Generative Models and Optimal Transport

Marco Cuturi

École nationale de la statistique et de l'administration économique

ENSAE ParisTech

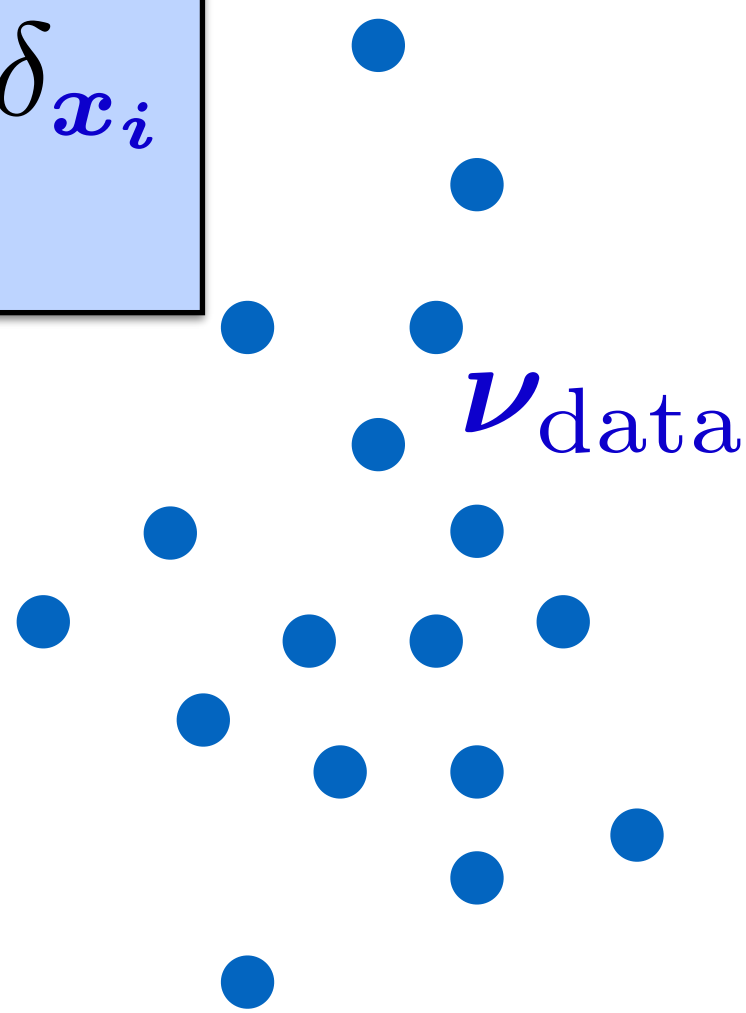
université PARIS-SACLAY

Joint work / work in progress with
G. Peyré, A. Genevay *(ENS)*, F. Bach *(INRIA)*,
G. Montavon, K-R Müller *(TU Berlin)*

Statistics 0.1 : Density Fitting

We collect data

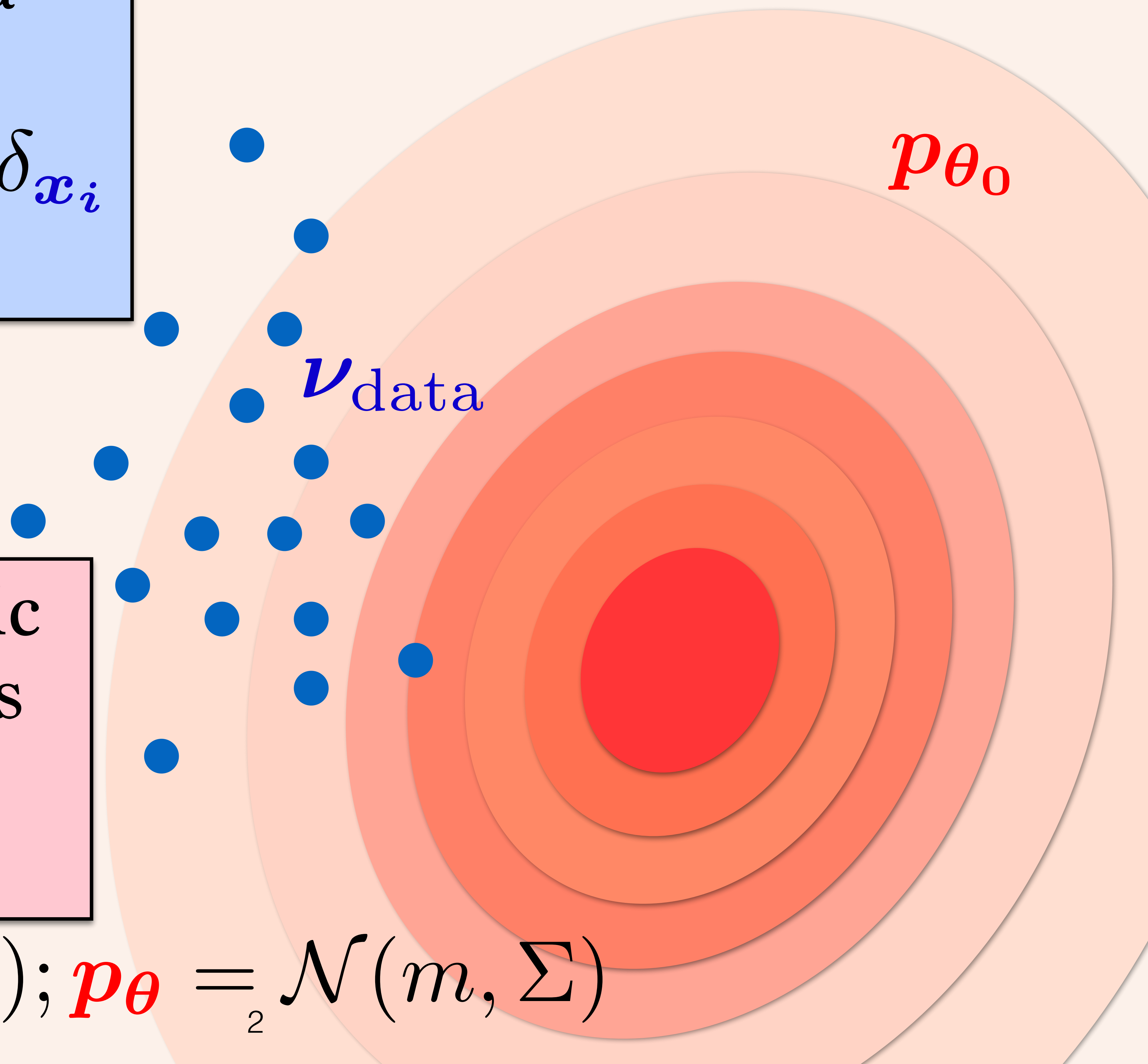
$$\boldsymbol{\nu}_{\mathrm{data}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{\boldsymbol{x_i}}$$



Statistics 0.1 : Density Fitting

We collect data

$$\boldsymbol{\nu}_{\text{data}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{\boldsymbol{x_i}}$$

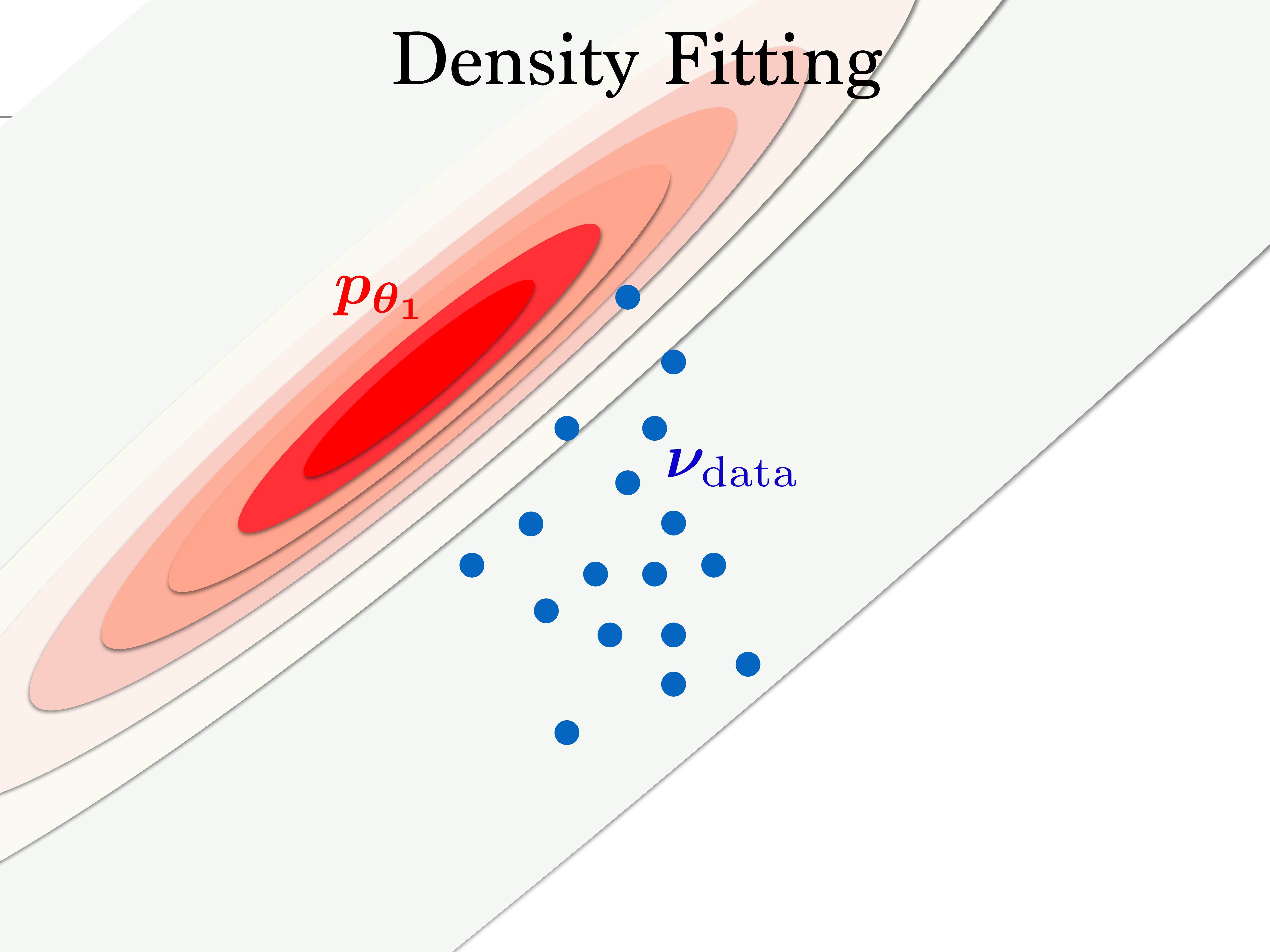


We fit a parametric family of densities

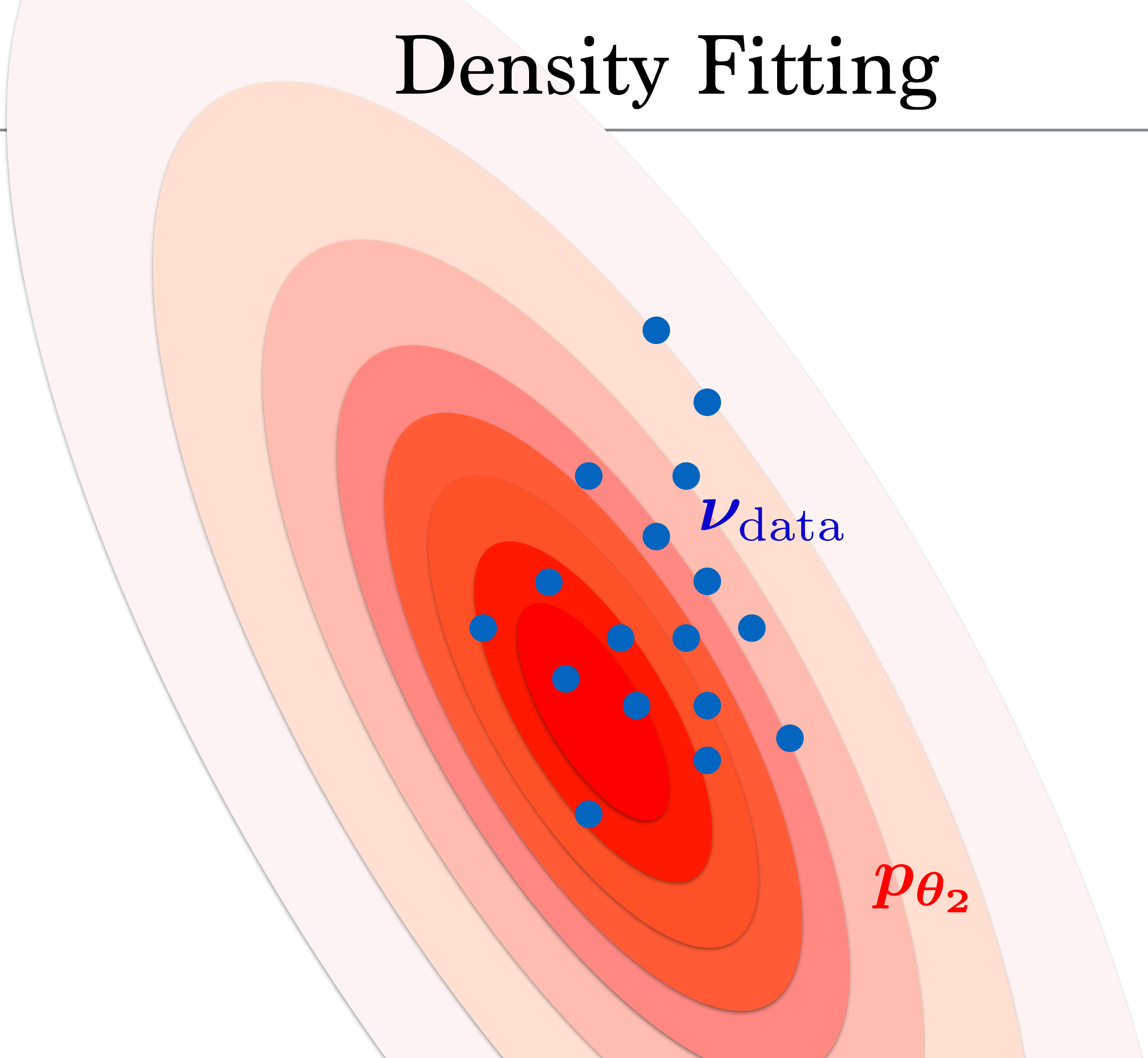
$$\{p_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

$$e.g.\ \boldsymbol{\theta} = (m, \Sigma); \boldsymbol{p_\theta} = \mathcal{N}(m, \Sigma)$$

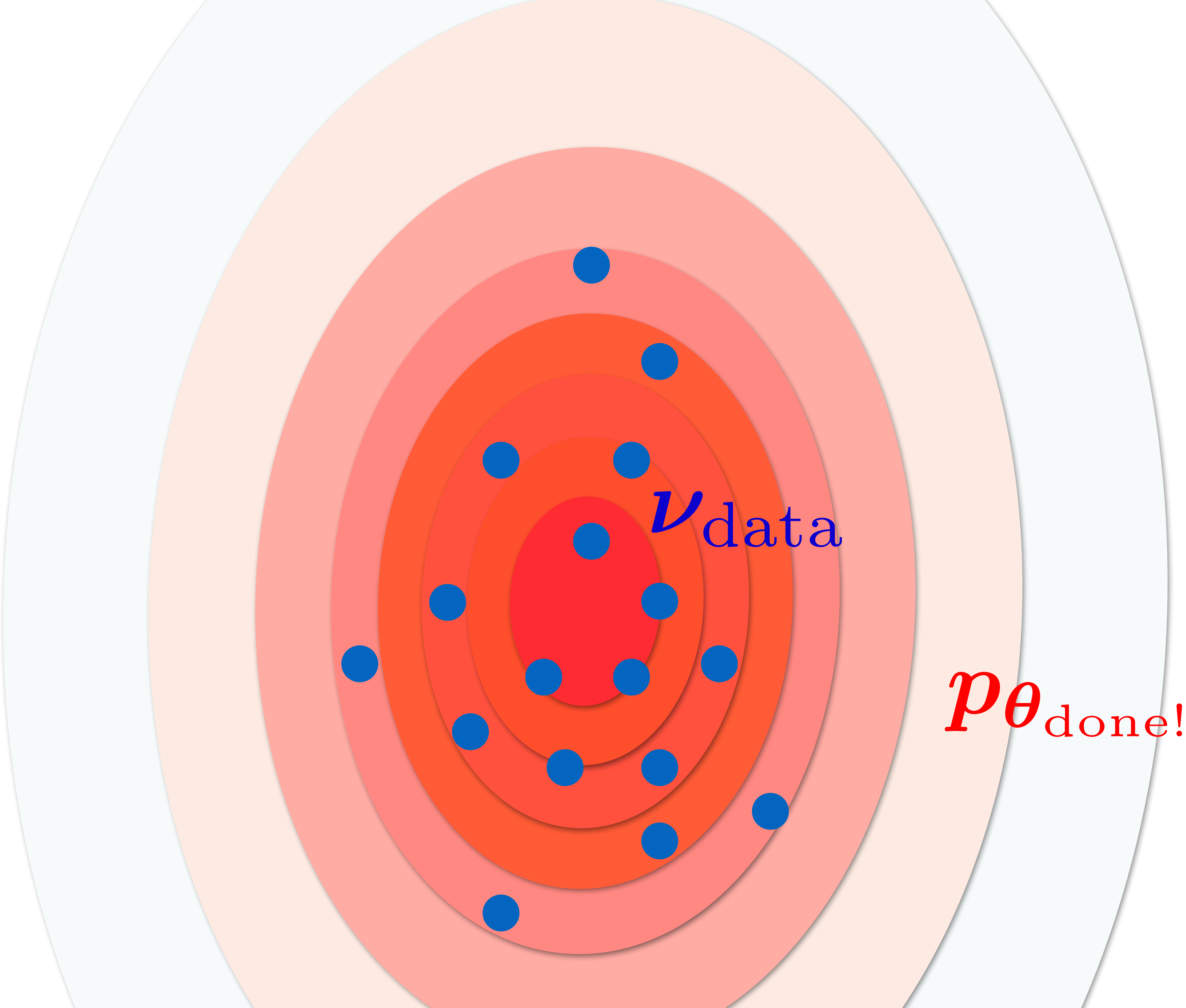
Density Fitting



Density Fitting

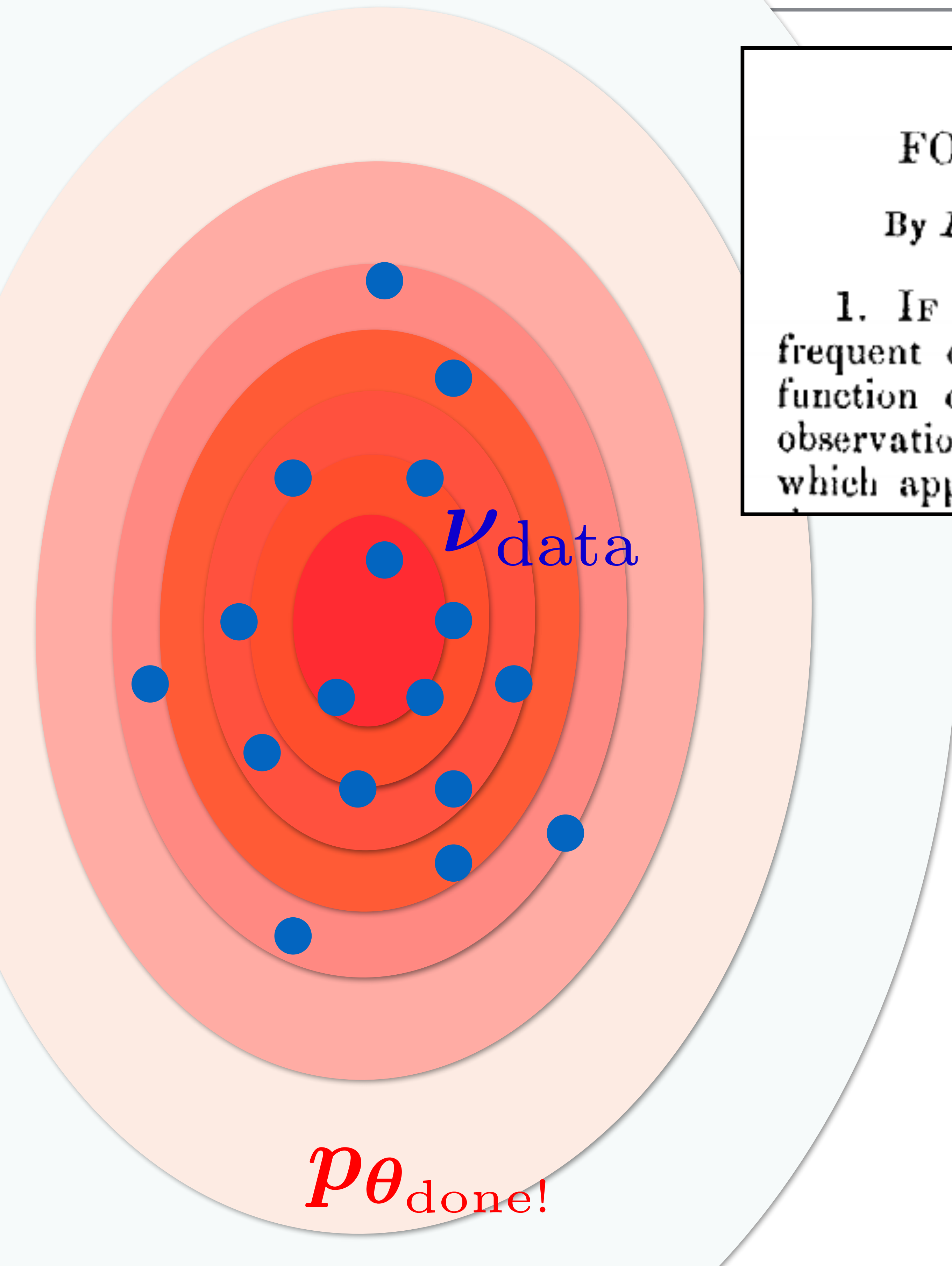


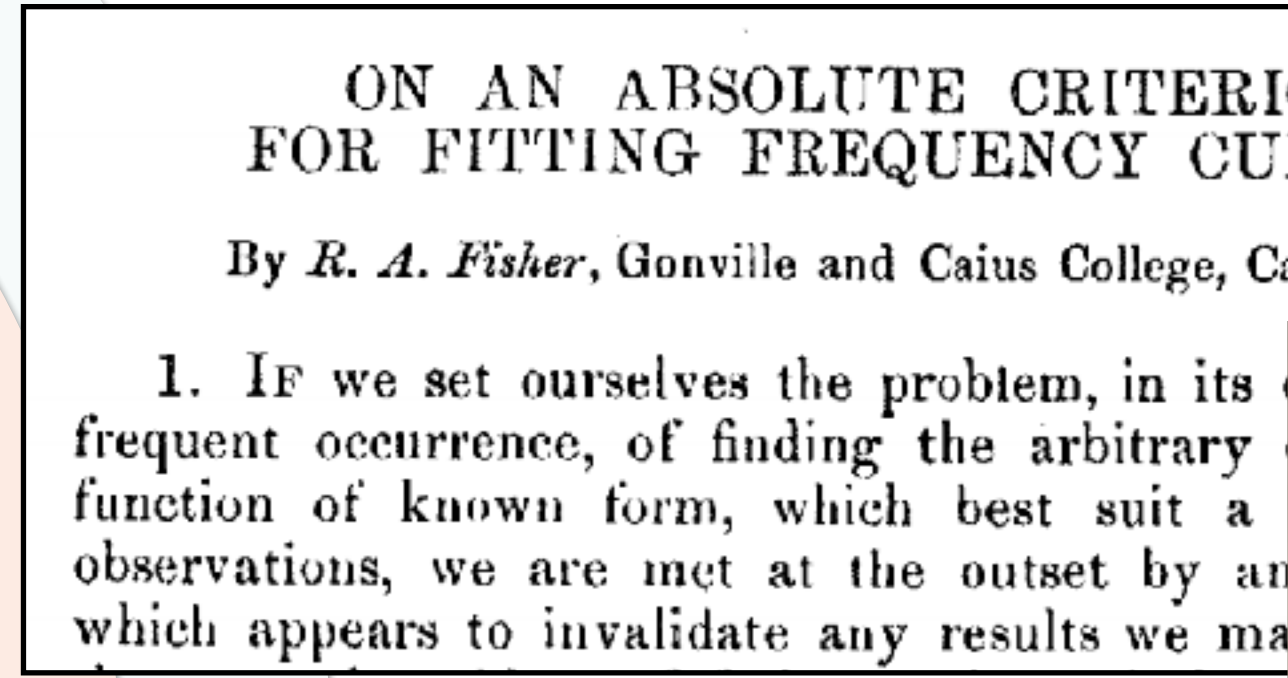
Density Fitting



We stop when there is a good fit.

Maximum Likelihood Estimation





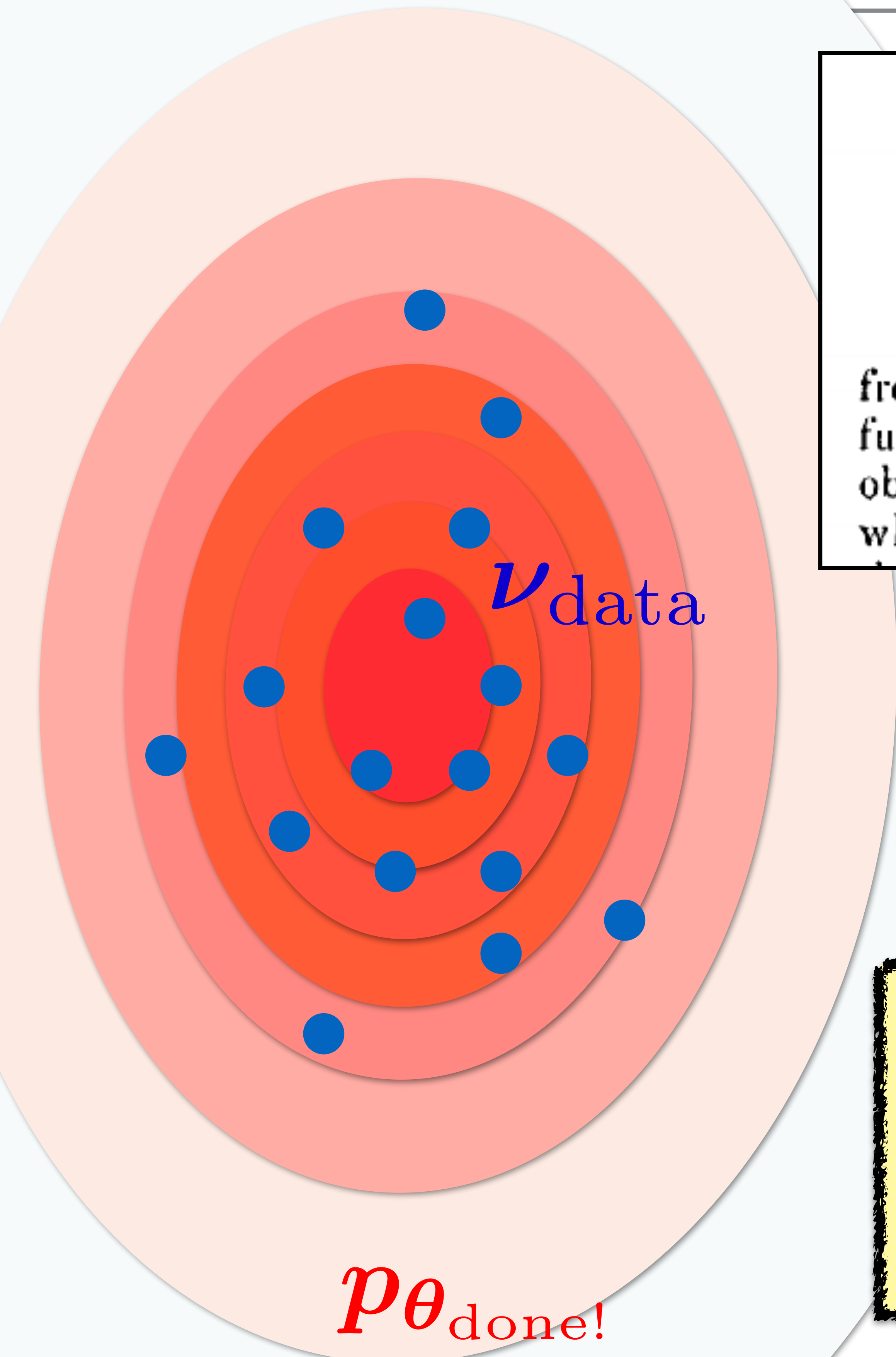
ON AN ABSOLUTE CRITERION
FOR FITTING FREQUENCY CURVES.

By *R. A. Fisher*, Gonville and Caius College, Cambridge.

1. If we set ourselves the problem, in its
frequent occurrence, of finding the arbitrary
function of known form, which best suit a
observations, we are met at the outset by an
which appears to invalidate any results we ma

$$\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$$

Maximum Likelihood Estimation



ON AN ABSOLUTE CRITERION
FOR FITTING FREQUENCY CURVES.

By *R. A. Fisher*, Gonville and Caius College, Cambridge.

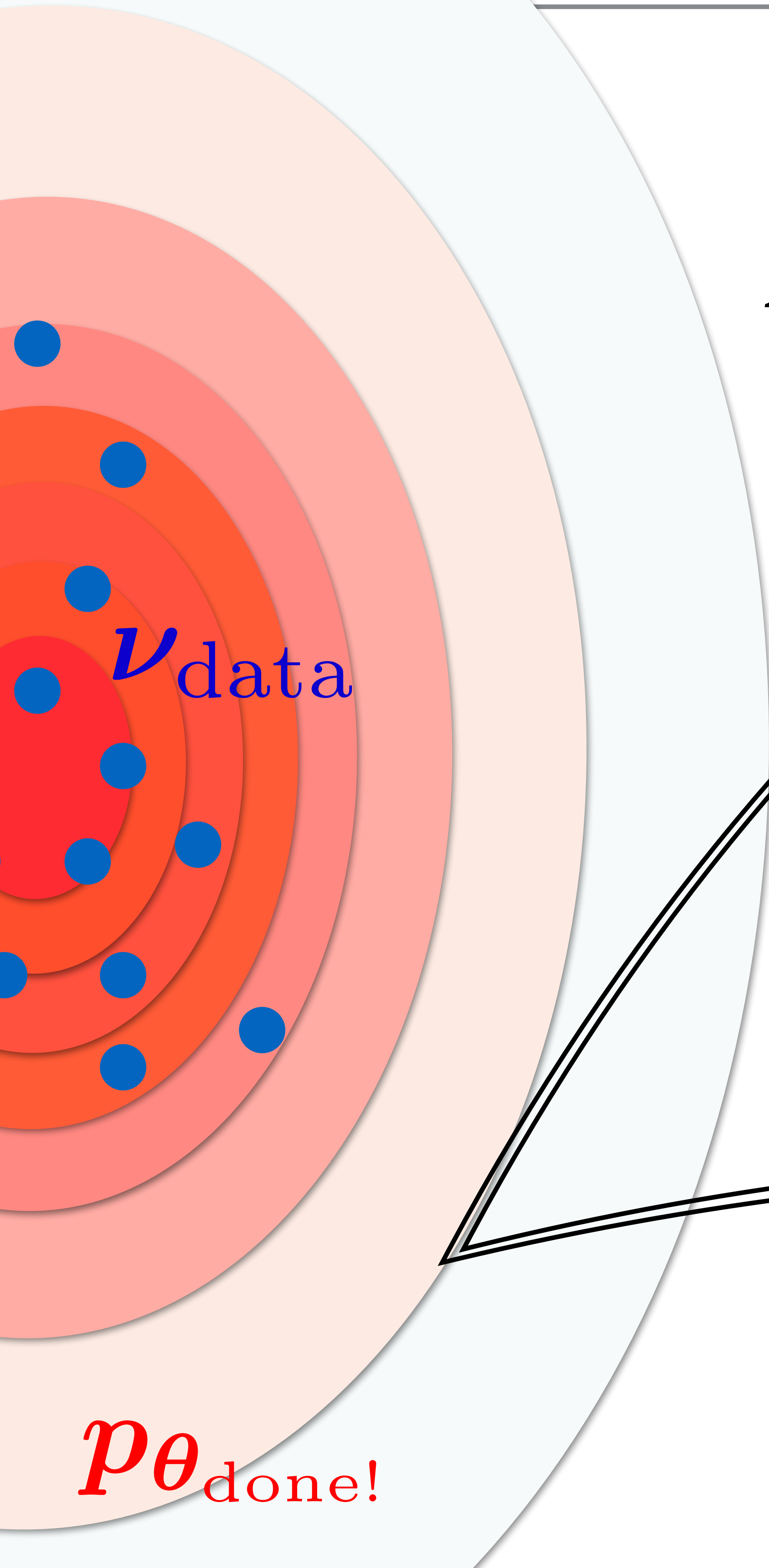
1. If we set ourselves the problem, in its
frequent occurrence, of finding the arbitrary
function of known form, which best suit a
observations, we are met at the outset by an
which appears to invalidate any results we ma

$$\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \log p_{\theta}(x_i)$$

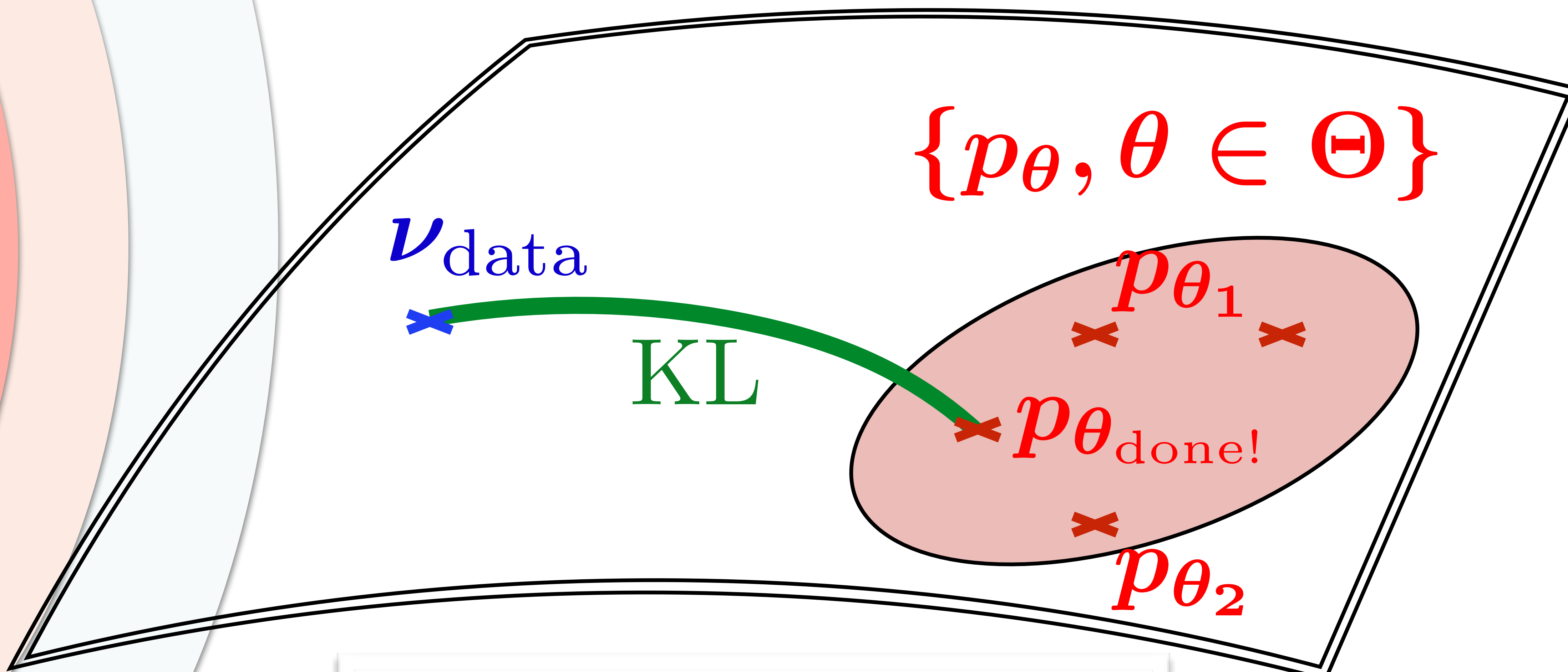
$$\log 0 = -\infty$$

$p_{\theta}(x_i)$ **must** be > 0

Maximum Likelihood Estimation

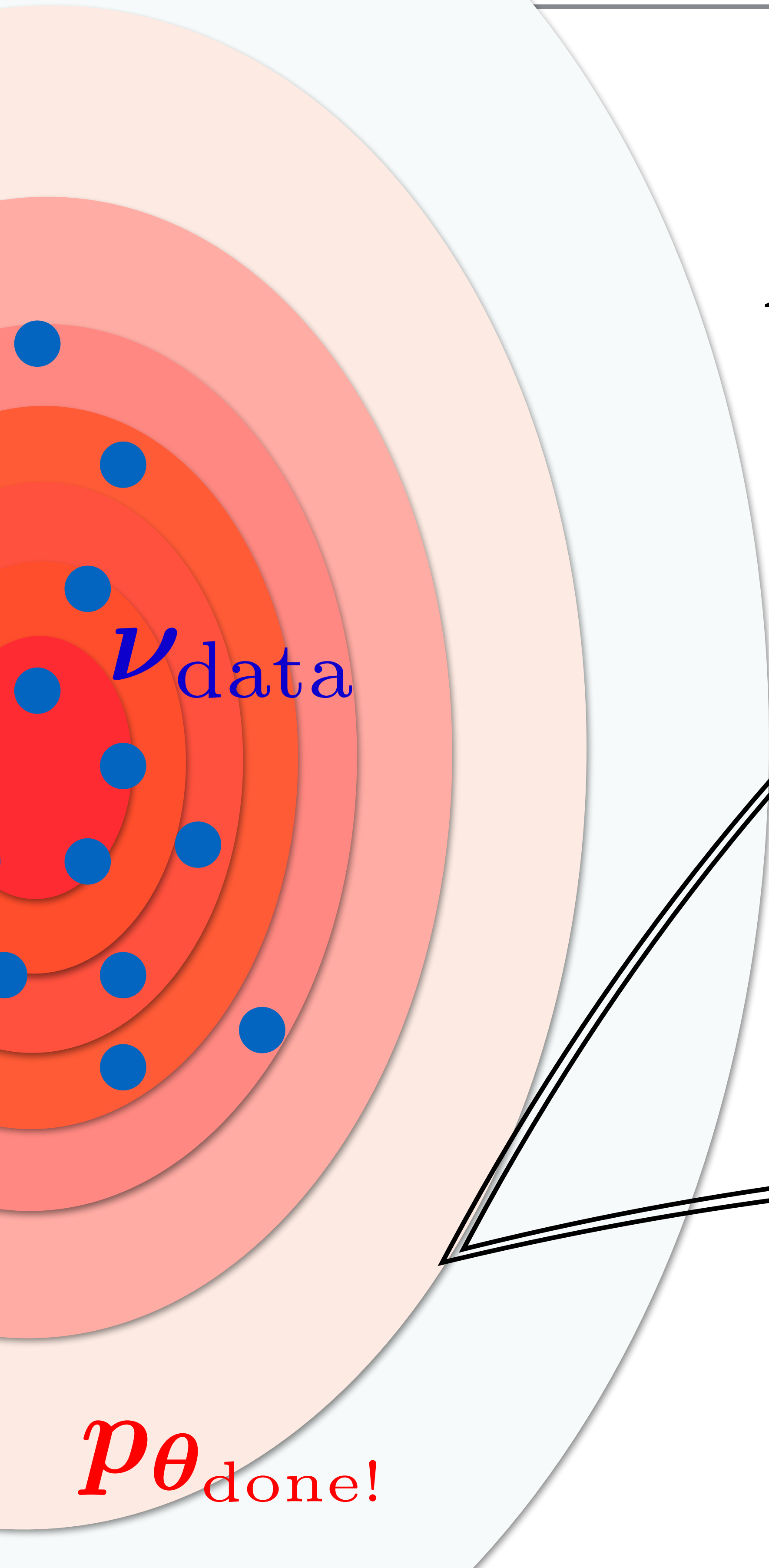


Equivalent to a KL projection in the space of probability measures

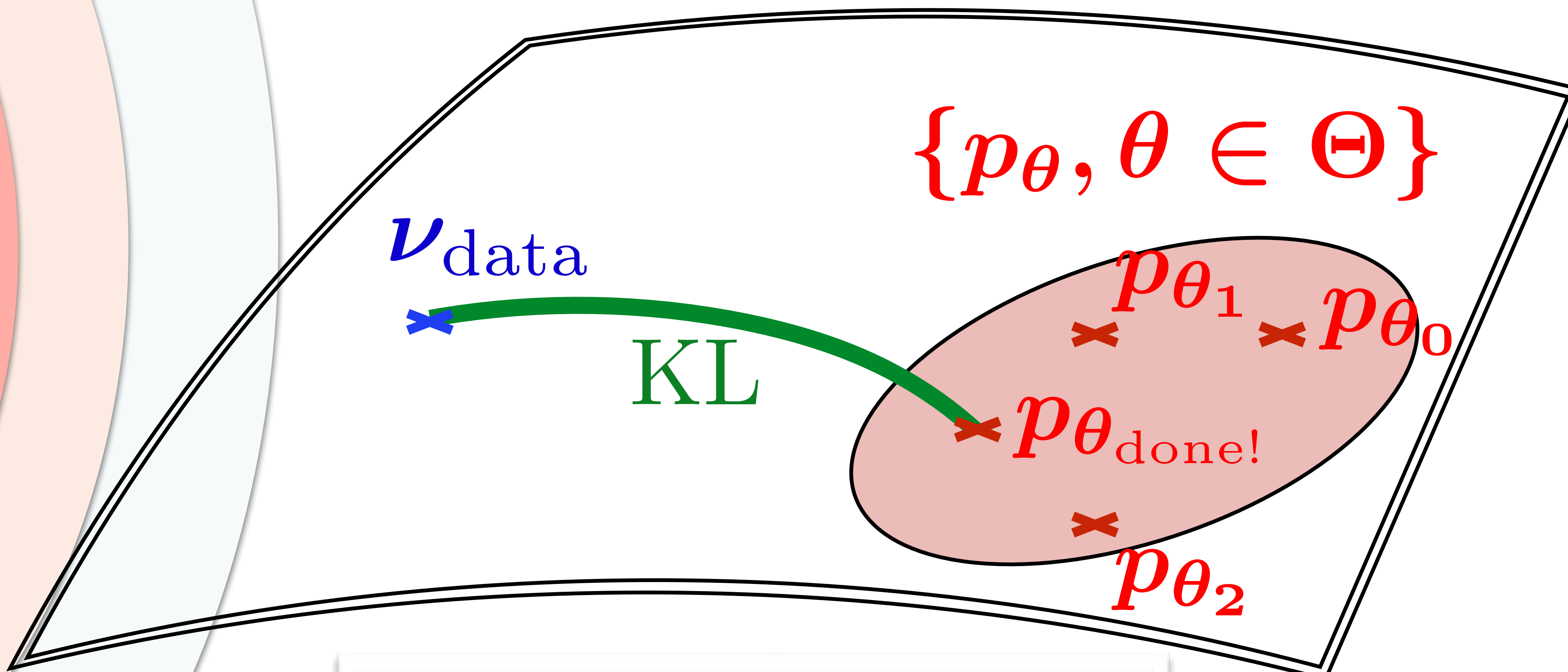


$$\min_{\boldsymbol{\theta} \in \Theta} \mathrm{KL}(\boldsymbol{\nu}_{\text{data}} \| \boldsymbol{p}_{\boldsymbol{\theta}})$$

Maximum Likelihood Estimation

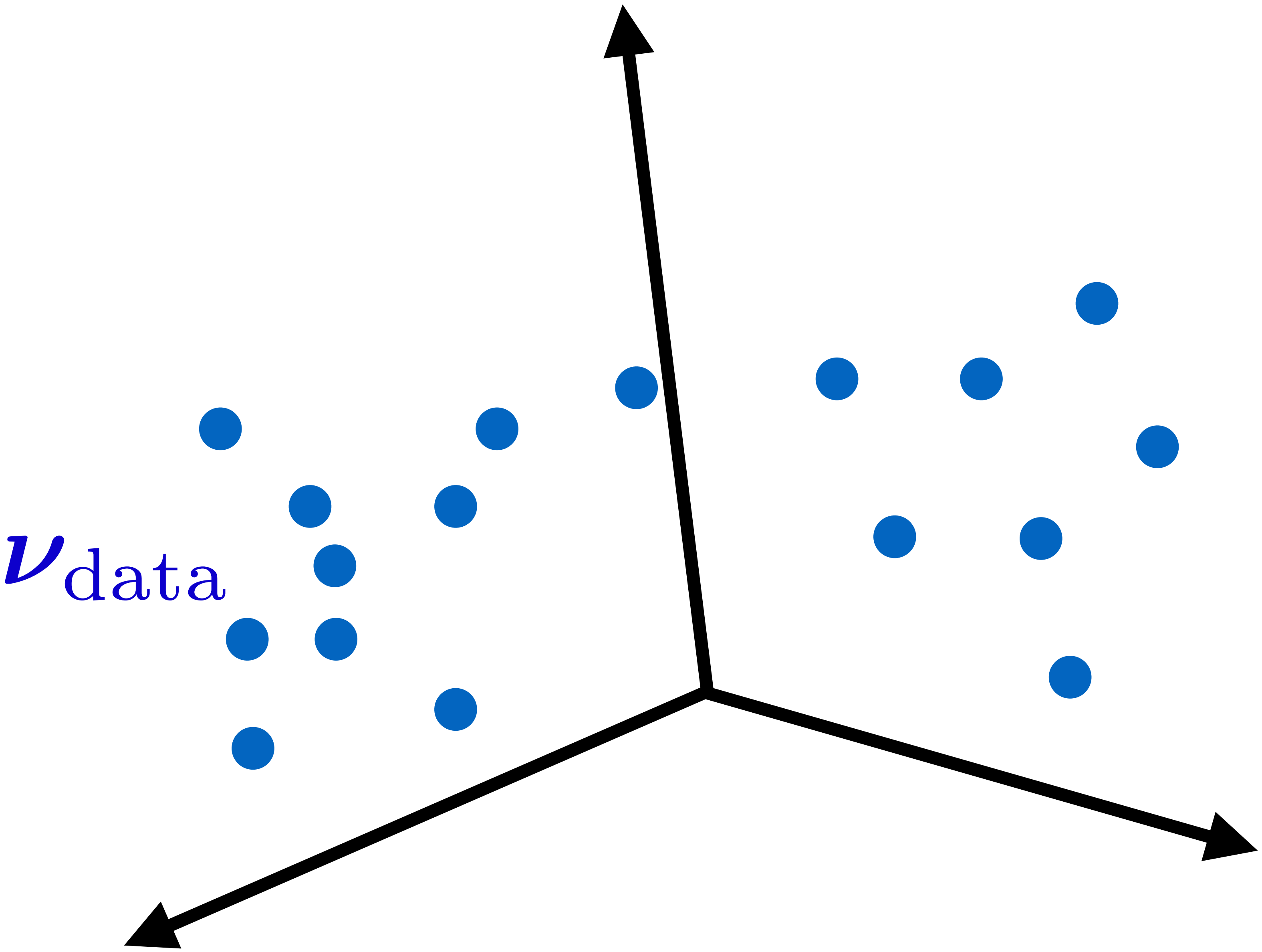


Equivalent to a KL projection in the space of probability measures



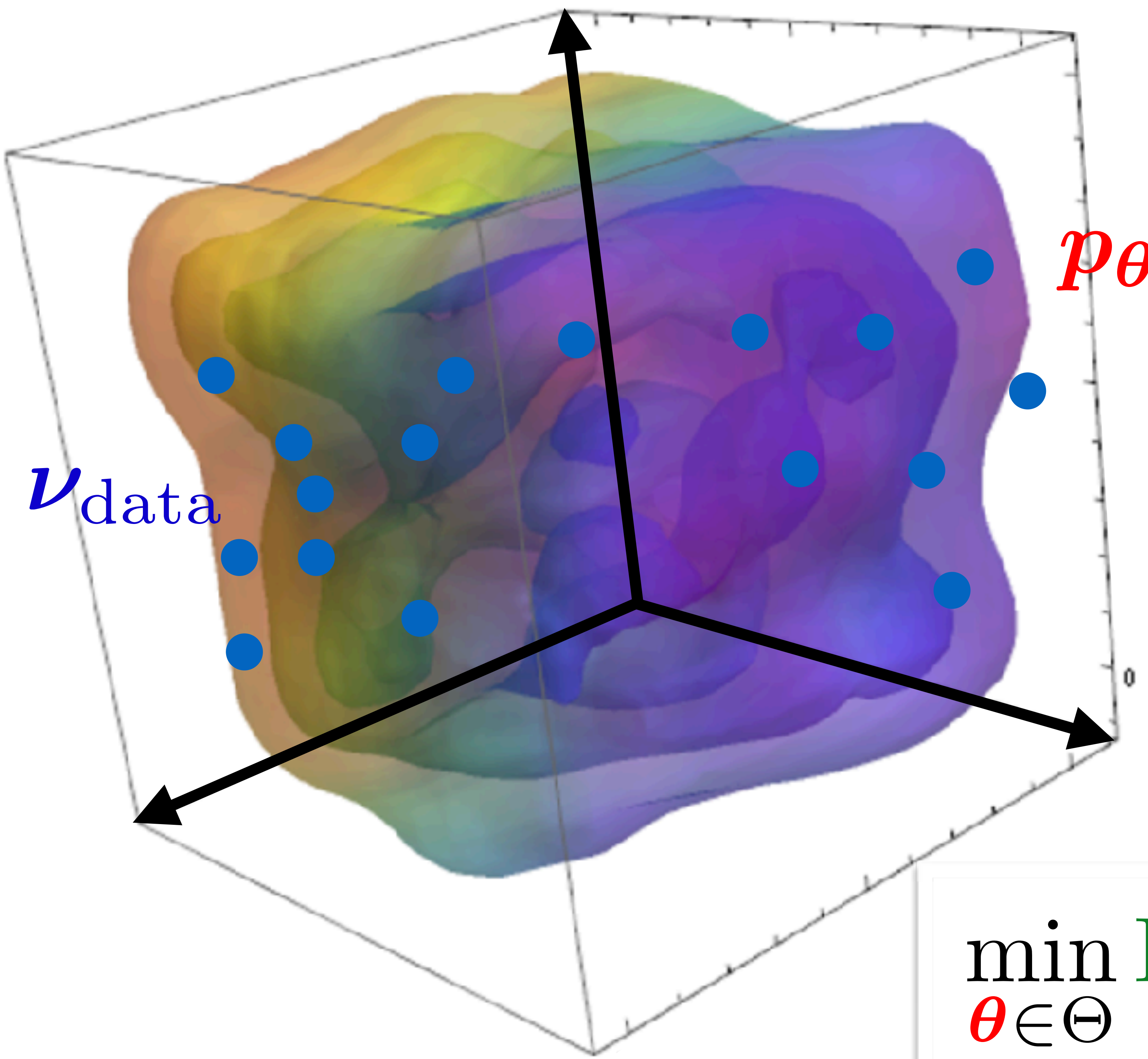
$$\min_{\theta \in \Theta} \mathrm{KL}(\nu_{\text{data}} \| p_\theta)$$

In higher dimensional spaces...



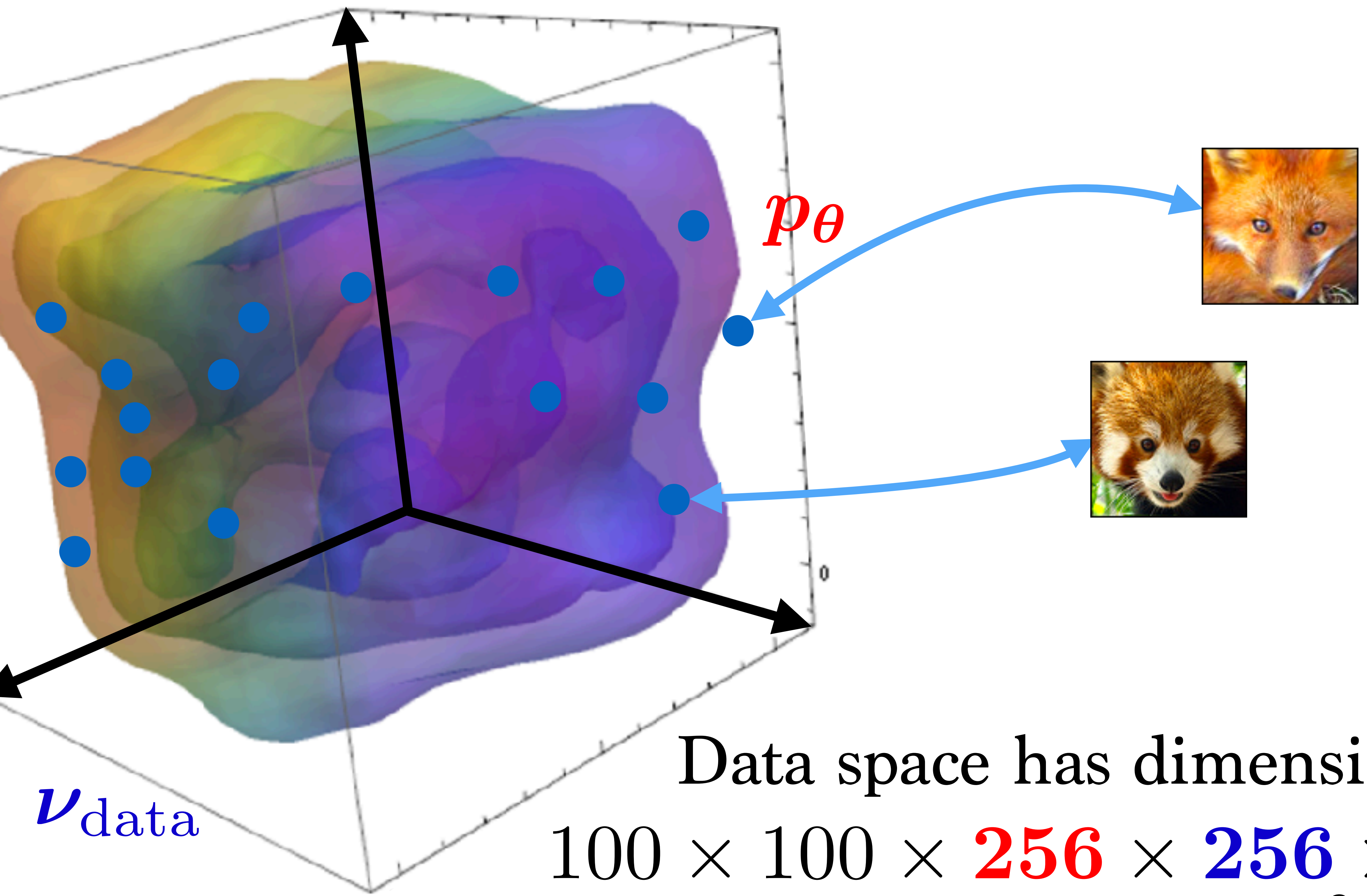
$$\min_{\theta \in \Theta} \mathrm{KL}(\nu_{\text{data}} \| p_{\theta})$$

In higher dimensional spaces...



$$\min_{\theta \in \Theta} \mathrm{KL}(\nu_{\text{data}} \| p_{\theta})$$

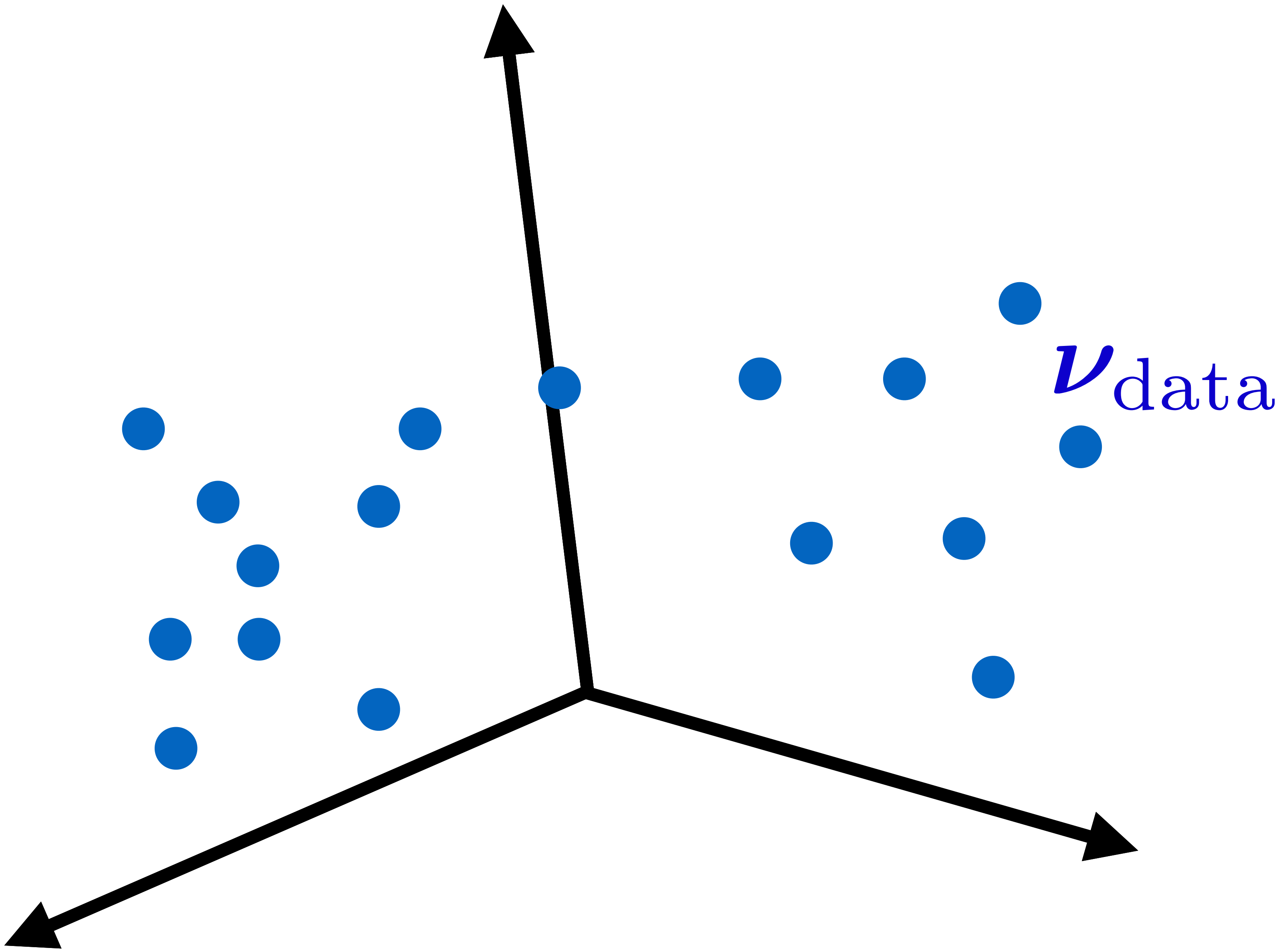
In higher dimensional spaces...



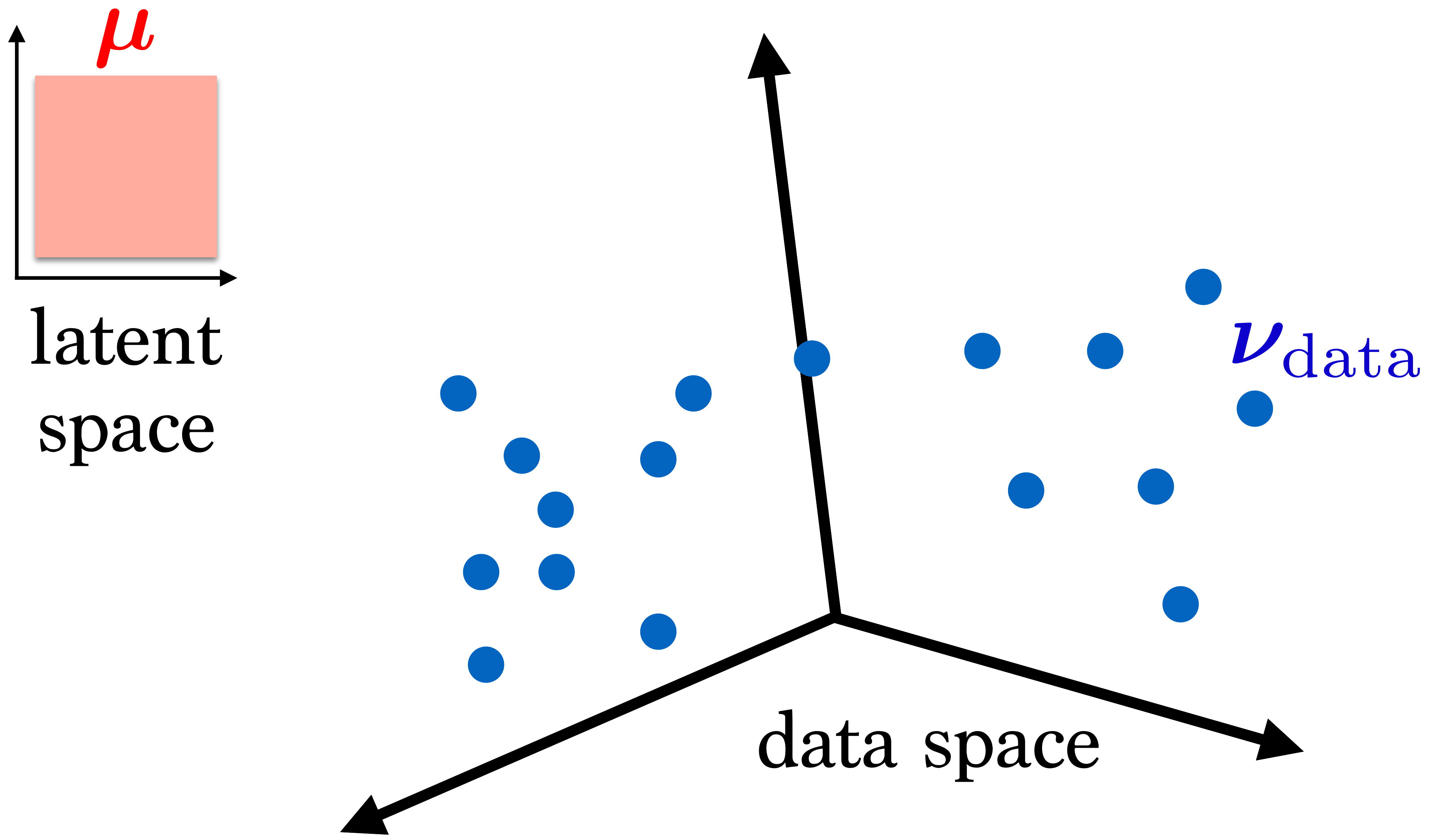
Data space has dimension

$$100 \times 100 \times \textcolor{red}{\mathbf{256}} \times \textcolor{blue}{\mathbf{256}} \times \textcolor{green}{\mathbf{256}}$$
$$\approx 167 \times 10^9$$

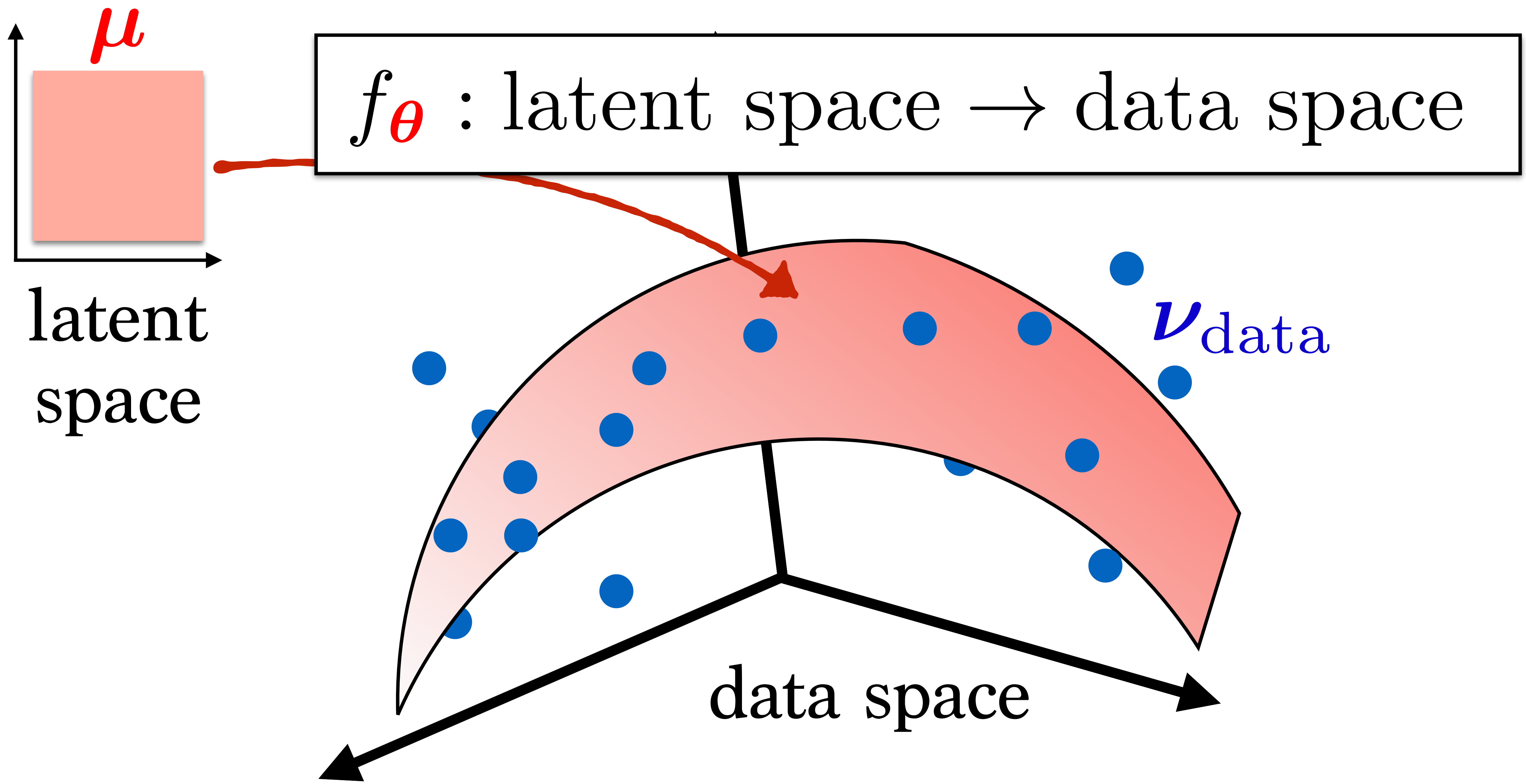
Generative Models



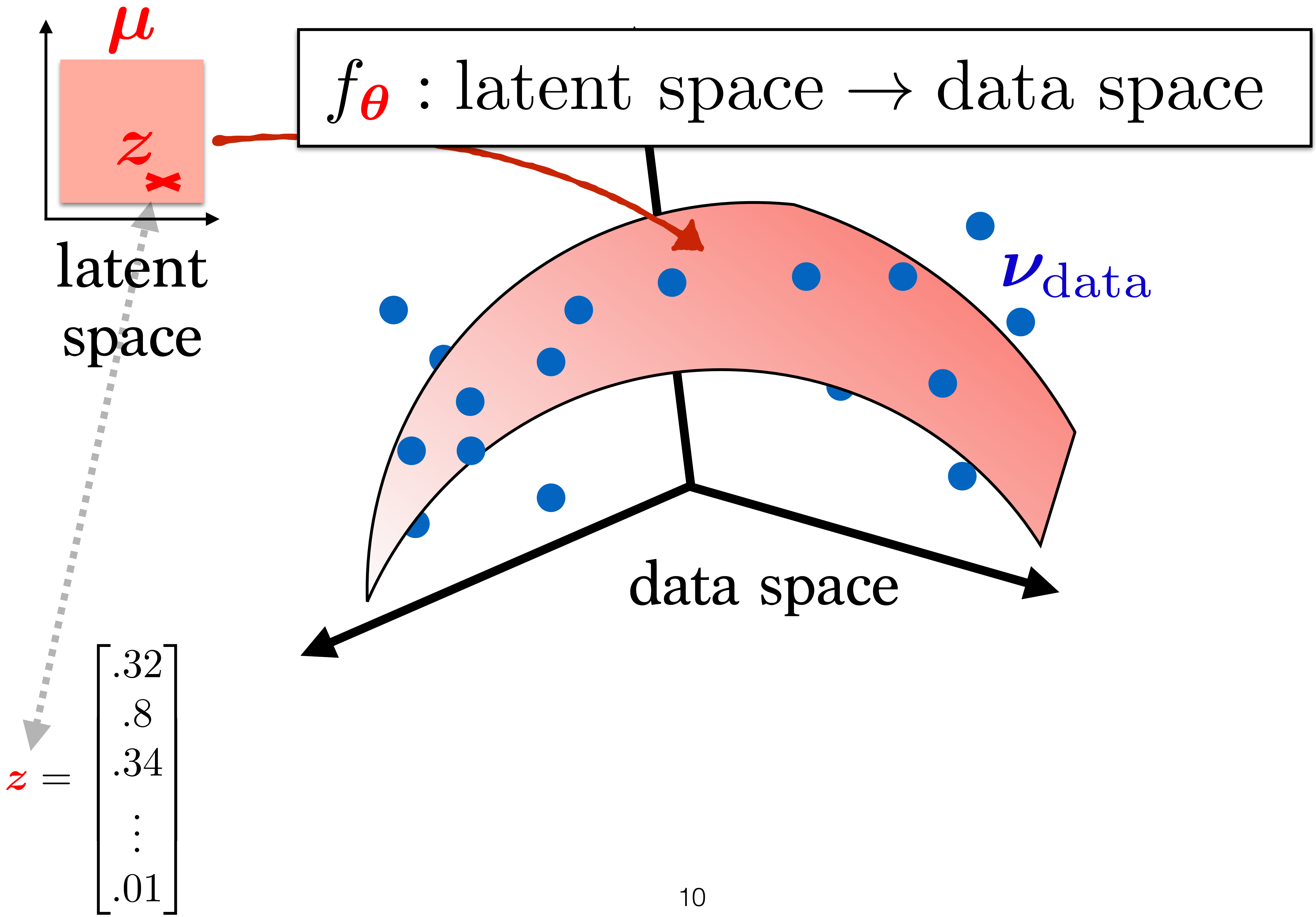
Generative Models



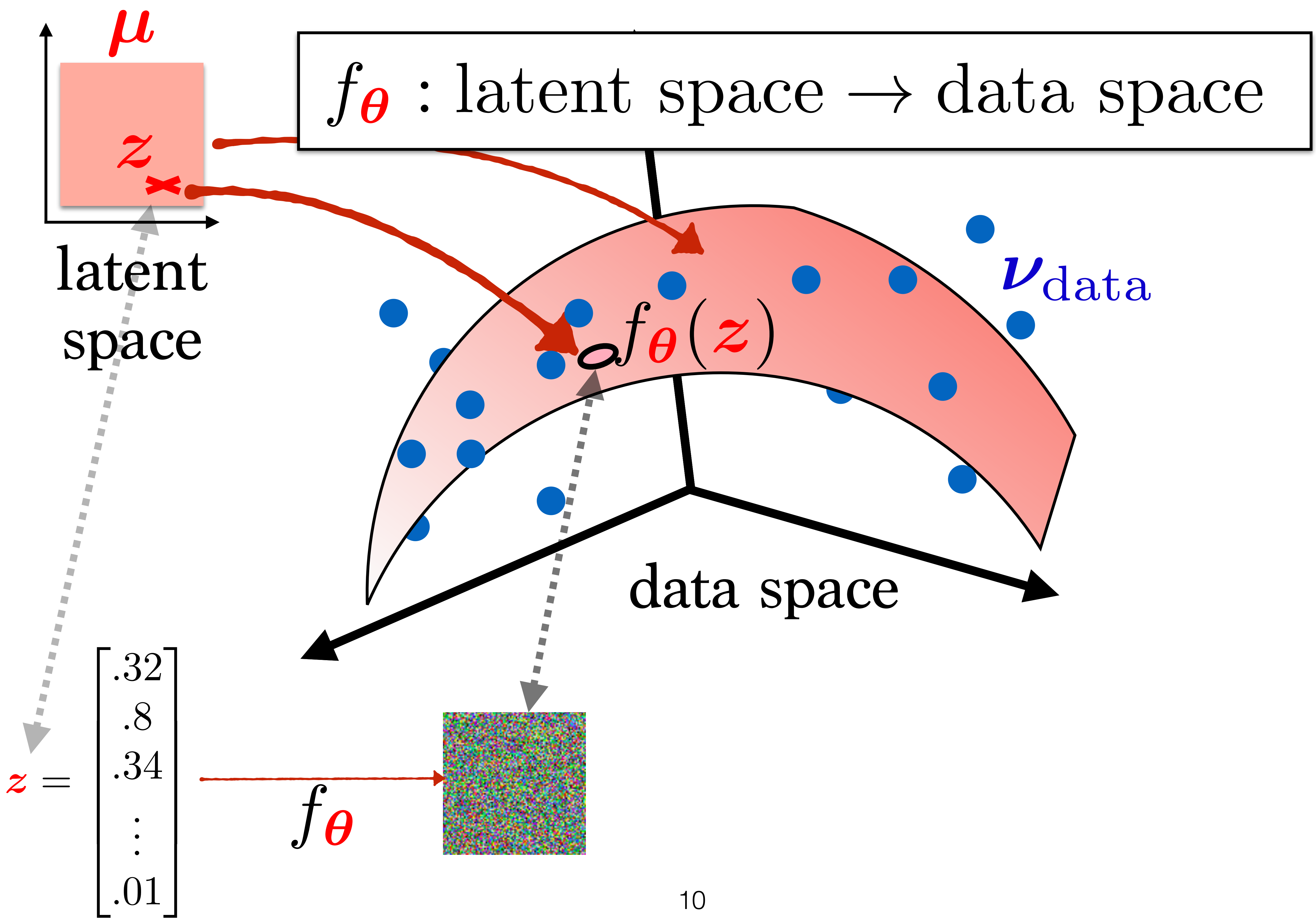
Generative Models



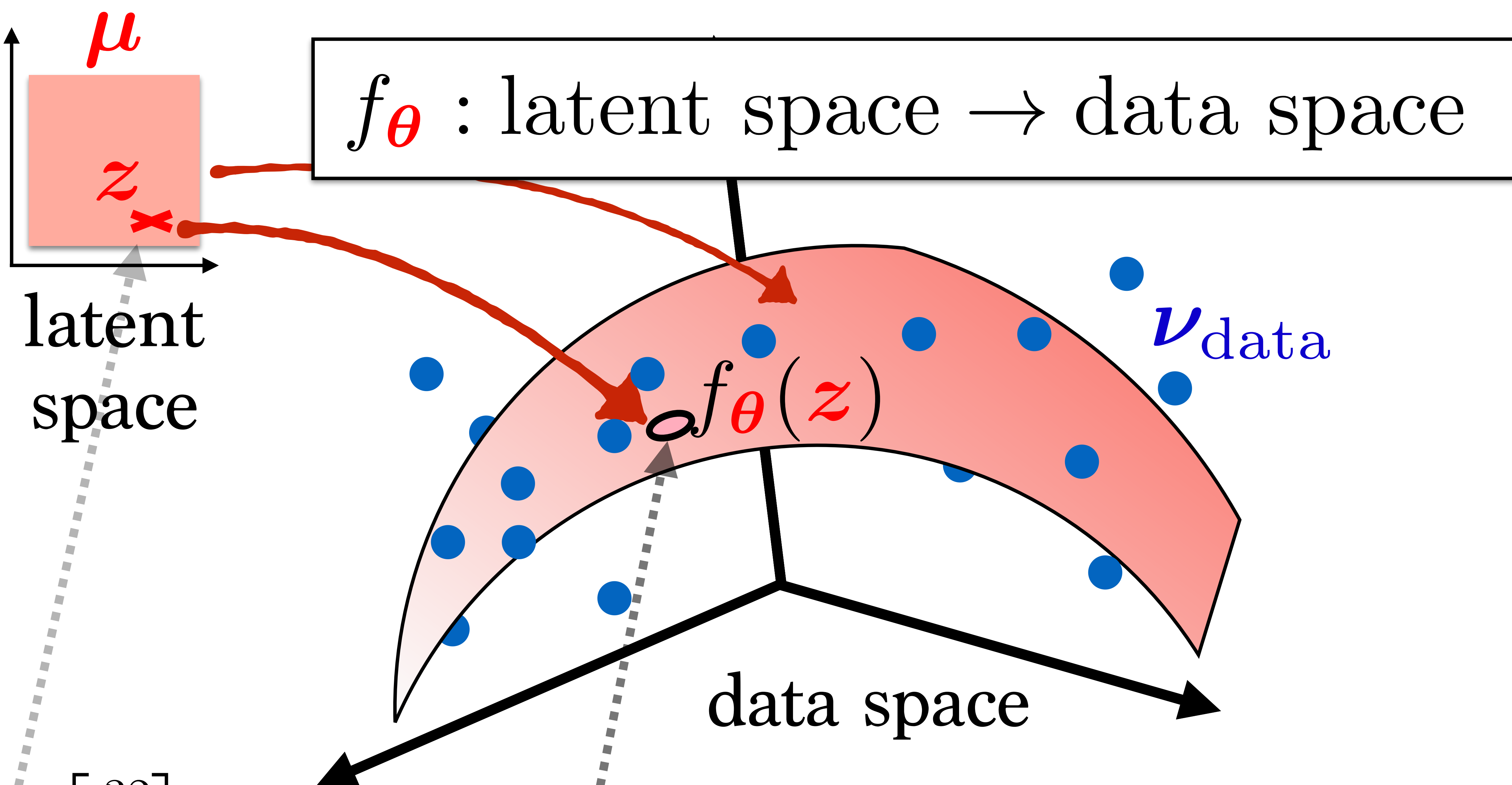
Generative Models



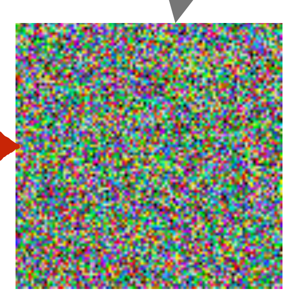
Generative Models

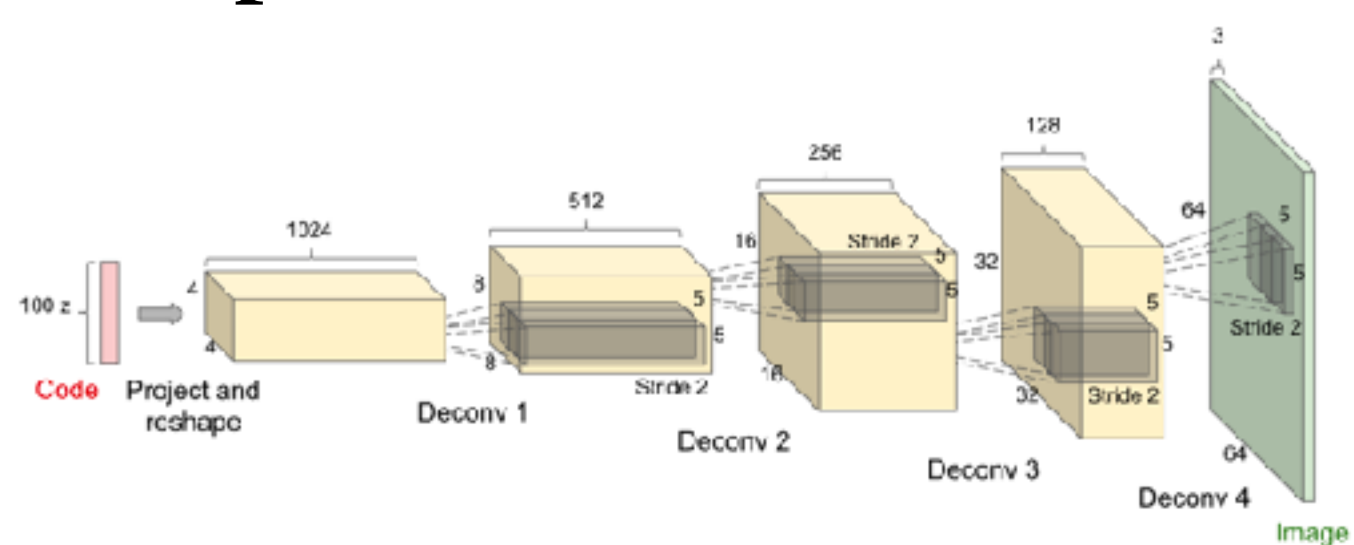


Generative Models

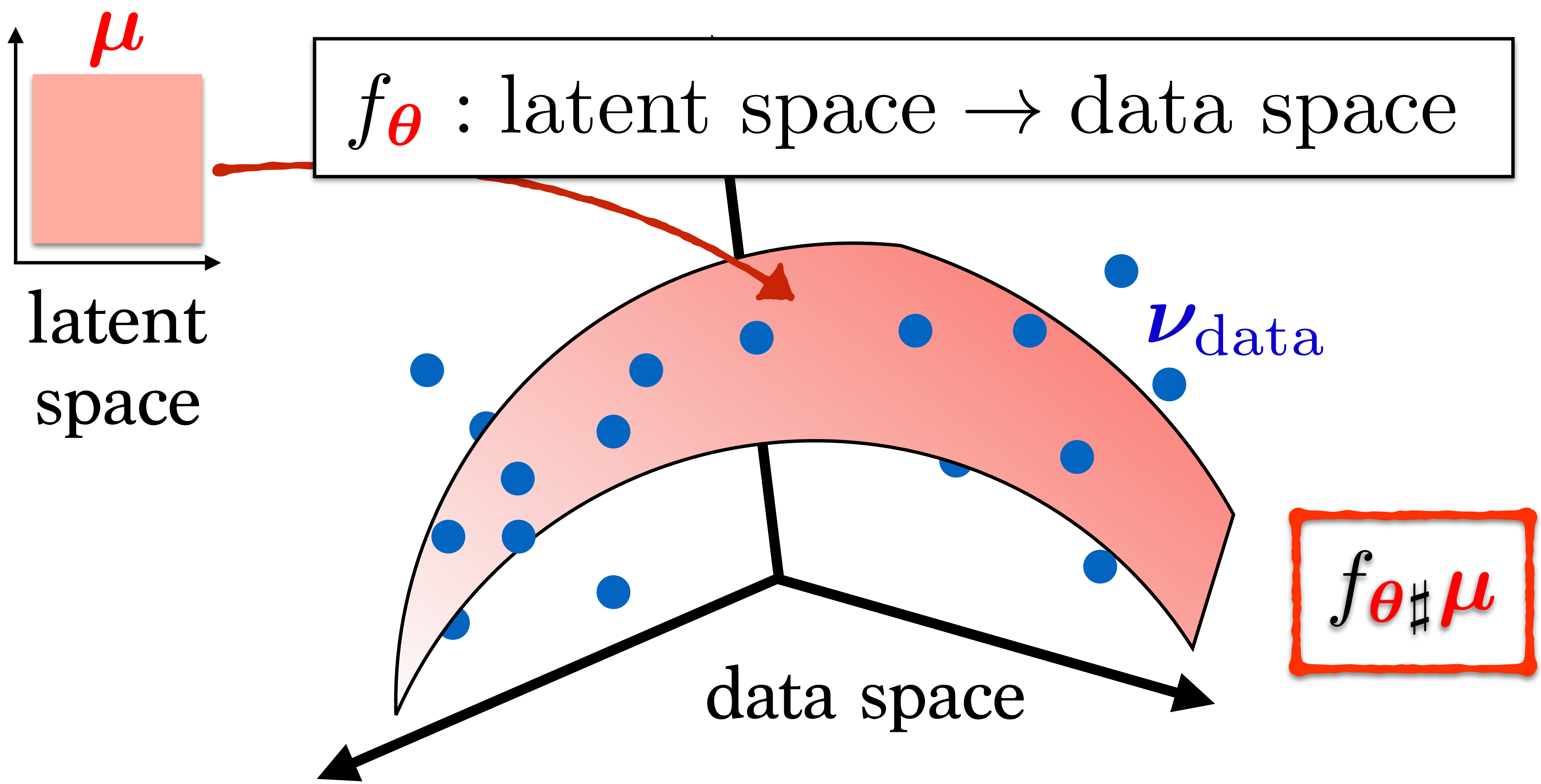


$$z = \begin{bmatrix} .32 \\ .8 \\ .34 \\ \vdots \\ .01 \end{bmatrix} \xrightarrow{f_\theta}$$

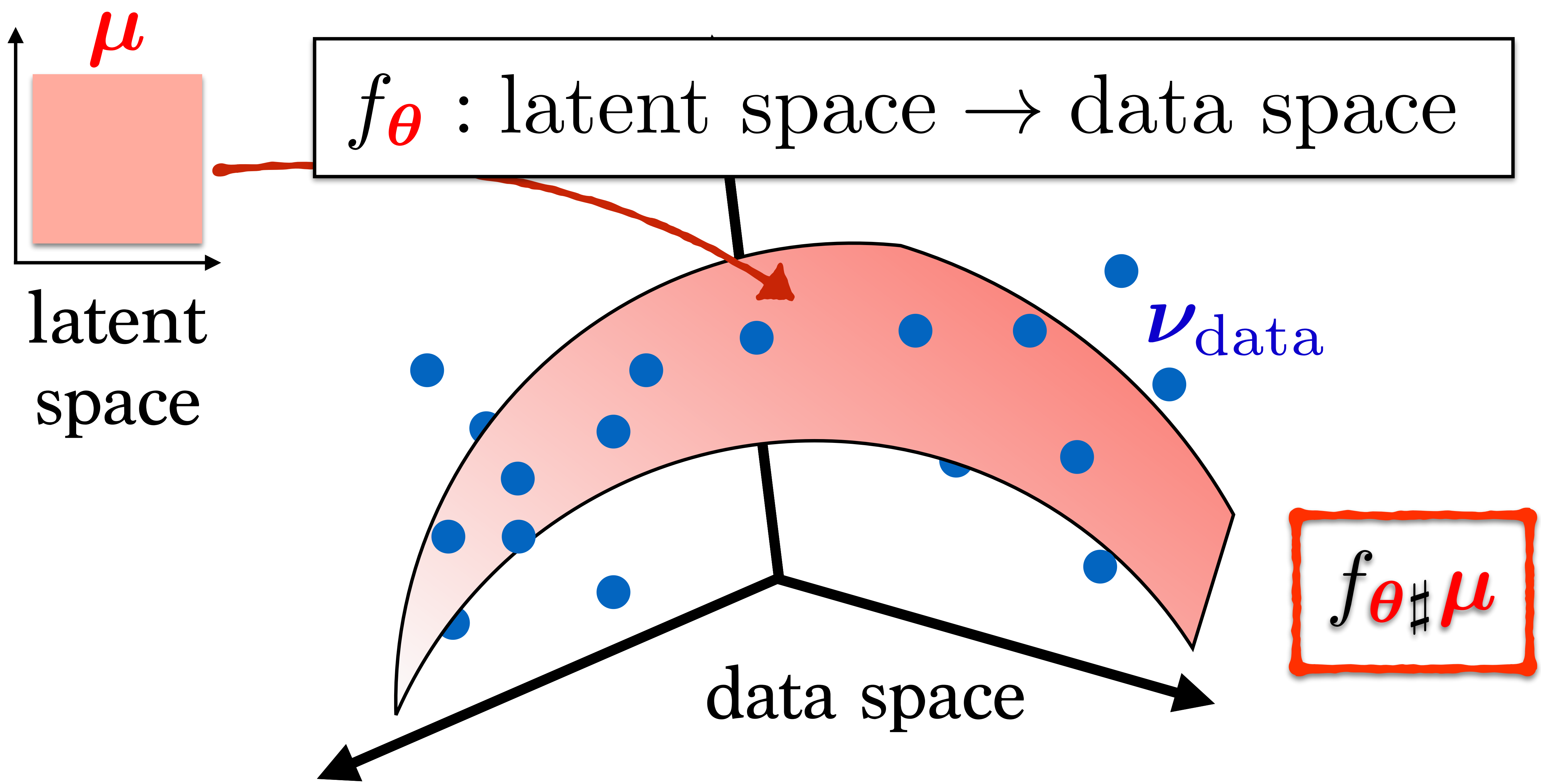




Generative Models

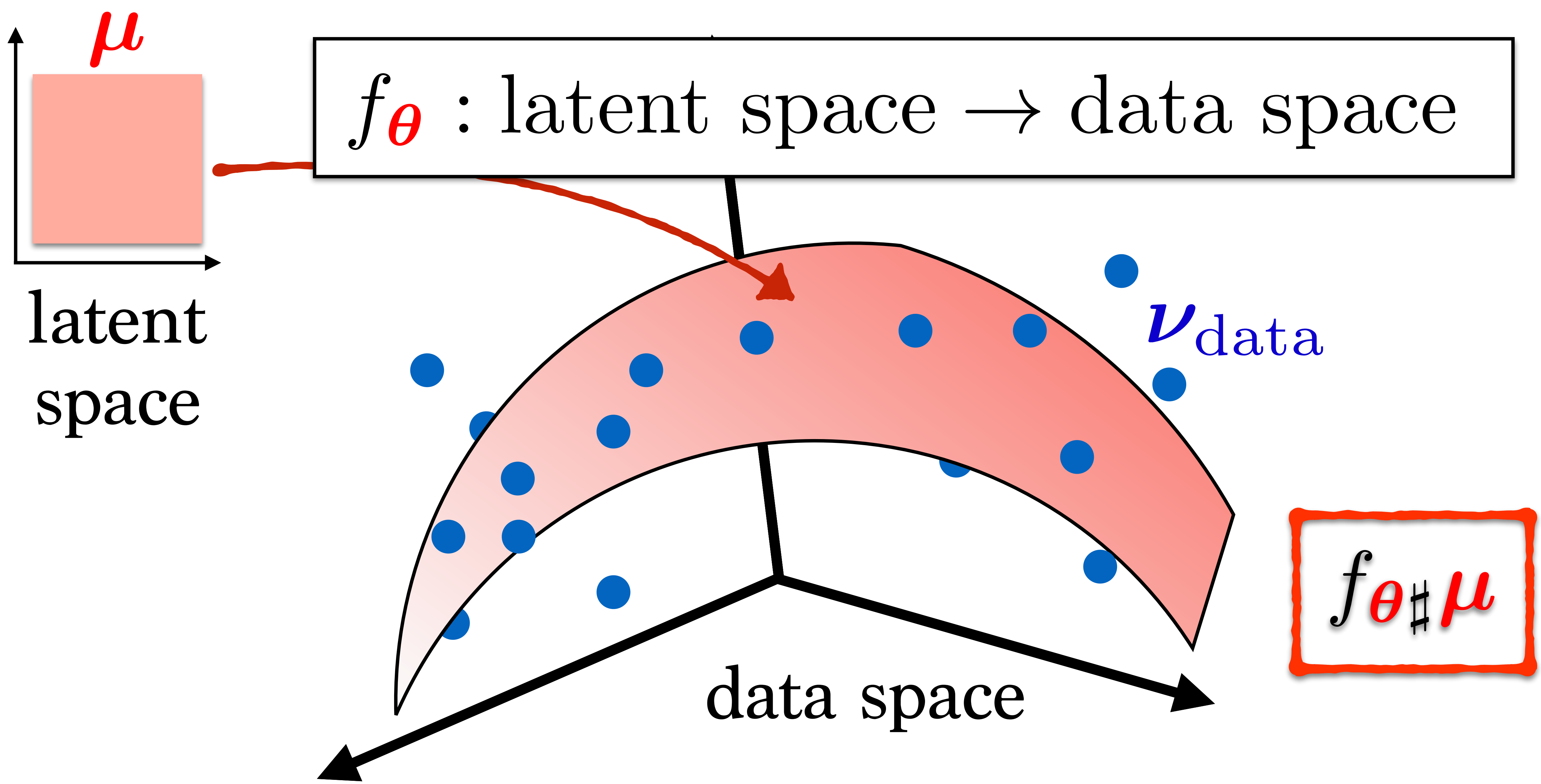


Generative Models



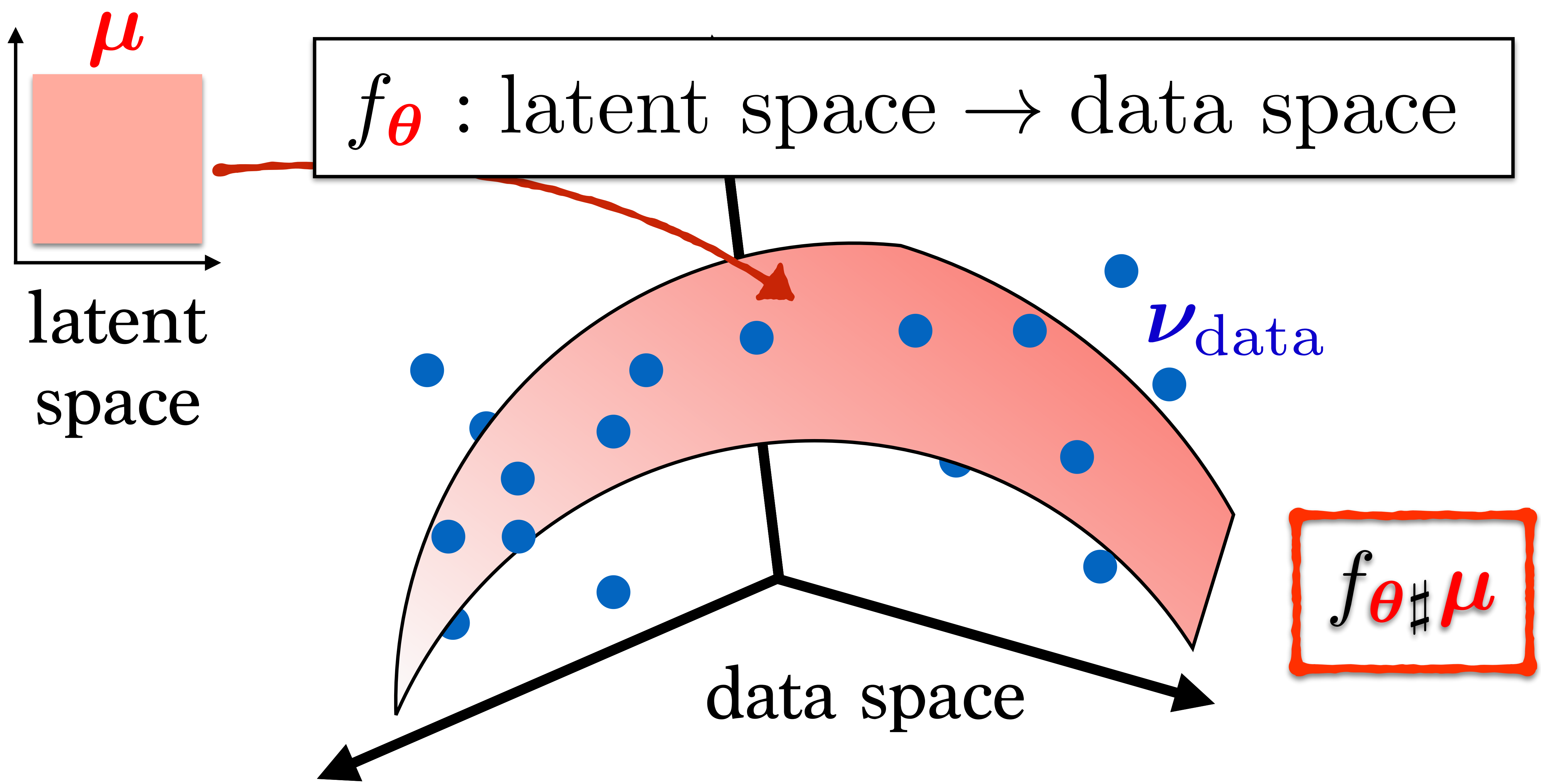
Push-forward: $\forall B \subset \Omega,\ f_\sharp\mu(B) := \mu(f^{-1}(B))$

Generative Models



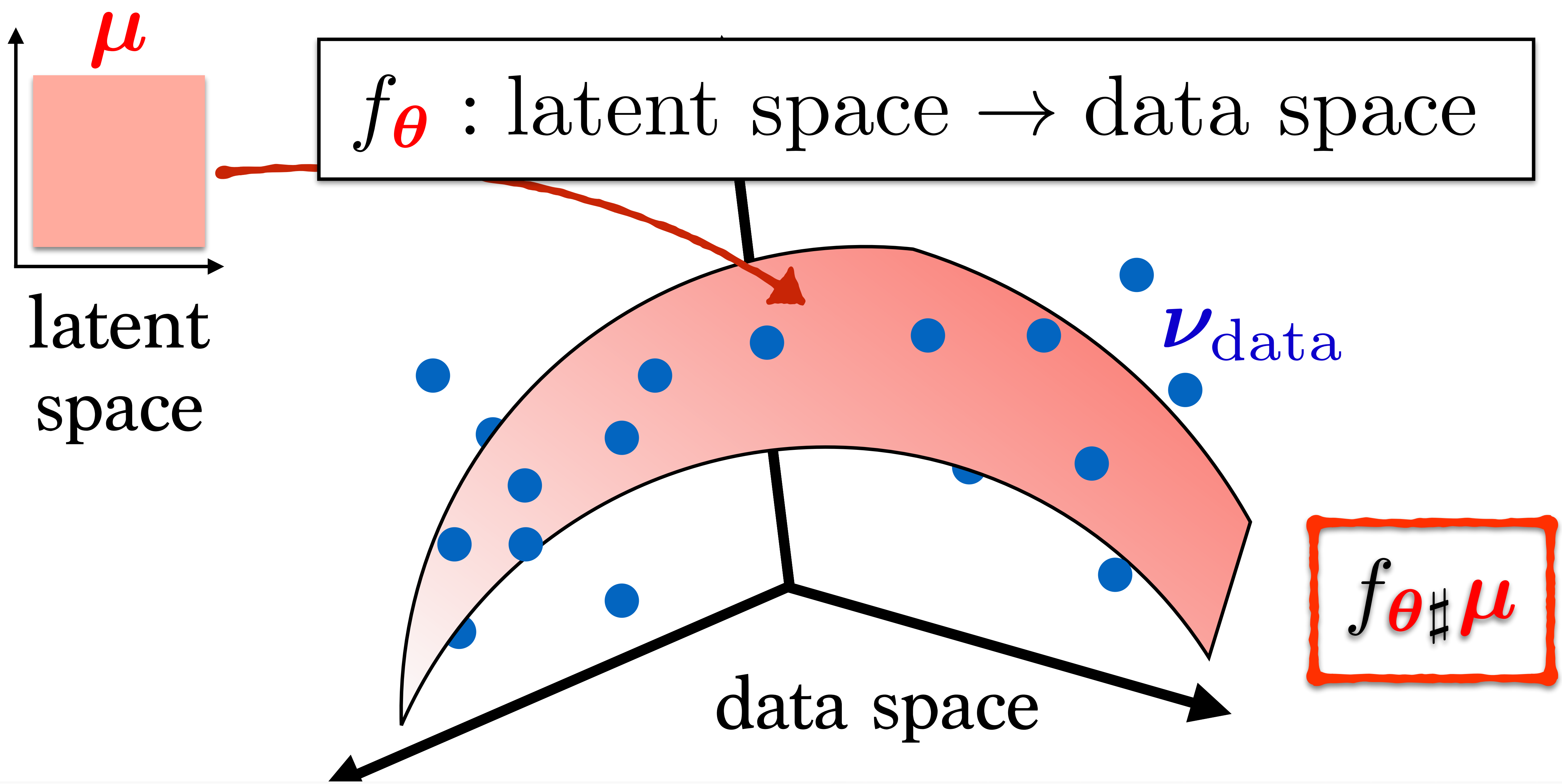
Goal: find $\boldsymbol{\theta}$ such that $f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu}$ fits $\boldsymbol{\nu}_{\text{data}}$

Generative Models



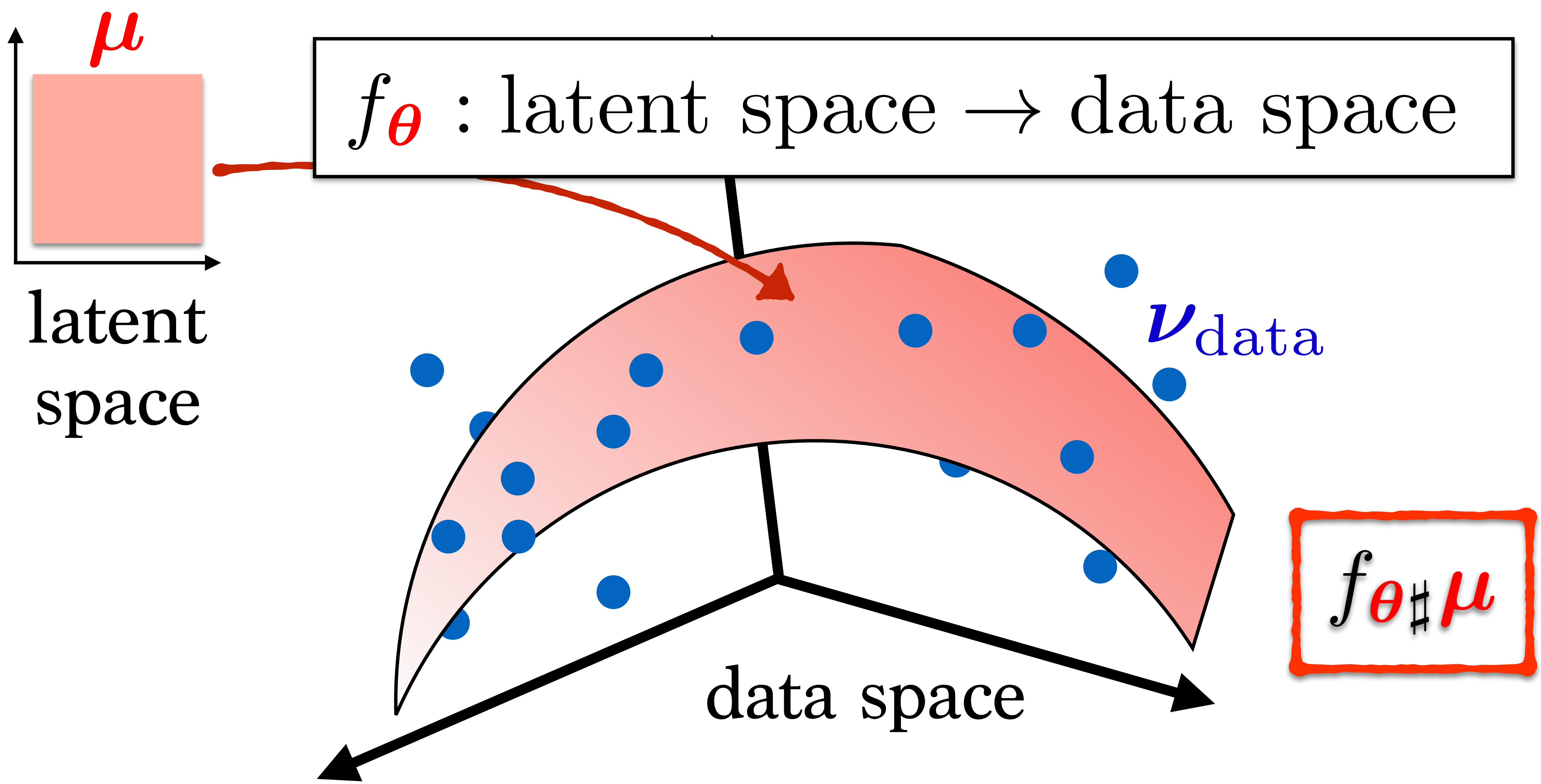
Goal: find θ such that $f_{\theta\sharp}\mu$ fits ν_{data}

Generative Models



Difference between fitting a push forward measure $f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu}$ *vs.* a density $p_{\boldsymbol{\theta}}$?

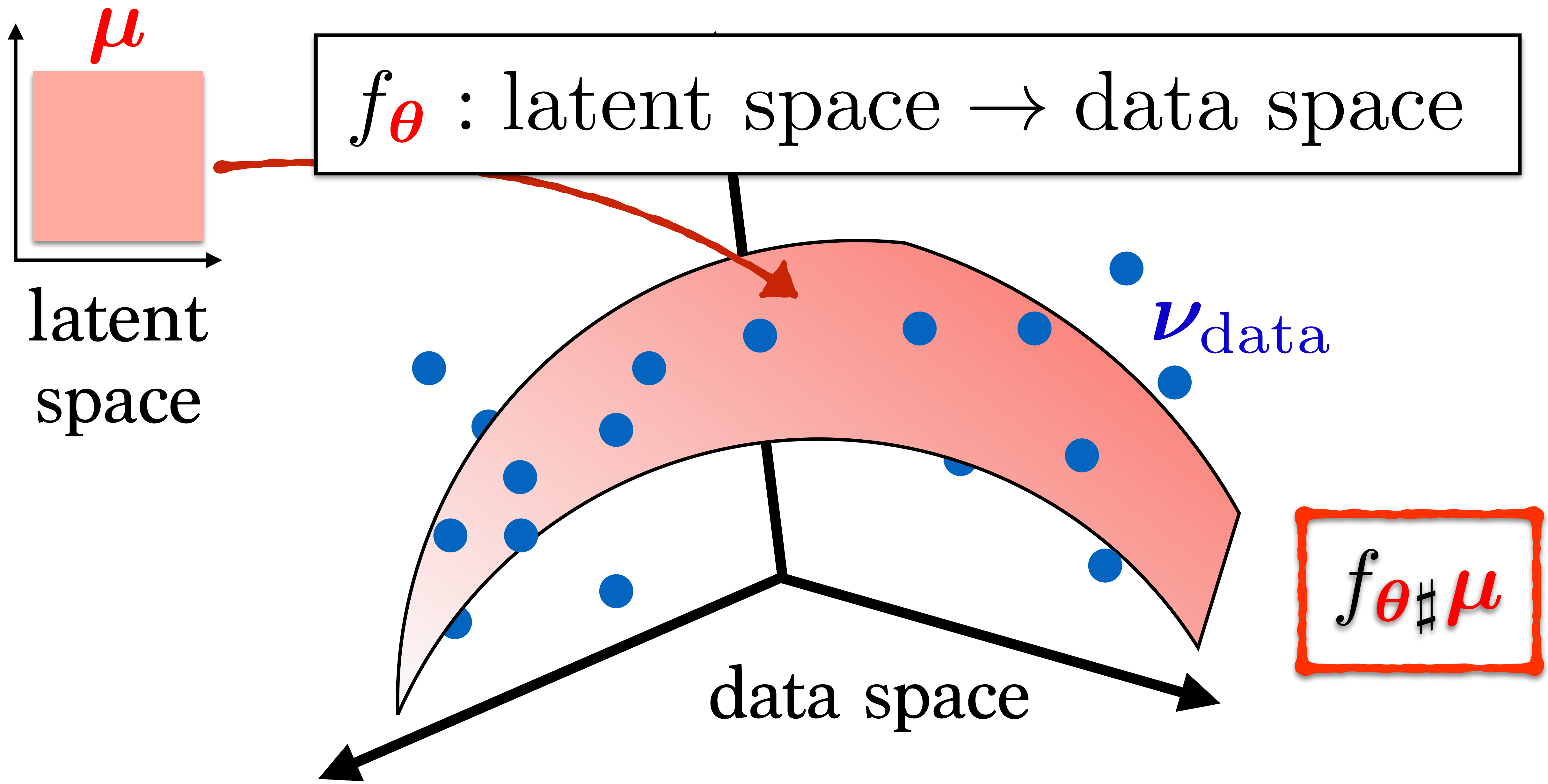
Generative Models



MLE

$$\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \log \boldsymbol{p}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \quad = \quad \min_{\boldsymbol{\theta}\in\Theta} \mathrm{KL}(\boldsymbol{\nu}_{\text{data}} \| \boldsymbol{p}_{\boldsymbol{\theta}})$$

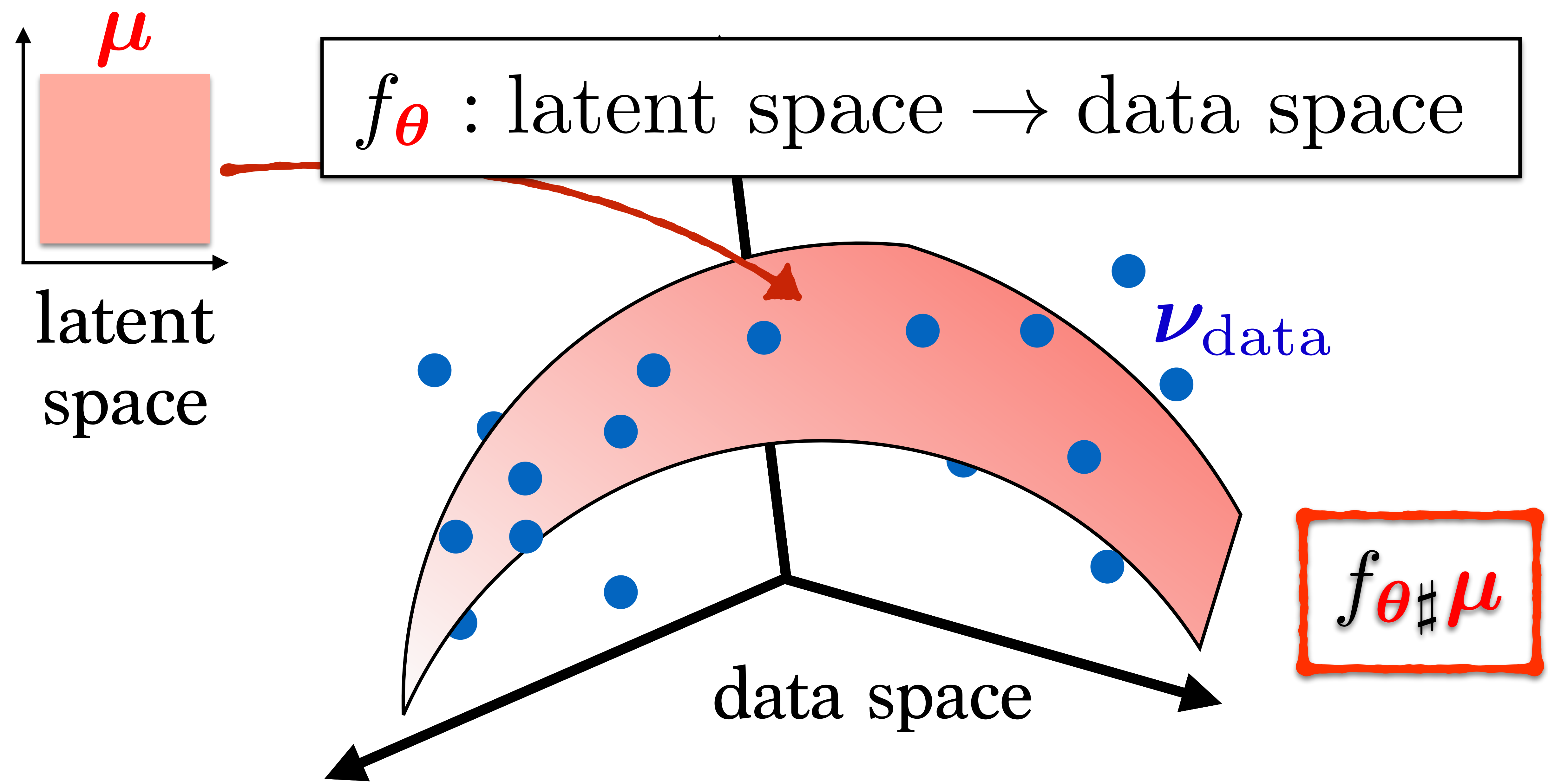
Generative Models



MLE

$$\max_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \log f_{\theta\sharp}\mu(x_i) \qquad \min_{\theta\in\Theta} \mathrm{KL}(\nu_{\text{data}} \| f_{\theta\sharp}\mu)$$

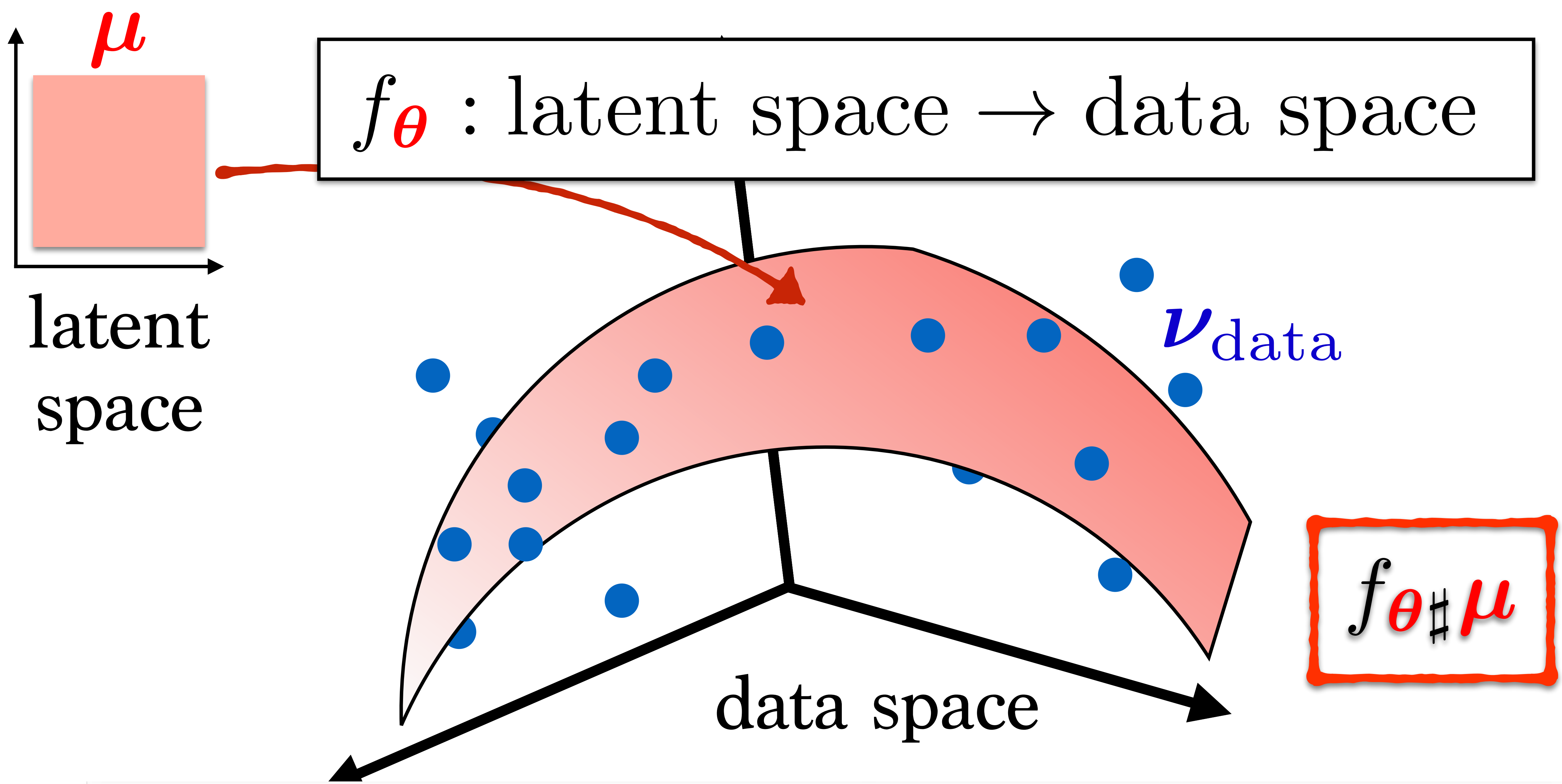
Generative Models



~~MLE~~

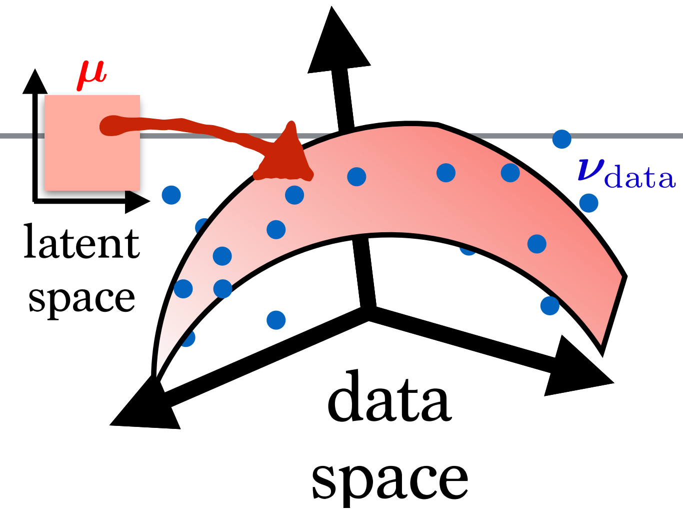
$$\max_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \log f_{\theta\sharp}\mu(x_i) \qquad \min_{\theta\in\Theta} \mathrm{KL}(\nu_{\text{data}} \| f_{\theta\sharp}\mu)$$

Generative Models



Need a more flexible **discrepancy function** to compare $\boldsymbol{\nu}_{\text{data}}$ and $f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu}$

Workarounds?



- Formulation as **adversarial problem** **[GPM...'14]**

$$\min_{\theta \in \Theta} \max_{\text{classifiers } g} \text{Accuracy}_g\left((f_{\theta \sharp}\mu, +1), (\nu_{\text{data}}, -1) \right)$$

- Use a **richer metric** Δ for probability measures, able to handle measures with non-overlapping supports:

$$\min_{\theta \in \Theta} \Delta(\nu_{\text{data}}, p_\theta), \quad \textbf{not} \min_{\theta \in \Theta} \mathrm{KL}(\nu_{\text{data}} \| p_\theta)$$

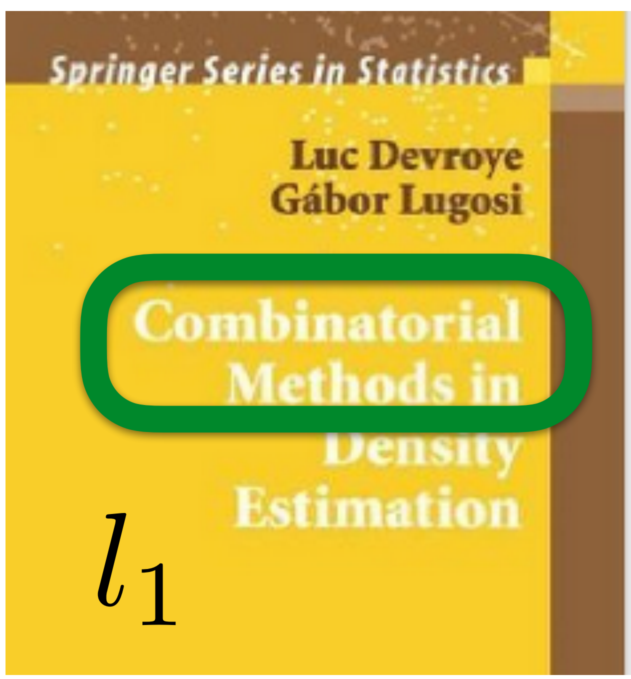
Minimum △ Estimation

The Annals of Statistics
1980, Vol. 8, No. 3, 457–487

MINIMUM CHI-SQUARE, NOT MAXIMUM LIKELIHOOD!

BY JOSEPH BERKSON
Mayo Clinic, Rochester, Minnesota

ELSEVIER
Computational Statistics & Data Analysis 29 (1998) 81–103
COMPUTATIONAL STATISTICS & DATA ANALYSIS



l_1

Minimum Hellinger distance estimation for Poisson mixtures

Dimitris Karlis, Evdokia Xekalaki*
Department of Statistics, Athens University of Economics and Business, 76 Patission Str., 104 34 Athens, Greece

ELSEVIER
Available online at www.sciencedirect.com
SCIENCE DIRECT
Statistics & Probability Letters 76 (2006) 1298–1302
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

On minimum Kantorovich distance estimators

Federico Bassetti[a], Antonella Bodini[b], Eugenio Regazzini[a,*]

Minimum Kantorovich Estimation

- Use optimal transport theory, namely *Wasserstein distances* to define discrepancy Δ.

$$\min_{\boldsymbol{\theta}\in\Theta} W(\boldsymbol{\nu}_{\text{data}}, f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu})$$

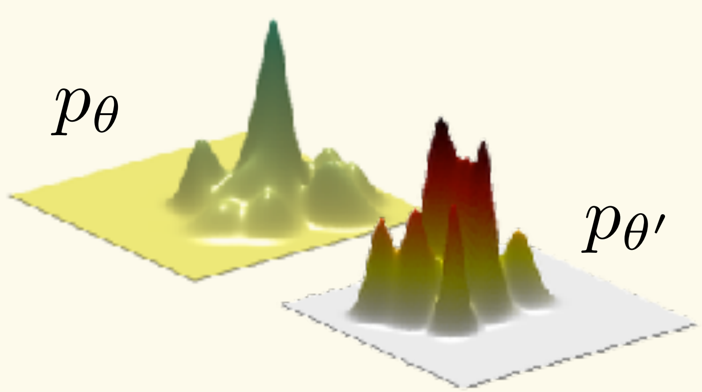
- Optimal transport? fertile field in mathematics.

Monge | Kantorovich | Koopmans | Dantzig | Brenier | Otto | McCann | Villani

Nobel '75 | Fields '10

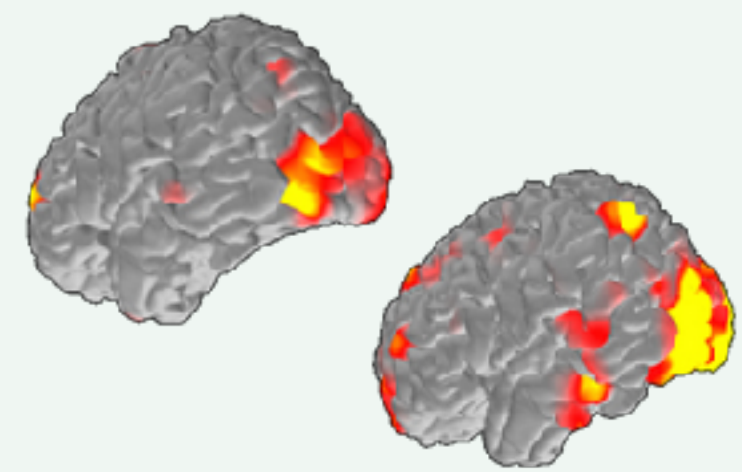
What is Optimal Transport?

A geometric toolbox to
compare **probability measures**
supported on a metric space.



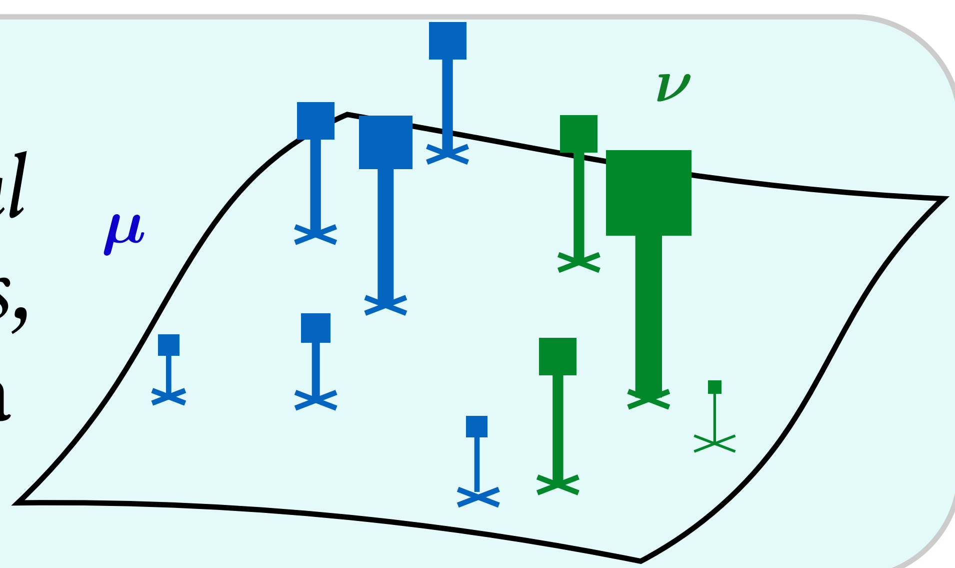
Statistical Models

Bags of features



Brain Activation Maps

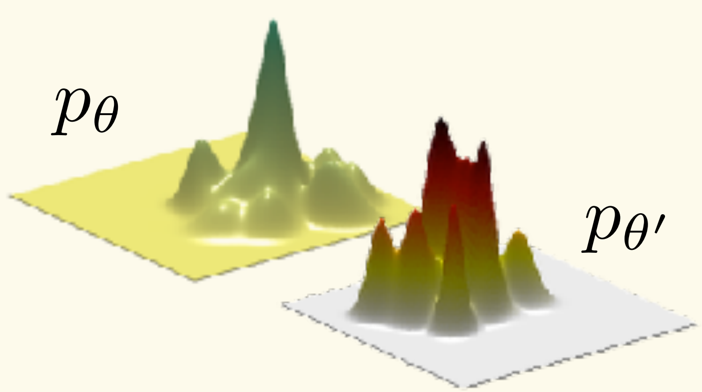
Empirical Measures, i.e. **data**



Color Histograms

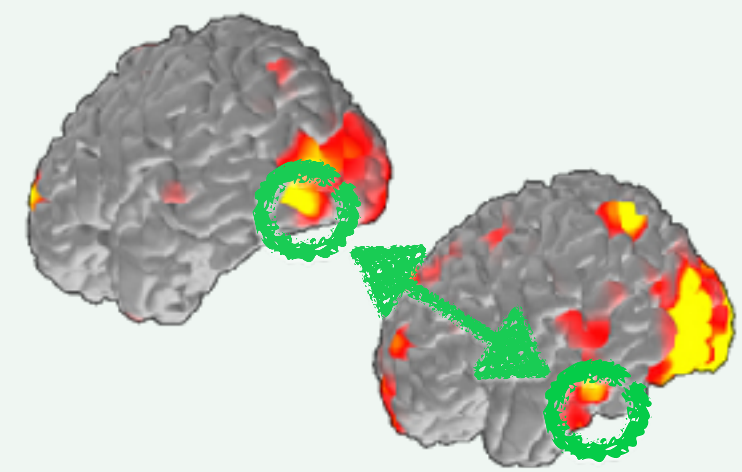
What is Optimal Transport?

A geometric toolbox to
compare probability measures
supported on a **metric space**.



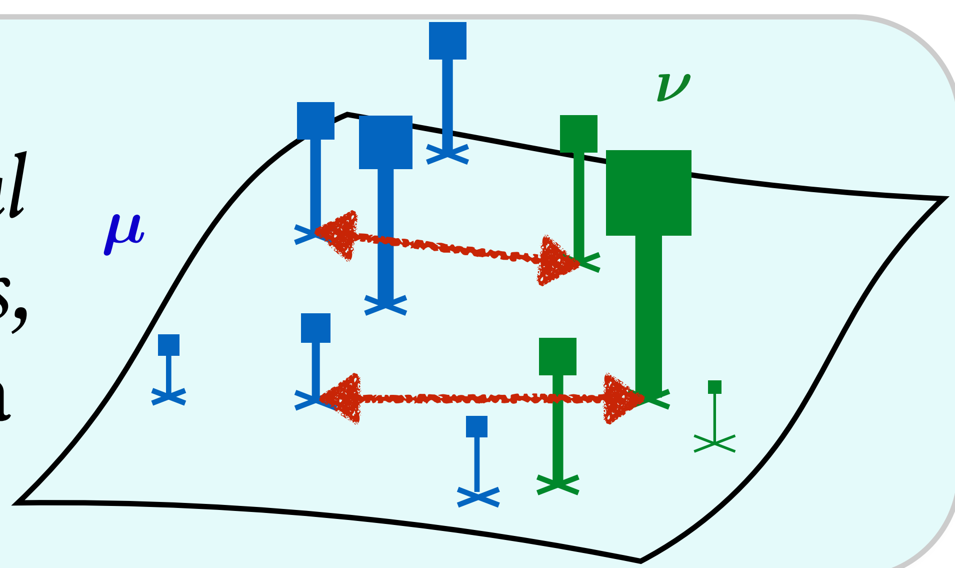
Statistical Models

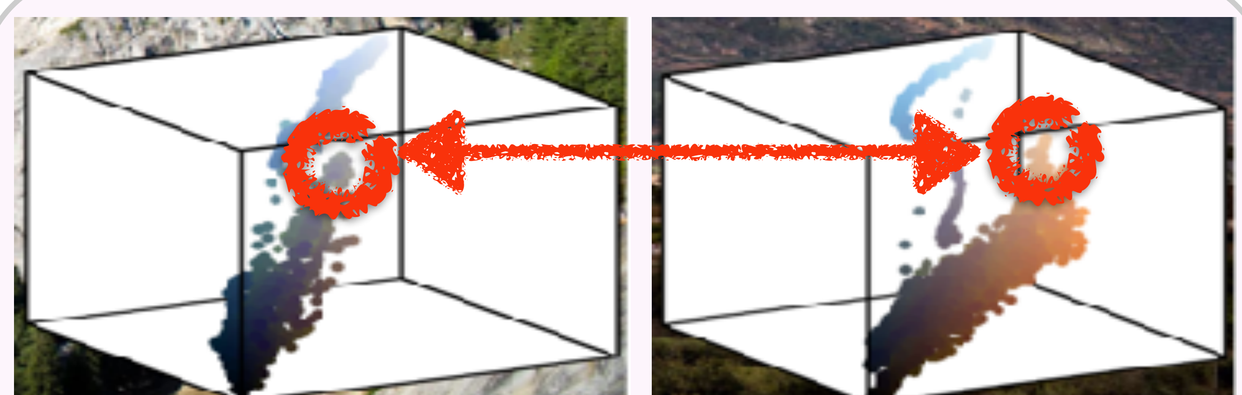
Bags of features



Brain Activation Maps

Empirical Measures, i.e. **data**





Color Histograms

What is Optimal Transport?

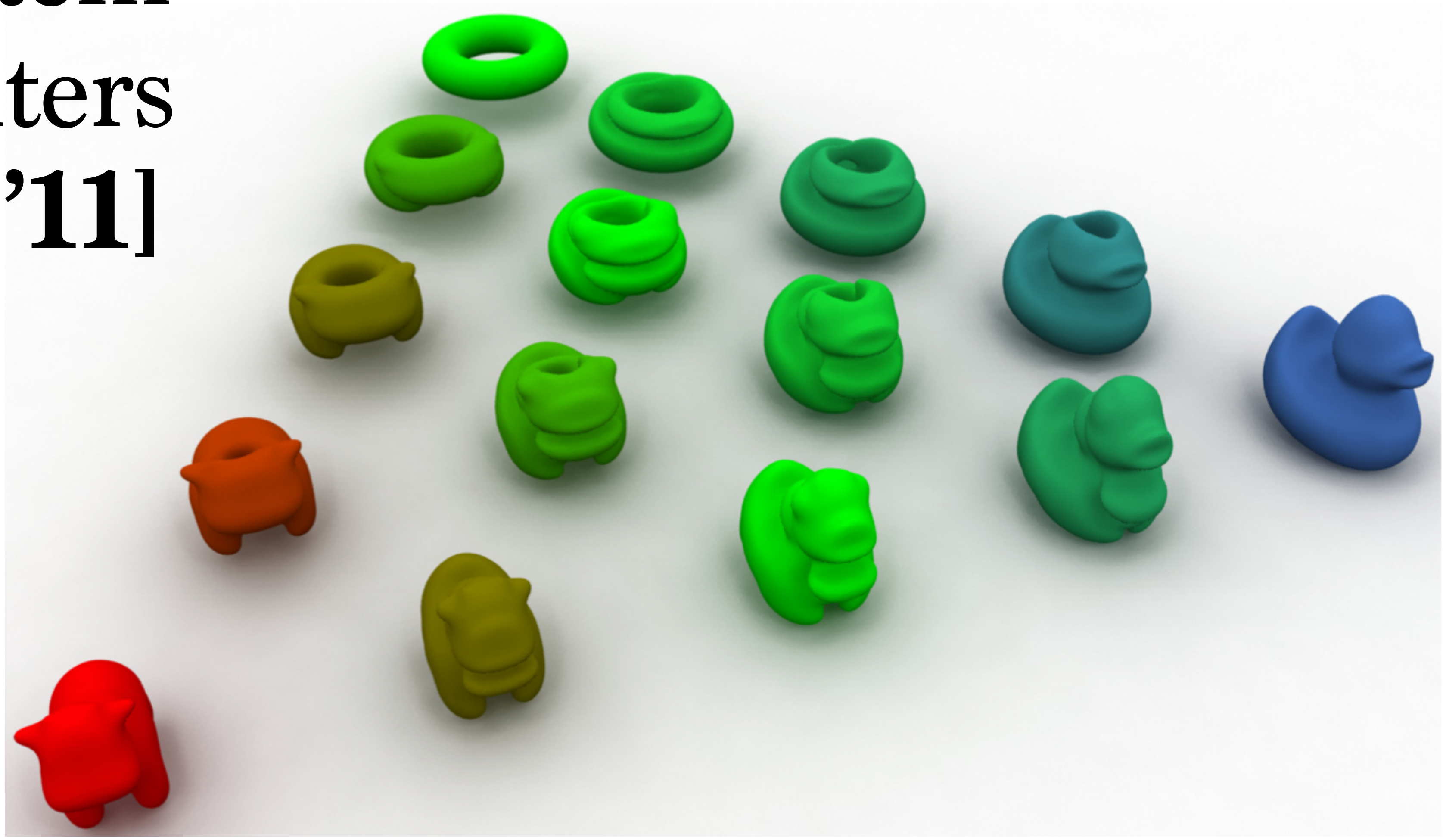
A powerful **geometric toolbox** to compare **probability measures.**

[SDPC..'15]

What is Optimal Transport?

A powerful **geometric toolbox** to compare **probability measures.**

Wasserstein Barycenters **[Agueh'11]**

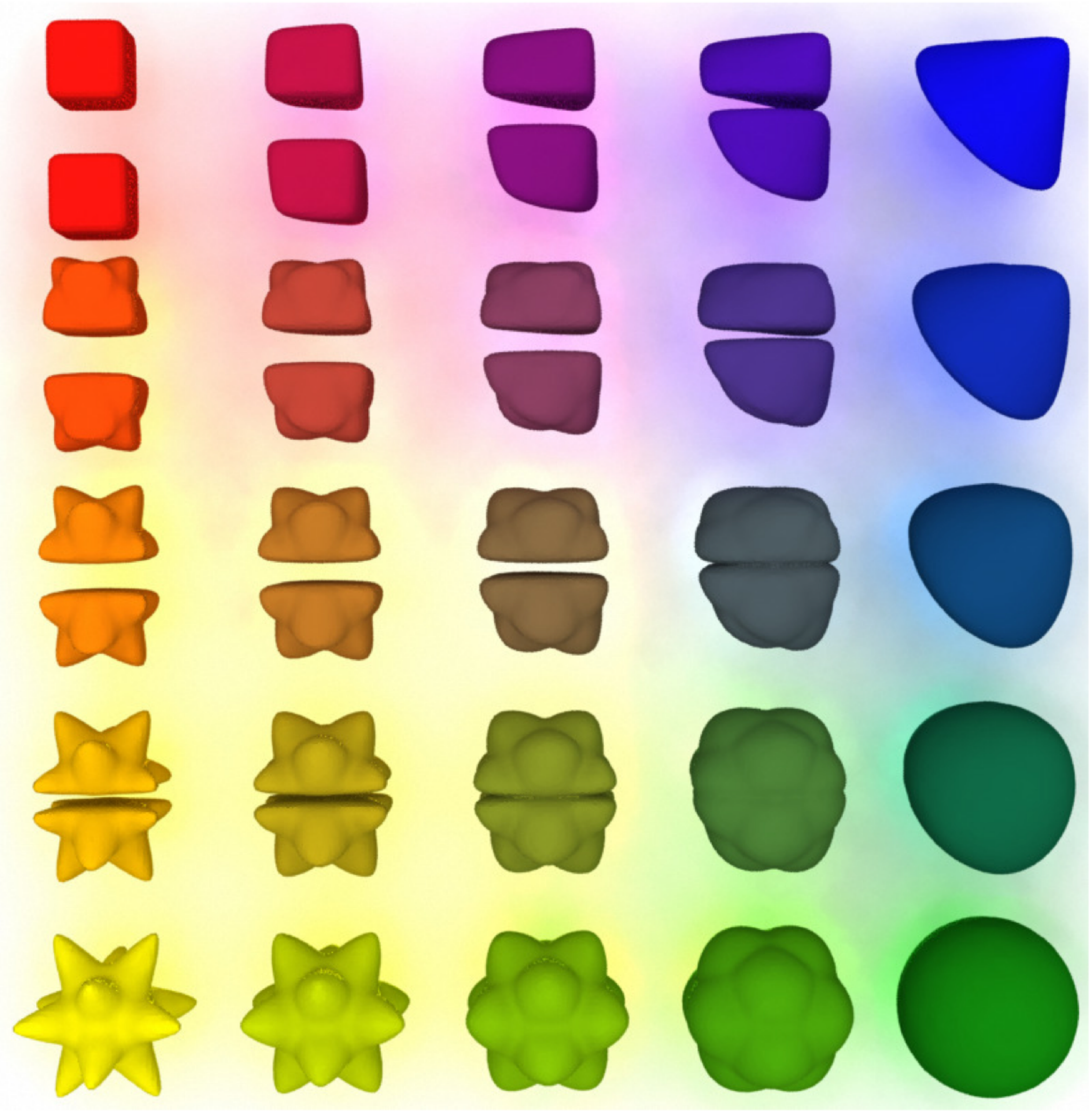


[SDPC..'15]

What is Optimal Transport?

A powerful **geometric toolbox** to compare **probability measures.**

Wasserstein Barycenters **[Agueh'11]**



[SDPC..'15]

Origins: Monge's Problem

MÉMOIRE SUR LA THÉORIE DES DÉBLAIS ET DES REMBLAIS.

Par M. MONGE.

LORSQU'ON doit tranſporter des terres d'un lieu dans un autre, on a coutume de donner le nom de *Déblai* au volume des terres que l'on doit tranſporter, & le nom de *Remblai* à l'eſpace qu'elles doivent occuper après le tranſport.

Origins: Monge's Problem

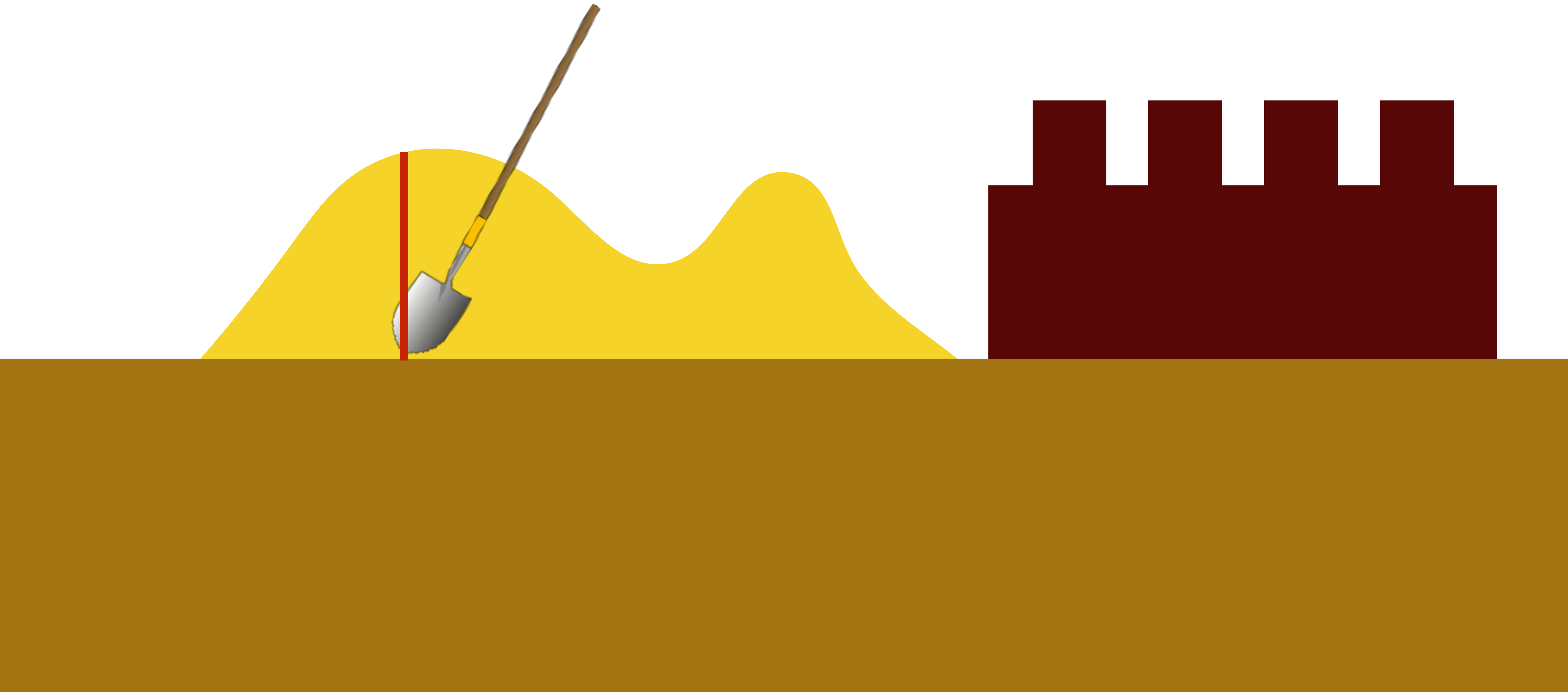
Origins: Monge's Problem

Origins: Monge's Problem

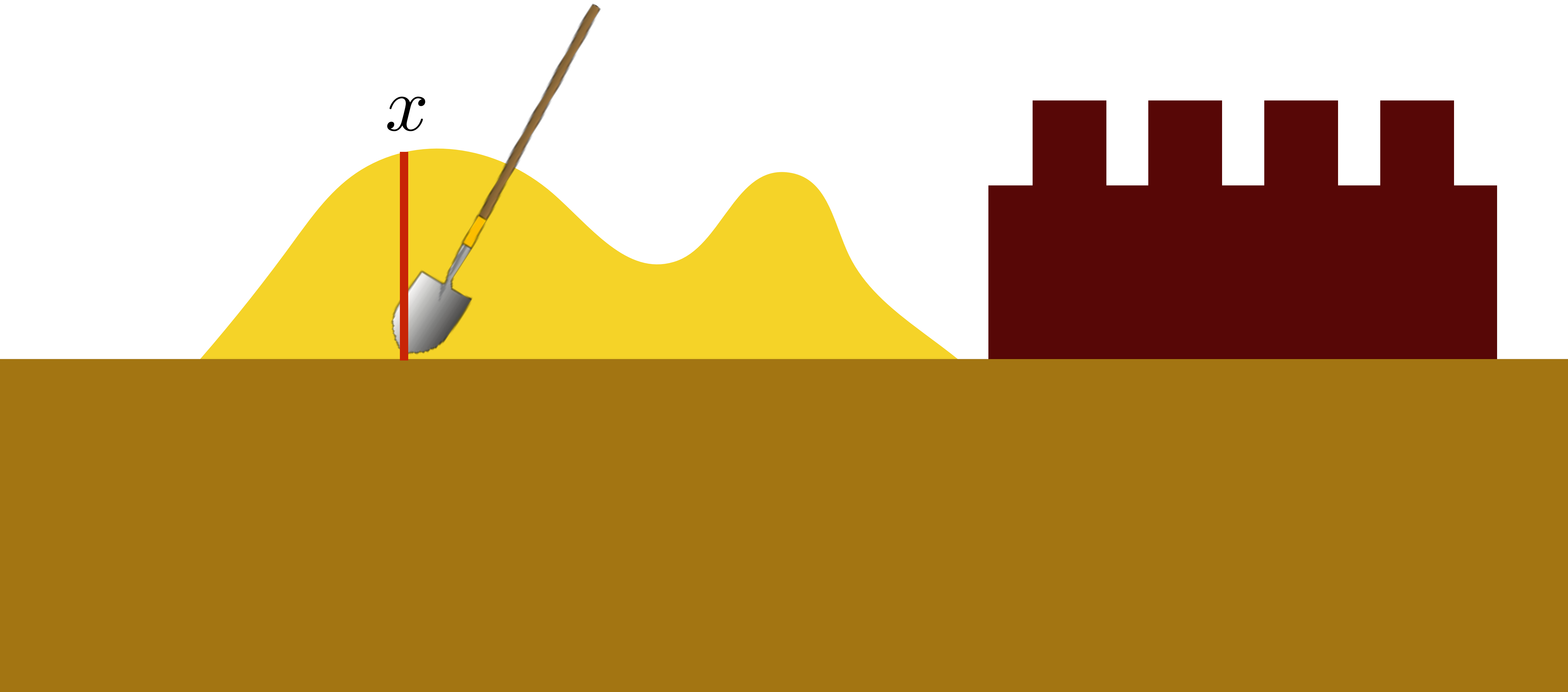
Origins: Monge's Problem

Origins: Monge's Problem

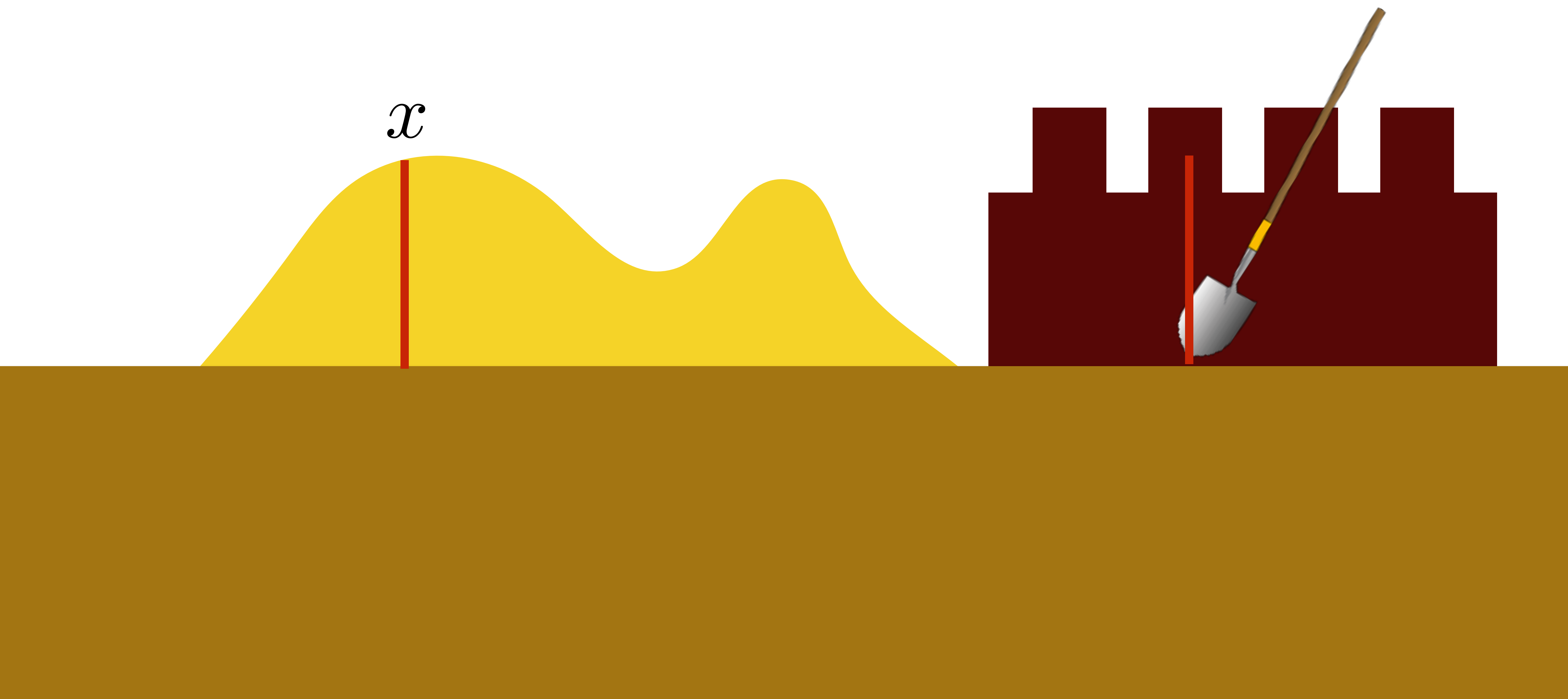
Origins: Monge's Problem



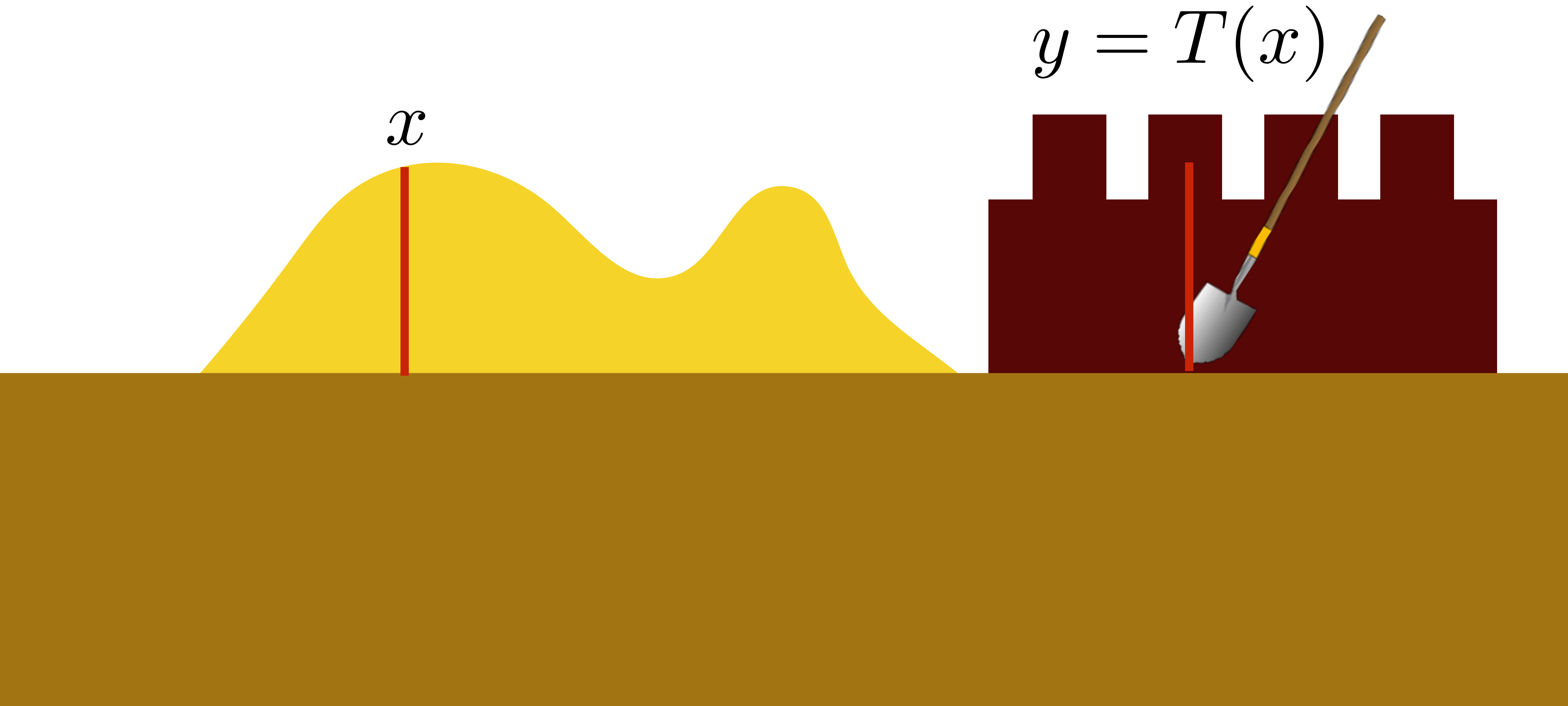
Origins: Monge's Problem



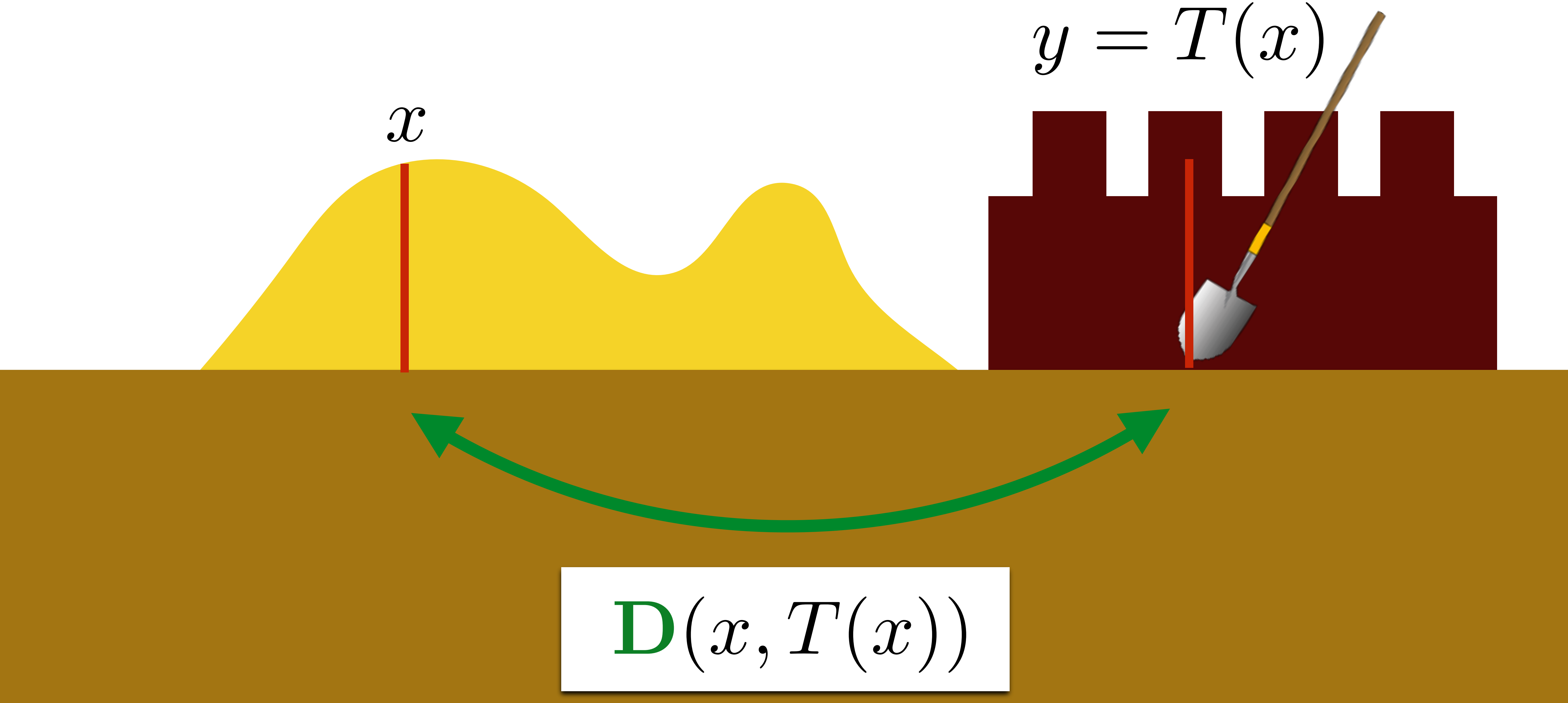
Origins: Monge's Problem



Origins: Monge's Problem



Origins: Monge's Problem



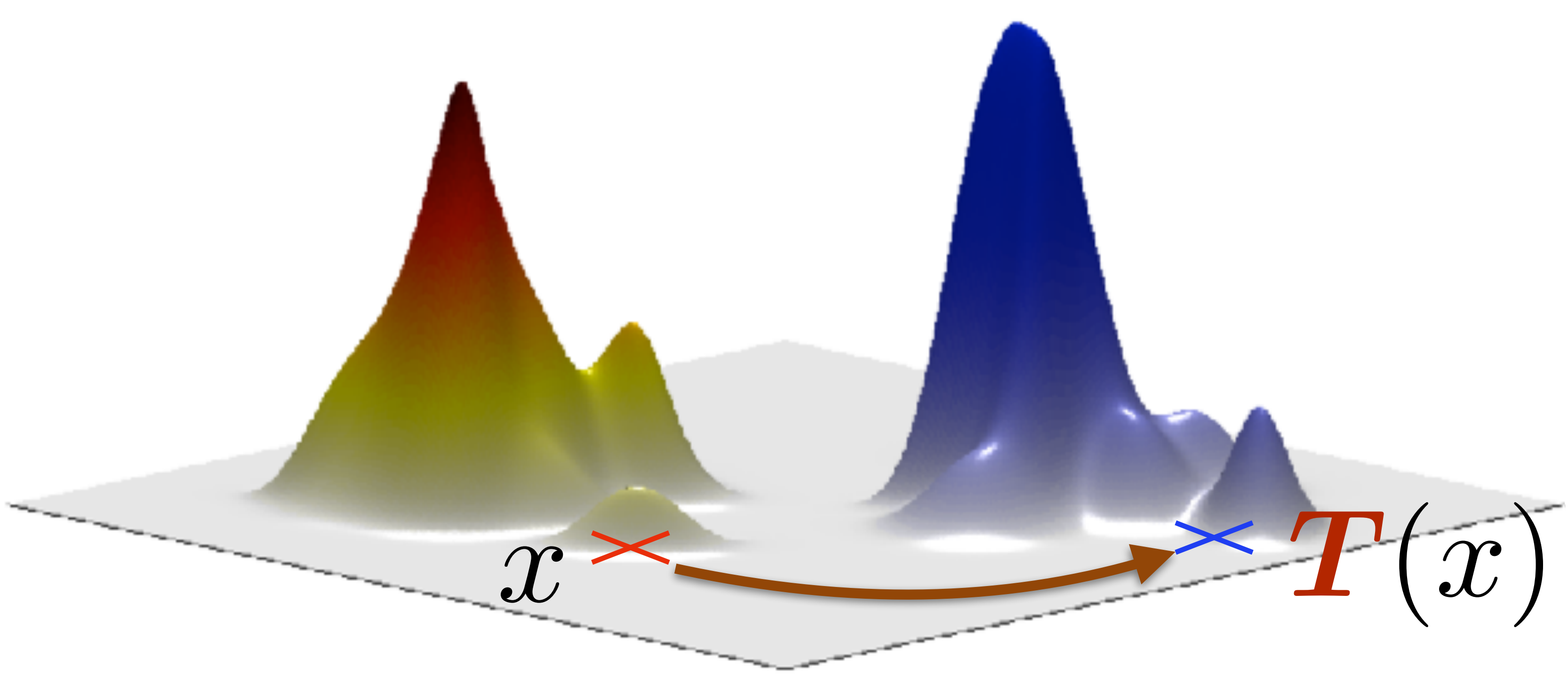
$\mathbf{D}(x, T(x))$

Origins: Monge's Problem

Ω a probability space, $\boldsymbol{c} : \Omega \times \Omega \to \mathbb{R}$.
$\boldsymbol{\mu}, \boldsymbol{\nu}$ two probability measures in $\mathcal{P}(\Omega)$.

[Monge'81] problem: find a map $\boldsymbol{T} : \Omega \to \Omega$

$$\inf_{\boldsymbol{T}_\sharp \boldsymbol{\mu} = \boldsymbol{\nu}} \int_\Omega \boldsymbol{c}(x, \boldsymbol{T}(x)) \boldsymbol{\mu}(dx)$$

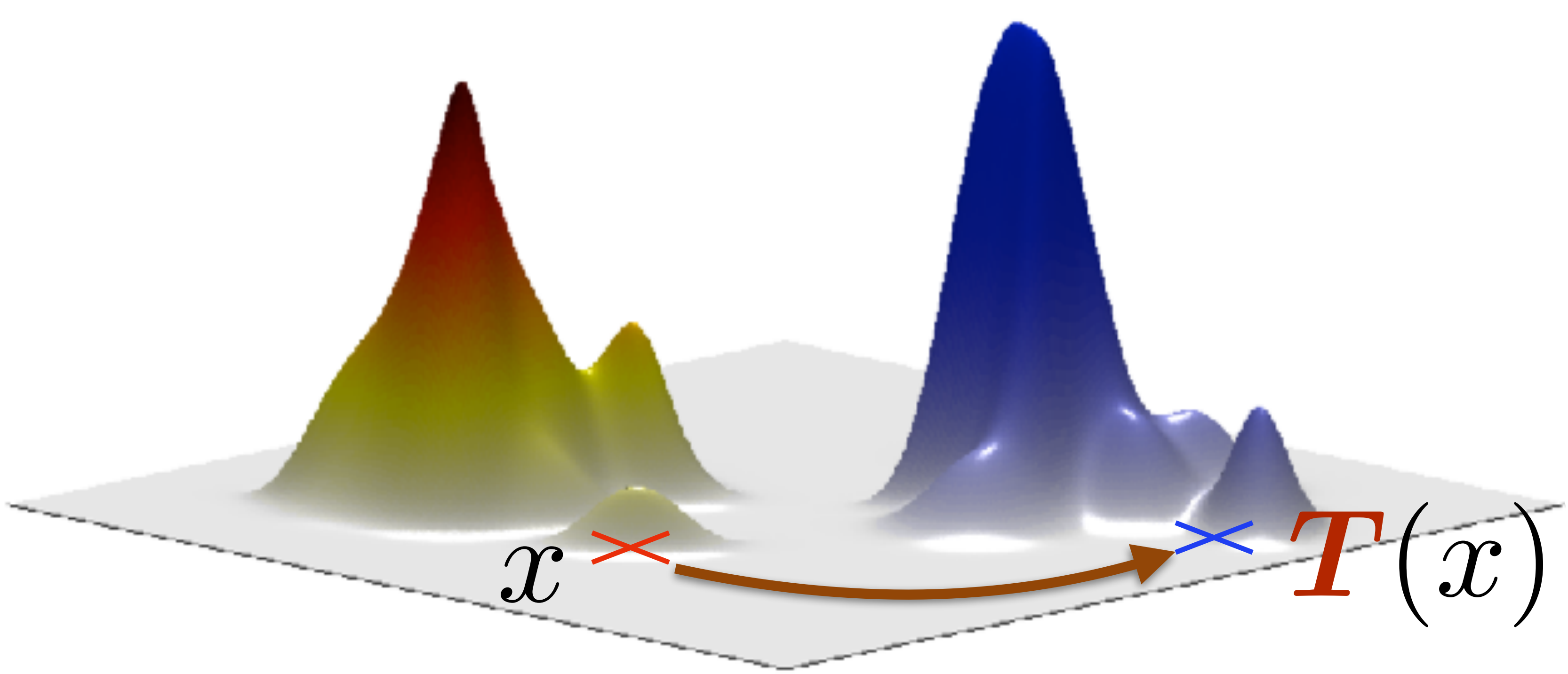


Origins: Monge's Problem

Ω a probability space, $\boldsymbol{c} : \Omega \times \Omega \to \mathbb{R}$.
$\boldsymbol{\mu}, \boldsymbol{\nu}$ two probability measures in $\mathcal{P}(\Omega)$.

[Monge'81] problem: find a map $\boldsymbol{T} : \Omega \to \Omega$

[Brenier'87] If $\Omega = \mathbb{R}^d$, $\boldsymbol{c} = \| \cdot - \cdot \|^2$,
$\boldsymbol{\mu}, \boldsymbol{\nu}$ a.c., then $\boldsymbol{T} = \nabla \boldsymbol{u}$, $\boldsymbol{u}$ **convex**.

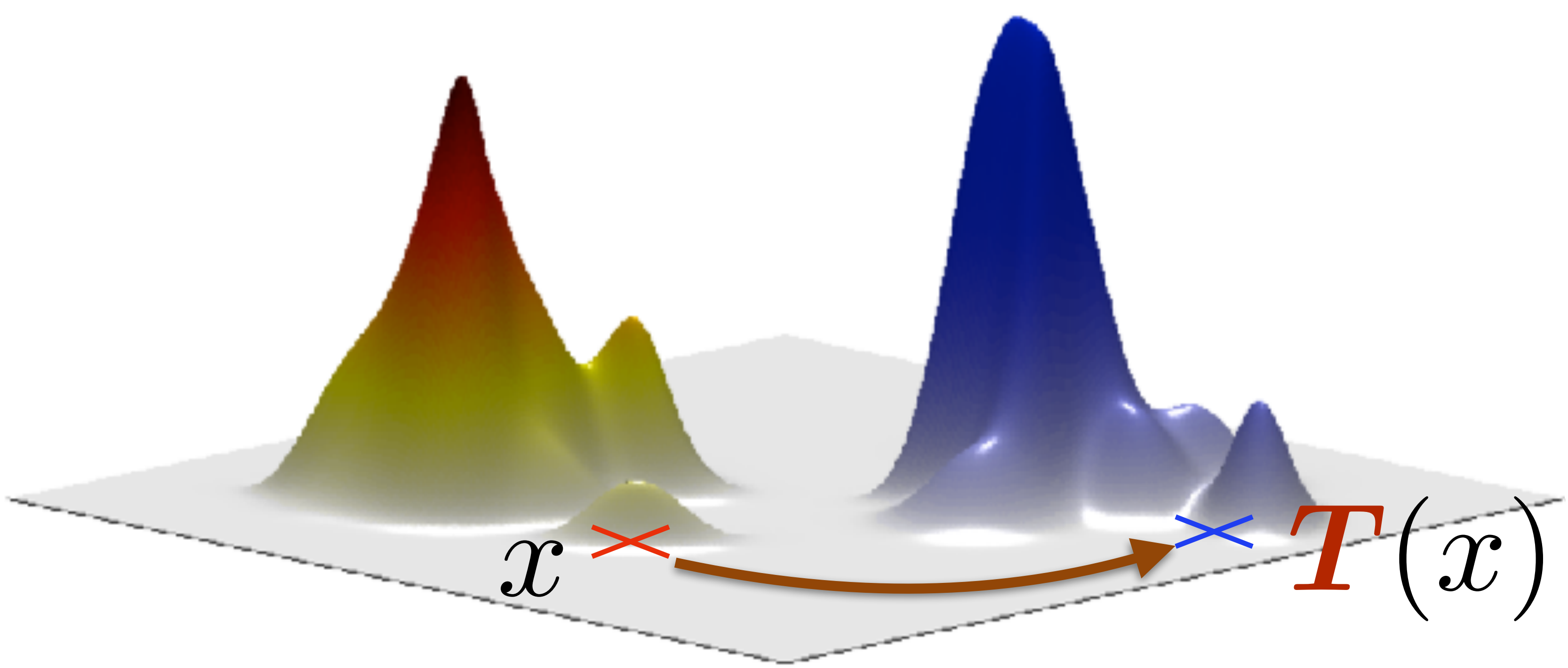


Monge's Problem

Ω a probability space, $\boldsymbol{c} : \Omega \times \Omega \to \mathbb{R}$.
$\boldsymbol{\mu}, \boldsymbol{\nu}$ two probability measures in $\mathcal{P}(\Omega)$.

[Monge'81] problem: find a map $\boldsymbol{T} : \Omega \to \Omega$

$$\inf_{\boldsymbol{T}_\sharp \boldsymbol{\mu} = \boldsymbol{\nu}} \int_\Omega \boldsymbol{c}(x, \boldsymbol{T}(x)) \boldsymbol{\mu}(dx)$$

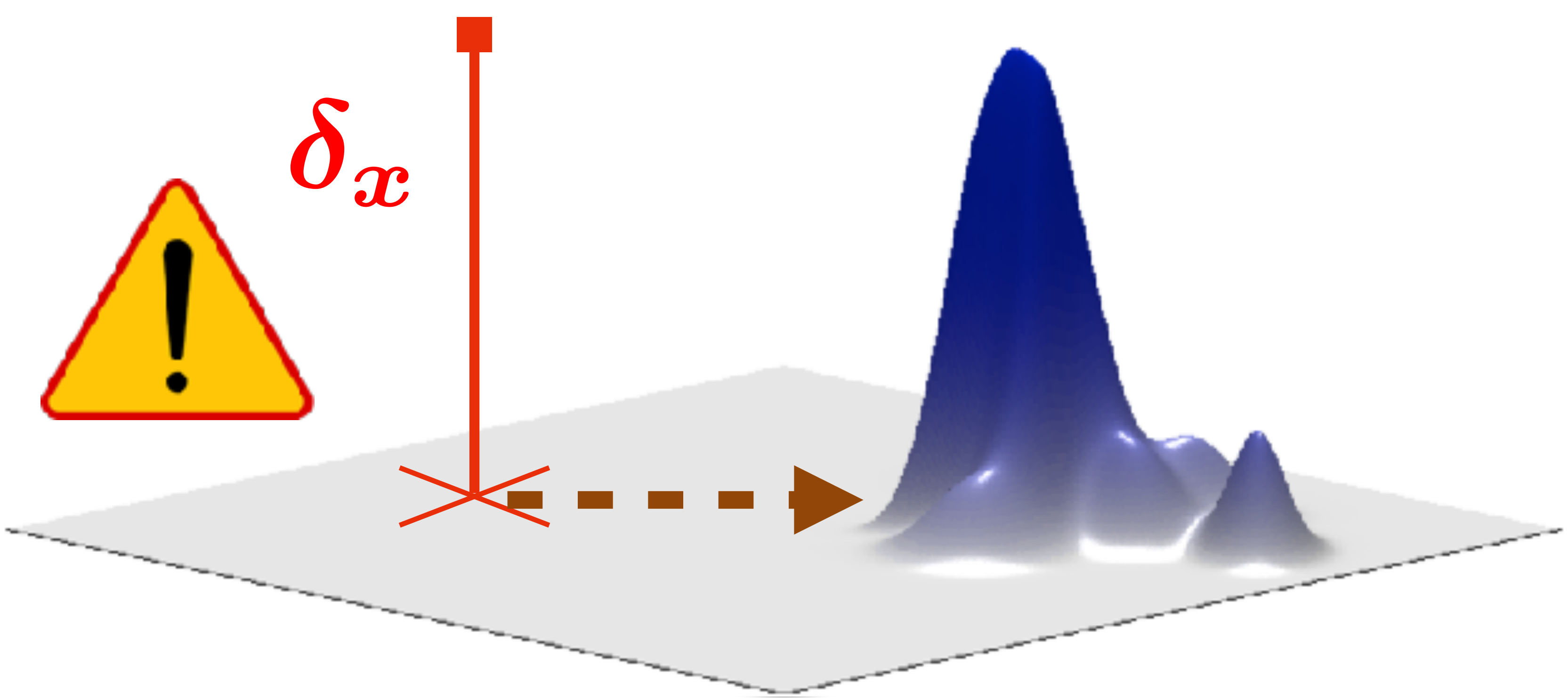


Monge's Problem

Ω a probability space, $\boldsymbol{c} : \Omega \times \Omega \to \mathbb{R}$.
$\boldsymbol{\mu}, \boldsymbol{\nu}$ two probability measures in $\mathcal{P}(\Omega)$.

[Monge'81] problem: find a map $\boldsymbol{T} : \Omega \to \Omega$

$$\inf_{\boldsymbol{T}_\sharp \boldsymbol{\mu} = \boldsymbol{\nu}} \int_\Omega \boldsymbol{c}(x, \boldsymbol{T}(x)) \boldsymbol{\mu}(dx)$$



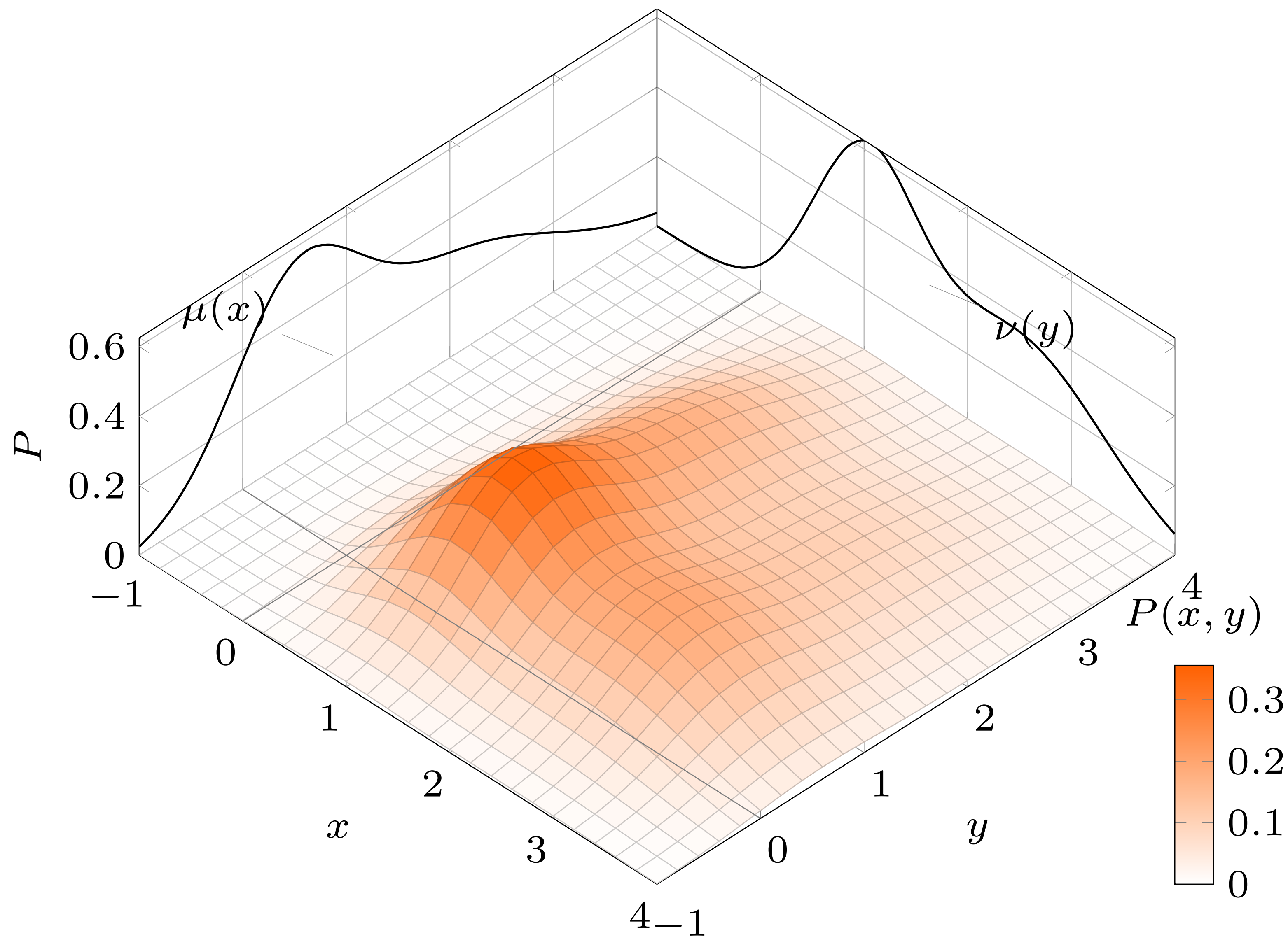
[Kantorovich'42] Relaxation

- Instead of maps $\boldsymbol{T} : \Omega \to \Omega$, consider probabilistic maps, i.e. **couplings** $\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega)$:

$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega) | \forall \boldsymbol{A}, \boldsymbol{B} \subset \Omega, \\ \boldsymbol{P}(\boldsymbol{A} \times \Omega) = \boldsymbol{\mu}(\boldsymbol{A}), \\ \boldsymbol{P}(\Omega \times \boldsymbol{B}) = \boldsymbol{\nu}(\boldsymbol{B})\}$$

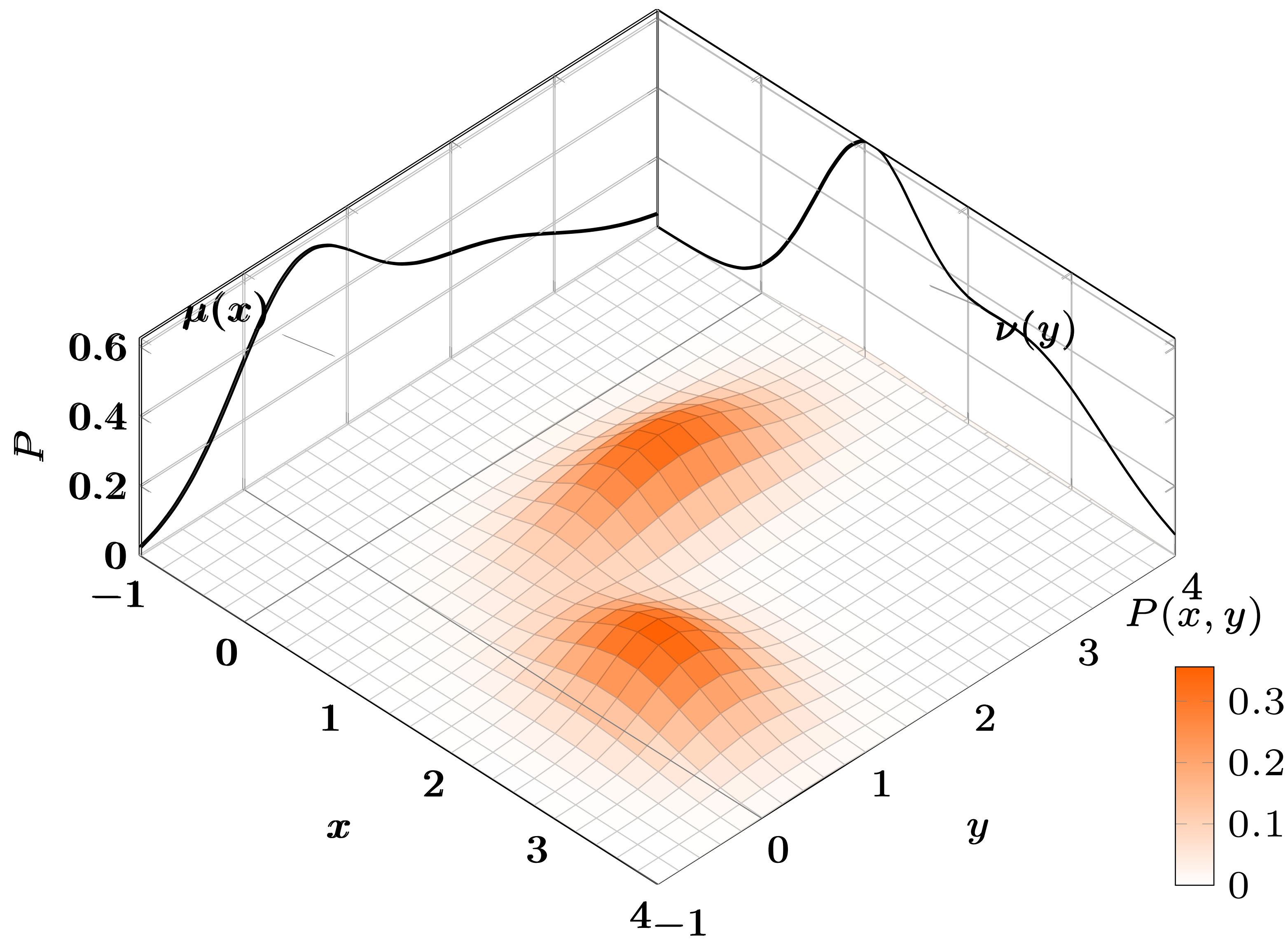
[Kantorovich'42] Relaxation

$$\Pi(\boldsymbol{\mu},\boldsymbol{\nu}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega) \,|\, \forall \boldsymbol{A},\boldsymbol{B} \subset \Omega,\; \boldsymbol{P}(\boldsymbol{A} \times \Omega) = \boldsymbol{\mu}(\boldsymbol{A}), \boldsymbol{P}(\Omega \times \boldsymbol{B}) = \boldsymbol{\nu}(\boldsymbol{B})\}$$



[Kantorovich'42] Relaxation

$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \overset{\text{def}}{=} \{\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega) | \forall \boldsymbol{A}, \boldsymbol{B} \subset \Omega, \boldsymbol{P}(\boldsymbol{A} \times \Omega) = \boldsymbol{\mu}(\boldsymbol{A}), \boldsymbol{P}(\Omega \times \boldsymbol{B}) = \boldsymbol{\nu}(\boldsymbol{B})\}$$



Wasserstein Distances

Def. For $p \geq 1$, the p-Wasserstein distance between $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$, defined by a metric $\boldsymbol{D}$ on Ω,

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})^p \boldsymbol{P}(dx, dy).$$

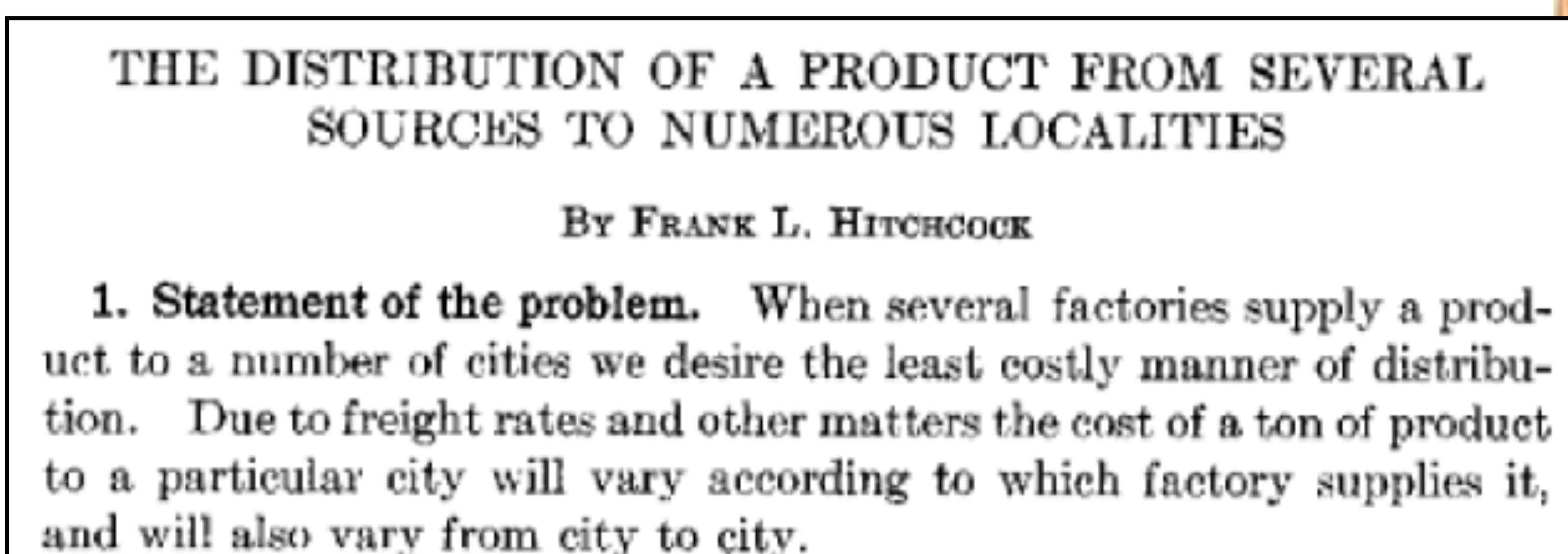
PRIMAL

Wasserstein Distances

Def. For $p \geq 1$, the p-Wasserstein distance between $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$, defined by a metric $\boldsymbol{D}$ on Ω,

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})^p \boldsymbol{P}(dx, dy).$$

PRIMAL


THE DISTRIBUTION OF A PRODUCT FROM SEVERAL SOURCES TO NUMEROUS LOCALITIES

By Frank L. Hitchcock

1. Statement of the problem. When several factories supply a product to a number of cities we desire the least costly manner of distribution. Due to freight rates and other matters the cost of a ton of product to a particular city will vary according to which factory supplies it, and will also vary from city to city.

Wasserstein Distances

Def. For $p \geq 1$, the p-Wasserstein distance between $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$, defined by a metric $\boldsymbol{D}$ on Ω,

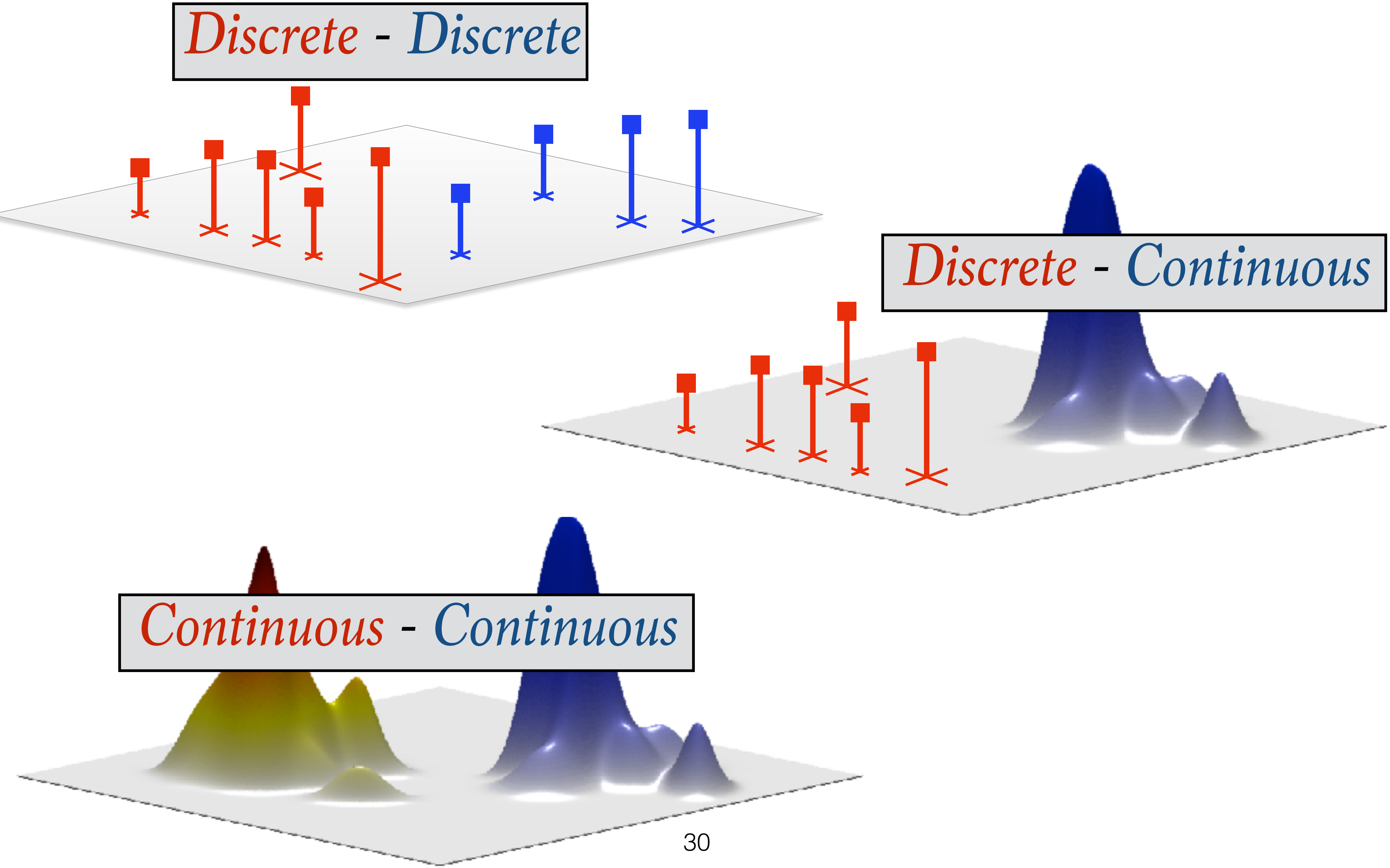
$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})^p \boldsymbol{P}(dx, dy).$$

PRIMAL

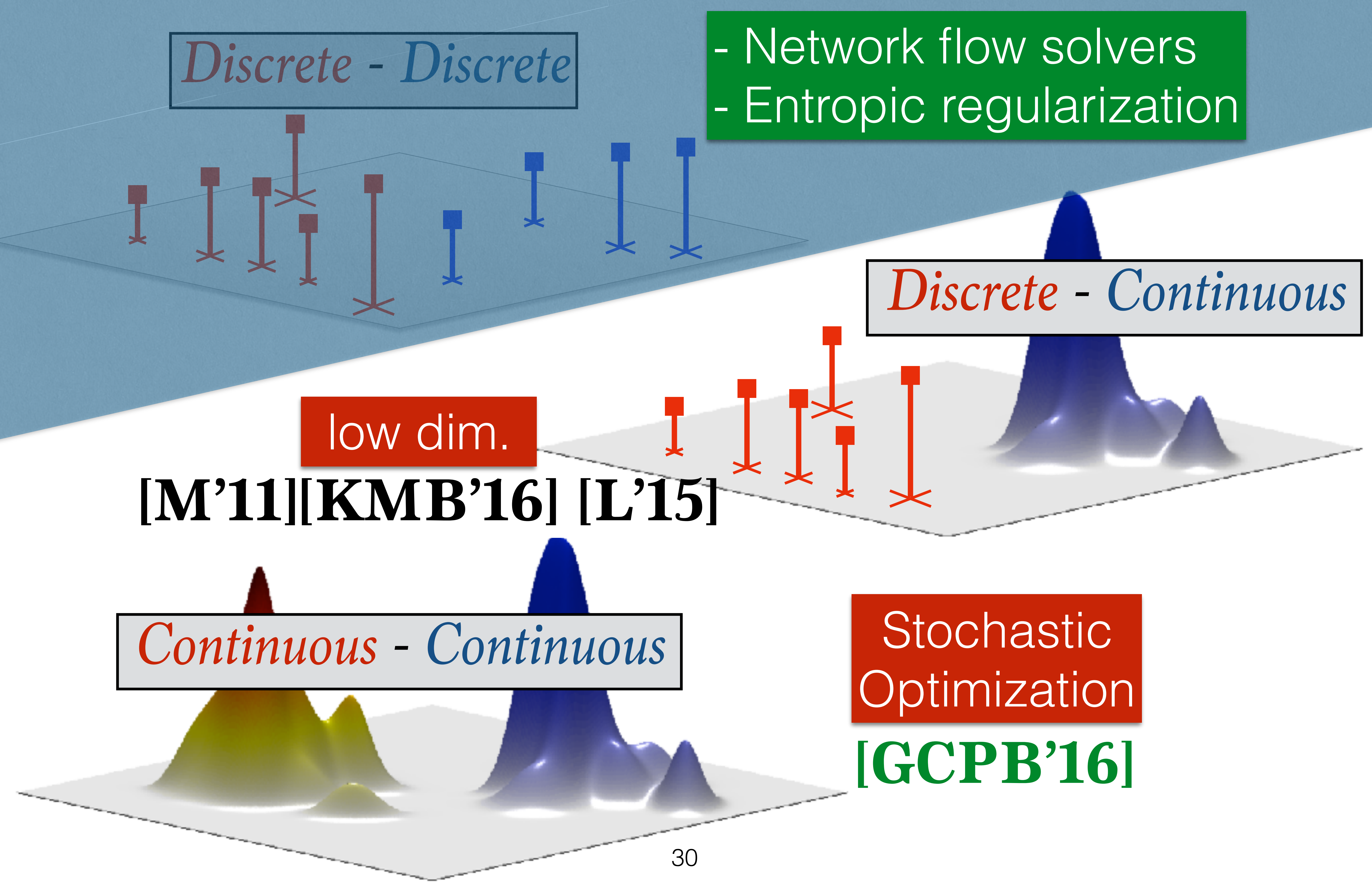
$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\substack{\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu}), \boldsymbol{\psi} \in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x) + \boldsymbol{\psi}(y) \leq \boldsymbol{D}^p(x,y)}} \int \boldsymbol{\varphi} \, d\boldsymbol{\mu} + \int \boldsymbol{\psi} \, d\boldsymbol{\nu}.$$

DUAL

W is versatile



W is versatile



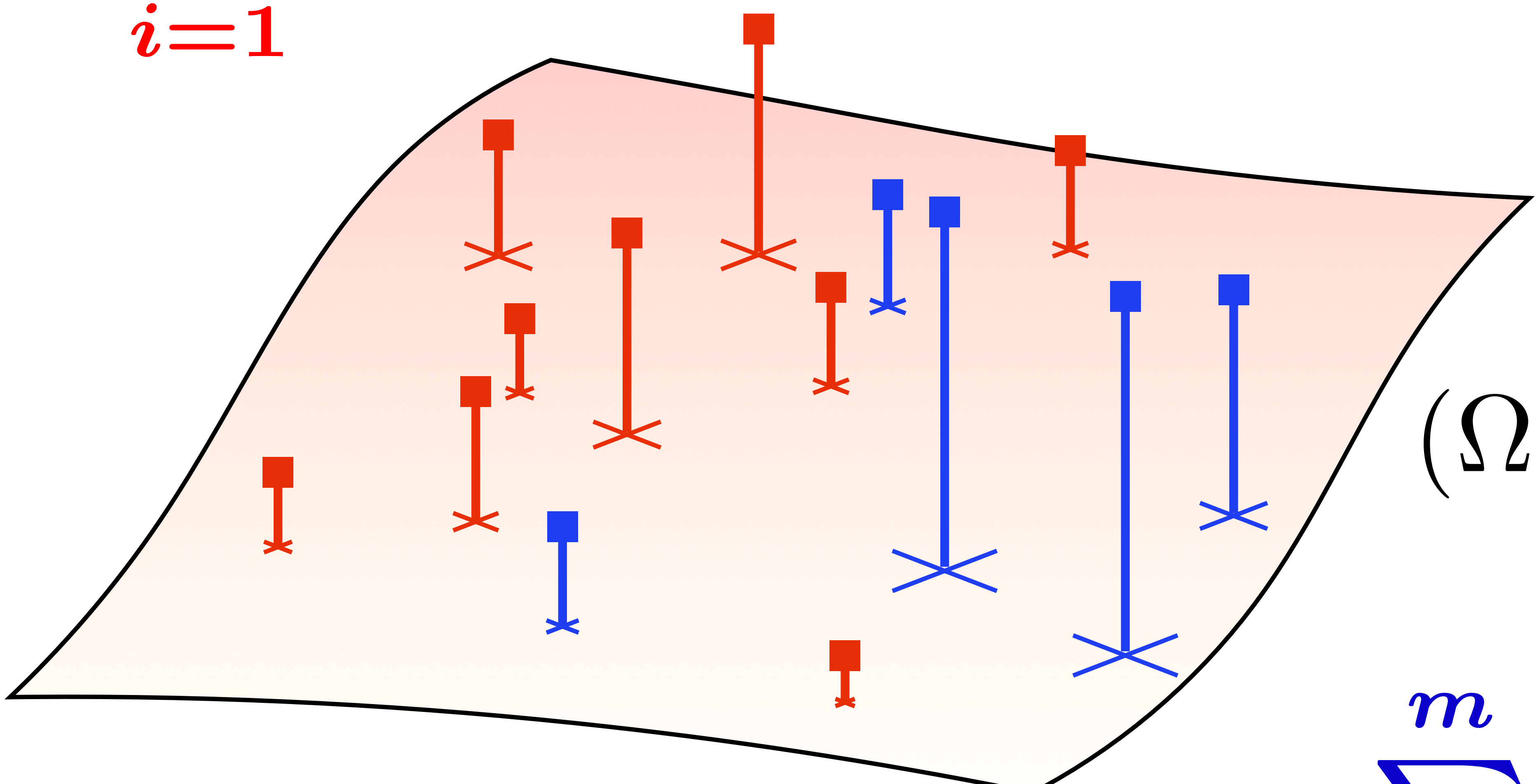
Minimum Kantorovich Estimators

$$\min_{\boldsymbol{\theta}\in\Theta} W(\boldsymbol{\nu}_{\text{data}}, f_{\boldsymbol{\theta}\sharp}\boldsymbol{\mu})$$

- **[Bassetti'06]** 1st reference discussing this approach.
- **[MMC'16]** use regularization in a finite setting.
- **[ACB'17]** *(WGAN)* **[BJGR'17]** *(Wasserstein ABC).*
- **Hot topics:** *approximate & differentiate W efficiently.*
- Today: ideas from our recent preprint **[GPC'17]**

Wasserstein on Empirical Measures

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



$$(\Omega, D)$$

$$\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

Wasserstein on Empirical Measures

Consider $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$.

$$M_{XY} \stackrel{\text{def}}{=} [D(x_i, y_j)^p]_{ij}$$

$$U(a, b) \stackrel{\text{def}}{=} \{P \in \mathbb{R}_+^{n \times m} | P\mathbf{1}_m = a, P^T\mathbf{1}_n = b\}$$

$$\begin{array}{cc} & \begin{array}{ccc} y_1 & \cdots & y_m \end{array} \\ \begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} & \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & D(x_i, y_j)^p & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \end{array} \qquad \begin{array}{cc} & \begin{array}{ccc} b_1 & \cdots & b_m \end{array} \\ \begin{array}{c} a_1 \\ \vdots \\ a_n \end{array} & \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & P\mathbf{1}_m = a & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \end{array}$$

Wasserstein on Empirical Measures

Consider $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$.

$$M_{\mathbf{XY}} \stackrel{\text{def}}{=} [D(\boldsymbol{x}_i, \boldsymbol{y}_j)^p]_{ij}$$

$$U(\boldsymbol{a}, \boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

$$\begin{array}{c c} & \begin{array}{ccc} \boldsymbol{y}_1 & \cdots & \boldsymbol{y}_m \end{array} \\ \begin{array}{c} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{array} & \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & D(\boldsymbol{x}_i, \boldsymbol{y}_j)^p & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \end{array} \qquad \begin{array}{c c} & \begin{array}{ccc} \boldsymbol{b}_1 & \cdots & \boldsymbol{b}_m \end{array} \\ \begin{array}{c} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{array} & \begin{bmatrix} \vdots & \vdots & \vdots \\ \vdots & \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b} & \vdots \\ \vdots & \vdots & \vdots \end{bmatrix} \end{array}$$

Wasserstein on Empirical Measures

Consider $\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{\delta}_{\boldsymbol{x}_i}$ and $\boldsymbol{\nu} = \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{\delta}_{\boldsymbol{y}_j}$.

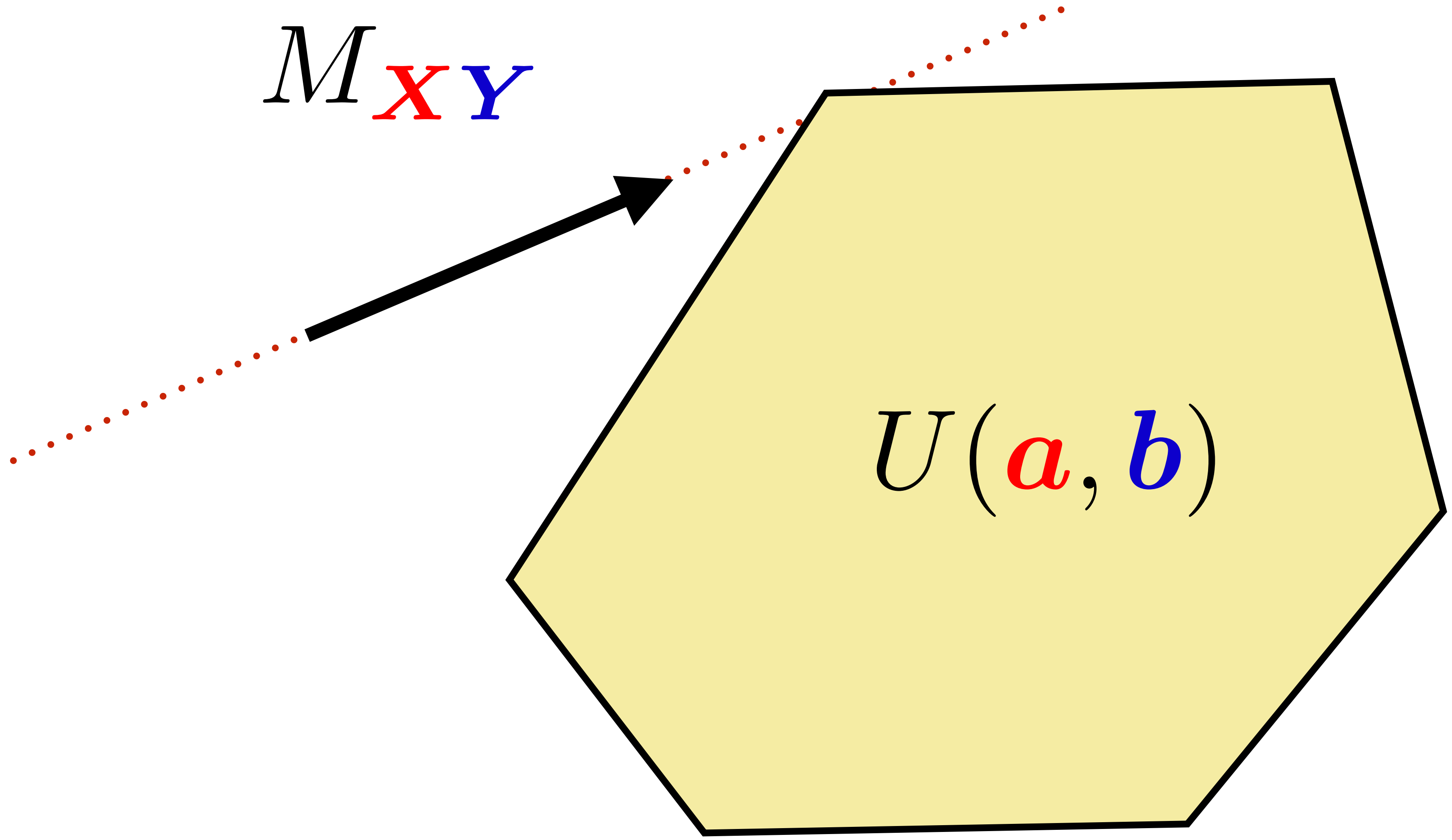
$$M_{\mathbf{XY}} \stackrel{\text{def}}{=} [D(\boldsymbol{x}_i, \boldsymbol{y}_j)^p]_{ij}$$

$$U(\boldsymbol{a}, \boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

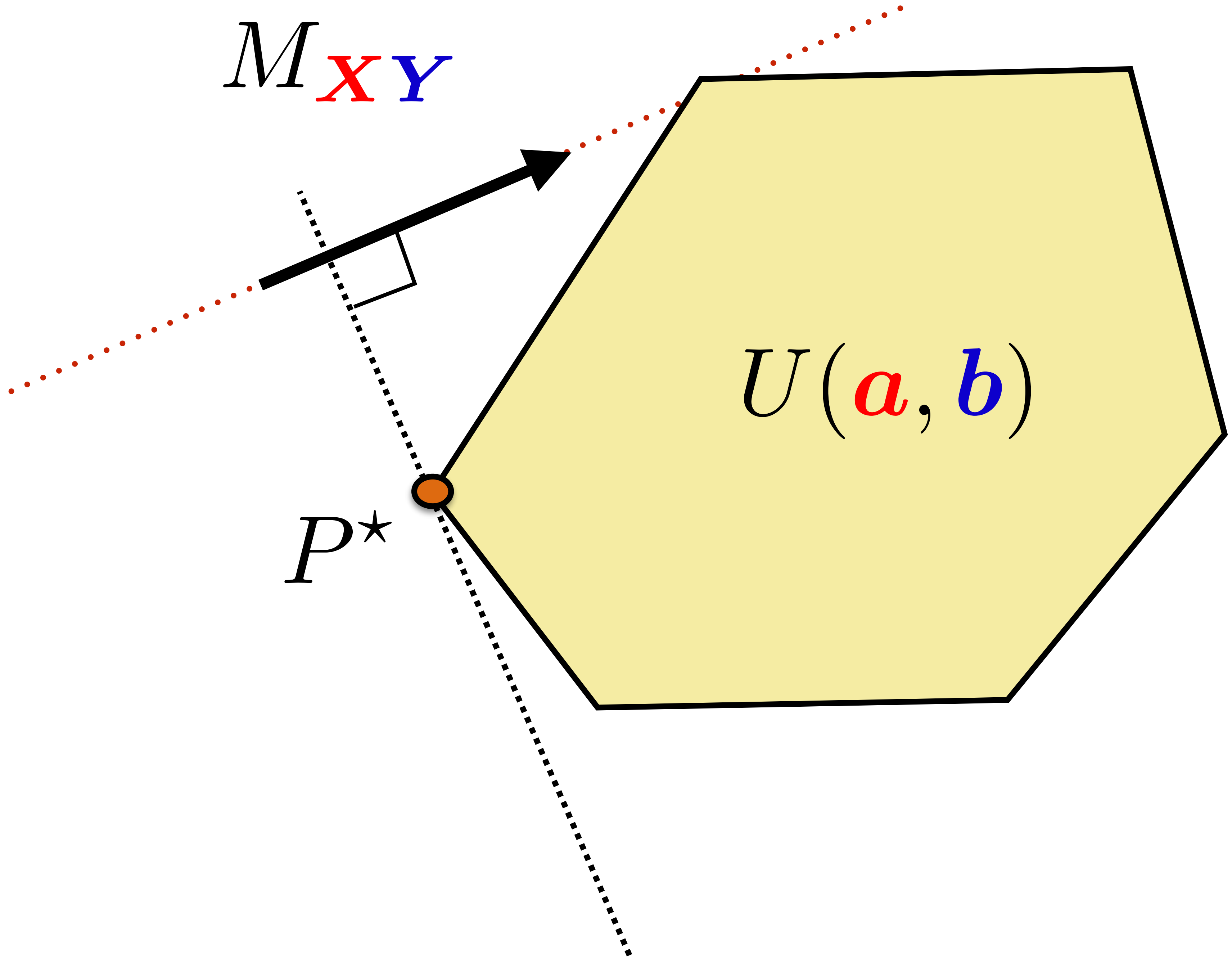
Def. Optimal Transport Problem

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\mathbf{XY}} \rangle$$

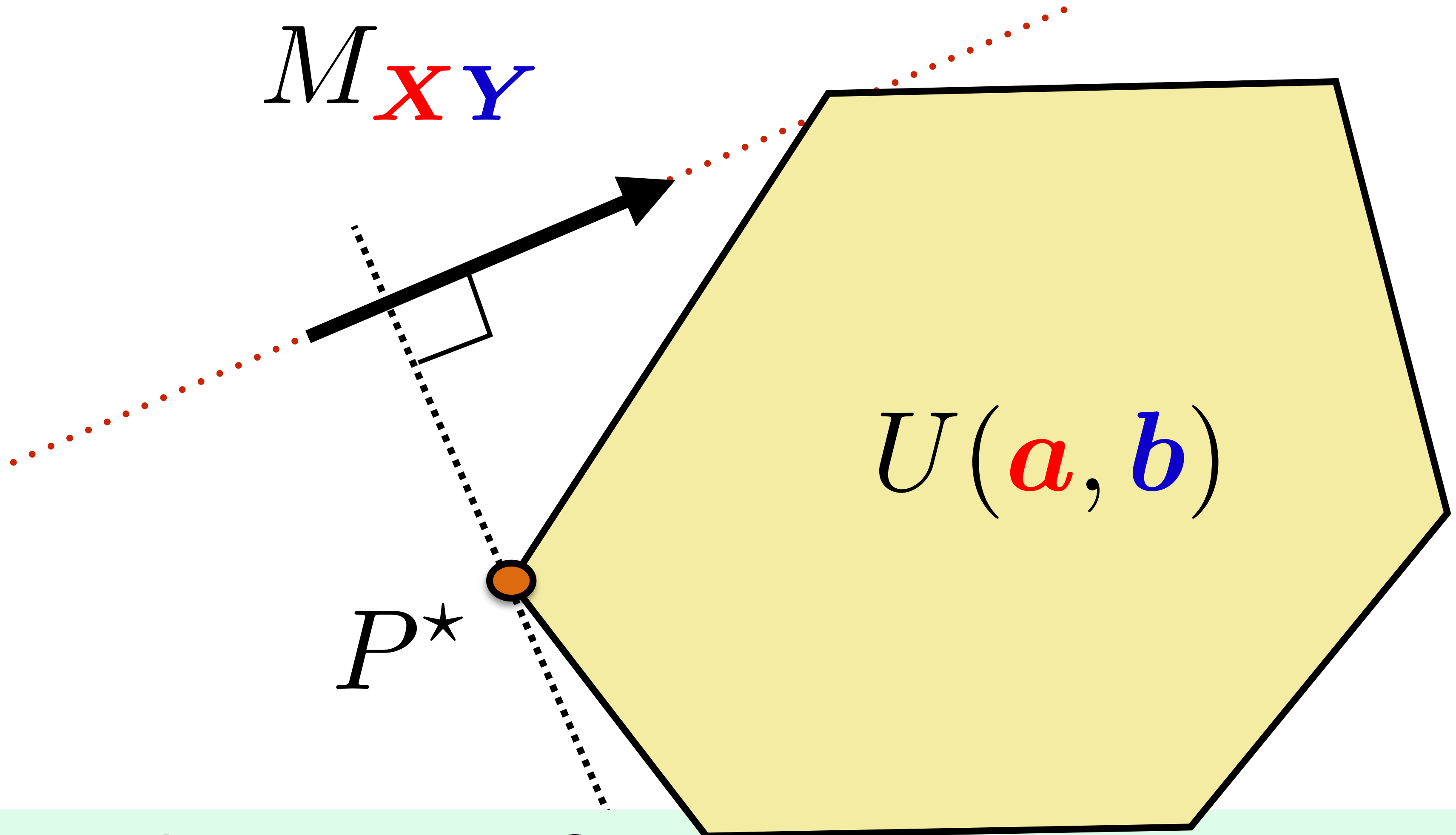
Discrete OT Problem



Discrete OT Problem



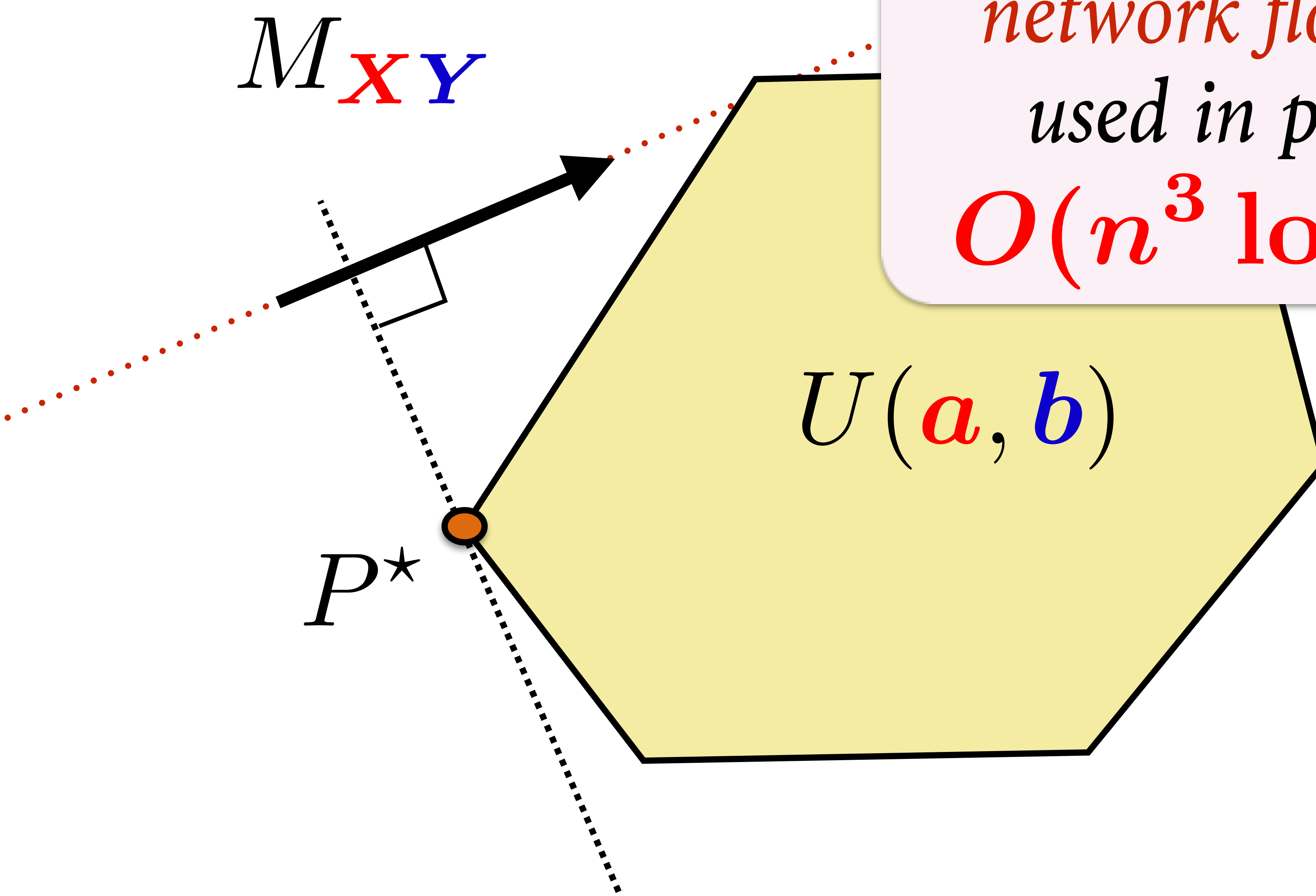
Discrete OT Problem



Def. Dual OT problem

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^m \\ \boldsymbol{\alpha}_{\boldsymbol{i}} + \boldsymbol{\beta}_{\boldsymbol{j}} \leq D(\boldsymbol{x}_{\boldsymbol{i}}, \boldsymbol{y}_{\boldsymbol{j}})^p}} \alpha^T \boldsymbol{a} + \beta^T \boldsymbol{b}$$

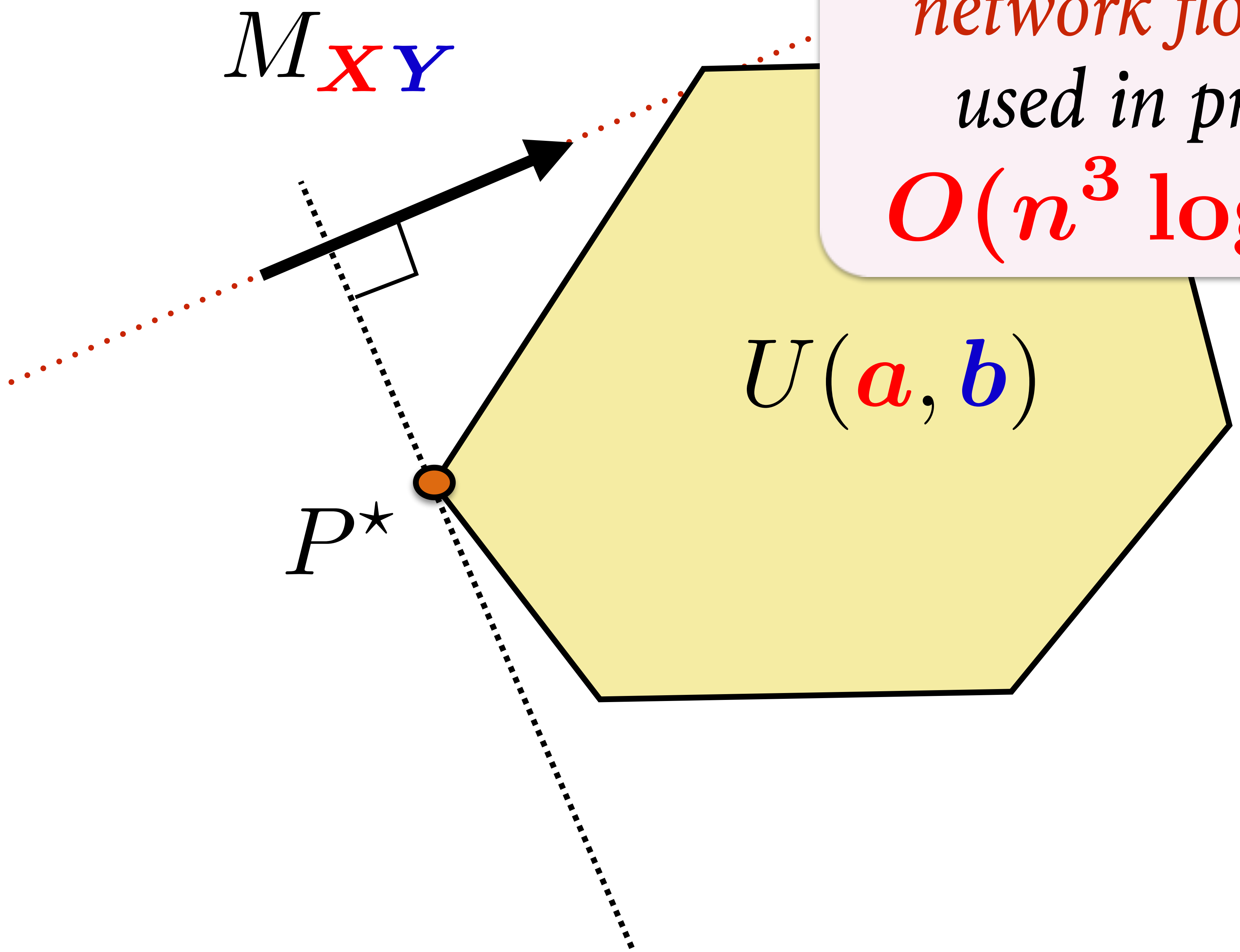
Discrete OT Problem



network flow solver
used in practice.
$\boldsymbol{O(n^3 \log(n))}$

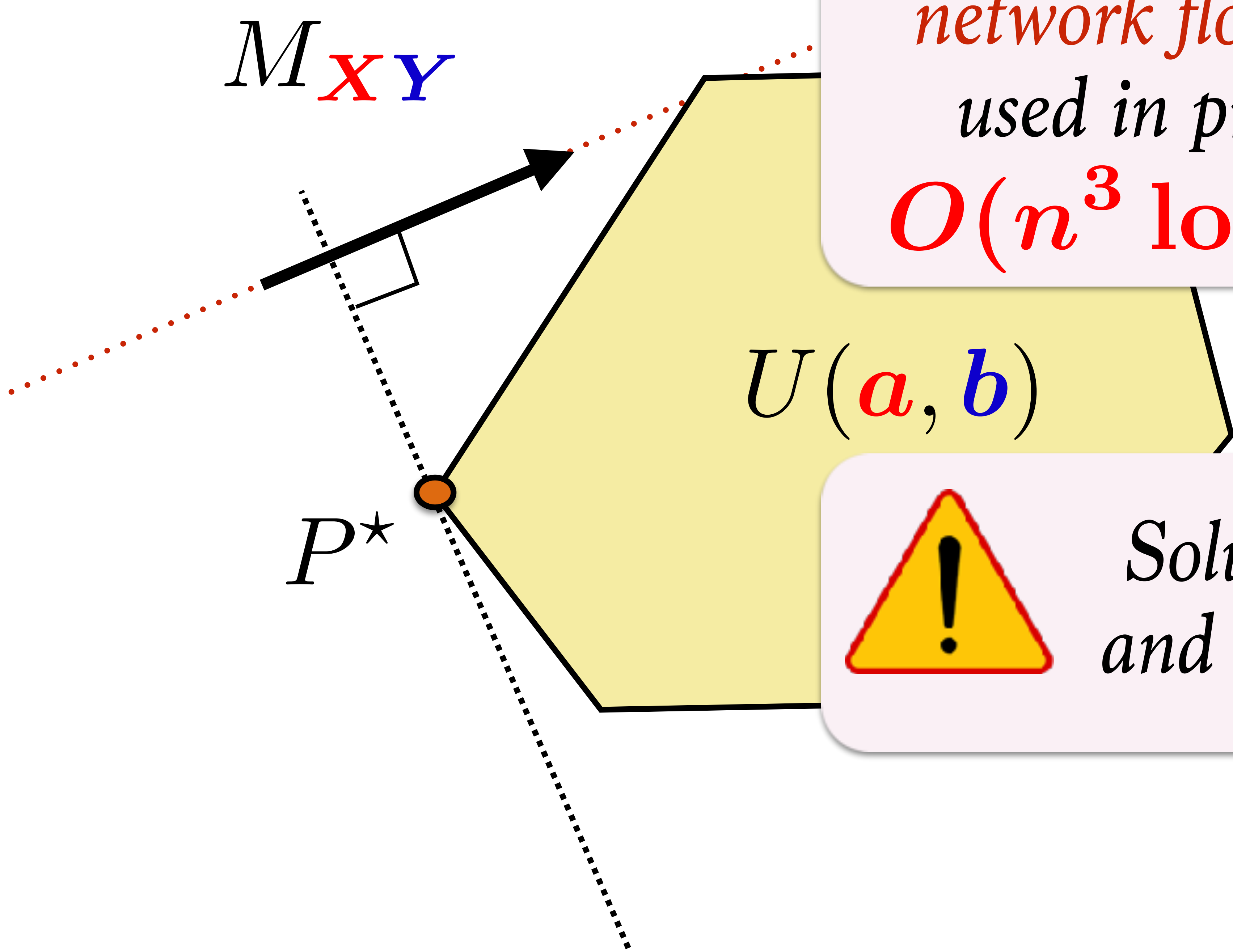
Note: flow/PDE formulations **[Beckman'61]/[Benamou'98]** can be used for *p=1/p=2* for a sparse-graph metric/Euclidean metric.

Discrete OT Problem



network flow solver
used in practice.
$\boldsymbol{O(n^3 \log(n))}$

Discrete OT Problem

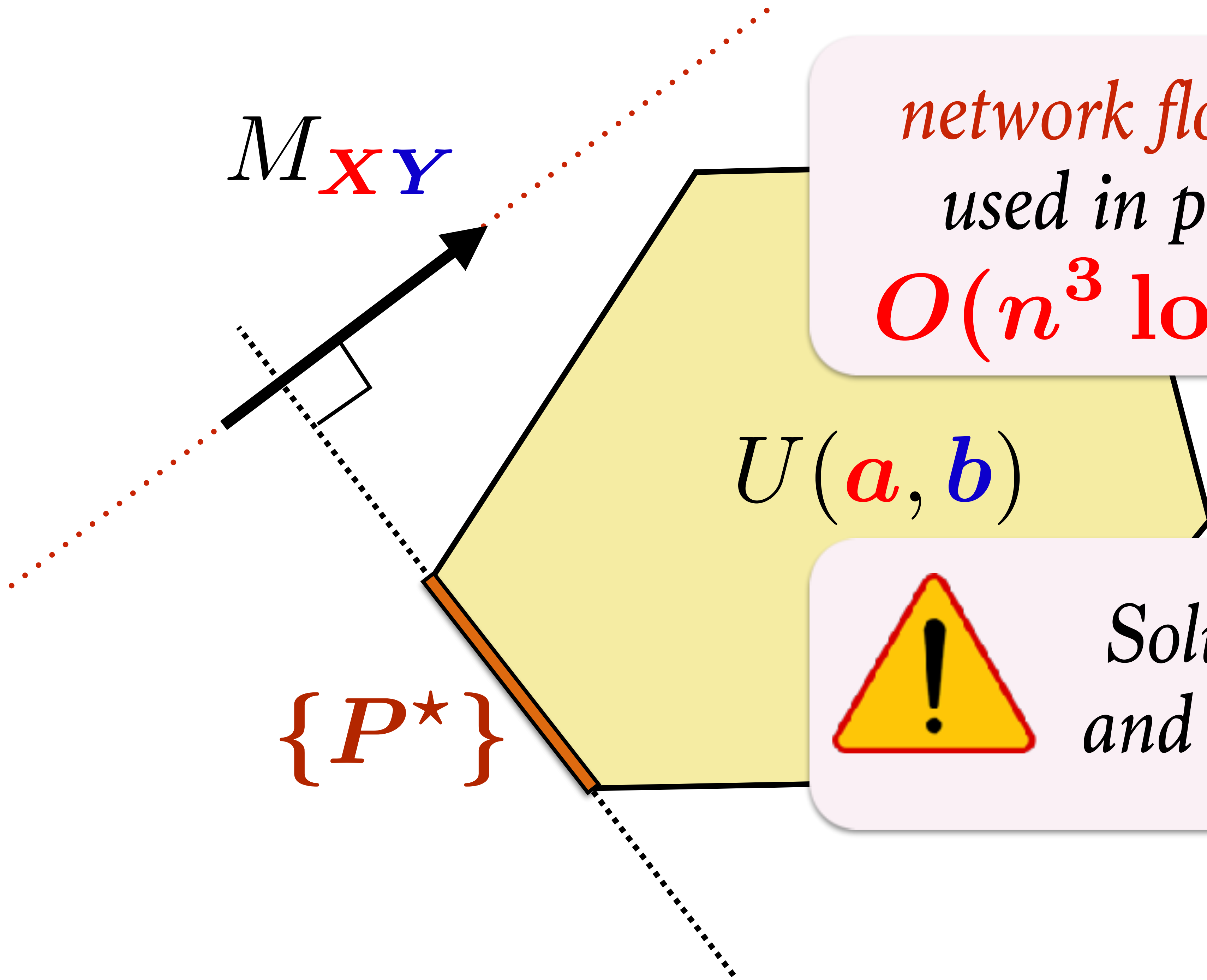


network flow solver *used in practice.*

$\boldsymbol{O(n^3 \log(n))}$

Solution $P^\star$ unstable and not always unique.

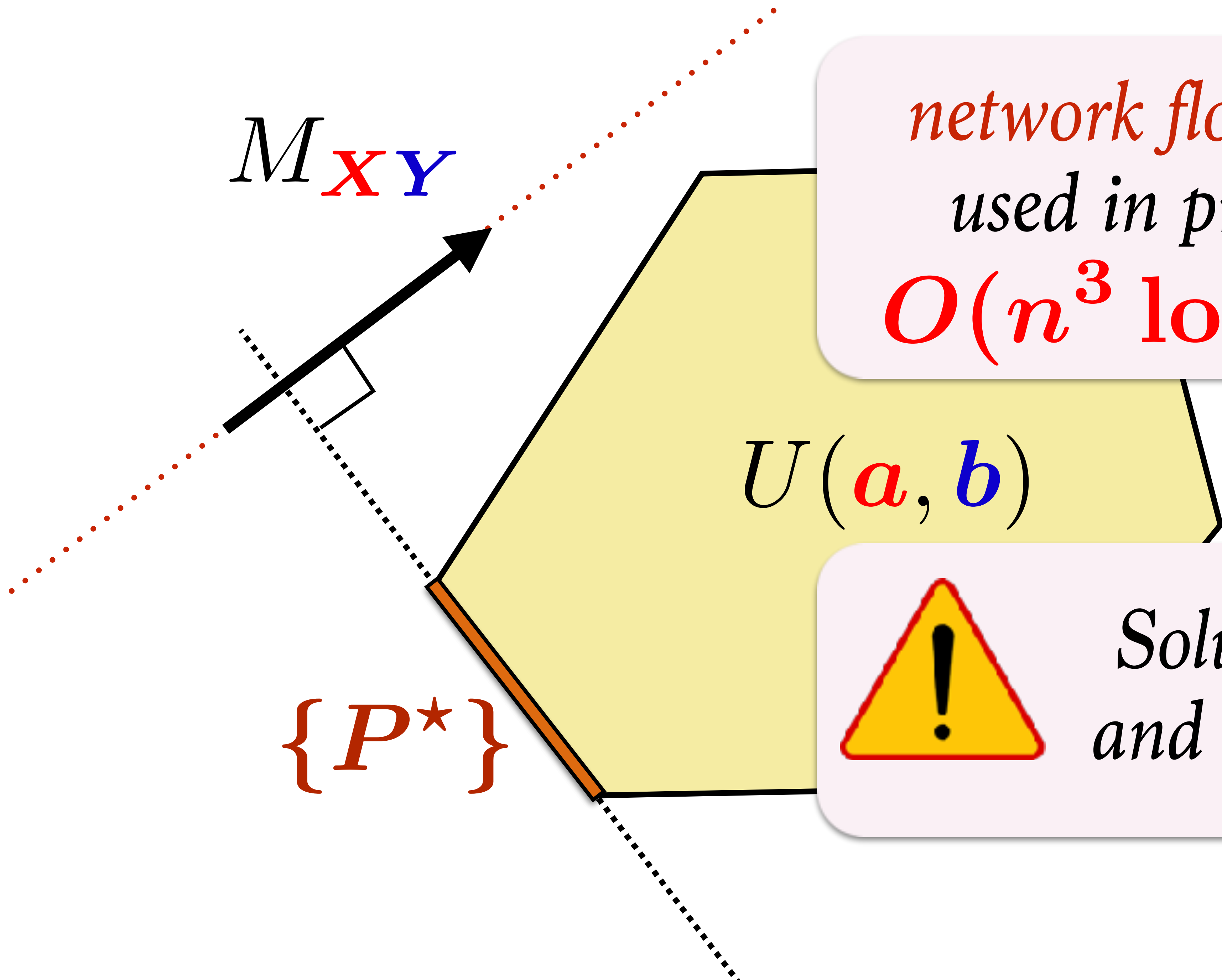
Discrete OT Problem



network flow solver *used in practice.*
$O(n^3 \log(n))$

Solution $P^\star$ unstable and not always unique.

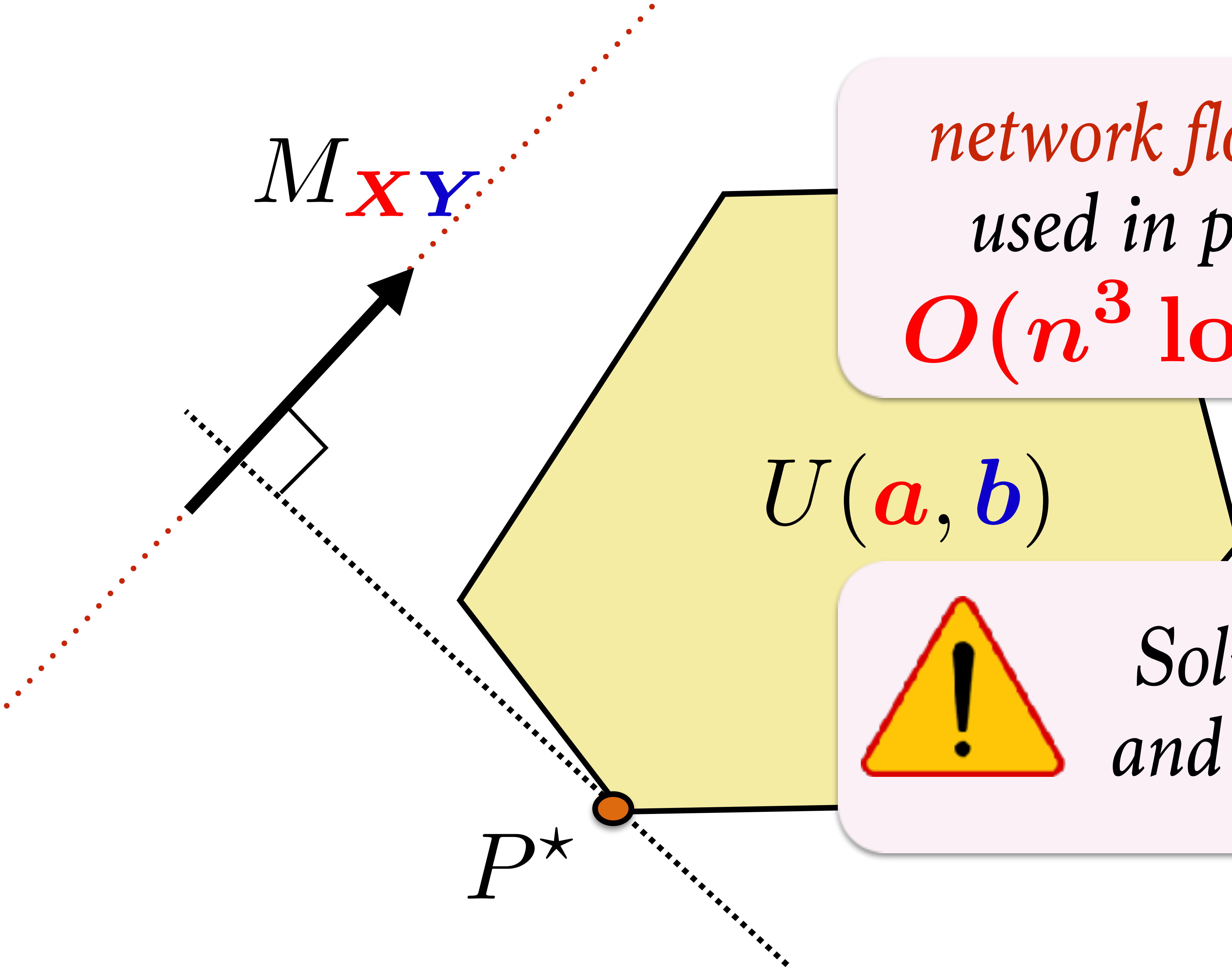
Discrete OT Problem



network flow solver *used in practice.*
$O(n^3 \log(n))$

Solution $P^\star$ unstable and not always unique.

Discrete OT Problem

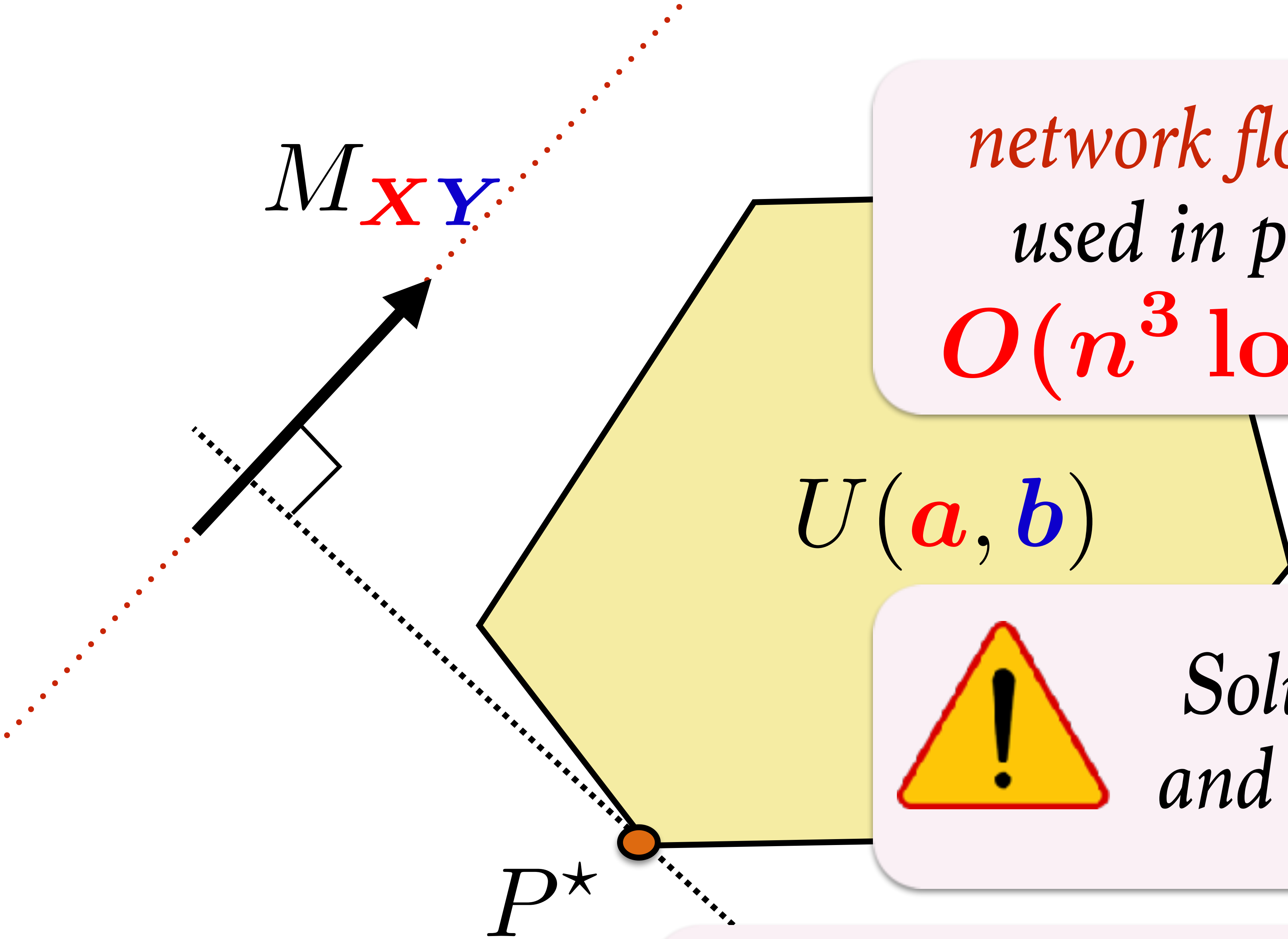


network flow solver *used in practice.*

$\boldsymbol{O(n^3 \log(n))}$

Solution $P^\star$ unstable and not always unique.

Discrete OT Problem



network flow solver used in practice. $O(n^3 \log(n))$

Solution $P^\star$ unstable and not always unique.

$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu})$ not differentiable.

Entropic Regularization [**Wilson'62**]

Def. Regularized Wasserstein, $\gamma \geq 0$

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$$

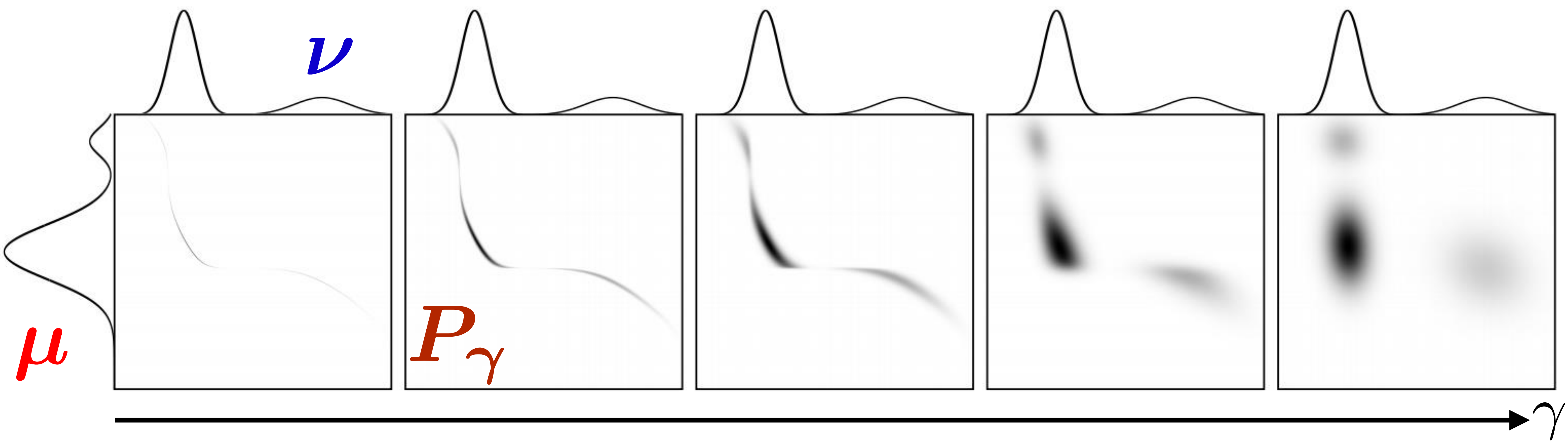
$$E(P) \stackrel{\text{def}}{=} - \sum_{i,j=1}^{nm} P_{ij} (\log P_{ij})$$

Note: Unique optimal solution because of strong concavity of Entropy

Entropic Regularization [**Wilson'62**]

Def. Regularized Wasserstein, $\gamma \geq 0$

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle - \gamma E(\boldsymbol{P})$$



Note: Unique optimal solution because of strong concavity of Entropy

Fast & Scalable Algorithm

Prop. If $P_\gamma \stackrel{\text{def}}{=} \underset{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$

then $\exists! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \stackrel{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

Fast & Scalable Algorithm

Prop. If $P_\gamma \stackrel{\text{def}}{=} \underset{\boldsymbol{P}\in U(\boldsymbol{a},\boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\mathbf{XY}} \rangle - \gamma E(\boldsymbol{P})$

then $\exists! \boldsymbol{u} \in \mathbb{R}^n_+, \boldsymbol{v} \in \mathbb{R}^m_+$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \stackrel{\text{def}}{=} e^{-M_{\mathbf{XY}}/\gamma}$$

$$L(P,\alpha,\beta) = \sum_{ij} P_{ij}M_{ij} + \gamma P_{ij}\log P_{ij} + \alpha^T(P\mathbf{1} - \boldsymbol{a}) + \beta^T(P^T\mathbf{1} - \boldsymbol{b})$$

$$\partial L/\partial P_{ij} = M_{ij} + \gamma(\log P_{ij} + 1) + \alpha_i + \beta_j$$

$$(\boldsymbol{\partial L/\partial P_{ij}} = 0) \Rightarrow P_{ij} = e^{\frac{\alpha_i}{\gamma}+\frac{1}{2}}\, e^{-\frac{M_{ij}}{\gamma}}\, e^{\frac{\beta_j}{\gamma}+\frac{1}{2}} = \boldsymbol{u_i}\, K_{ij} \boldsymbol{v_j}$$

Fast & Scalable Algorithm

Prop. If $P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a},\boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\mathbf{XY}} \rangle - \gamma E(\boldsymbol{P})$

then $\exists ! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \textbf{diag}(\boldsymbol{u}) K \textbf{diag}(\boldsymbol{v}), \quad K \overset{\text{def}}{=} e^{-M_{\mathbf{XY}}/\gamma}$$

- **[Sinkhorn'64]** fixed-point iterations for $(\boldsymbol{u}, \boldsymbol{v})$

$$\boldsymbol{u} \leftarrow \boldsymbol{a} / K\boldsymbol{v}, \quad \boldsymbol{v} \leftarrow \boldsymbol{b} / K^T \boldsymbol{u}$$

- $O(nm)$ complexity, GPGPU parallel **[C'13]** .
- $O(n^{d+1})$ if $\Omega = \{1, \ldots, n\}^d$ and D^p separable.

[S..C..'15]

Sinkhorn Divergence

Def. For $\gamma > 0$, let $W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \langle \boldsymbol{P}_{\boldsymbol{\gamma}}, M_{\boldsymbol{XY}} \rangle$

Prop. $W_\gamma(\boldsymbol{\mu}, \boldsymbol{\mu}) > 0$

Def. Normalized Sinkhorn Divergence

$$\bar{W}_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) - \frac{1}{2}\left(W_\gamma(\boldsymbol{\mu}, \boldsymbol{\mu}) + W_\gamma(\boldsymbol{\nu}, \boldsymbol{\nu})\right)$$

Prop. If $p = 1$, $\bar{W}_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \underset{\gamma \to \infty}{\longrightarrow} \mathrm{ED}(\boldsymbol{\mu}, \boldsymbol{\nu})$

Algorithmic Formulation

Def. For $L \geq 1$, define

$$W_L(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \langle \boldsymbol{P_L}, M_{\mathbf{XY}} \rangle,$$

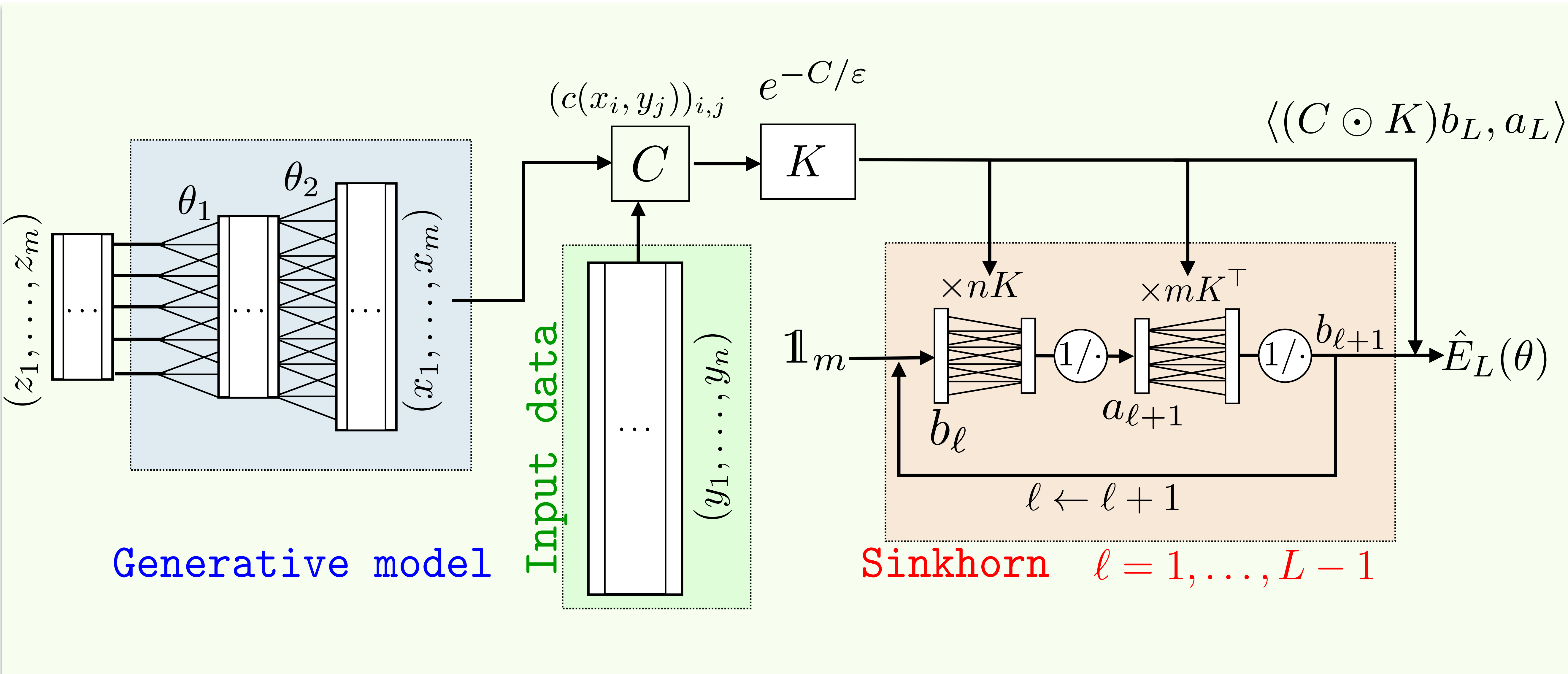
where $\boldsymbol{P_L} \stackrel{\text{def}}{=} \mathbf{diag}(\boldsymbol{u_L}) K \mathbf{diag}(\boldsymbol{v_L})$,

$\boldsymbol{v_0} = \mathbf{1}_m; l \geq 0, \boldsymbol{u_l} \stackrel{\text{def}}{=} \boldsymbol{a}/K\boldsymbol{v_l}, \boldsymbol{v_{l+1}} \stackrel{\text{def}}{=} \boldsymbol{b}/K^T\boldsymbol{u_l}.$

Prop. $\frac{\partial W_L}{\partial \mathbf{X}}, \frac{\partial W_L}{\partial \boldsymbol{a}}$ can be computed recursively, in $O(L)$ kernel $K \times$vector products.

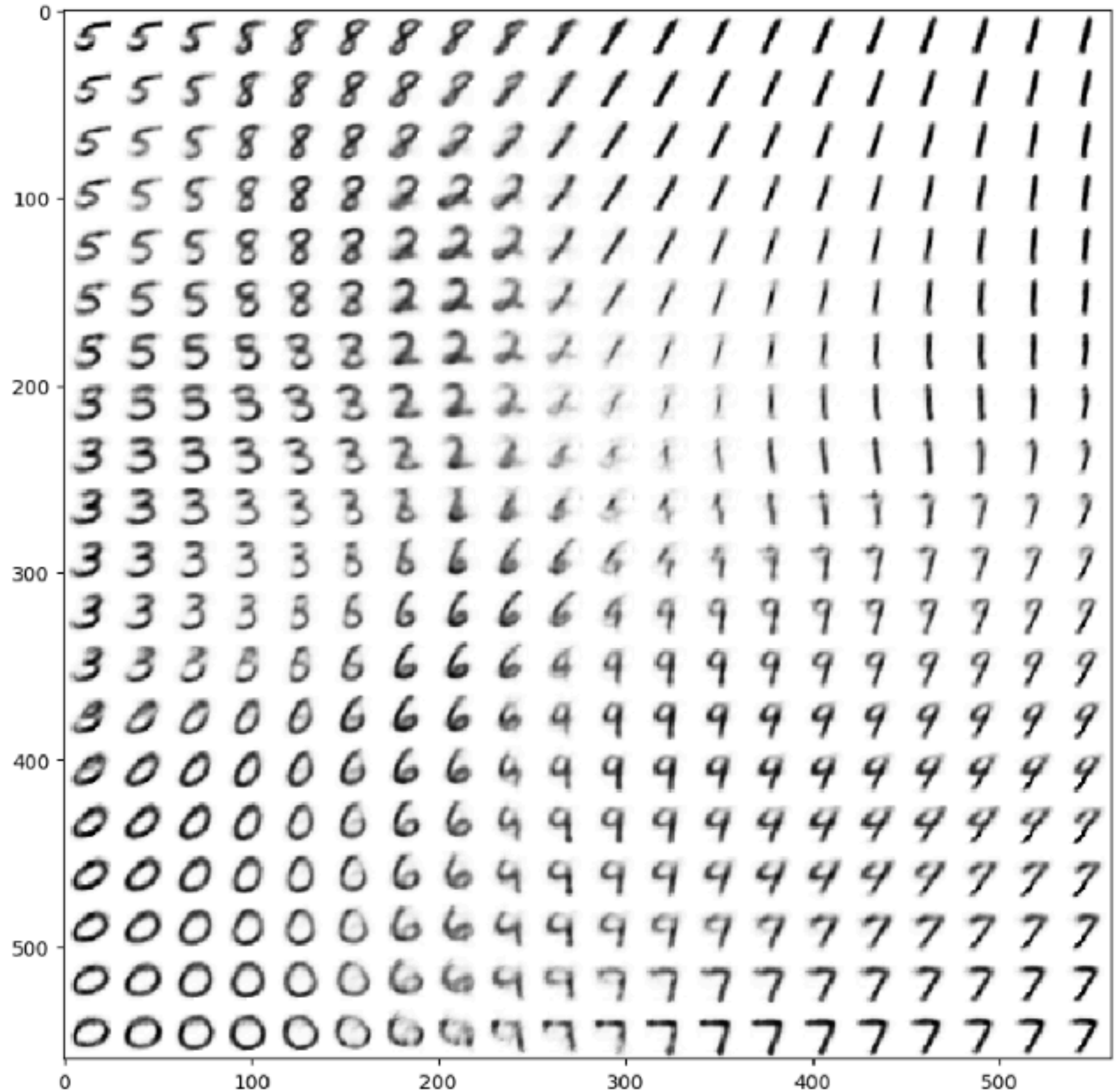
Proposal: Autodiff OT using Sinkhorn

Approximate W loss by the transport cost $\bar{W}_L$ after L Sinkhorn iterations.

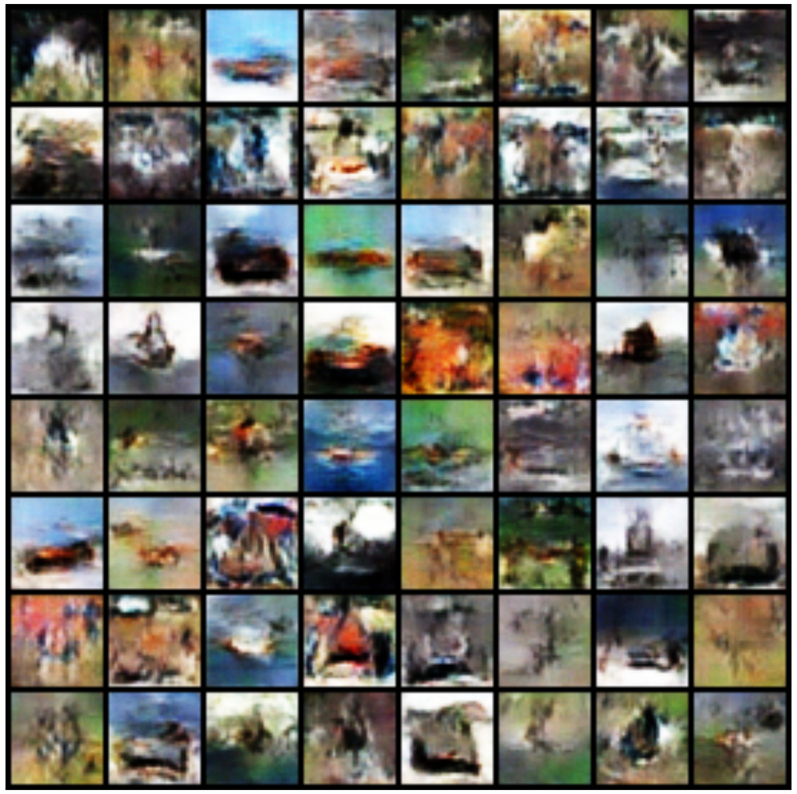


[GPC'17]

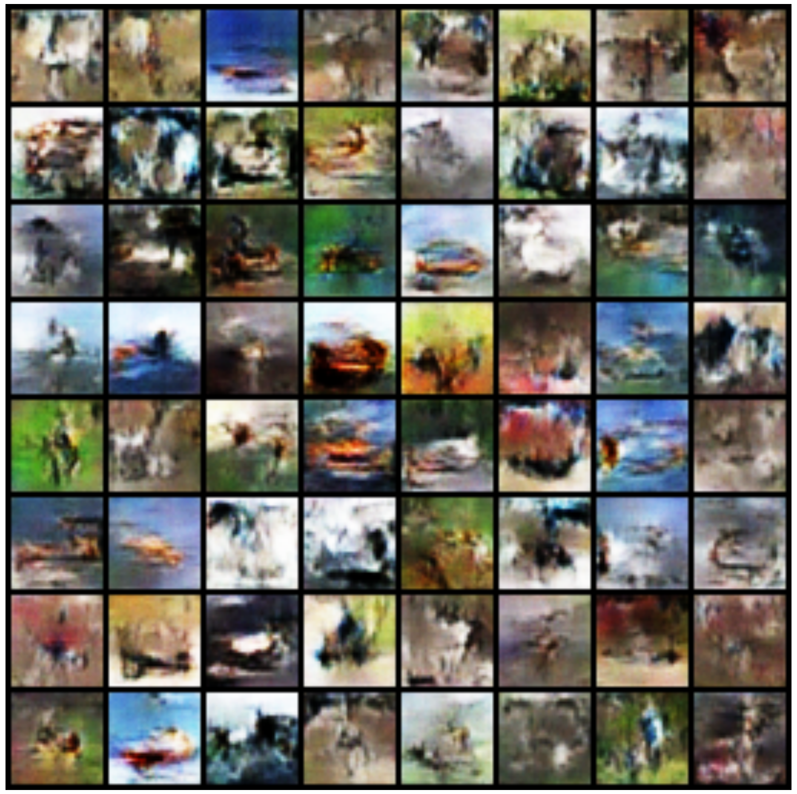
Example: MNIST, Learning f_{θ}



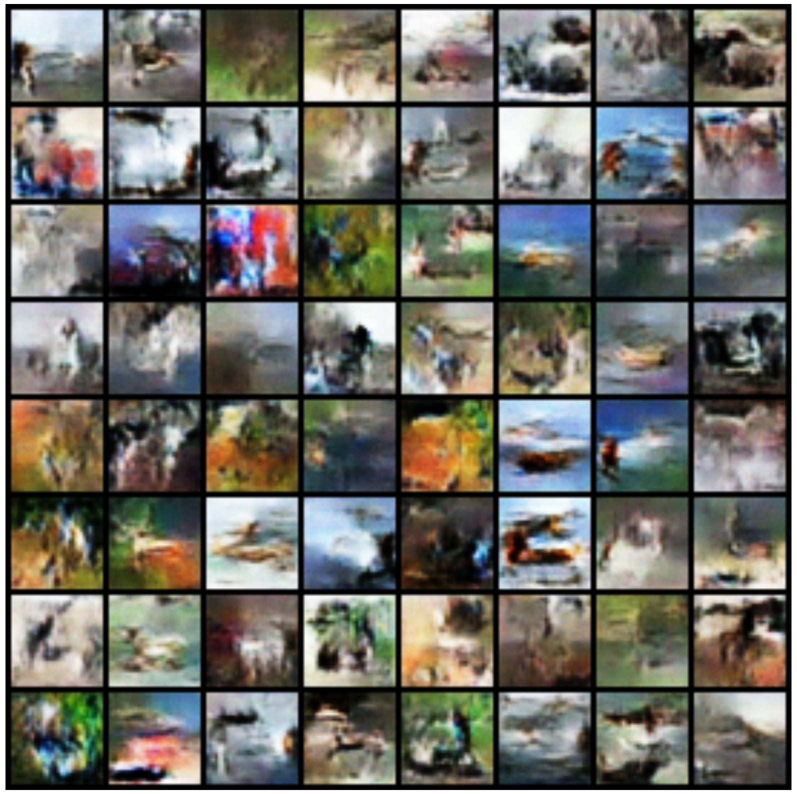
Example: Generation of Images



MMD-GAN



$\tau = 1000$



$\tau=10$

- CIFAR 10 images
- In these examples the cost function is also learned adversarially, as a NN mapping onto feature vectors.

Concluding Remarks

- *Regularized* OT is much faster than OT.
- *Regularized* OT can interpolate between W and the *MMD / Energy distance* metrics.
- The solution of *regularized OT* is *"auto-differentiable"*.
- **Many open problems remain!**

Sat Dec 9th 08:00 AM -- 06:30 PM @ None
Workshop
Optimal Transport and Machine Learning
Olivier Bousquet · Marco Cuturi · Gabriel Peyré · Fei Sha · Justin Solomon
NIPS'17 WORKSHOP

Dates n/a. @ TBA
Tutorial
A Primer on Optimal Transport
Marco Cuturi · Justin M Solomon
NIPS'17 TUTORIAL