + \mathbf{sim}(O_i, O_j) + \varepsilon_{ij})$
- $\uparrow \alpha_{O_i}$ means $\uparrow$ image i's similarity to every j
  - But this would require moving $\vec{E}_j$ for every $j \neq i$!

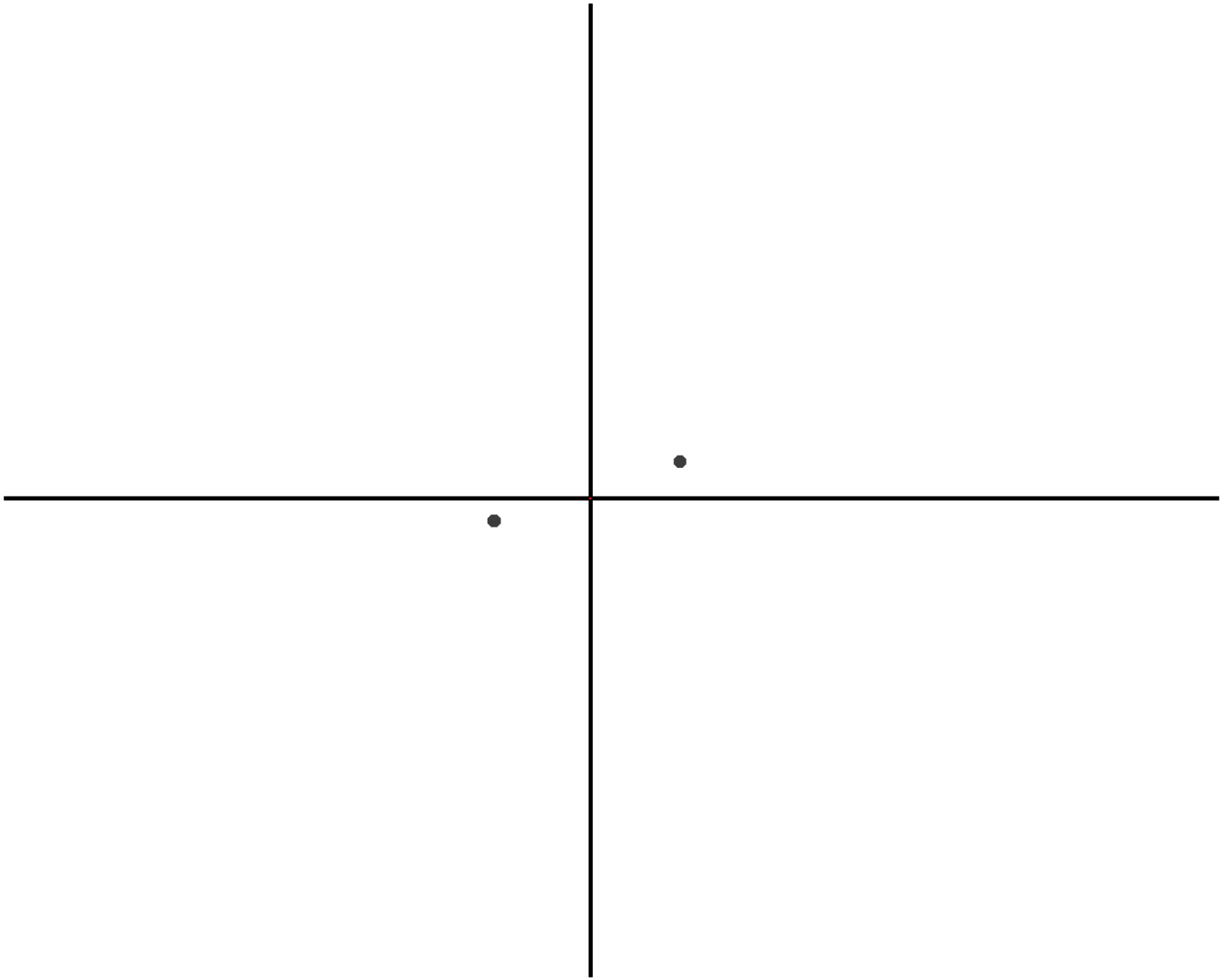
Some brief comments

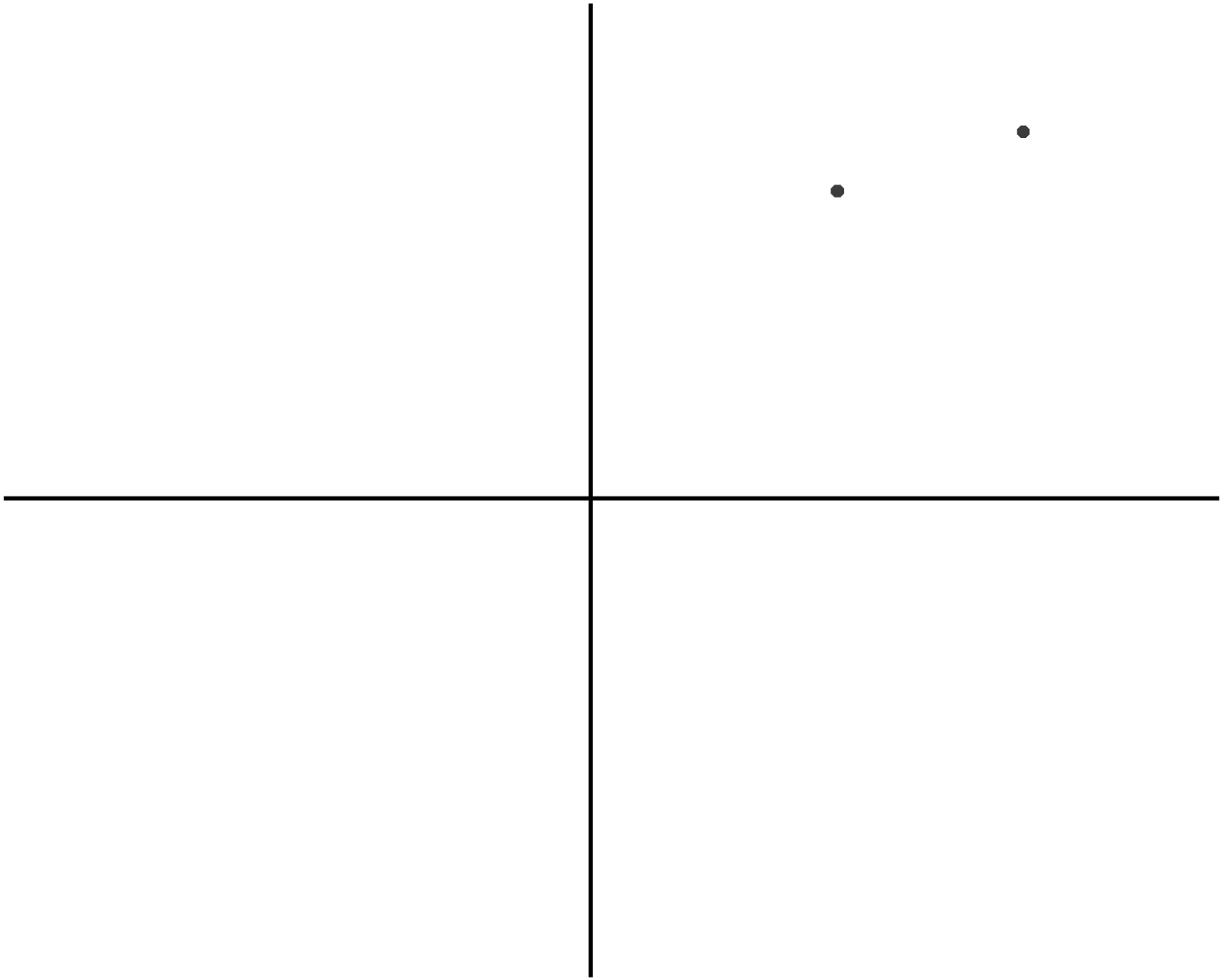
- What is real and what is not?
  - Which assumptions define the theorized "true" DGP?
  - Which are modeling conveniences that you are agnostic to?
- **Choice 1:** the existence of a "true" embedding space
  - $d_i \in \mathbb{R}^{768}$ treated as perfectly measured attribute of i
- **Choice 2:** the additive & multiplicative effects model
  - $\mathbf{sim}(d_i, d_j) = \mathbf{tanh}(\mu + \alpha_{O_i} + \alpha_{O_j} + \mathbf{sim}(O_i, O_j) + \varepsilon_{ij})$
- $\uparrow \alpha_{O_i}$ means $\uparrow$ image i's similarity to every j
  - But this would require moving $\vec{E}_j$ for every $j \neq i$!
- $\uparrow \varepsilon_{ij}$ means $\uparrow \mathbf{sim}(d_i, d_j)$, requires moving $\vec{E}_i$ or $\vec{E}_j$

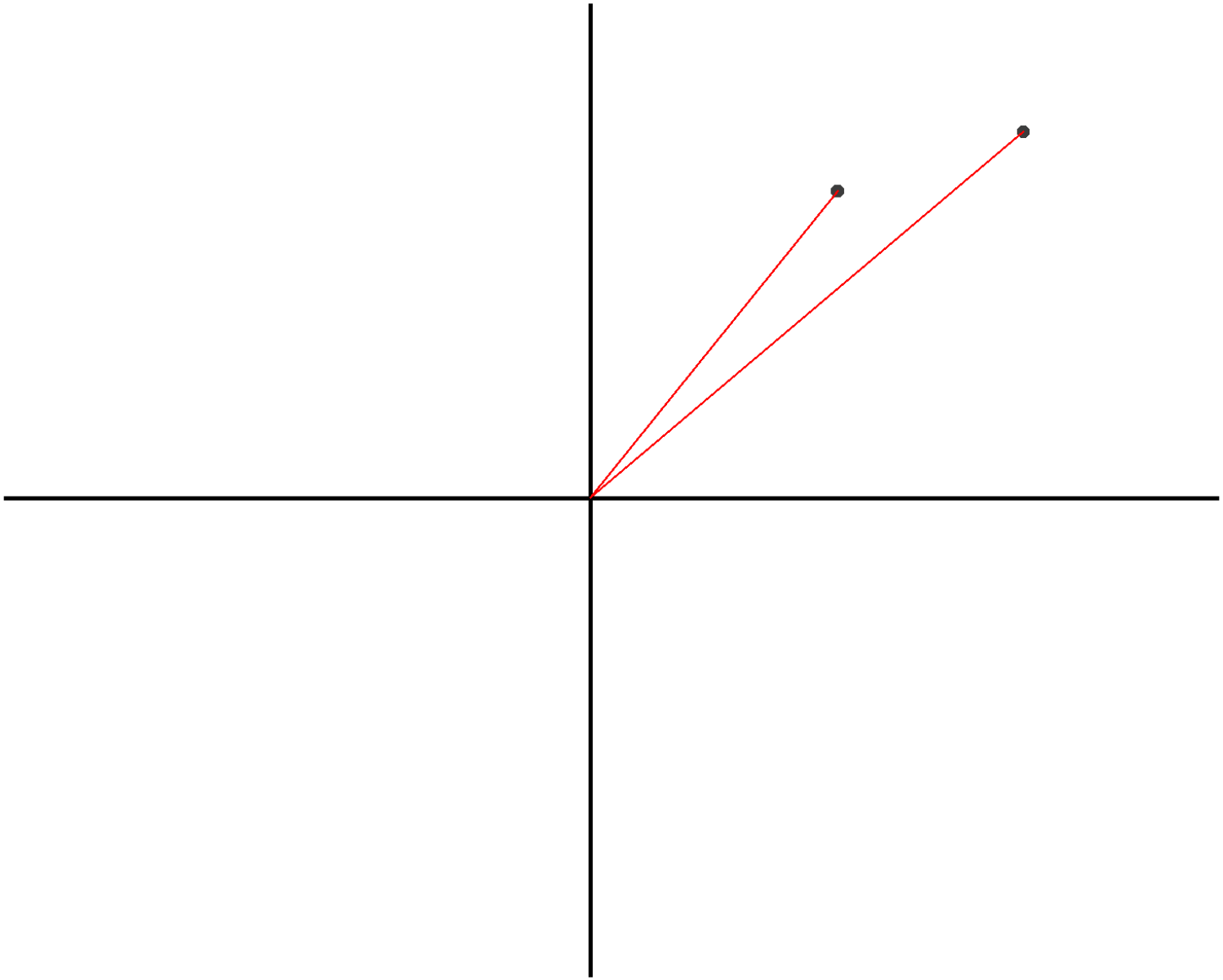
Some brief comments

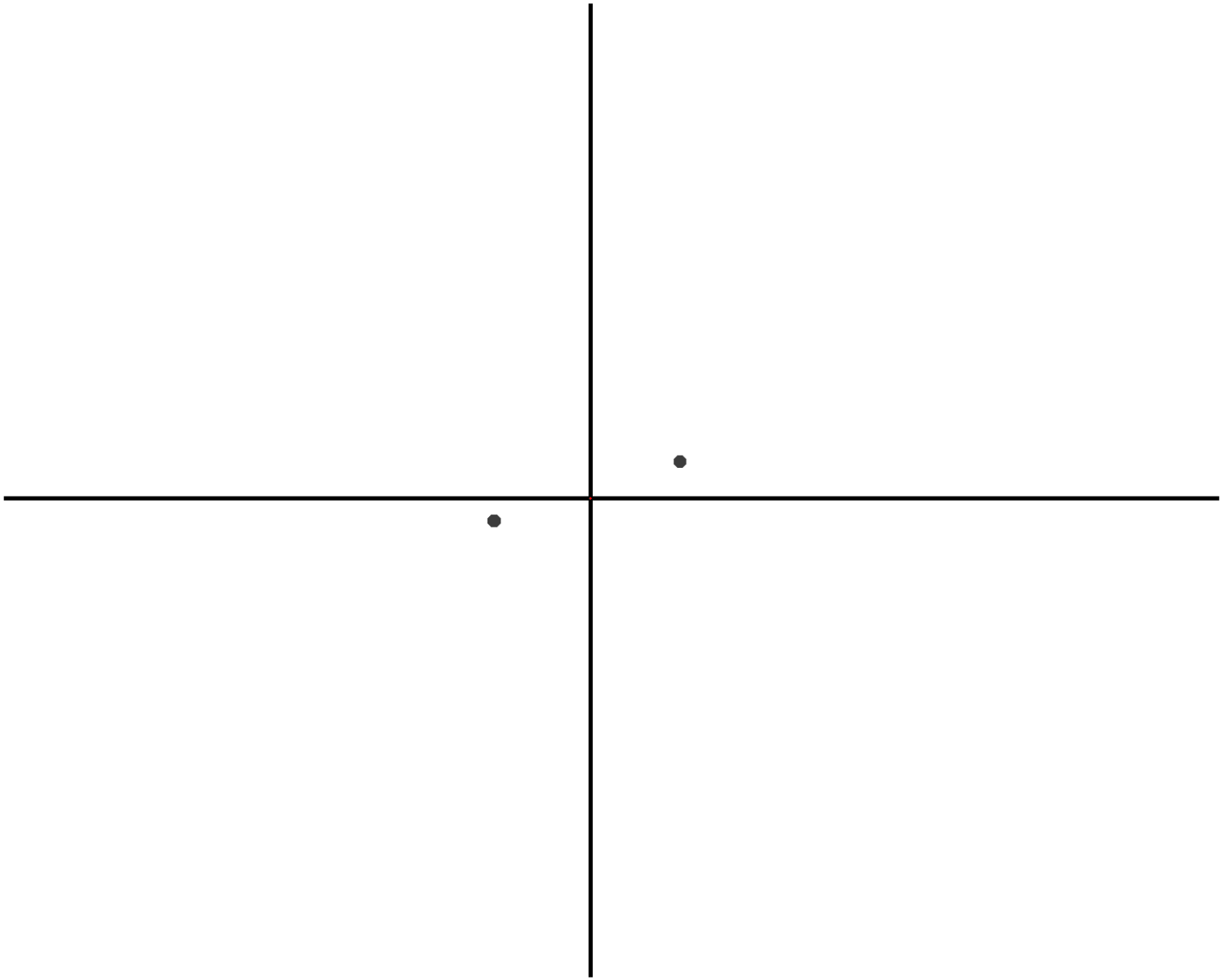
- What is real and what is not?
  - Which assumptions define the theorized "true" DGP?
  - Which are modeling conveniences that you are agnostic to?
- **Choice 1:** the existence of a "true" embedding space
  - $d_i \in \mathbb{R}^{768}$ treated as perfectly measured attribute of i
- **Choice 2:** the additive & multiplicative effects model
  - $\mathbf{sim}(d_i, d_j) = \mathbf{tanh}(\mu + \alpha_{O_i} + \alpha_{O_j} + \mathbf{sim}(O_i, O_j) + \varepsilon_{ij})$
- $\uparrow \alpha_{O_i}$ means $\uparrow$ image i's similarity to every j
  - But this would require moving $\vec{E}_j$ for every $j \neq i$!
- $\uparrow \varepsilon_{ij}$ means $\uparrow \mathbf{sim}(d_i, d_j)$, requires moving $\vec{E}_i$ or $\vec{E}_j$
  - But this means changing all $\mathbf{sim}(d_i, d_k)$ or $\mathbf{sim}(d_j, d_k)$!

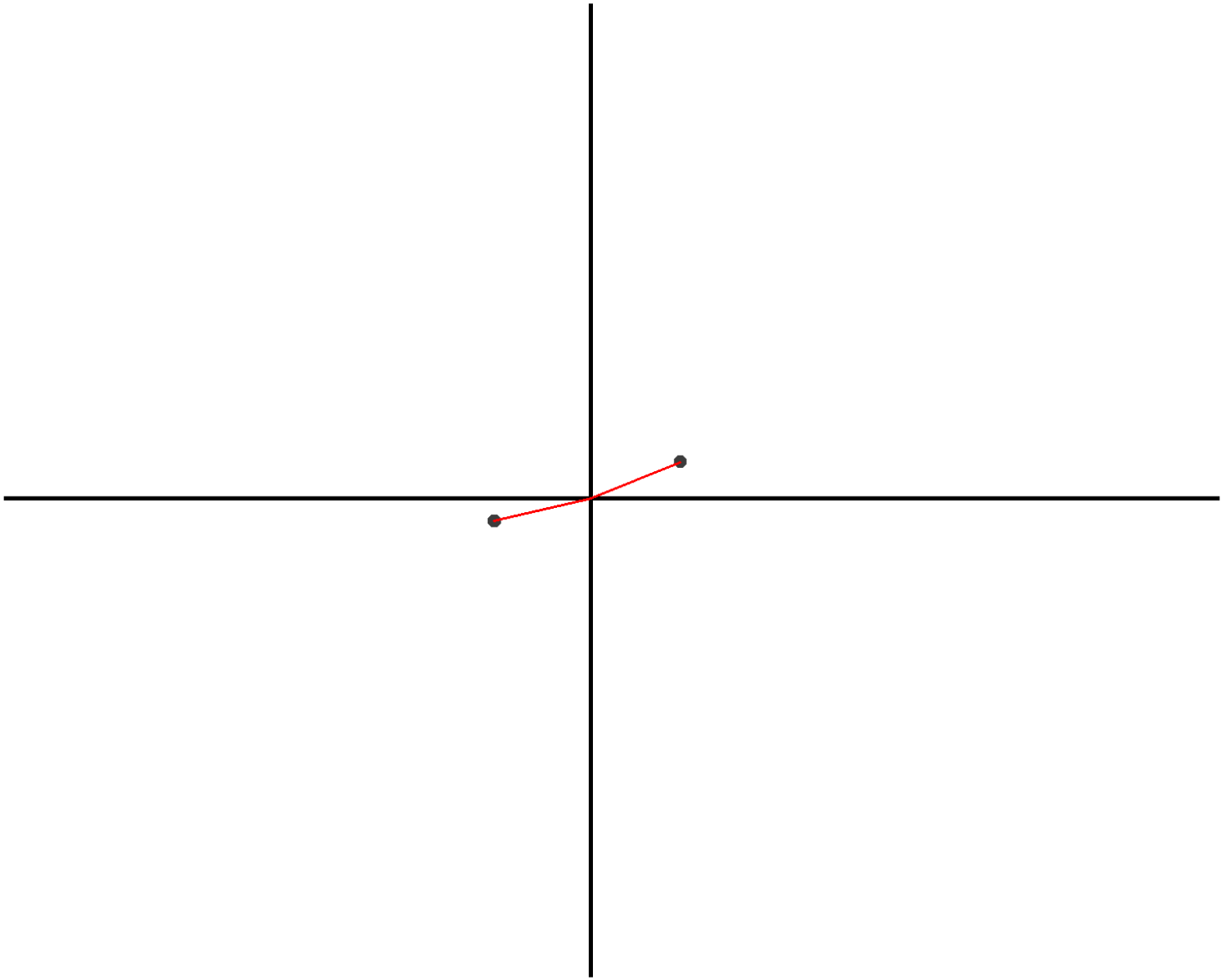












Fin