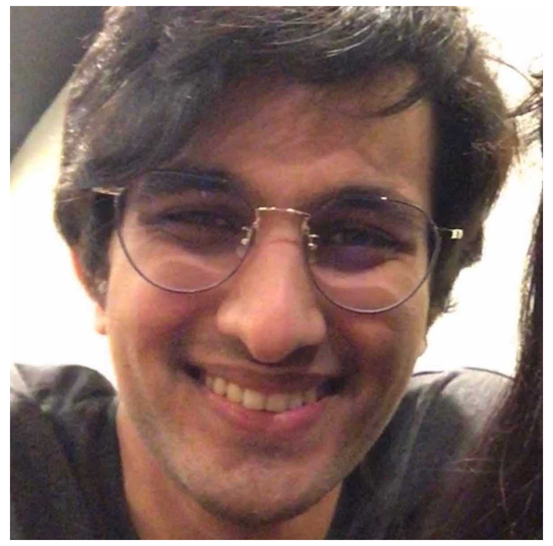
Identifiability Conditions for Domain Adaptation



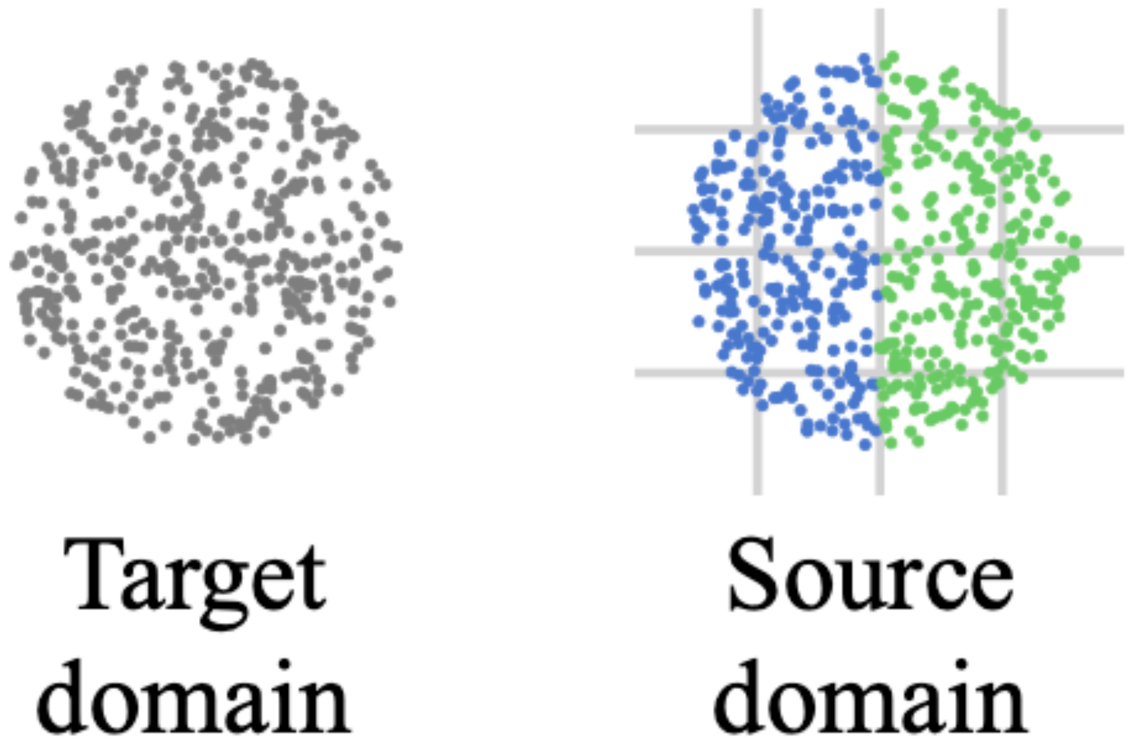
Ishaan
Gulrajani

Tatsunori B.
Hashimoto

Stanford University

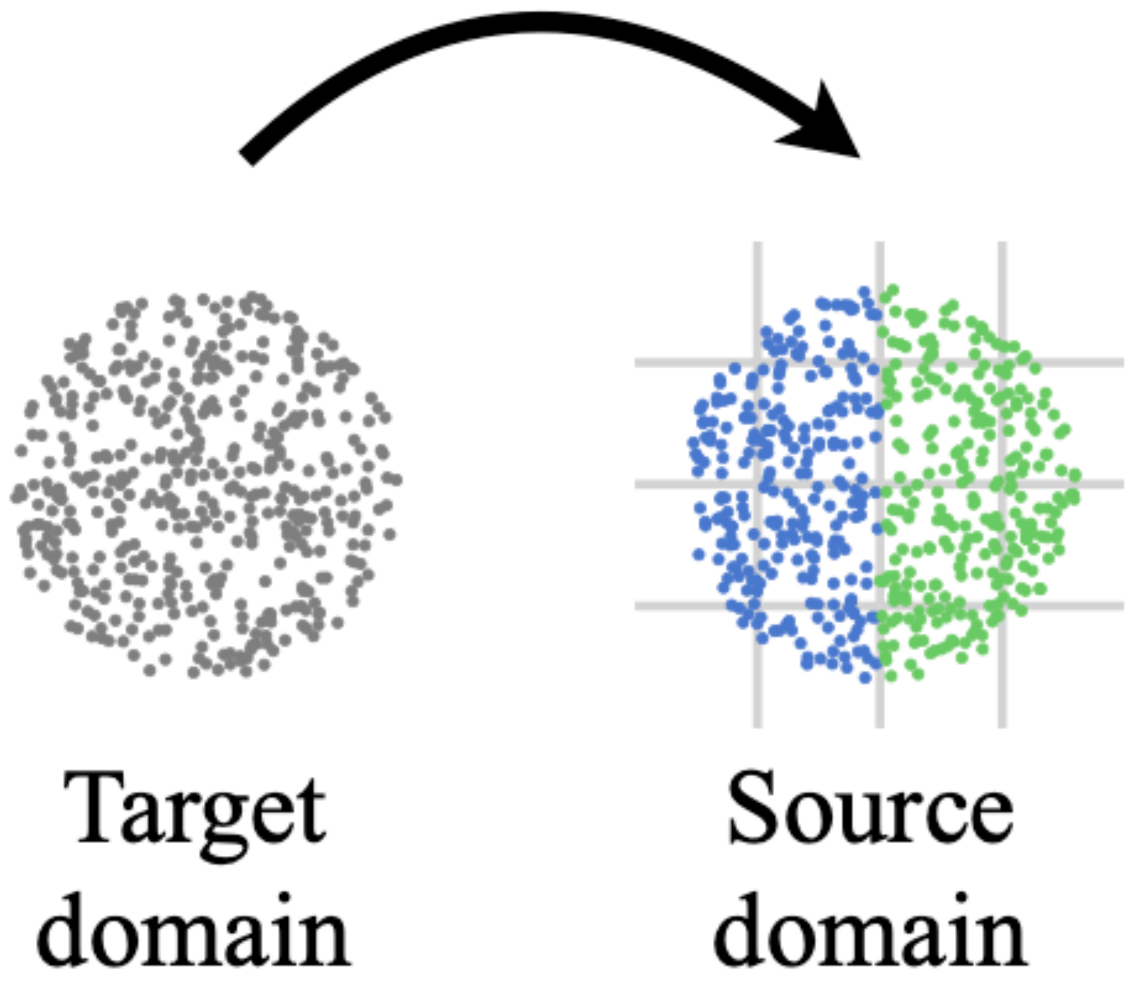
Unsupervised Domain Adaptation

- Labeled **source domain** + unlabeled **target domain**



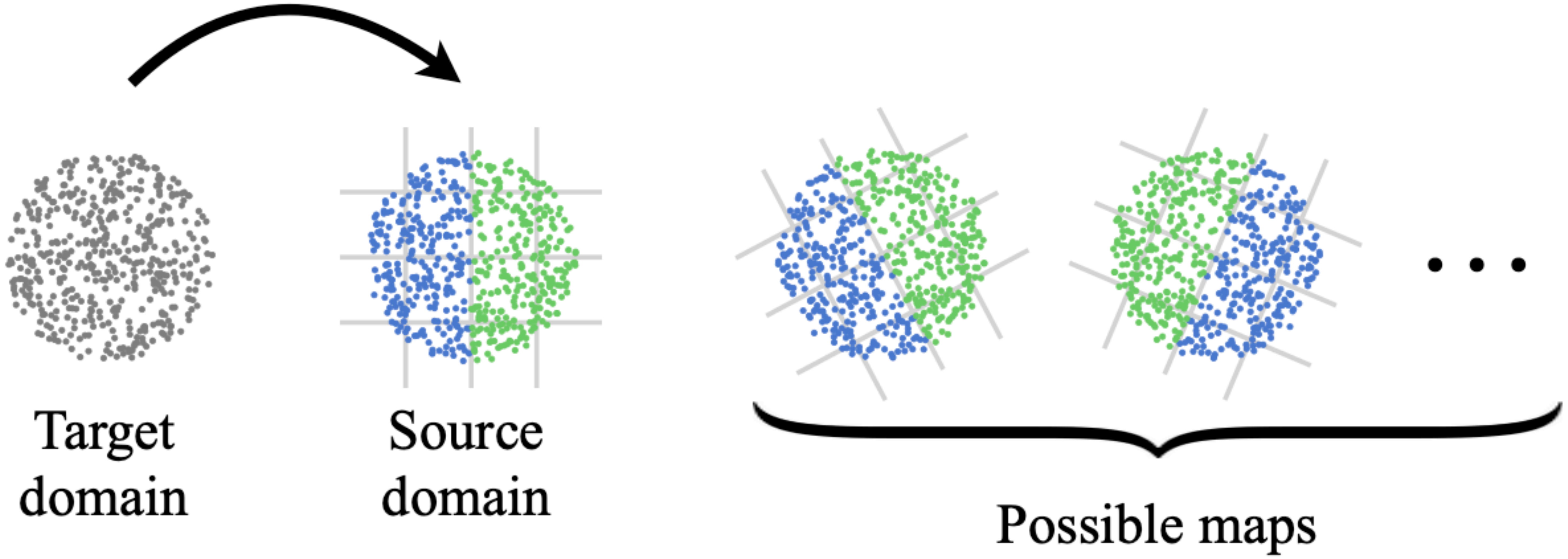
Unsupervised Domain Adaptation

- Labeled **source domain** + unlabeled **target domain**
- **Domain Mapping**: Learn target→source map by matching input distributions



Unsupervised Domain Adaptation

- Labeled **source domain** + unlabeled **target domain**
- **Domain Mapping**: Learn target→source map by matching input distributions
- **Underspecification:** Many “spurious maps” which yield wrong predictions *despite zero held-out loss.*



When are domain maps identifiable?

Theory + Algorithms

Orthogonal Linear Maps

Idea: spurious maps correspond to **symmetries** in the distribution.

Orthogonal Linear Maps

Idea: spurious maps correspond to **symmetries** in the distribution.

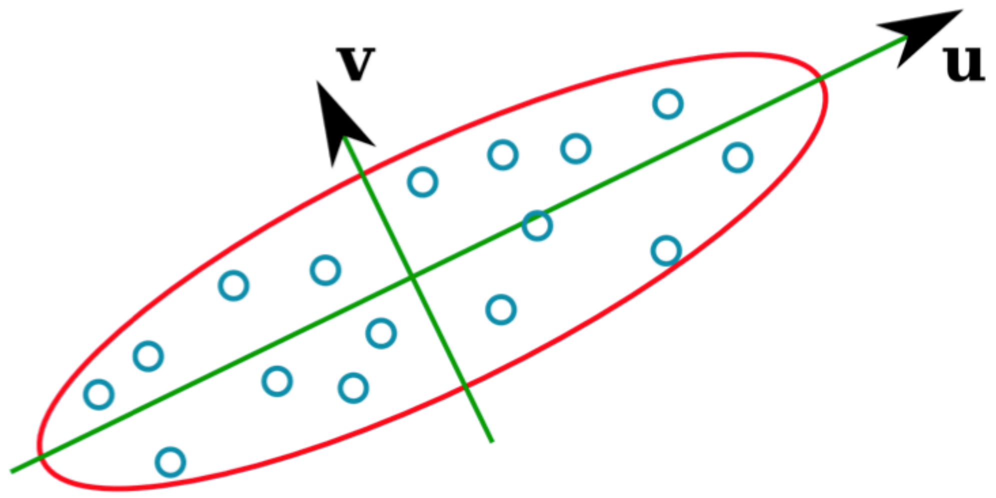
We can prove asymmetry using properties associated with the **second moment matrix**:

Orthogonal Linear Maps

Idea: spurious maps correspond to **symmetries** in the distribution.

We can prove asymmetry using properties associated with the **second moment matrix**:

#1: **Distinct eigenvalues** ⇒ rotational asymmetry



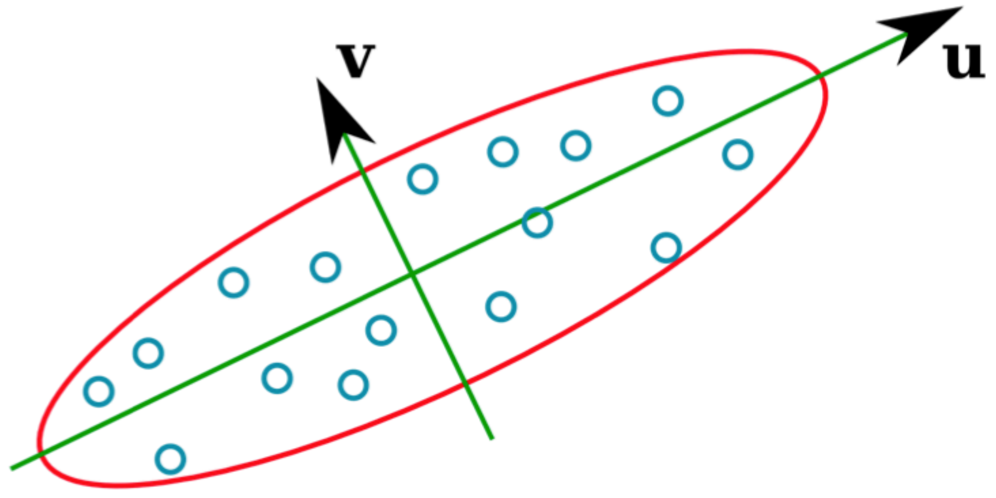
Orthogonal Linear Maps

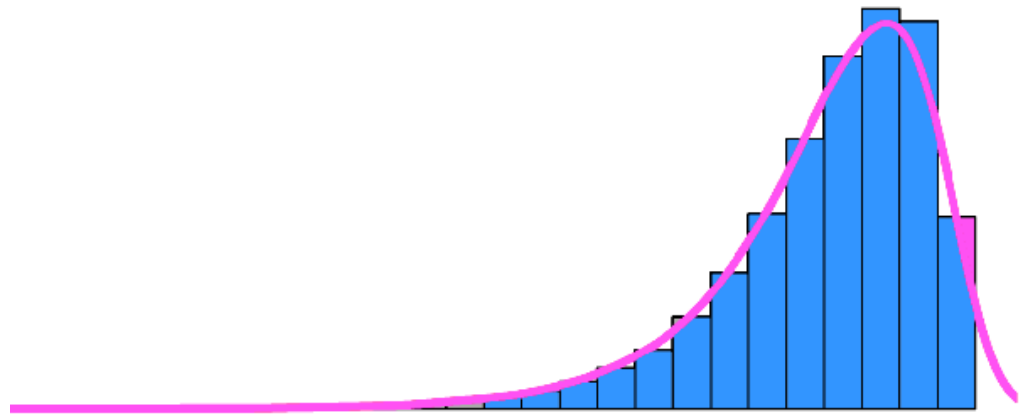
Idea: spurious maps correspond to **symmetries** in the distribution.

We can prove asymmetry using properties associated with the **second moment matrix**:

#1: **Distinct eigenvalues** ⇒ rotational asymmetry

#2: **Skewed marginals** along eigenvectors ⇒ reflection asymmetry





General Linear Maps

General Linear Maps

Whitening reduces general linear maps to orthogonal maps…

General Linear Maps

Whitening reduces general linear maps to orthogonal maps…

… but second moment conditions no longer hold.

General Linear Maps

Whitening reduces general linear maps to orthogonal maps…

… but second moment conditions no longer hold.

We derive analogous conditions on the **third moment tensor** of the whitened distribution**:**

Unique CP decomposition with **no repeated weights** (analogous to eigenvalues) $\Rightarrow$ General linear asymmetry

Identifiability Guarantees From Data

Moment conditions are **hard to verify based on a dataset alone**.

Identifiability Guarantees From Data

Moment conditions are **hard to verify based on a dataset alone**.

Intuition:

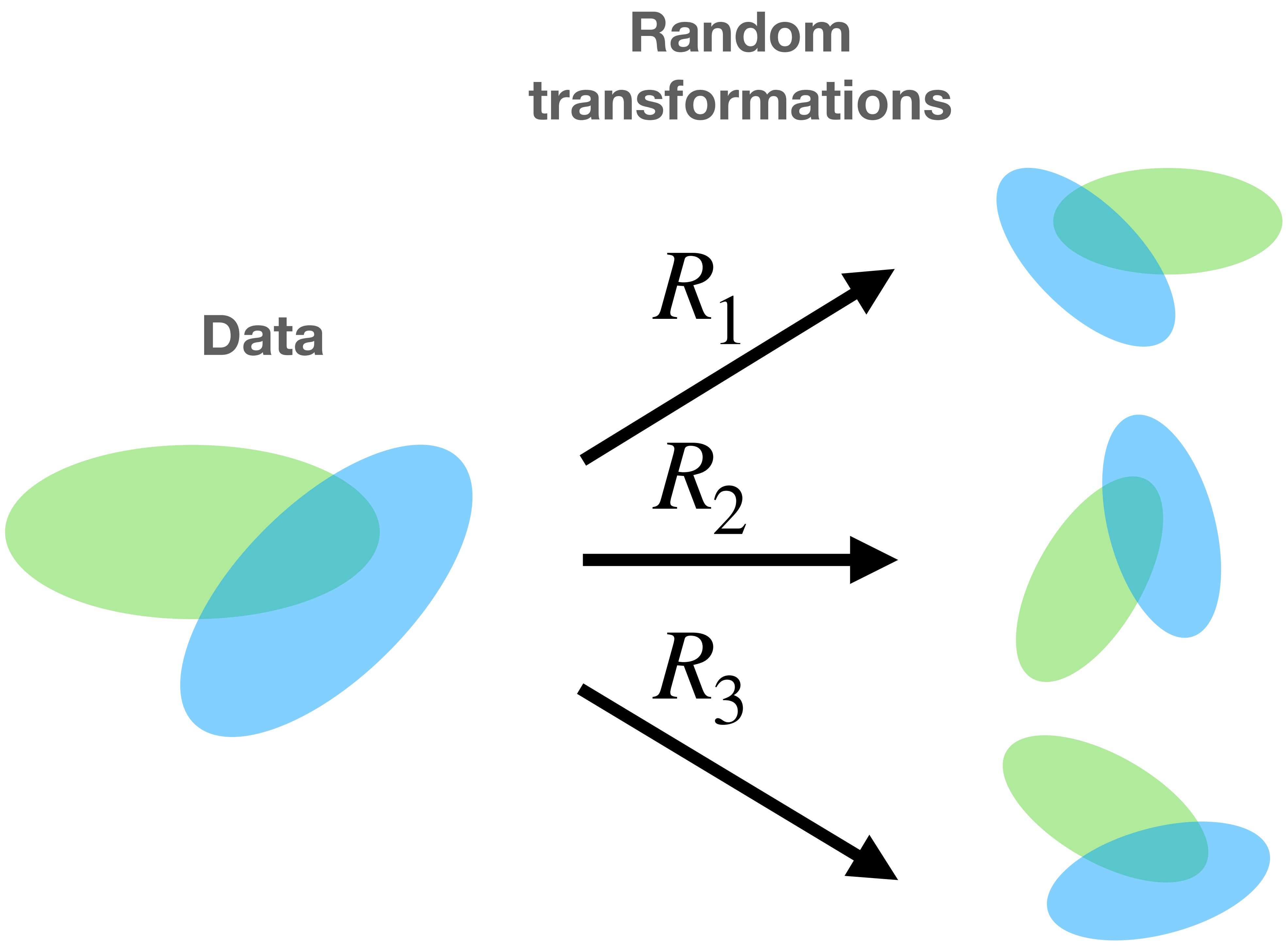
1. An “unbiased” mapping algorithm chooses randomly from possible maps
2. Random orthogonal transformations can make any mapping algorithm “unbiased”

Identifiability Guarantees From Data

Algorithm:

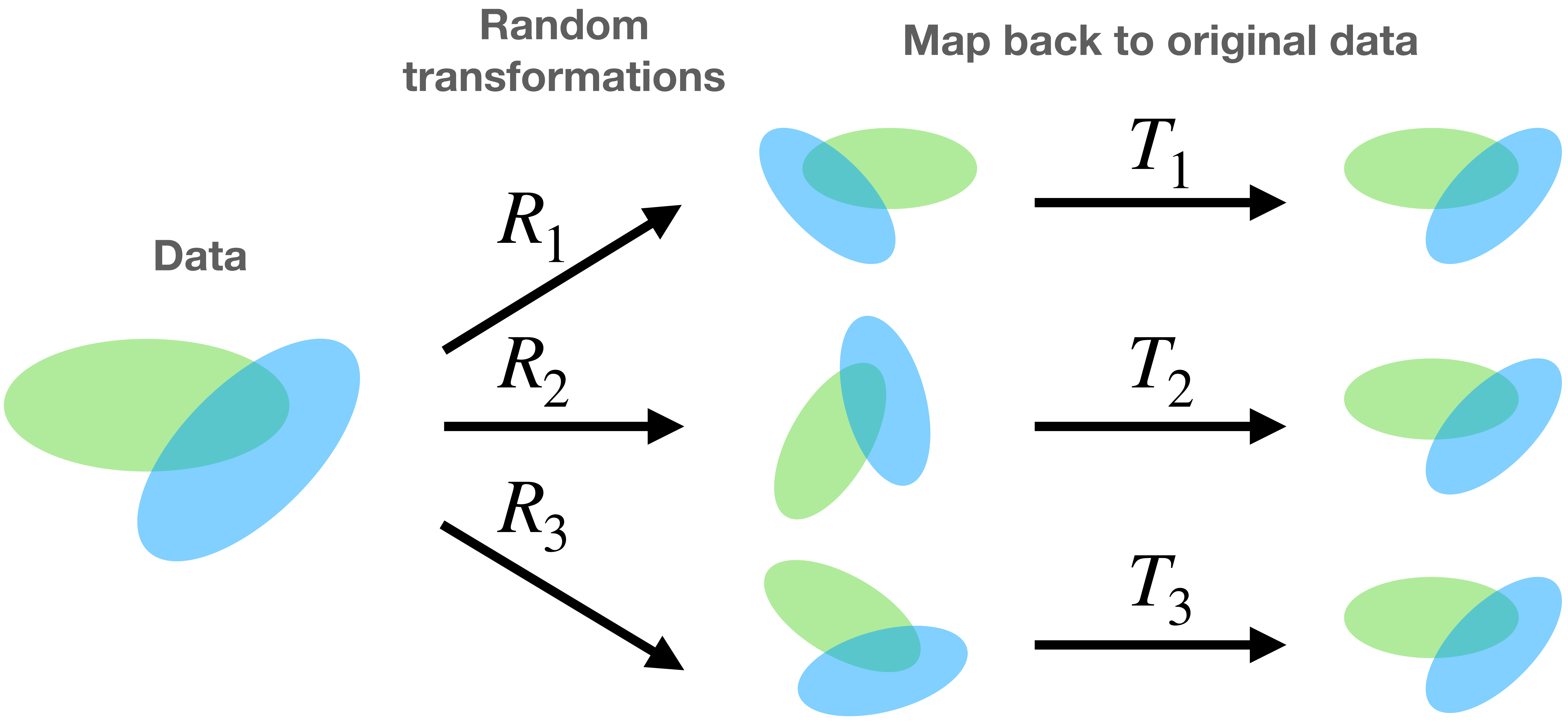
Identifiability Guarantees From Data

Algorithm:



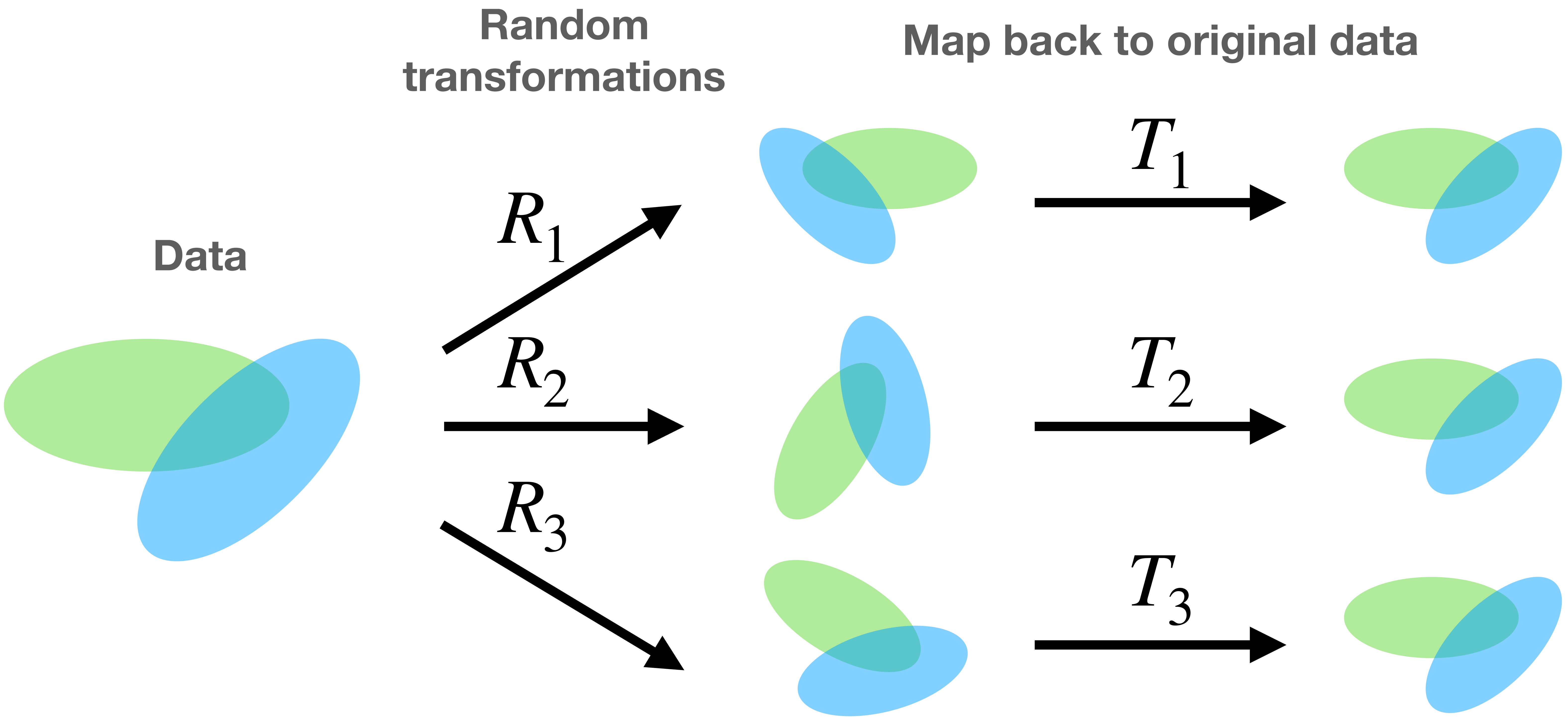
Identifiability Guarantees From Data

Algorithm:



Identifiability Guarantees From Data

Algorithm:



Test:

$T_i = R_i^{-1}$ **for all** i**?**

Beyond The Linear Case

Idea: Bound the worst-case error over the set of possible maps.

$$\mathscr{L}_T(h) \leq \mathscr{L}_S(h_s) + \sup_{T \in \tilde{\mathscr{T}}} \mathbb{E}_{P^t} \left[\ell(h(x), h_s(T(x))) \right]$$

Beyond The Linear Case

Idea: Bound the worst-case error over the set of possible maps.

$$\mathscr{L}_T(h) \leq \mathscr{L}_S(h_s) + \sup_{T \in \tilde{\mathscr{T}}} \mathbb{E}_{P^t}\left[\ell(h(x), h_s(T(x)))\right]$$

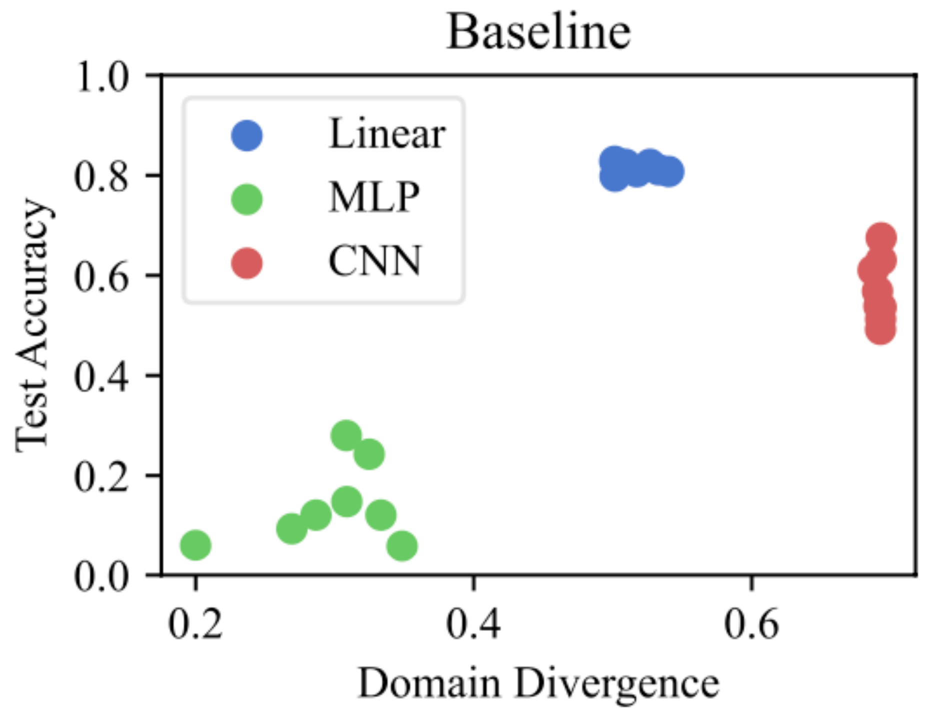
Heuristic: Approximate $\tilde{\mathscr{T}}$ (the set of possible maps) by a few random restarts of a mapping algorithm.

This leads to a **loss function for domain mapping.**

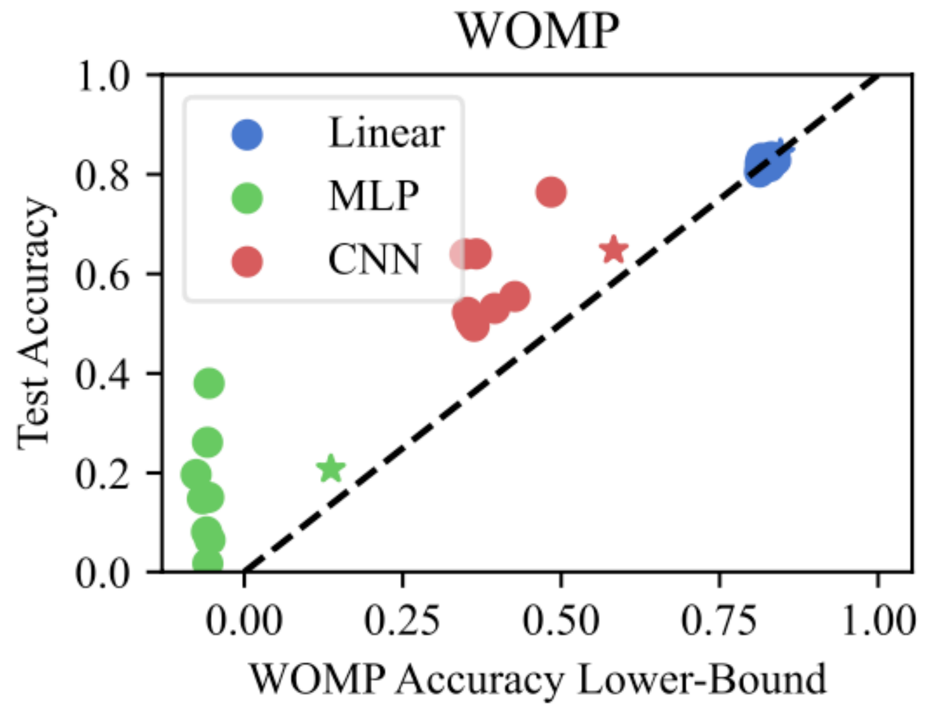
Beyond The Linear Case

Predicting target-domain accuracy without target-domain labels:

Baseline



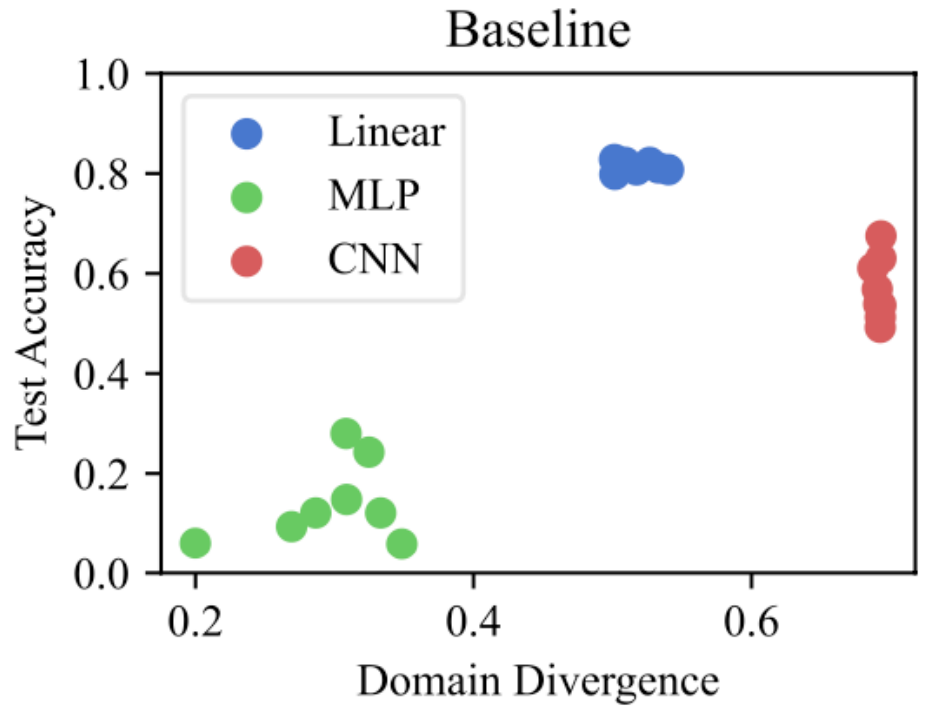
Our Method

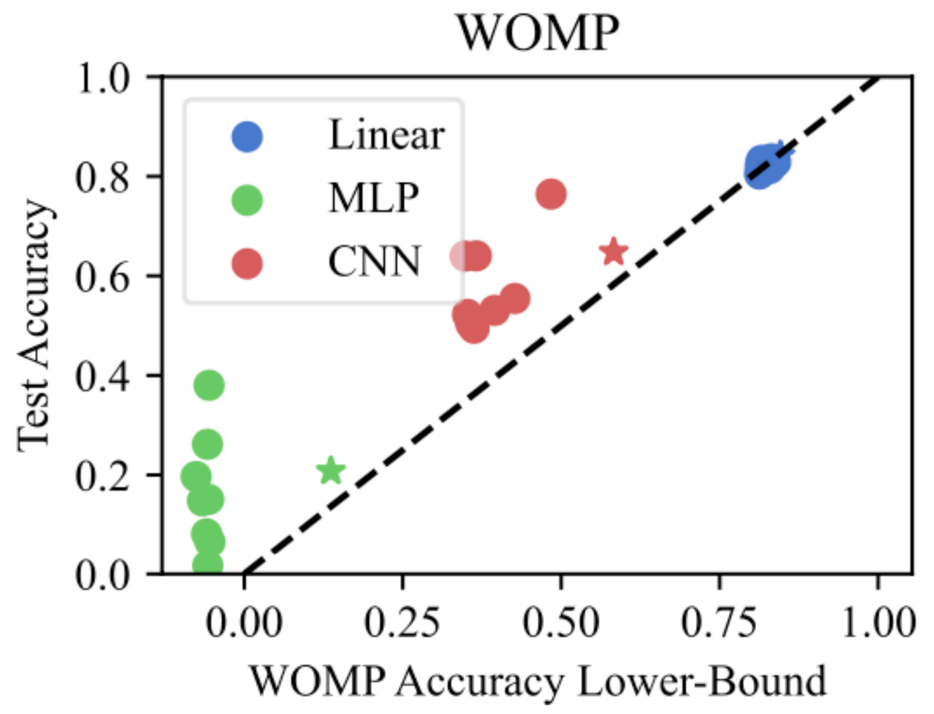


Beyond The Linear Case

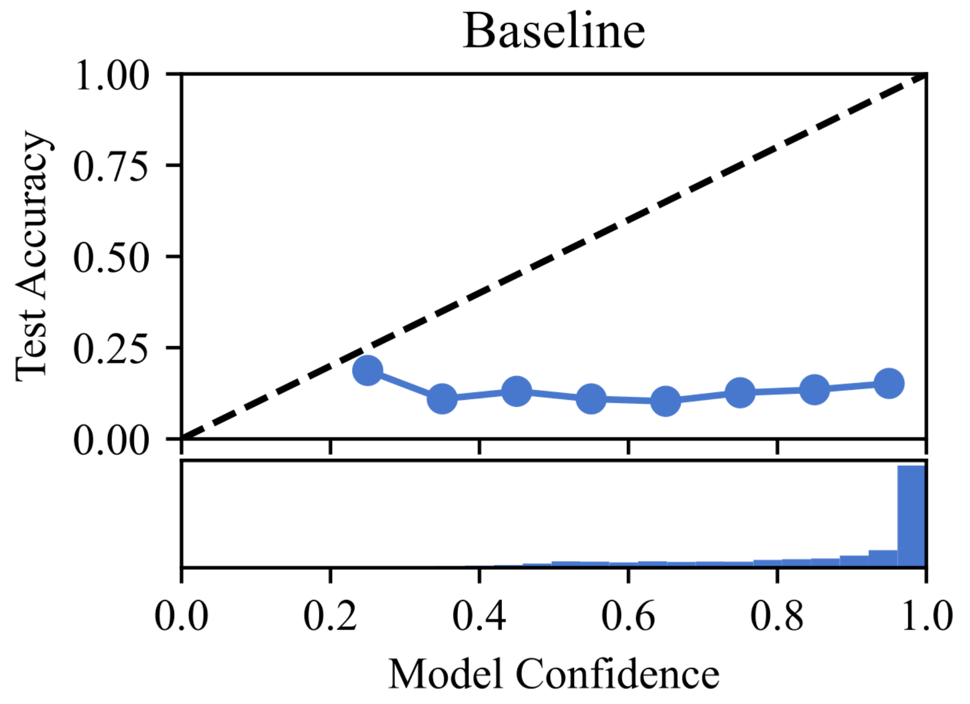
Baseline **Our Method**

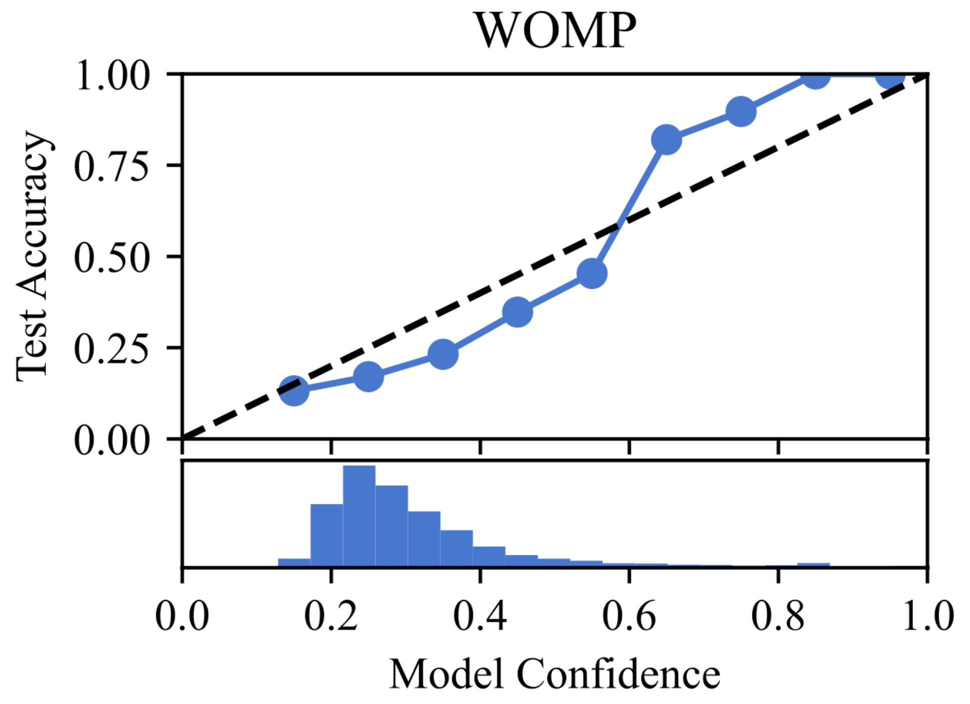
Predicting target-domain accuracy without target-domain labels:





Learning uncertainty-aware target-domain classifiers:





Thank you!