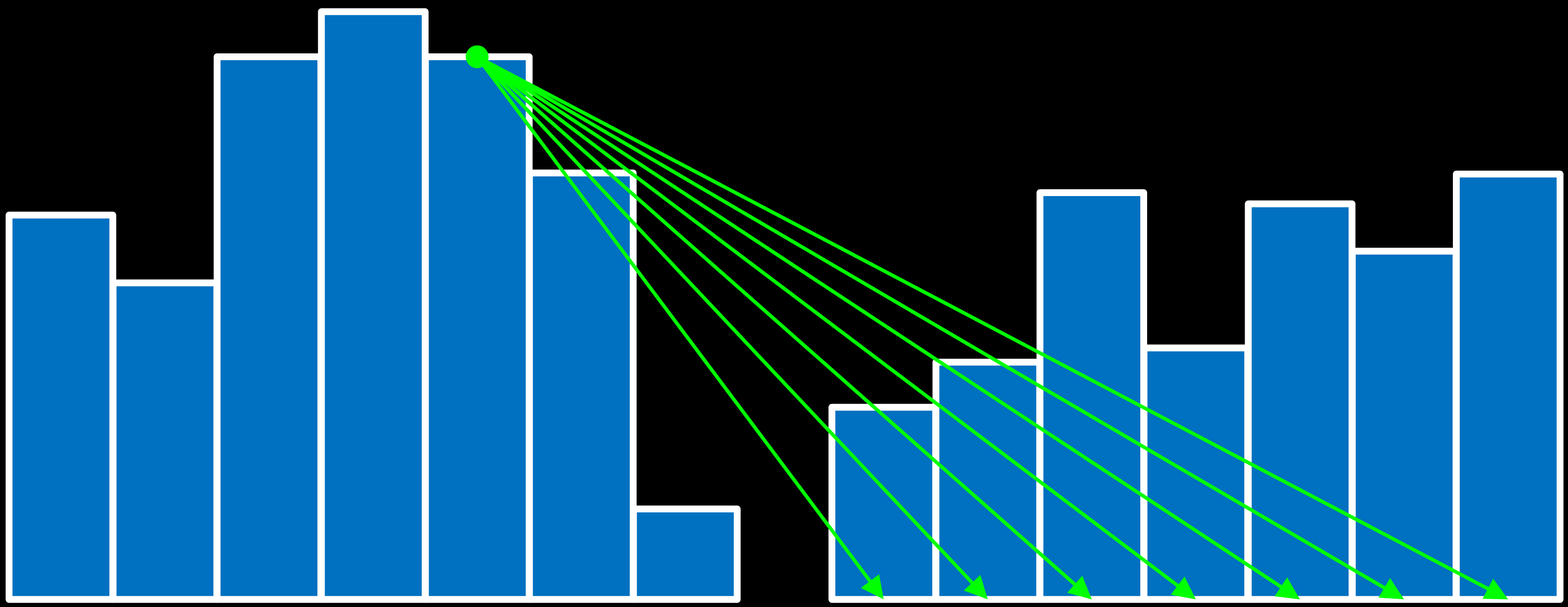


Optimal Transport

Justin Solomon
MIT, Spring 2019

Back to comfortable ground!

…toward my own research!

Big Idea

Understand geometry from a

“softened” probabilistic

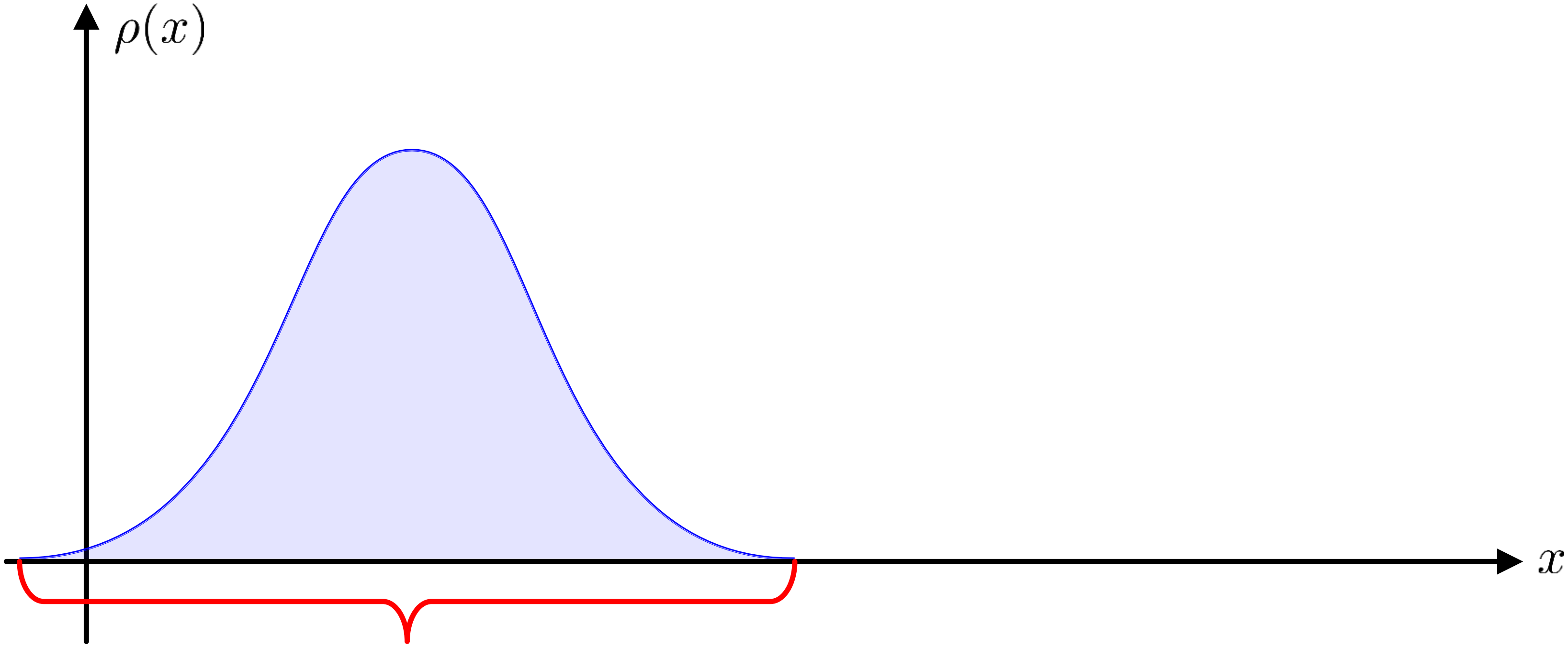
standpoint.

Secondary goal:

Application of machinery from previous lectures

(vector fields, geodesics, metric spaces, optimization…)

Probabilistic Geometry

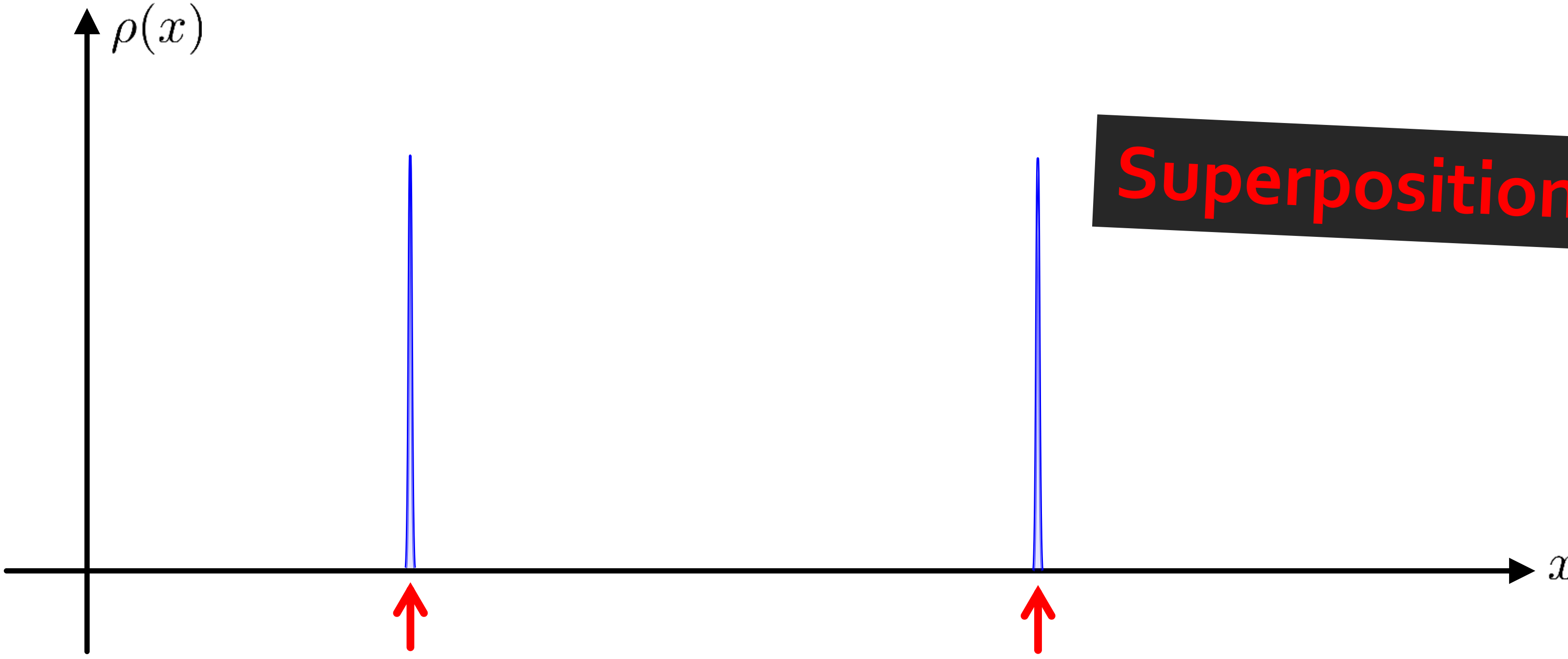


"Somewhere over here."

Probabilistic Geometry

"Exactly here."

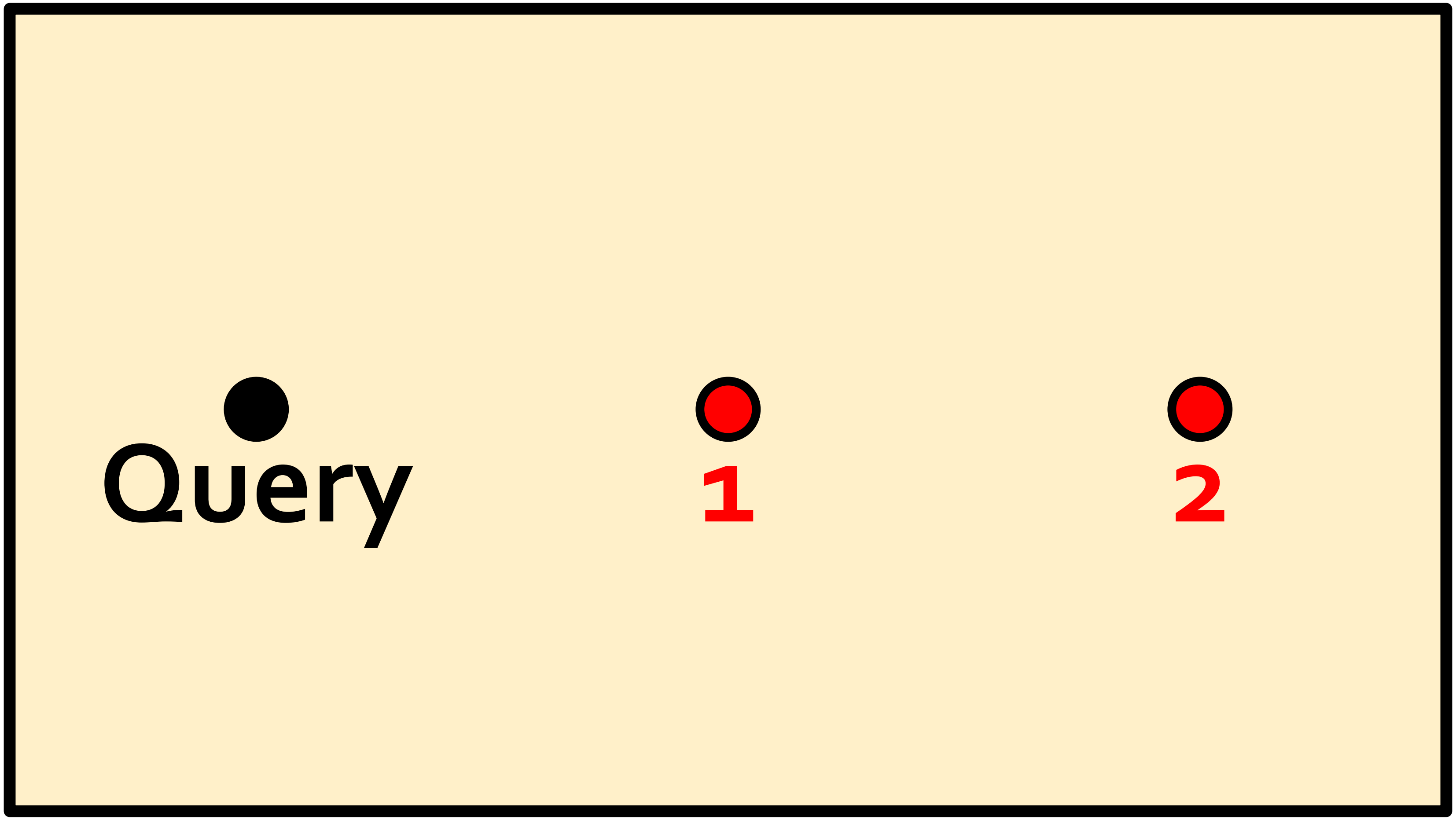
Probabilistic Geometry



Superposition

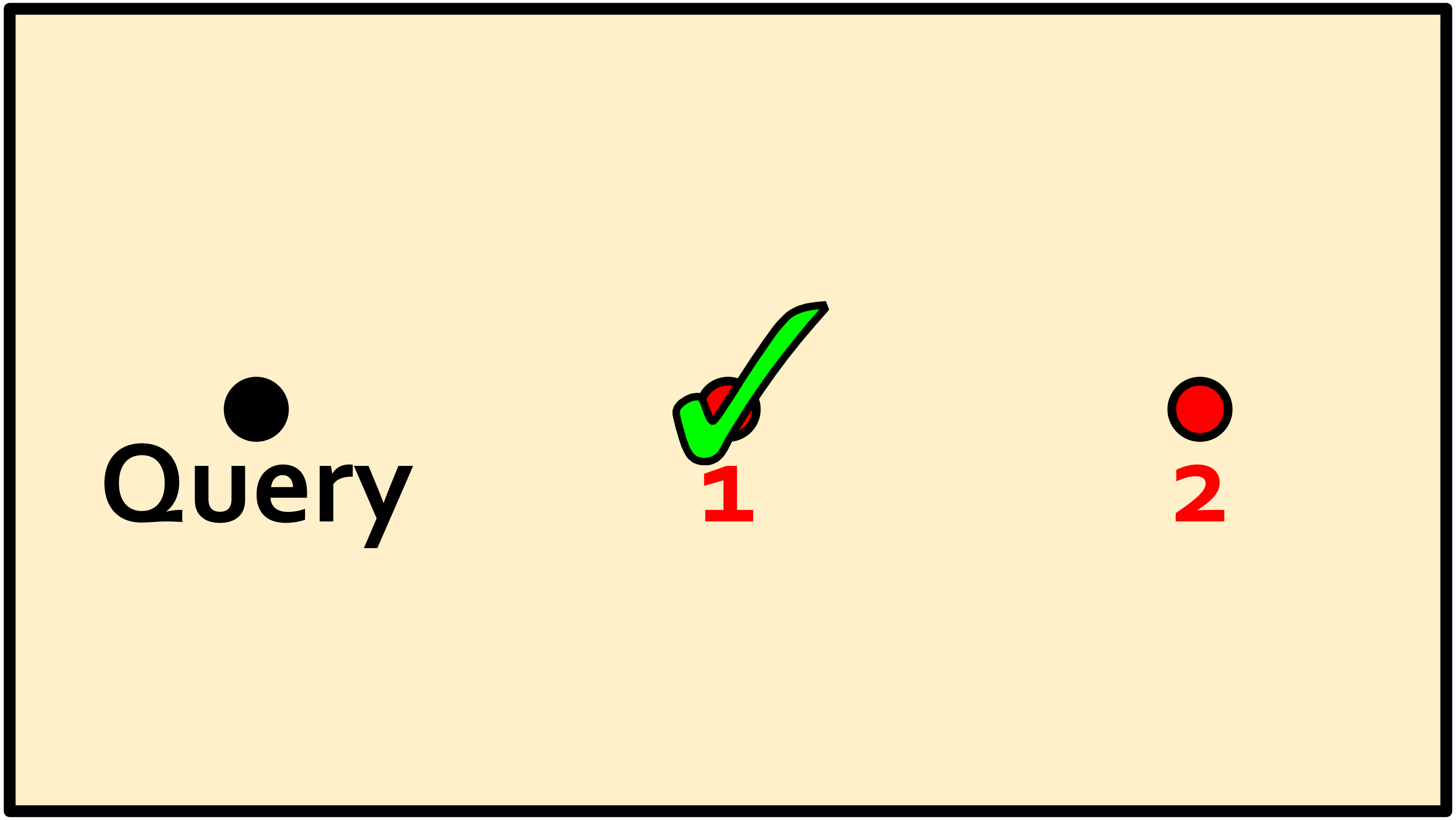
"One of these two places."

Motivating Question



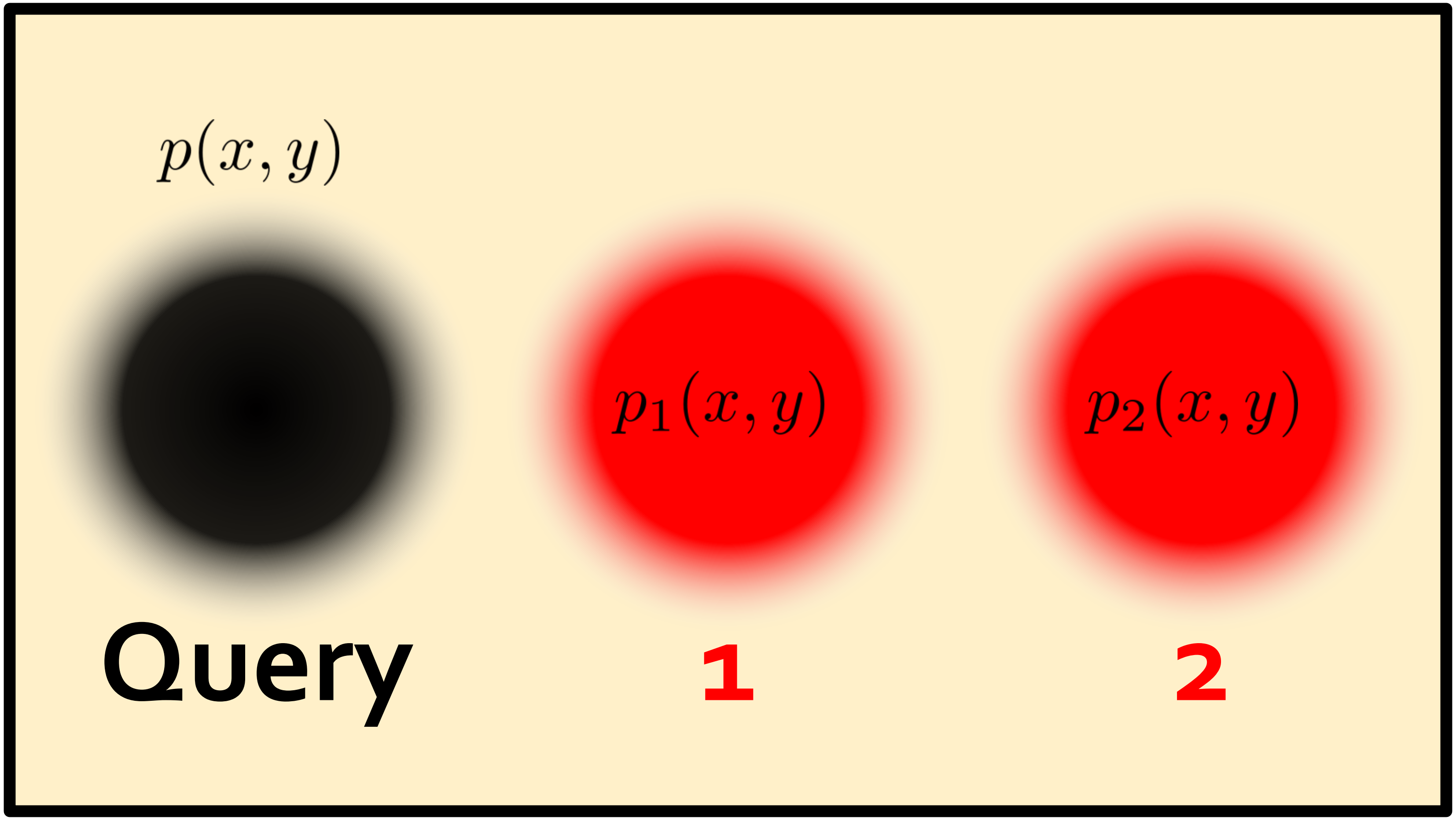
Which is closer, 1 or 2?

Motivating Question



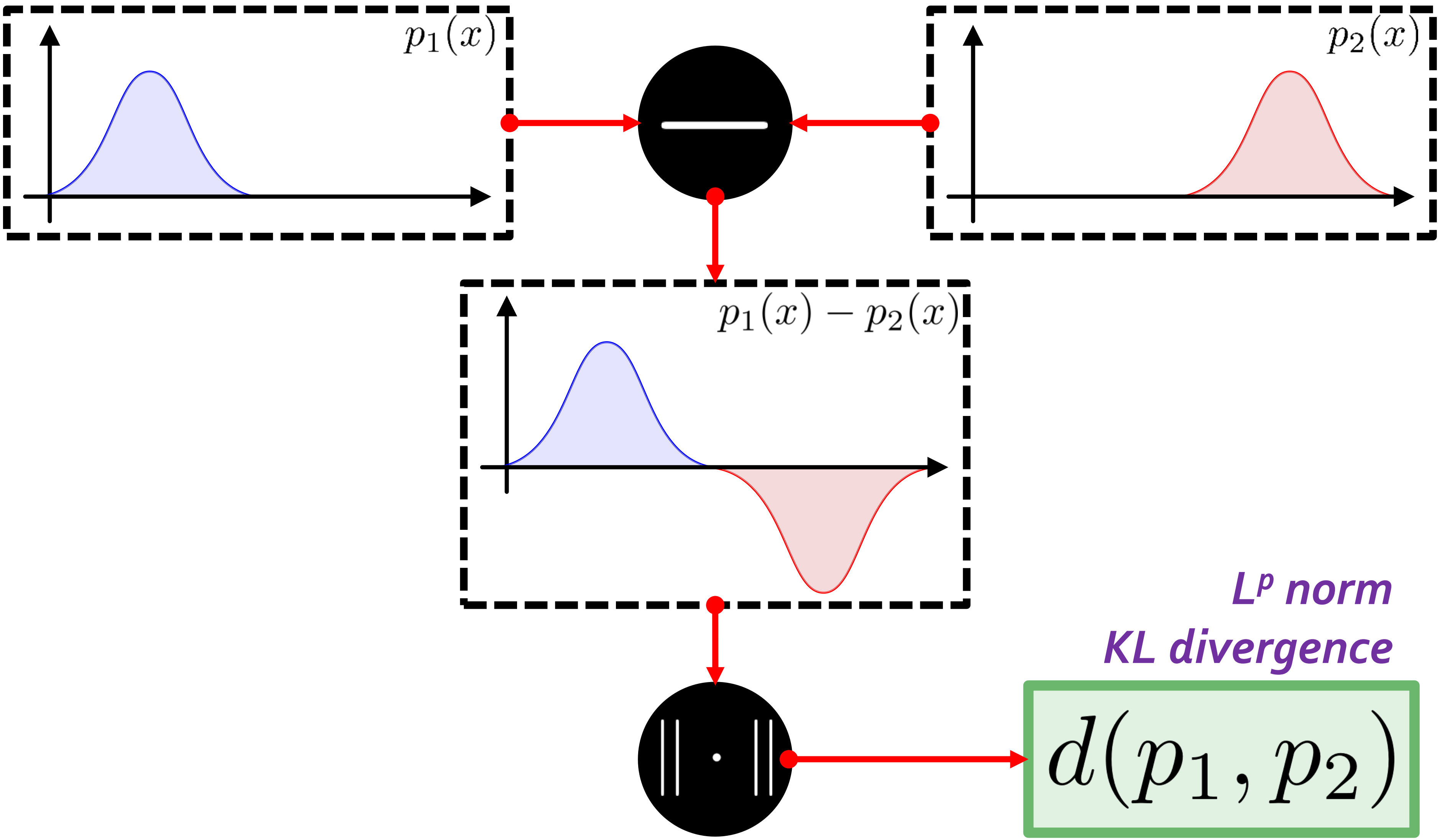
Which is closer, 1 or 2?

Fuzzy Version

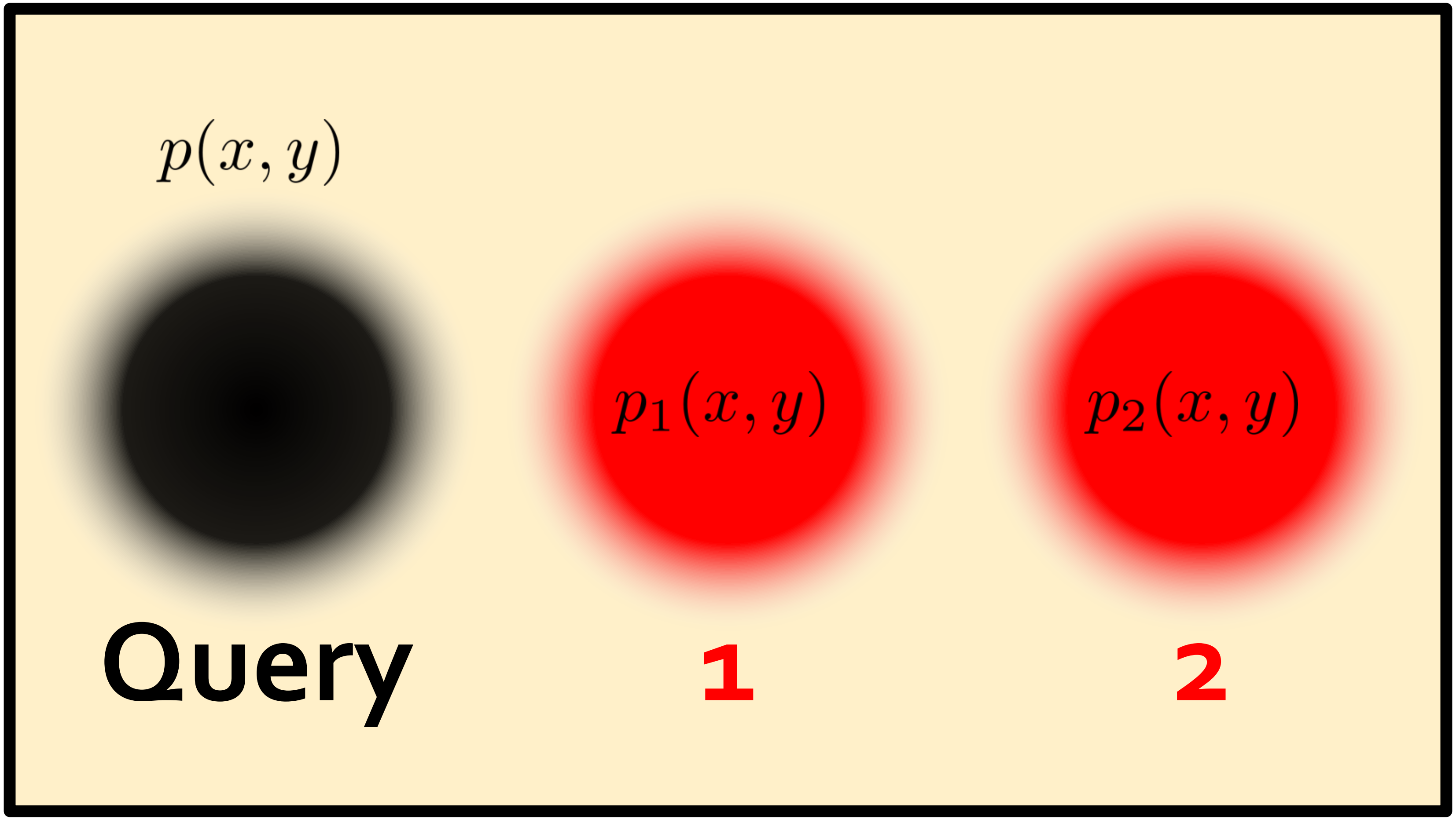


Which is closer, 1 or 2?

Typical Measurement

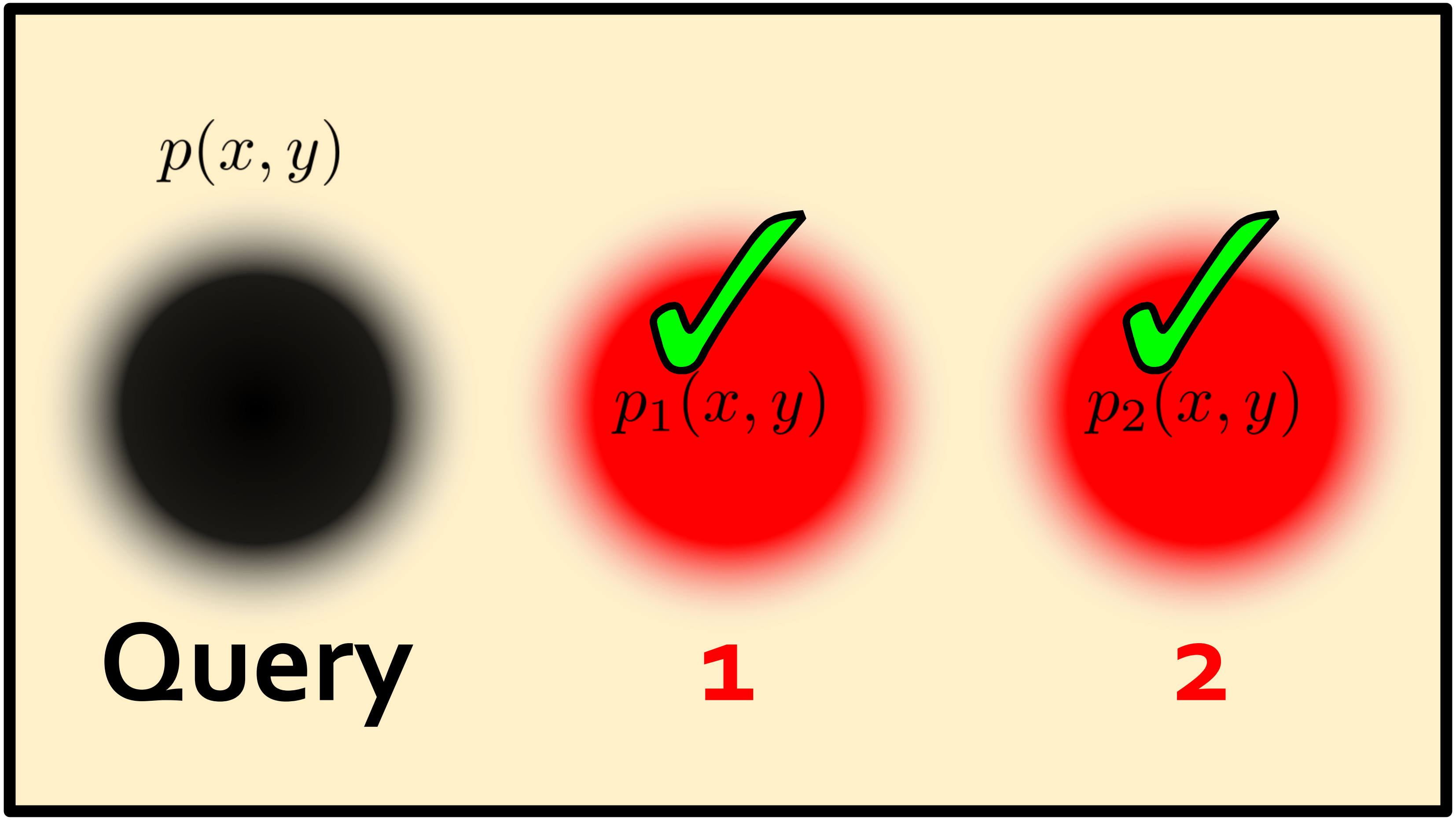


Returning to the Question



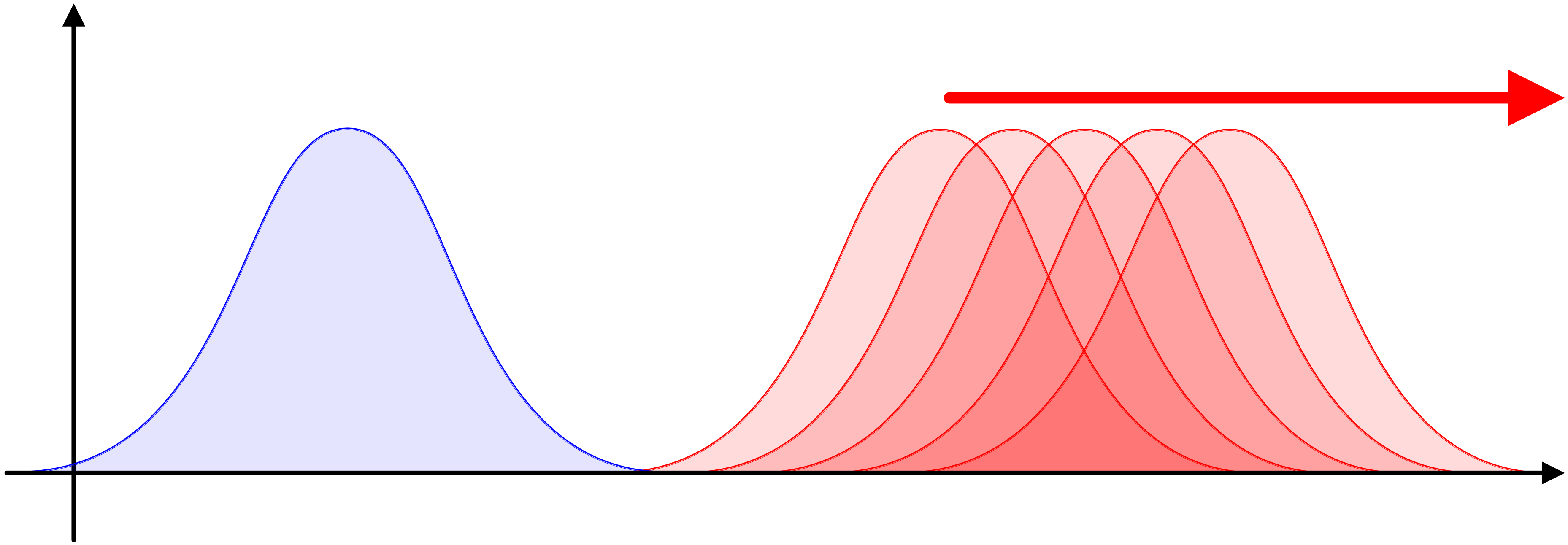
Which is closer, 1 or 2?

Returning to the Question



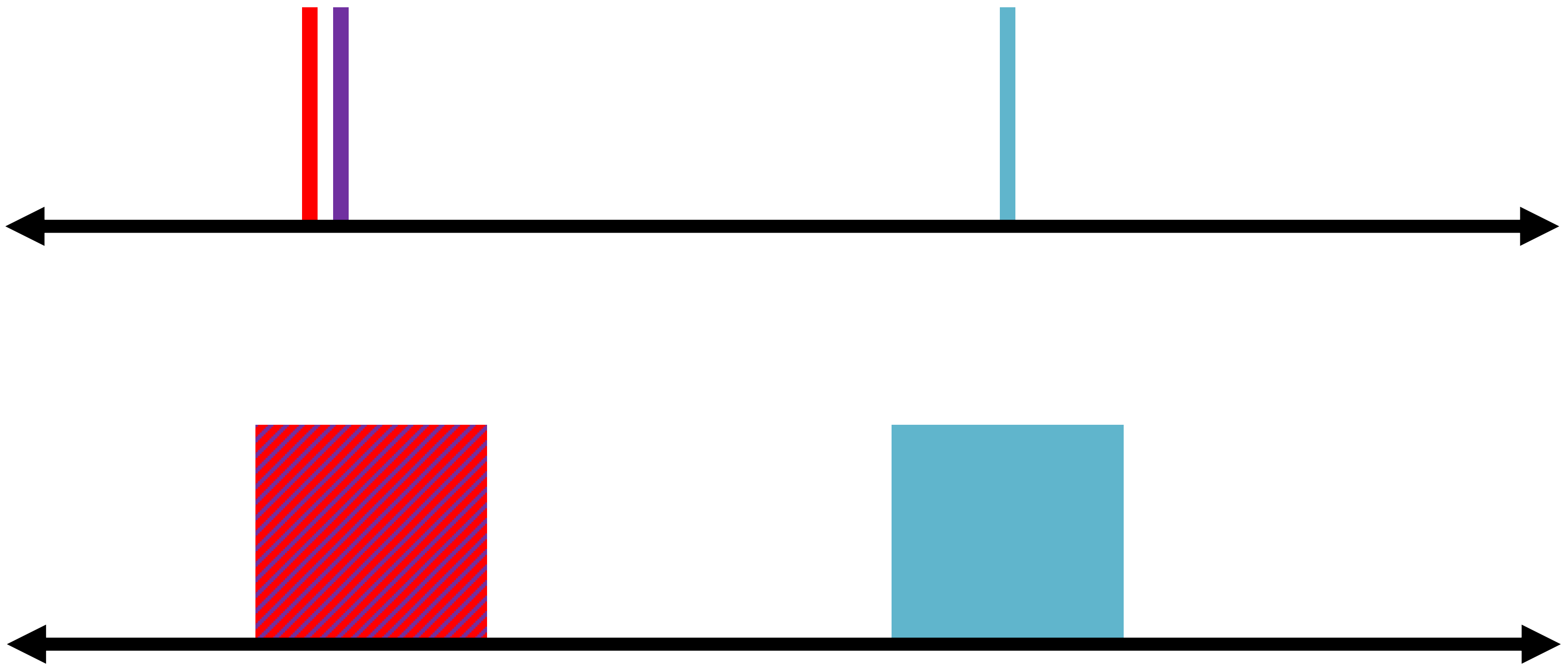
Neither! Equidistant.

What's Wrong?



Measured overlap, not displacement.

Related Issue



Smaller bins worsen histogram distances

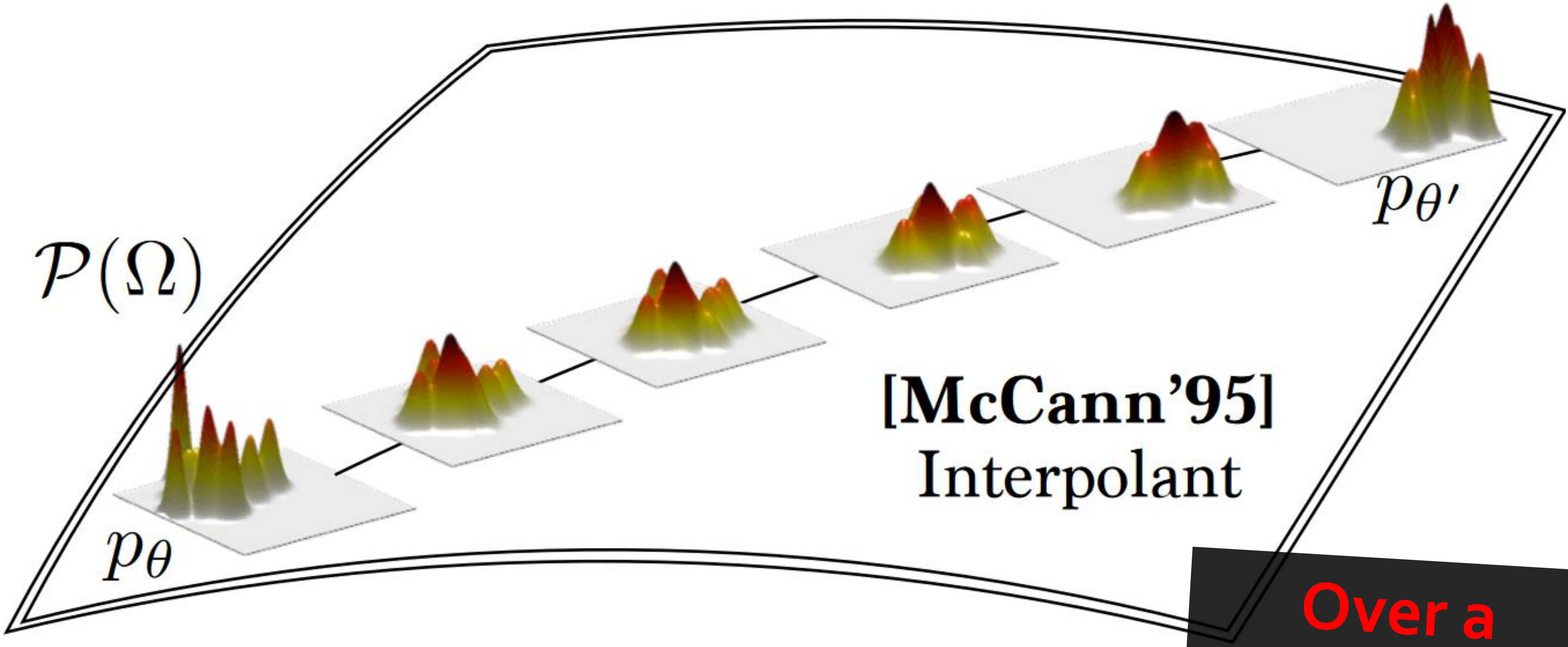
The Root Cause

Permuting histogram bins has

no effect

on these distances.

Optimal Transport

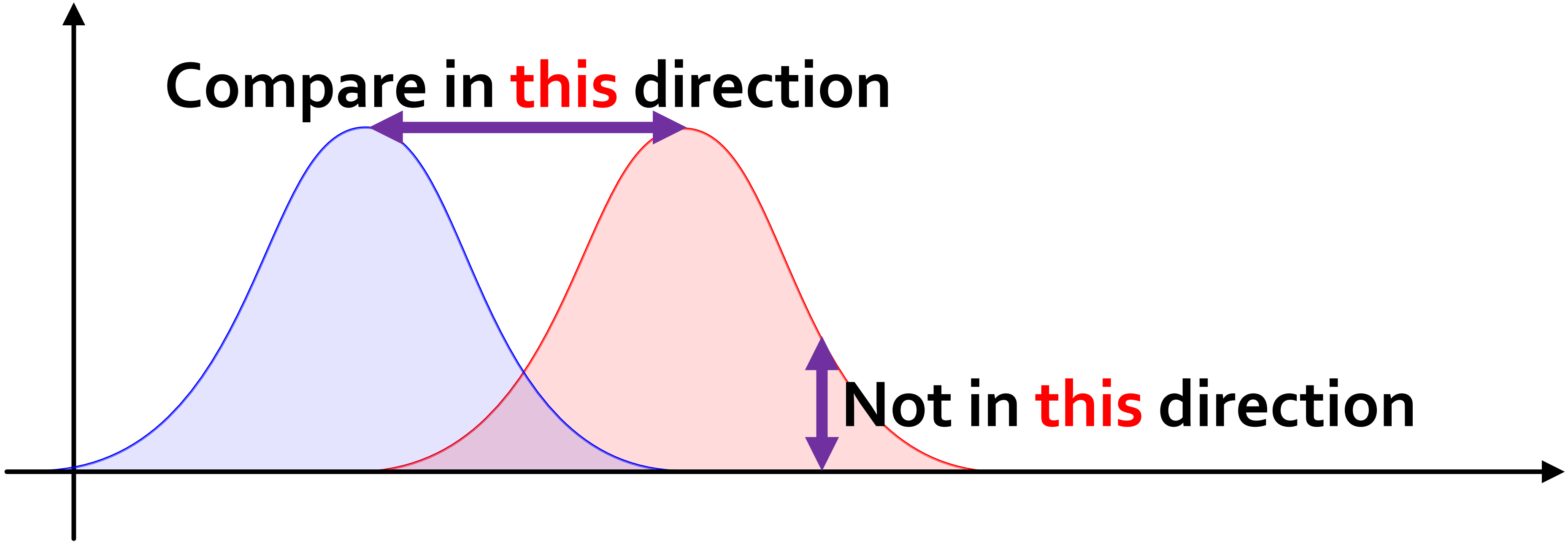


Over a metric space

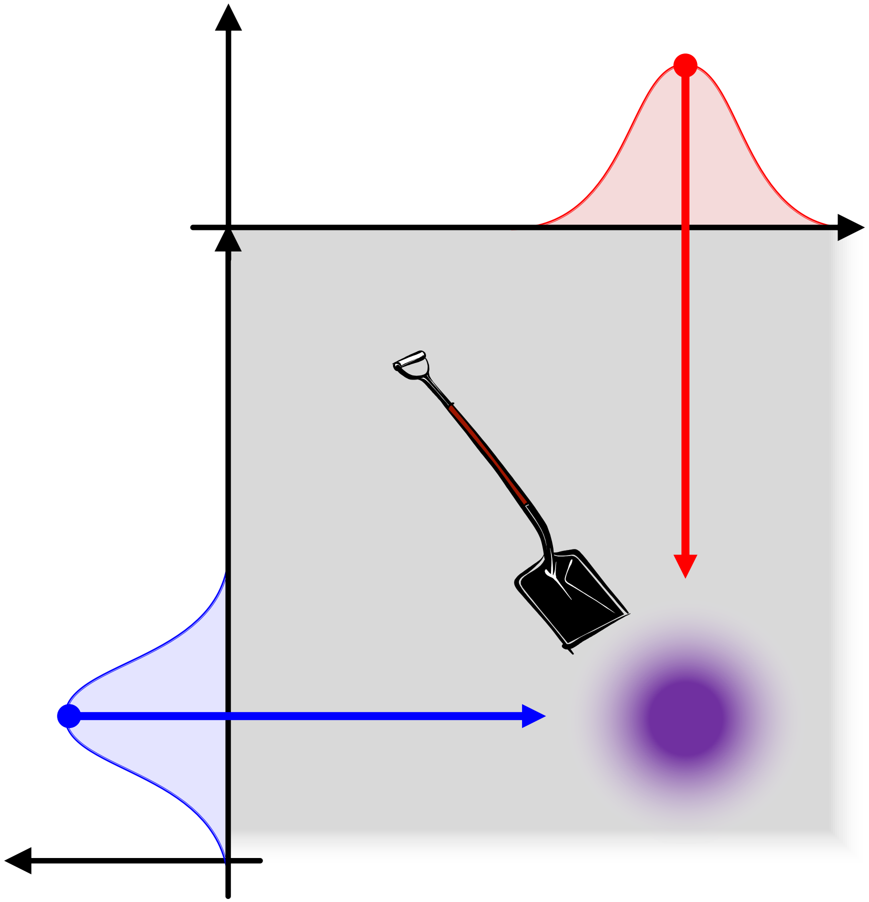
Image courtesy M. Cuturi

Geometric theory of probability

Alternative Idea

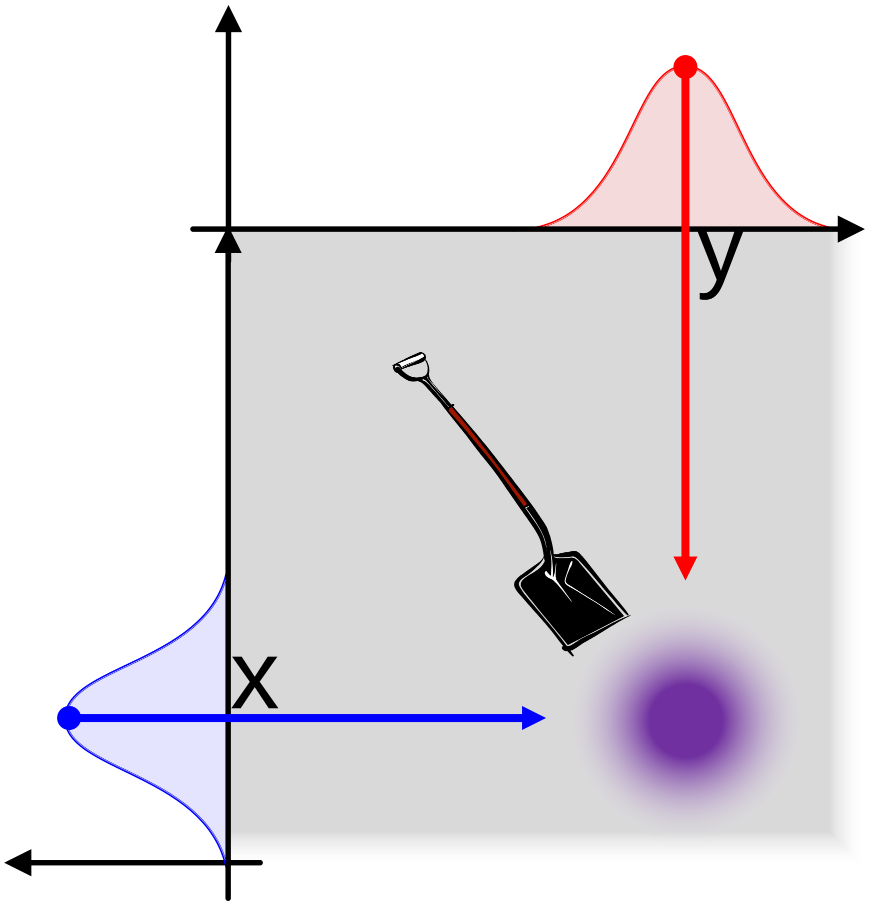


Alternative Idea



Match mass from the distributions

Earth Mover's Distance



Cost to move mass m from x to y:

$$\boldsymbol{m \cdot d(x, y)}$$

Match mass from the distributions

Transportation Matrix

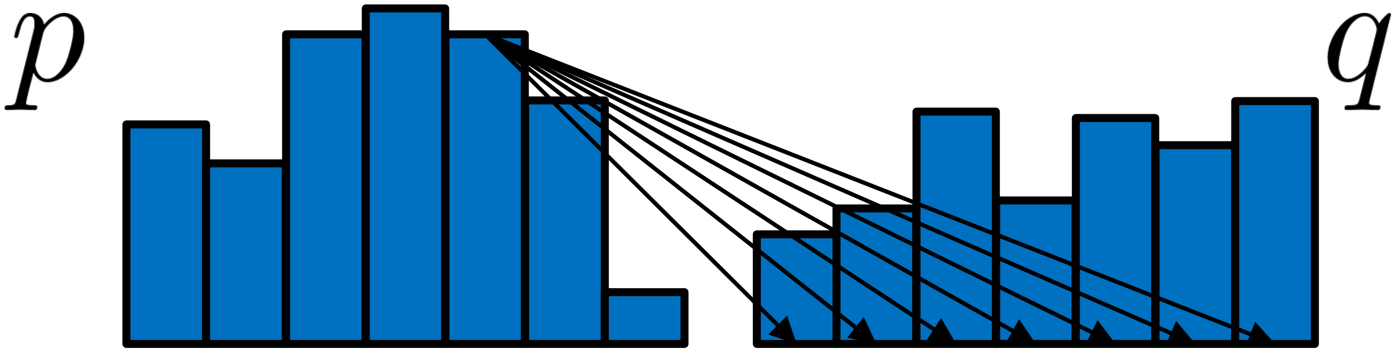
- **Supply** **distribution** $\boldsymbol{p_0}$
- **Demand** **distribution** $\boldsymbol{p_1}$

$$T \geq 0$$

$$T\mathbf{1} = p_0$$

$$T^\top \mathbf{1} = p_1$$

Earth Mover's Distance



$$\min_T \sum_{ij} T_{ij} d(x_i, x_j)$$ $\boldsymbol{m \cdot d(x, y)}$

$$\text{s.t.} \sum_j T_{ij} = p_i$$ **Starts at $\boldsymbol{p}$**

$$\sum_i T_{ij} = q_j$$ **Ends at $\boldsymbol{q}$**

$$T \geq 0$$ **Positive mass**

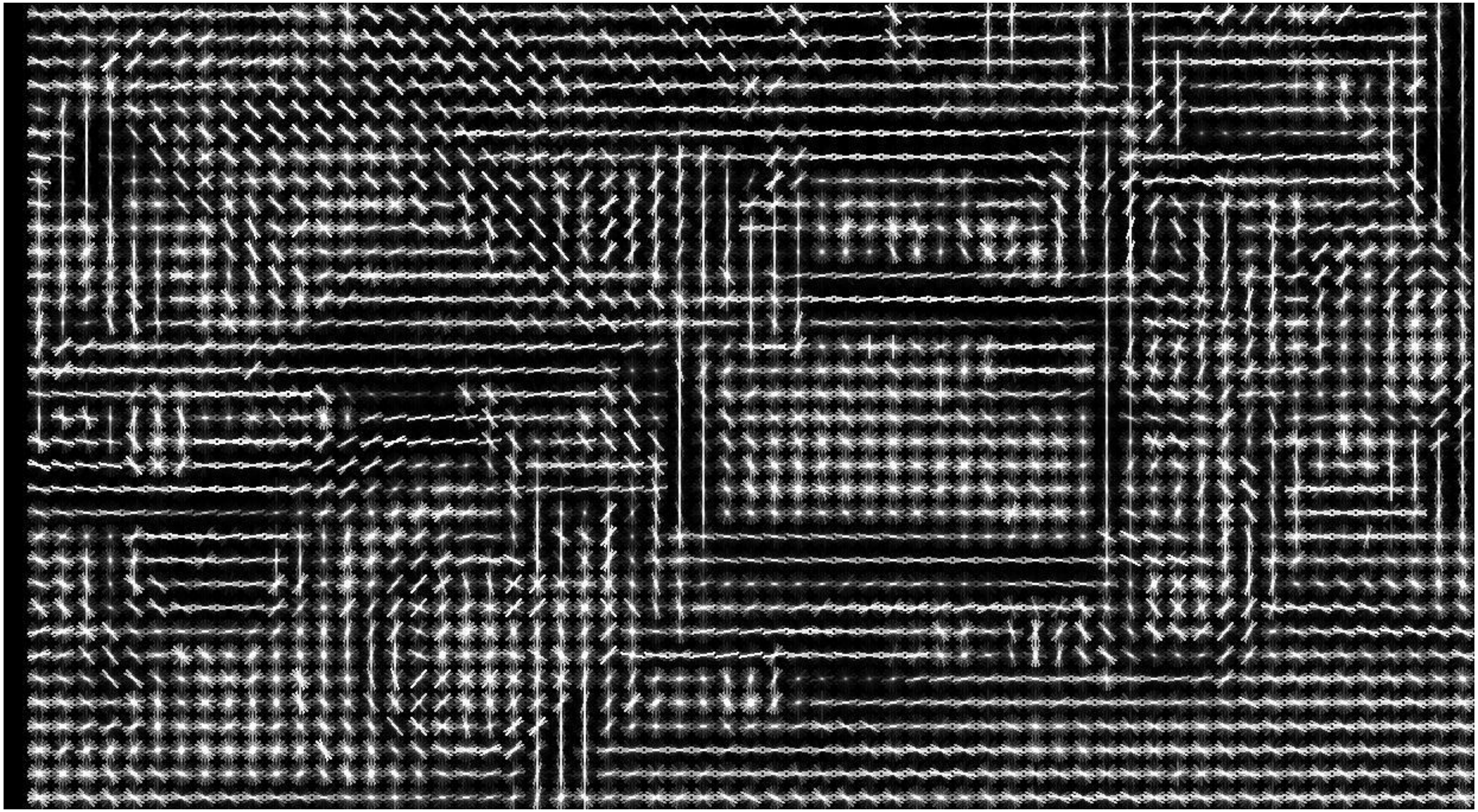
Important Theorem

EMD is a metric when d(x,y) satisfies the triangle inequality.

"The Earth Mover's Distance as a Metric for Image Retrieval"
Rubner, Tomasi, and Guibas; IJCV 40.2 (2000): 99—121.

Revised in:
"Ground Metric Learning"
Cuturi and Avis; JMLR 15 (2014)

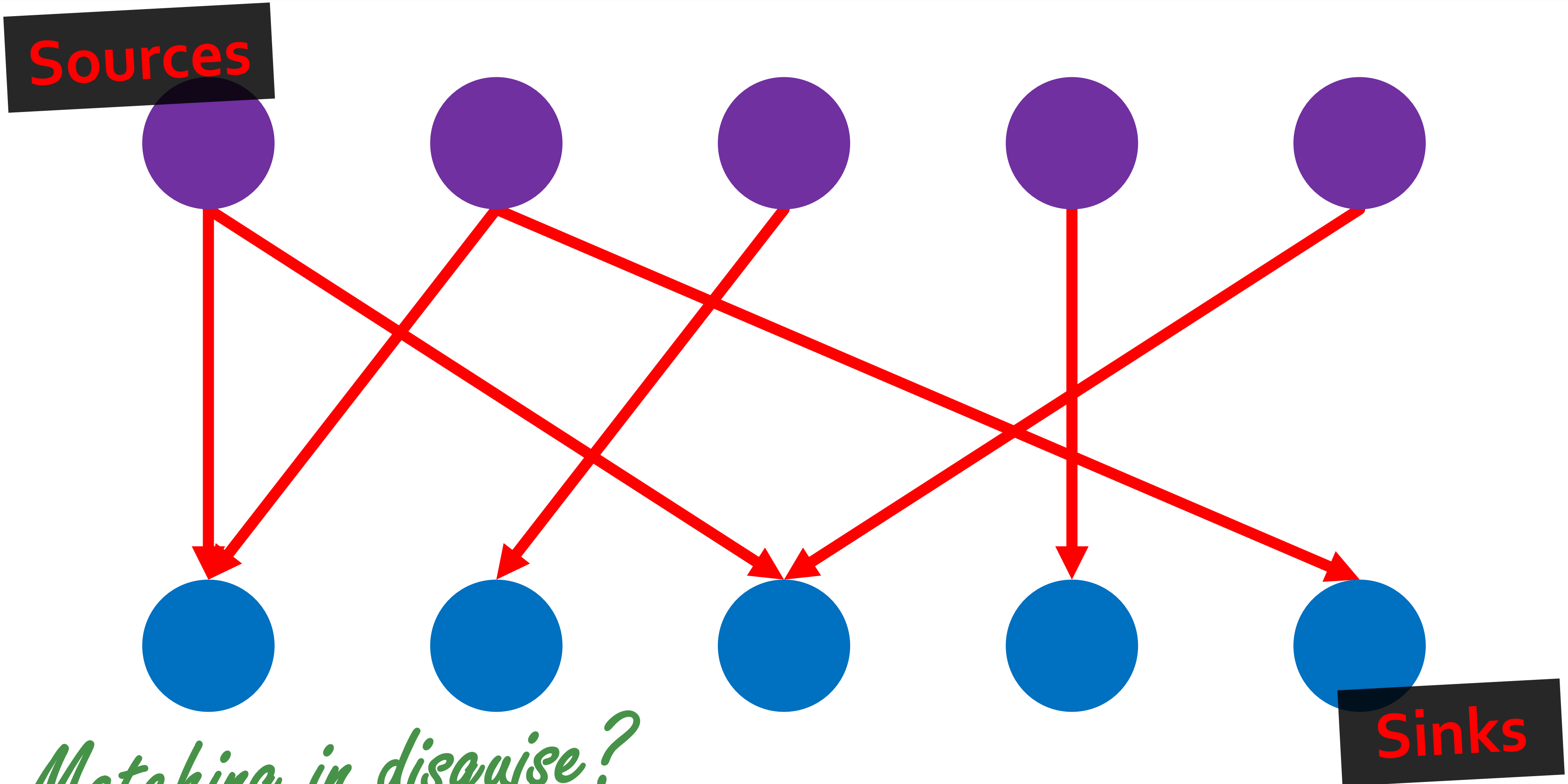
Basic Application



http://web.mit.edu/vondrick/ihog/

Comparing histogram descriptors

Discrete Perspective



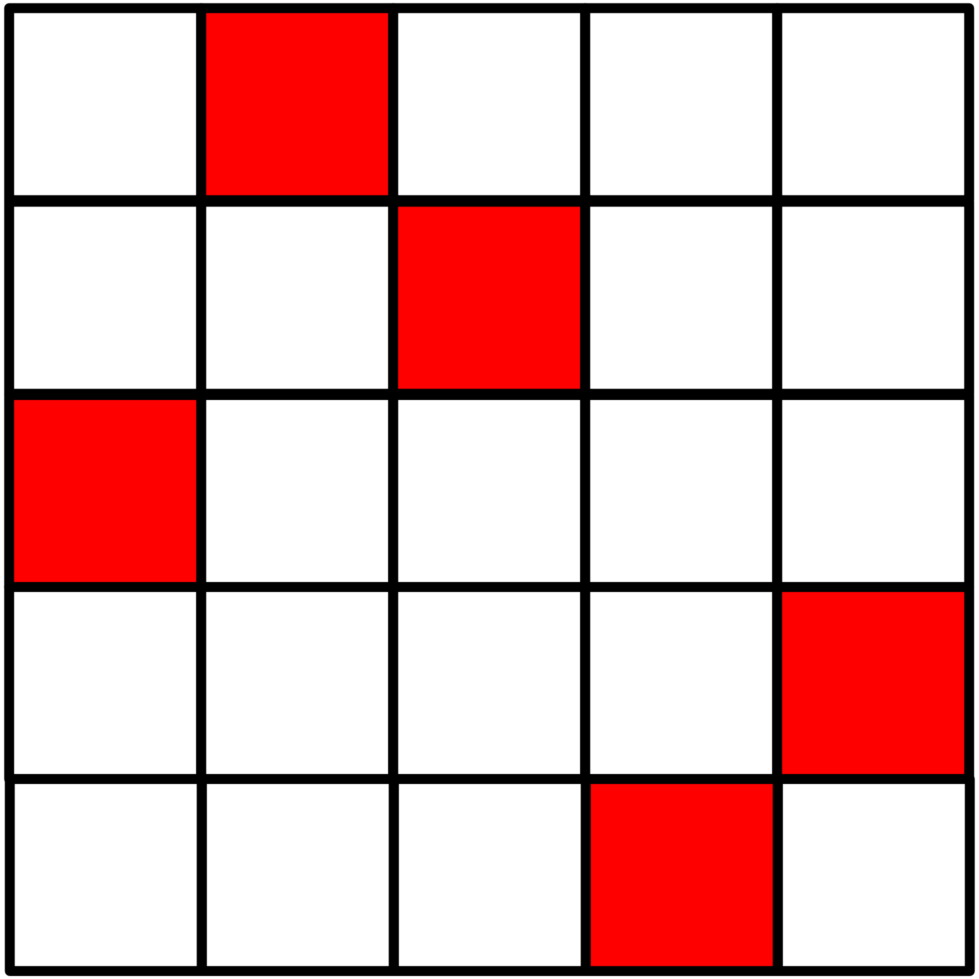
Matching in disguise?

Min-cost flow

Algorithm for Small-Scale Problems

- **Step 1:** Compute D_{ij}
- **Step 2:** Solve linear program
  - Simplex
  - Interior point
  - Hungarian algorithm
  - ...

Transportation Matrix Structure



Matches bins

Underlying map!

Discrete Perspective

Useful conclusions:

1. Practical

Can do better than generic solvers.

Min-cost flow

Discrete Perspective

Useful conclusions:

1. Practical

Can do better than generic solvers.

2. Theoretical

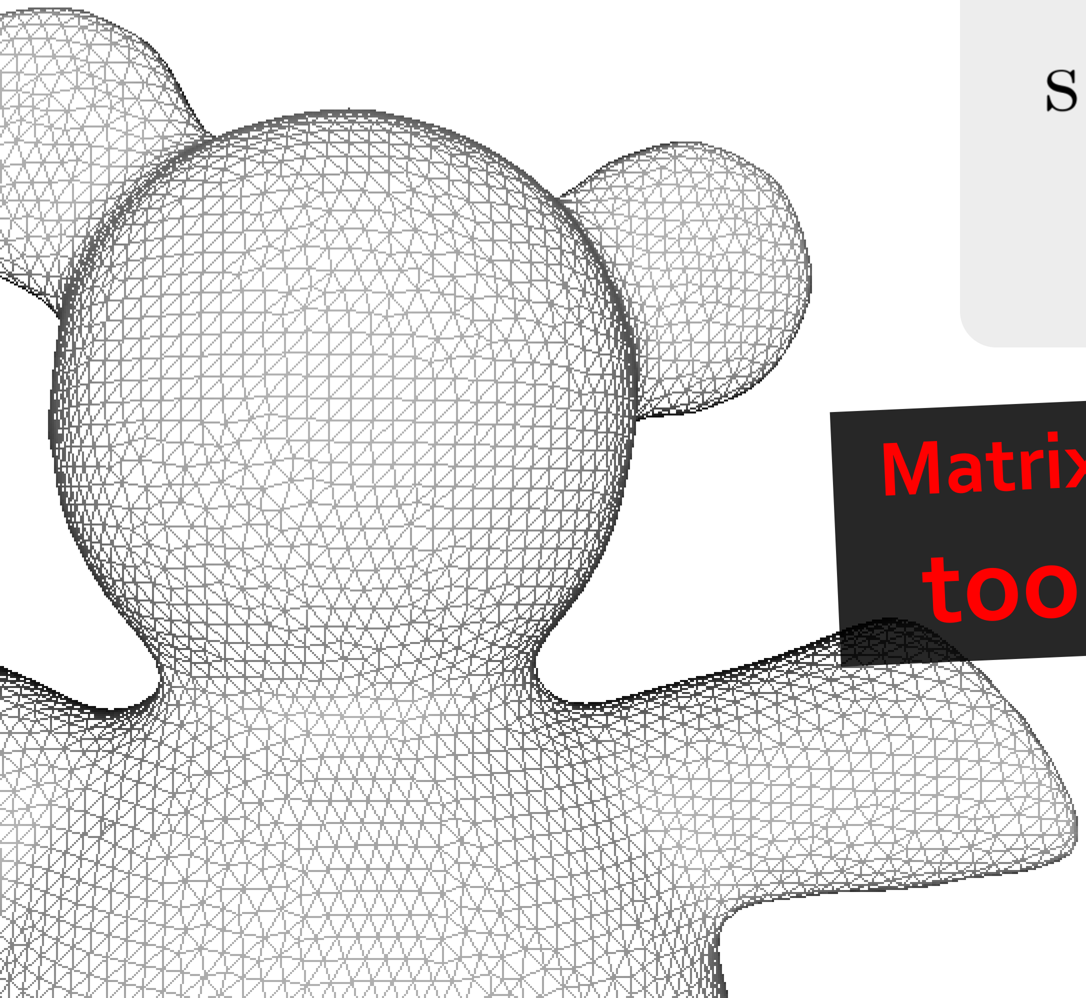
"Complementary slackness"

$T \in [0,1]^{n\times n}$ usually contains $O(n)$ nonzeros.

Min-cost flow

Challenge for Large-Scale Problems

$$
\begin{aligned}
\min_T \quad & \textstyle\sum_{ij} T_{ij} d(x_i, x_j) \\
\text{s.t.} \quad & \textstyle\sum_j T_{ij} = p_i \\
& \textstyle\sum_i T_{ij} = q_j \\
& T \geq 0
\end{aligned}
$$



Matrix T_{ij} is too big!

Precompute $d(x_i, x_j)$ for all i, j!

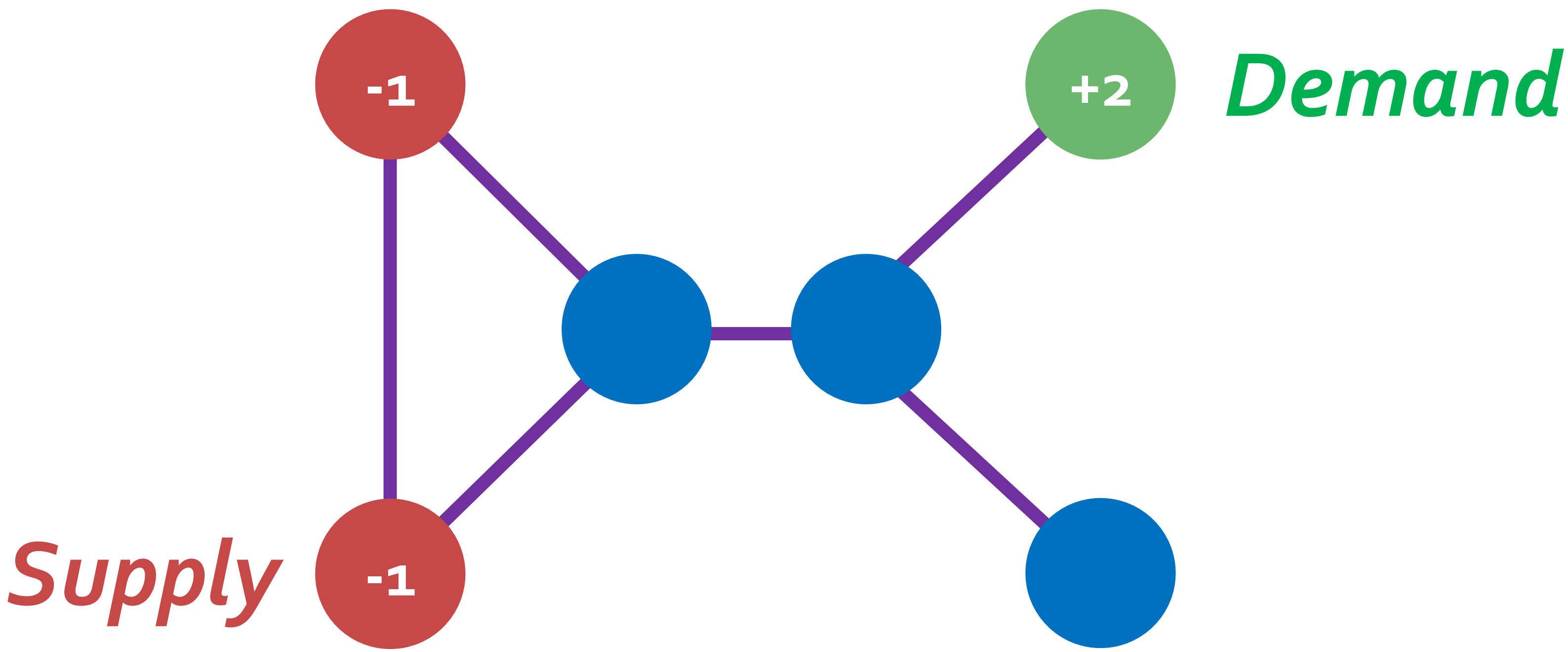
Today’s Questions

- Can we optimize faster?
- Is there a continuum interpretation?
- What properties does this model exhibit?

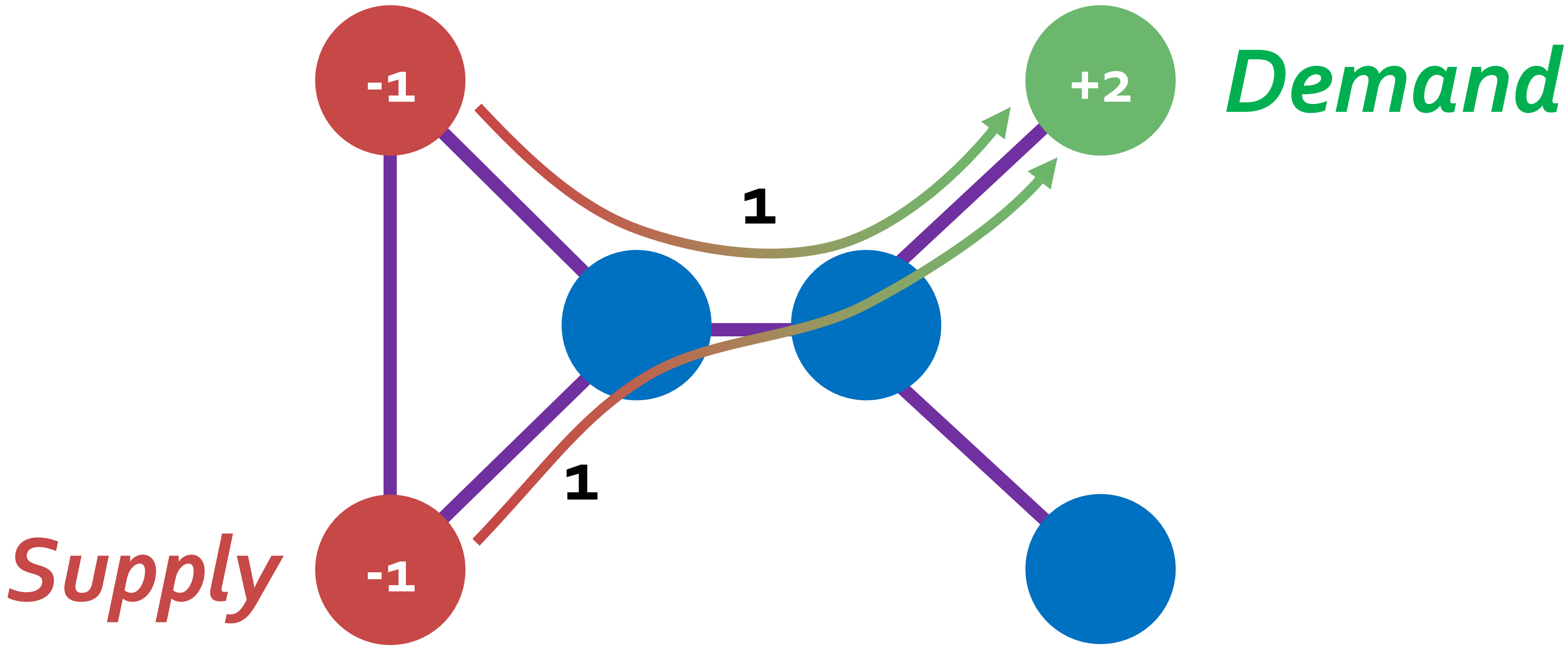
We’ll answer them in parallel!

First example:

Linear Transportation on Graphs



Linear Transportation on Graphs



Routing Supply to Meet Demand

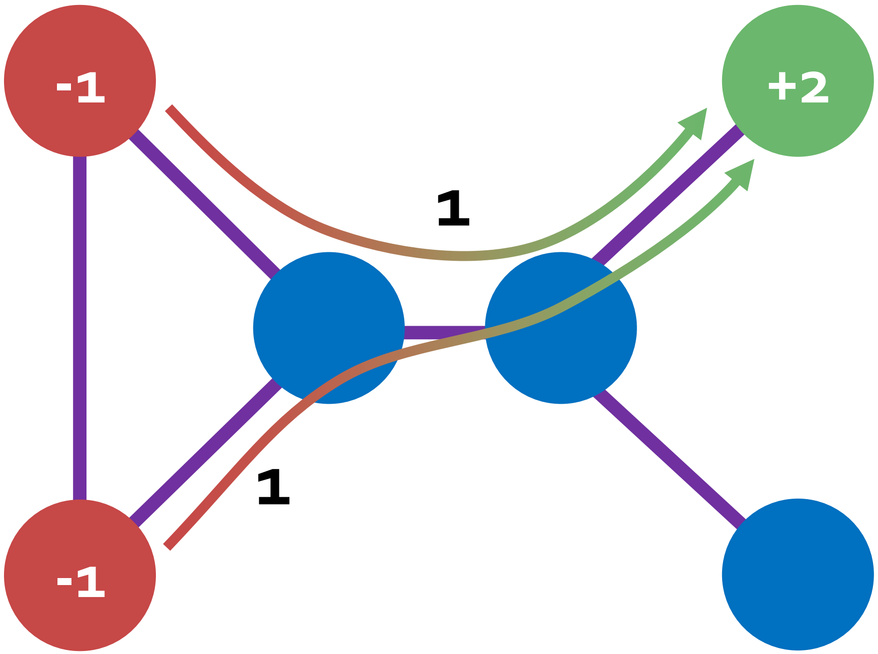
$$\min_{T} \langle T, D \rangle$$

$$\text{s.t. } T \geq 0$$

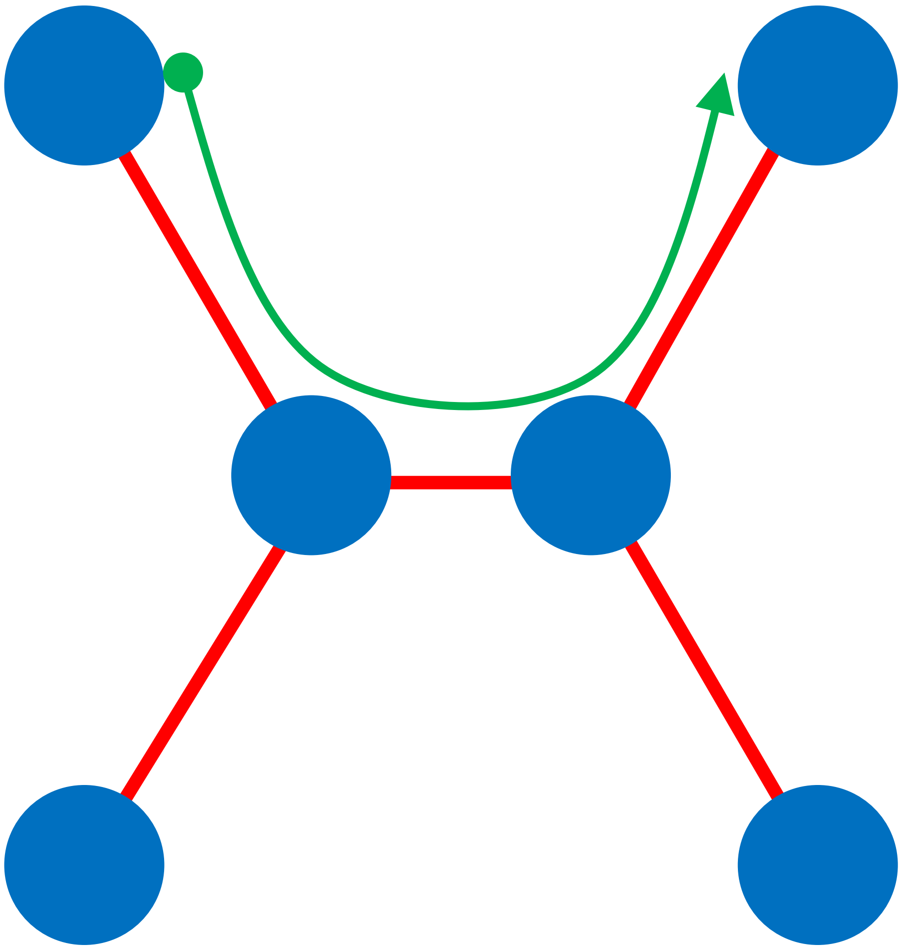
$$T\mathbf{1} = p_0$$

$$T^\top \mathbf{1} = p_1$$

D contains shortest path length.

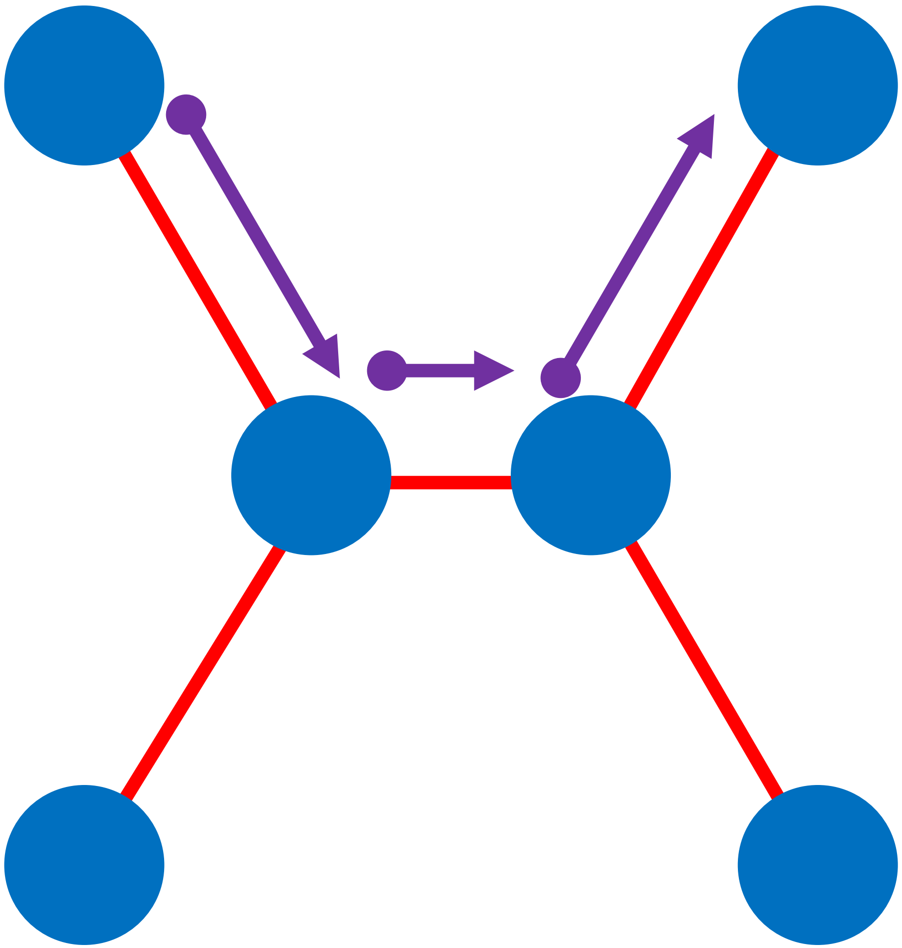


Simplification for Linear Cost



$$
\begin{aligned}
\min_T \quad & \textstyle\sum_{ij} T_{ij} d(x_i, x_j) \\
\text{s.t.} \quad & \textstyle\sum_j T_{ij} = p_i \\
& \textstyle\sum_i T_{ij} = q_j \\
& T \geq 0
\end{aligned}
$$

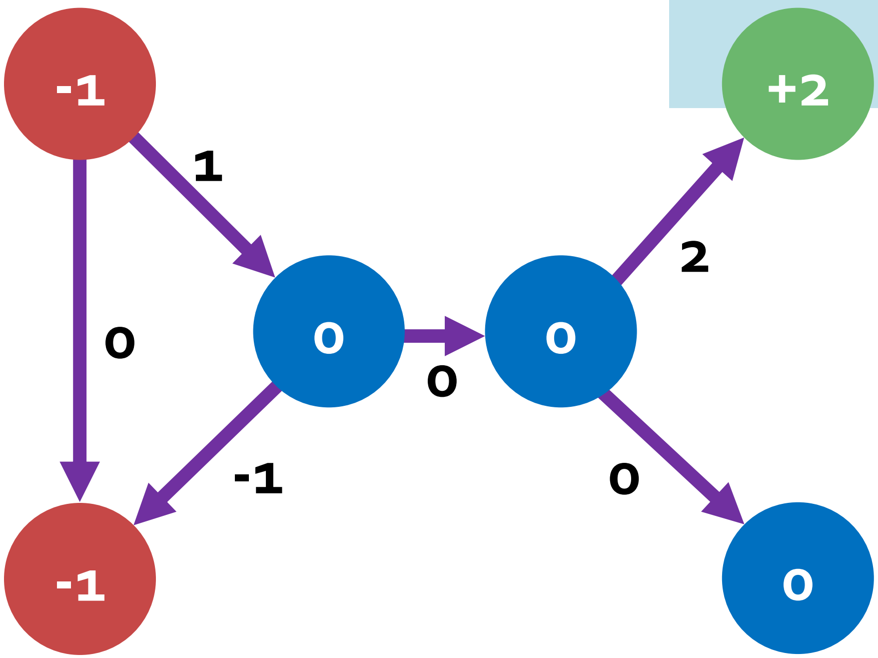
Simplification for Linear Cost



$$
\begin{aligned}
\min_T \quad & \textstyle\sum_{ij} T_{ij} d(x_i, x_j) \\
\text{s.t.} \quad & \textstyle\sum_j T_{ij} = p_i \\
& \textstyle\sum_i T_{ij} = q_j \\
& T \geq 0
\end{aligned}
$$

Differencing Operator

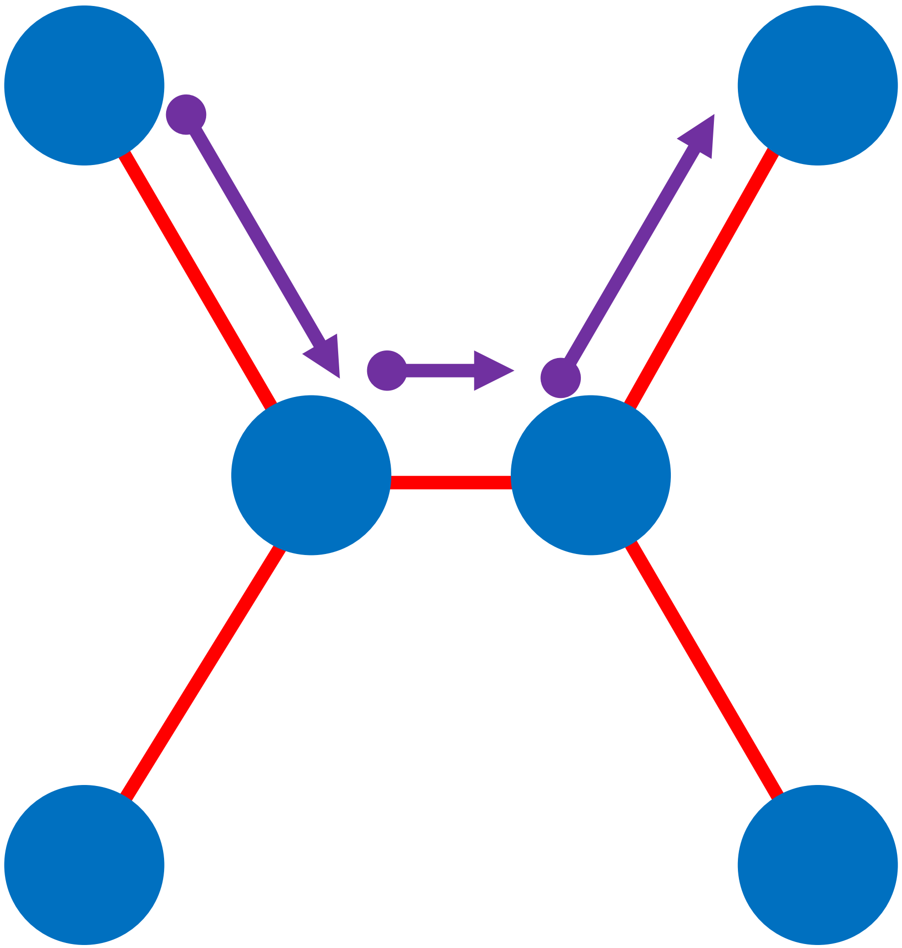
$$D_{ev} := \begin{cases} -1 & \text{if } E_{e1} = v \\ 1 & \text{if } E_{e2} = v \\ 0 & \text{otherwise} \end{cases}$$



Orient edges arbitrarily

Beckmann Formulation

Better scaling for sparse graphs!



$$\begin{aligned} \min_T \quad & \textstyle\sum_e c_e |J_e| \\ \text{s.t.} \quad & D^\top J = \underbrace{p_1 - p_0}_{f} \end{aligned}$$

In computer science:

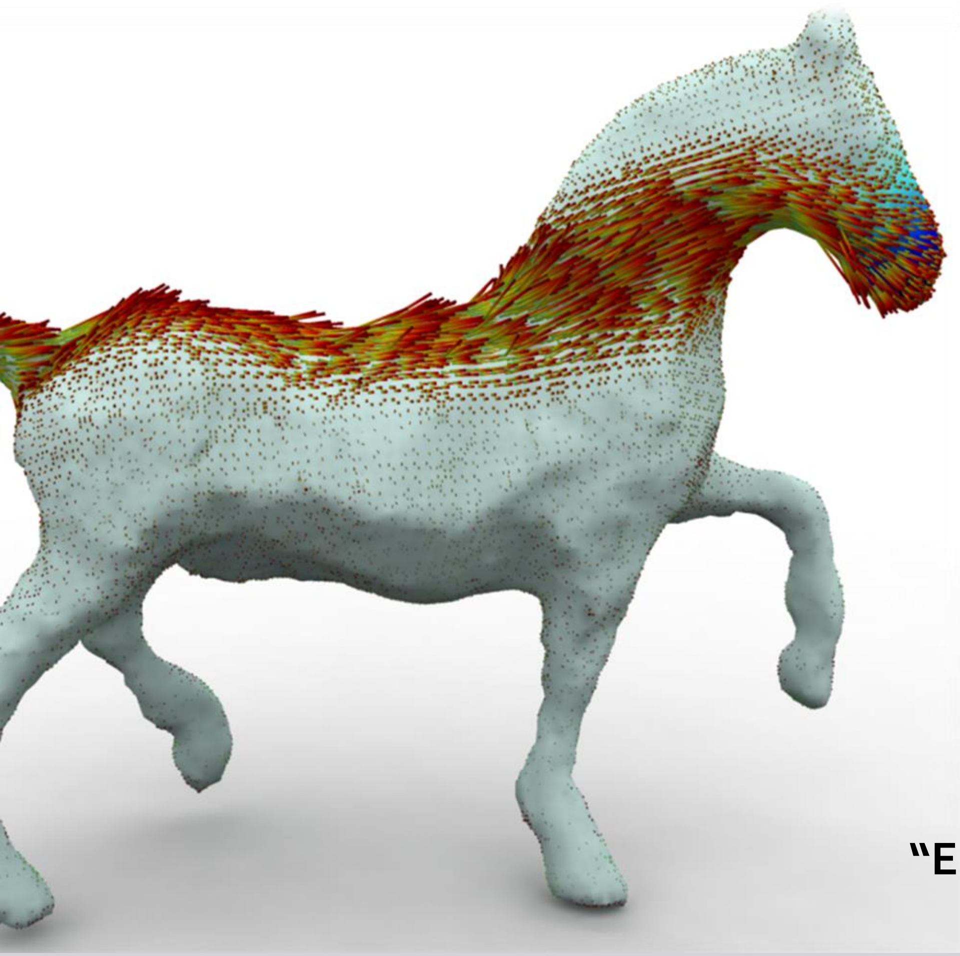
Network flow problem

What Happened?

We used the structure of *D*.

$$\begin{aligned} \min_T \quad & \textstyle\sum_{ij} T_{ij} d(x_i, x_j) \\ \text{s.t.} \quad & \textstyle\sum_j T_{ij} = p_i \\ & \textstyle\sum_i T_{ij} = q_j \\ & T \geq 0 \end{aligned}$$

Continuous Analog?



Probabilities *advect* along the surface

"Eulerian"

Application of our vector field lectures!

Solomon, Rustamov, Guibas, and Butscher. "Earth Mover's Distances on Discrete Surfaces." SIGGRAPH 2014

Think of probabilities like a fluid

Alternative Formulation for W_1

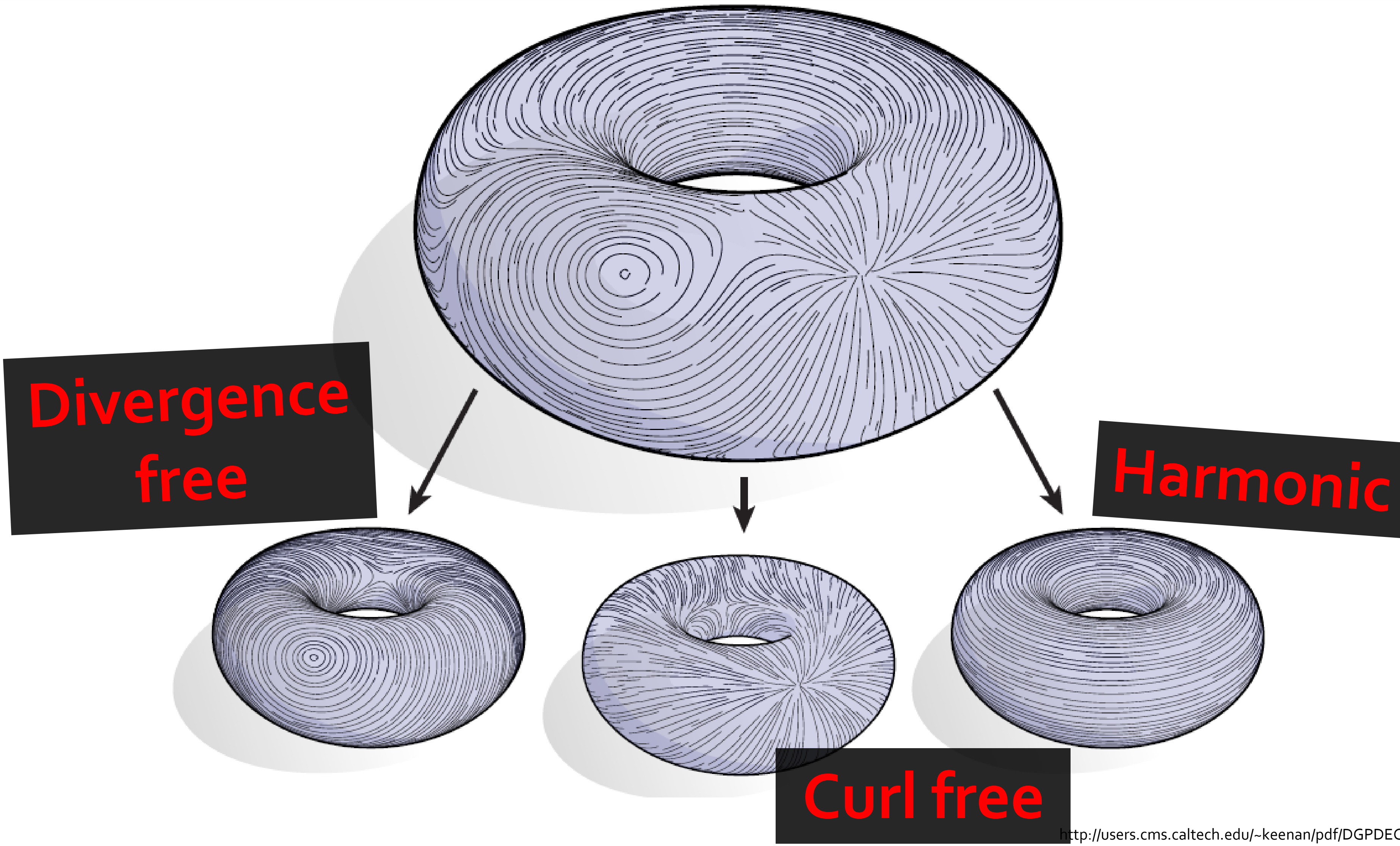
"Beckmann problem"

$$\mathcal{W}_1(\rho_0,\rho_1)=\begin{cases}\overbrace{\inf_J \int_M \|J(x)\|\,dx}^{\text{Total work}} \\ \text{s.t. } \underbrace{\begin{array}{l}\nabla\cdot J(x)=\rho_1(x)-\rho_0(x)\\ J(x)\cdot n(x)=0\ \forall x\in\partial M\end{array}}_{\text{Advects from } \rho_0 \text{ to } \rho_1}\end{cases}$$

Scales linearly

Recall:

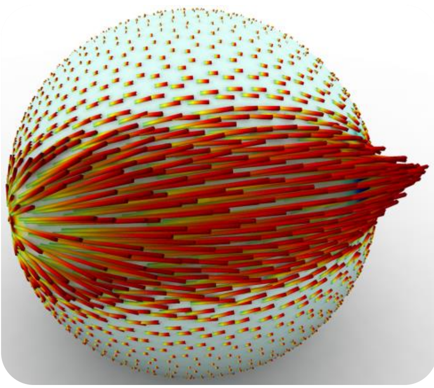
Helmholtz-Hodge Decomposition



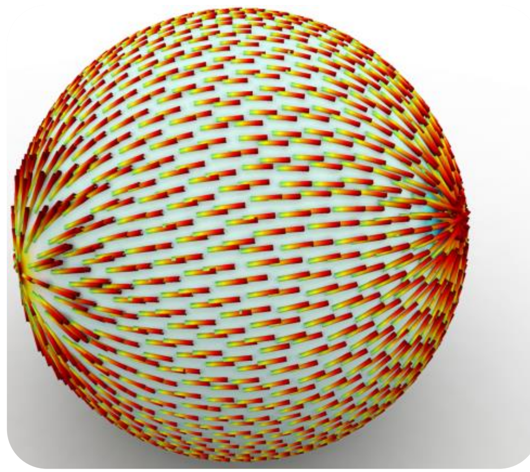
http://users.cms.caltech.edu/~keenan/pdf/DGPDEC.pdf

Hodge Decomposition of *J*

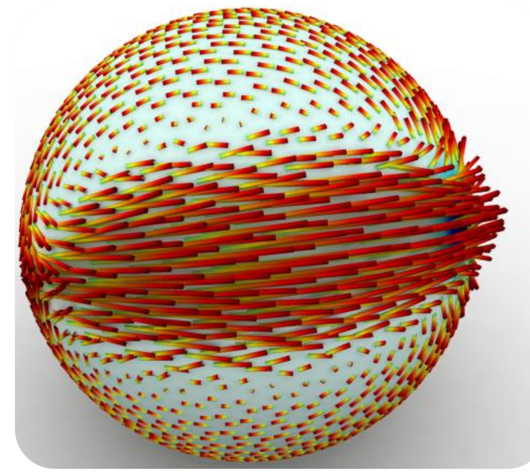
$$J(x) = \nabla f(x) + \mathcal{R}\nabla g(x)$$



Ignoring harmonic part!



Curl-free



Div-free

$$\nabla \cdot J = \Delta f = \rho_1 - \rho_0$$

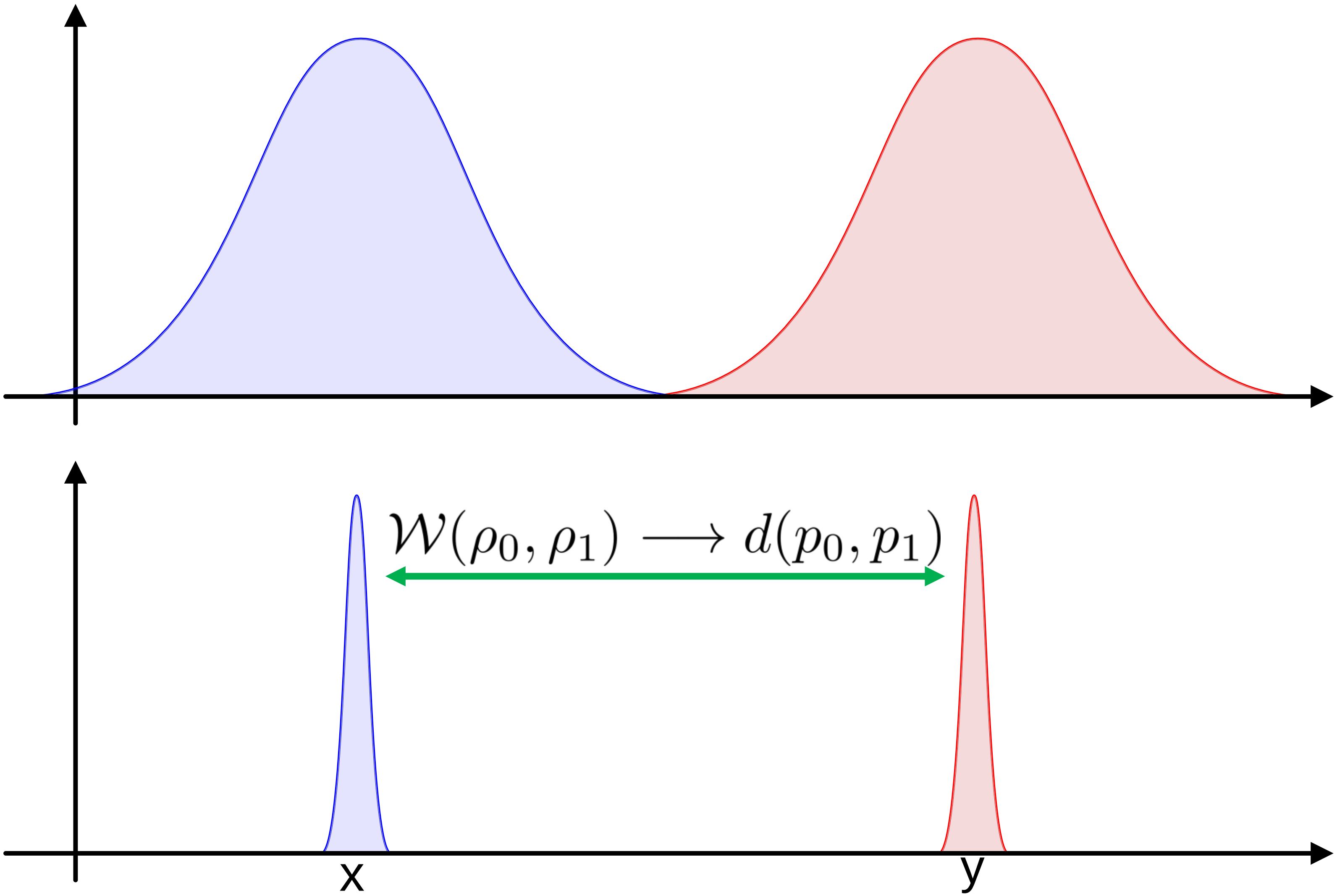
Fast Optimization

1. $\Delta f = \rho_1 - \rho_0$ **Sparse SPD linear solve for *f***

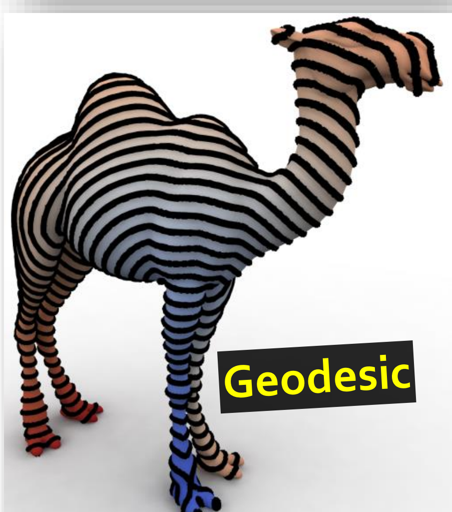
2. $$\inf_g \int_M \|\nabla f(x) + \mathcal{R} \cdot \nabla g(x)\| \, dx$$

Unconstrained and convex optimization for *g*

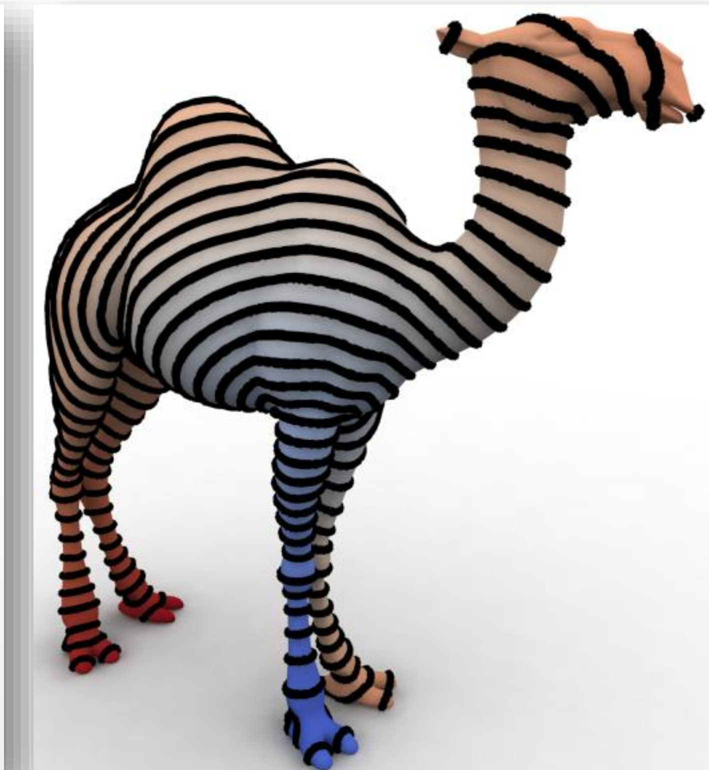
Pointwise Distance



Pointwise Distance



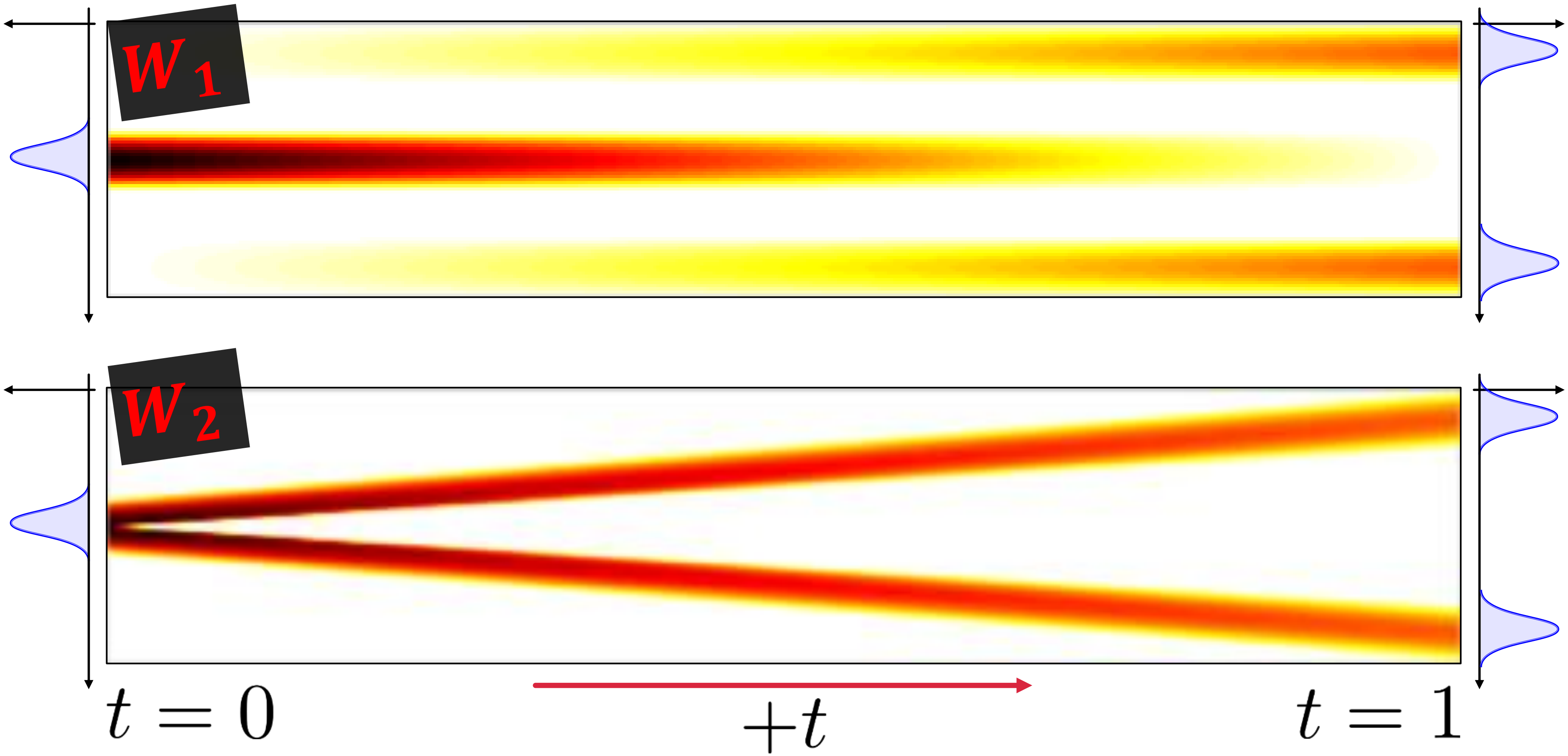
0 eigenfunctions



100 eigenfunctions

Proposition: Satisfies triangle inequality.

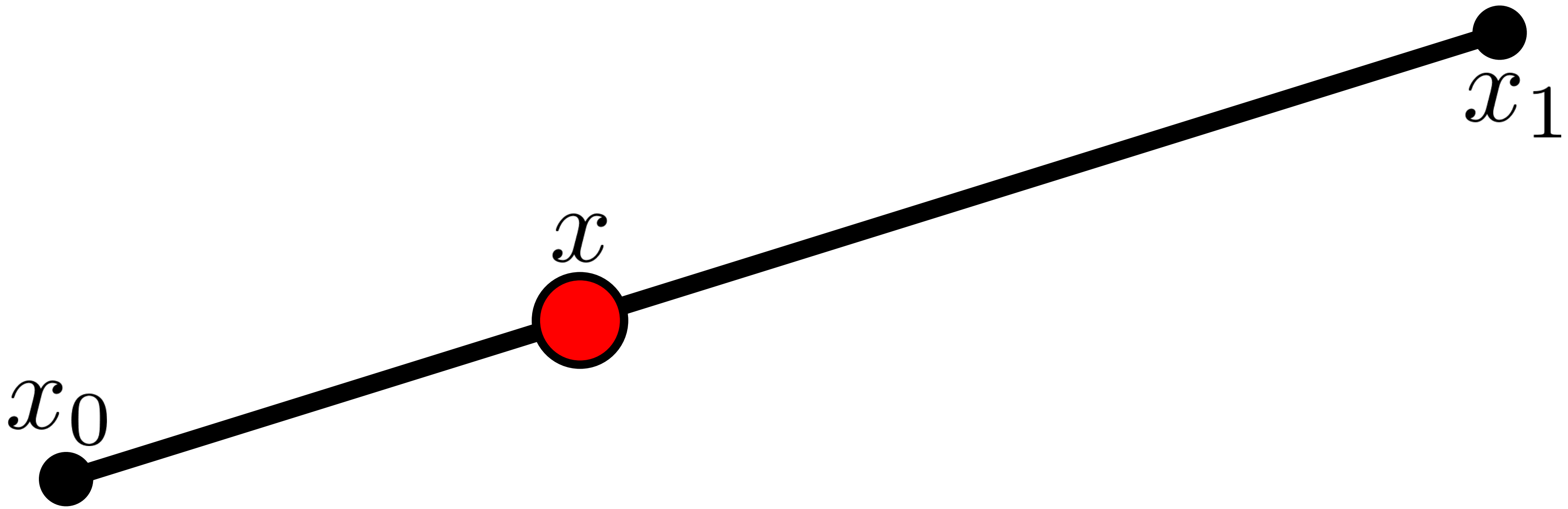
What's the Catch?



McCann. "A Convexity Principle for Interacting Gases." Advances in Mathematics 128 (1997).

No "displacement interpolation"

What Goes Wrong: Median Problems

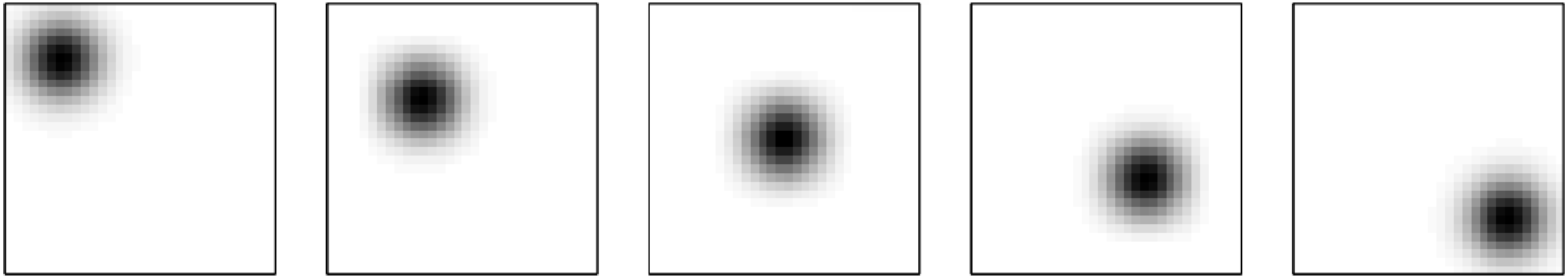


$$\min_{x \in \mathbb{R}^2} \left(\|x - x_0\|_2 + \|x - x_1\|_2 \right)$$

W_1 ineffective for averaging tasks

Displacement Interpolation

$$\mathcal{W}_2$$



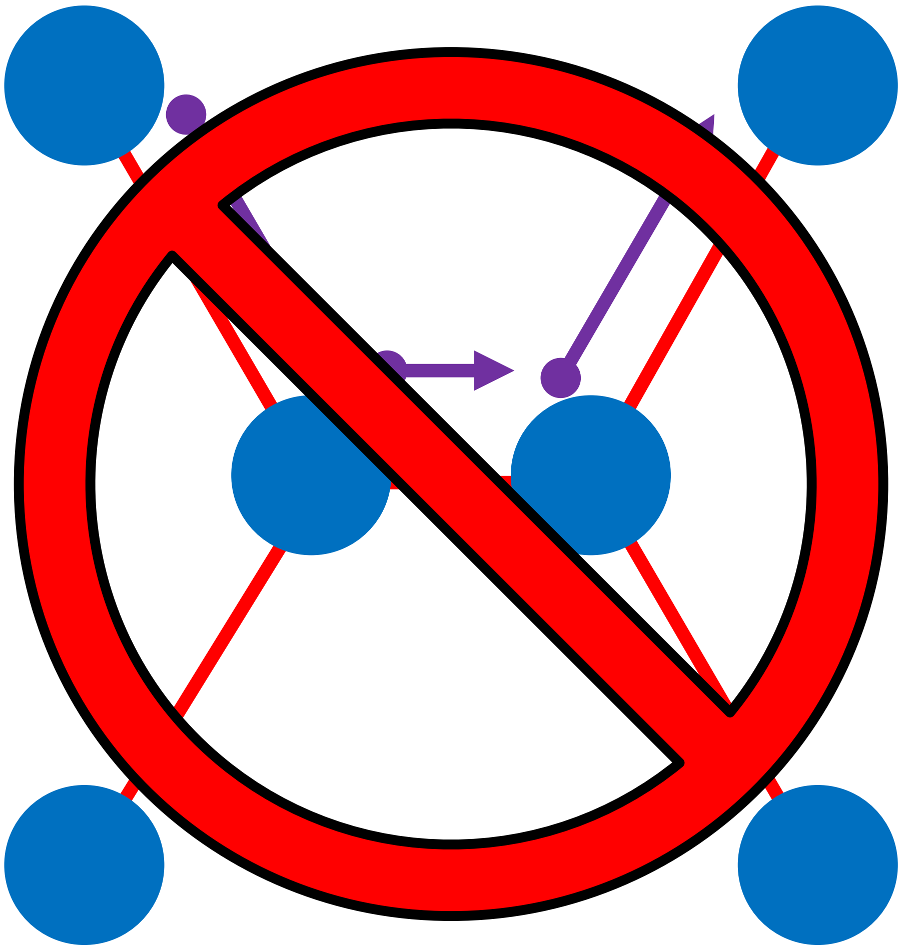
$t=0$ $\quad$ $t=1/4$ $\quad$ $t=1/2$ $\quad$ $t=3/4$ $\quad$ $t=1$

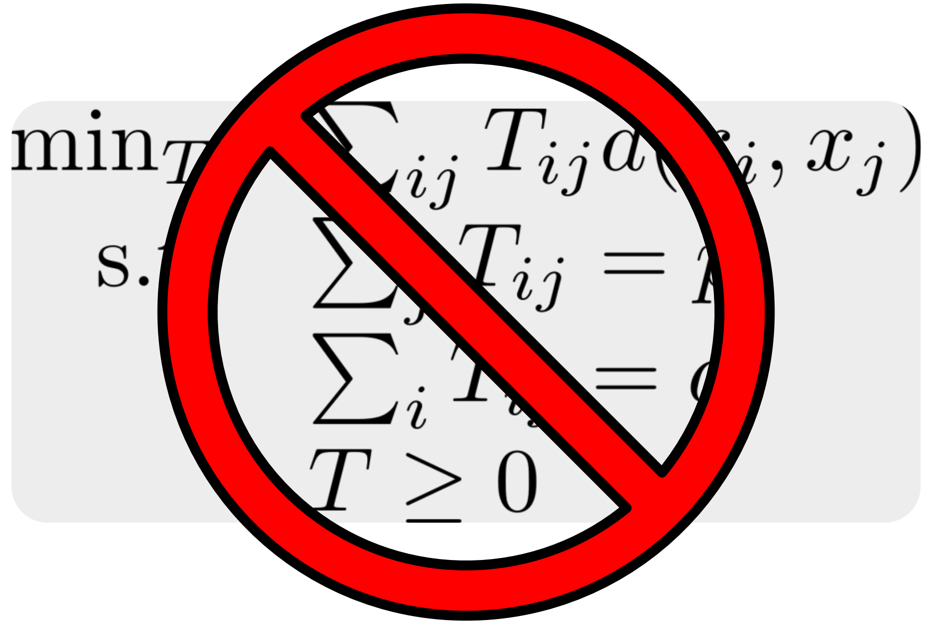
"Explains" shortest path.

Image from "Optimal Transport with Proximal Splitting" (Papadakis, Peyré, and Oudet)

Mass moves along shortest paths

Frustrating Issue





More General Formulation

Monge-Kantorovich Problem

Probability Measure

$$\mu(X) = 1$$

$$\mu(S \subseteq X) \in [0,1]$$

$$\mu\left(\cup_{i \in I} E_i\right) = \sum_{i \in I} \mu(E_i)$$

when E_i disjoint,

I countable

X is the domain

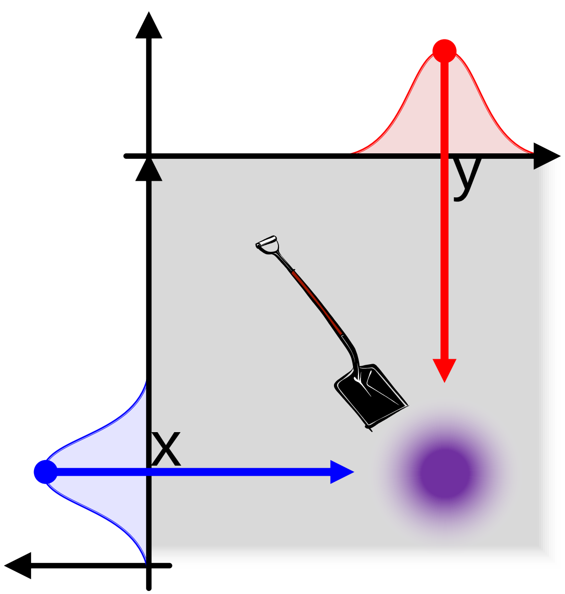
"$\text{Prob}(X)$"

Function from sets to probability

Measure Coupling

$$\mu, \nu \in \mathrm{Prob}(X)$$

$$\Pi(\mu,\nu) := \left\{ \pi \in \mathrm{Prob}(X \times X) : \begin{pmatrix} \pi(U \times X) = \mu(U) \\ \pi(X \times V) = \nu(V) \end{pmatrix} \right\}$$



Analog of transportation matrix

p-Wasserstein Distance

$$\mathcal{W}_p(\mu,\nu) \equiv \min_{\pi\in\Pi(\mu,\nu)} \left(\iint_{X\times X} d(x,y)^p \, d\pi(x,y) \right)^{1/p}$$

Shortest path distance

Expectation

General cost:
"Monge-Kantorovich problem"

http://www.sciencedirect.com/science/article/pii/S152407031200029X#

Continuous analog of EMD

Dual Formulation

$$\mathcal{W}_p(\mu,\nu) \equiv \max_{f,g} \left(\int_X f(x)\, d\mu(x) + \int_X g(y)\, d\nu(y) \right)^{1/p}$$

Cost to drop off

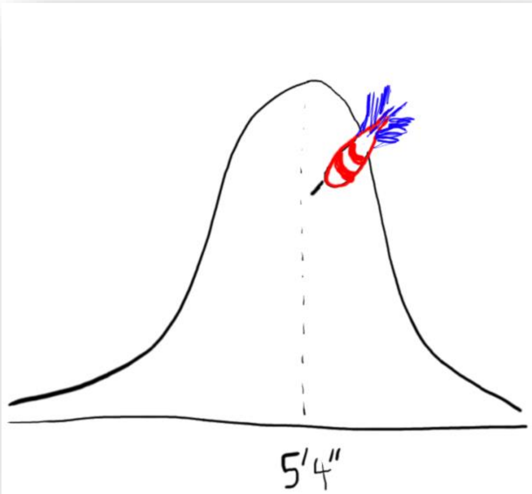
Cost to pick up

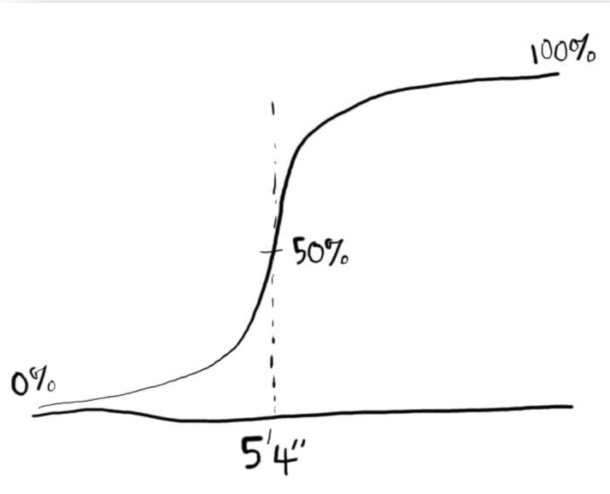
Can't be cheaper to drive

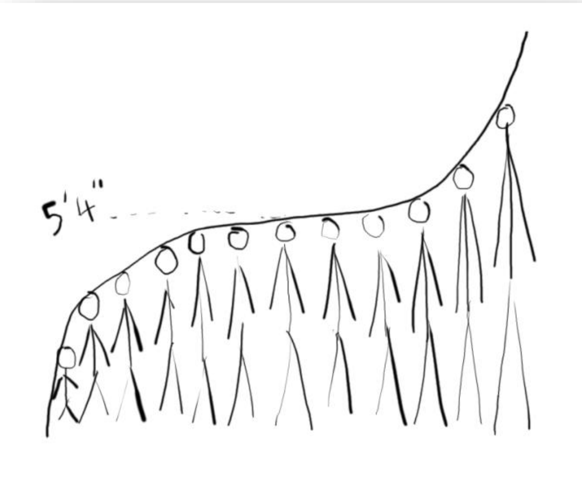
$$f(x) + g(y) \leq d(x,y)^p$$

Continuous analog of EMD

In One Dimension







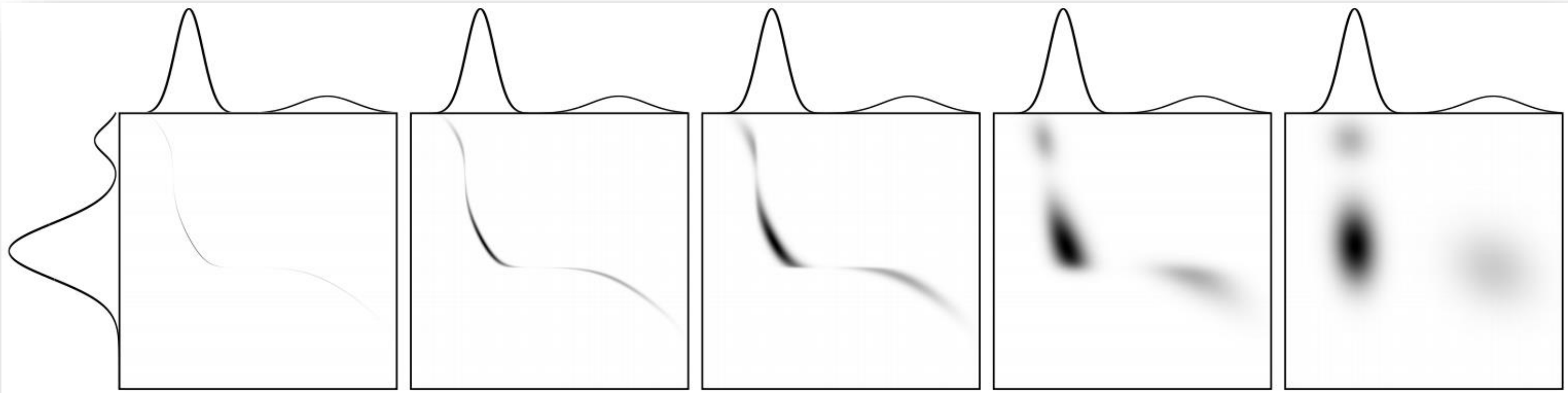
PDF ··········► **[CDF]** ··········► **CDF^{-1}**

$$\mathcal{W}_1(\mu,\nu) = \|\mathrm{CDF}(\mu) - \mathrm{CDF}(\nu)\|_1$$

$$\mathcal{W}_2(\mu,\nu) = \|\mathrm{CDF}^{-1}(\mu) - \mathrm{CDF}^{-1}(\nu)\|_2$$

http://realgl.blogspot.com/2013/01/pdf-cdf-inv-cdf.html

Entropic Regularization



$$\begin{aligned} \min_T \quad & \sum_{ij} T_{ij} d(x_i, x_j) - \gamma H(T) \\ \text{s.t.} \quad & \sum_j T_{ij} = p_i \\ & \sum_i T_{ij} = q_j \\ & T \geq 0 \end{aligned}$$

$$H(T) := -\sum_{ij} T_{ij} \log T_{ij}$$

Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport" (NIPS 2013)

Key Lemma

Prove on the board:

$$T = \operatorname{diag}(u) K \operatorname{diag}(v),$$

$$\text{where } K_{ij} := e^{-D_{ij}/\gamma}$$

$$\begin{aligned} \min_T \quad & \textstyle\sum_{ij} T_{ij} d(x_i, x_j) - \gamma H(T) \\ \text{s.t.} \quad & \textstyle\sum_j T_{ij} = p_i \\ & \textstyle\sum_i T_{ij} = q_j \\ & T \geq 0 \end{aligned}$$

$$H(T) := -\sum_{ij} T_{ij} \log T_{ij}$$

Sinkhorn Algorithm

$$T = \mathrm{diag}(u) K \mathrm{diag}(v),$$

$$\text{where } K_{ij} := e^{-D_{ij}/\gamma}$$

$$u \leftarrow p / Kv$$

$$v \leftarrow q / K^\top u$$

Sinkhorn & Knopp. "Concerning nonnegative matrices and doubly stochastic matrices". Pacific J. Math. 21, 343–348 (1967).

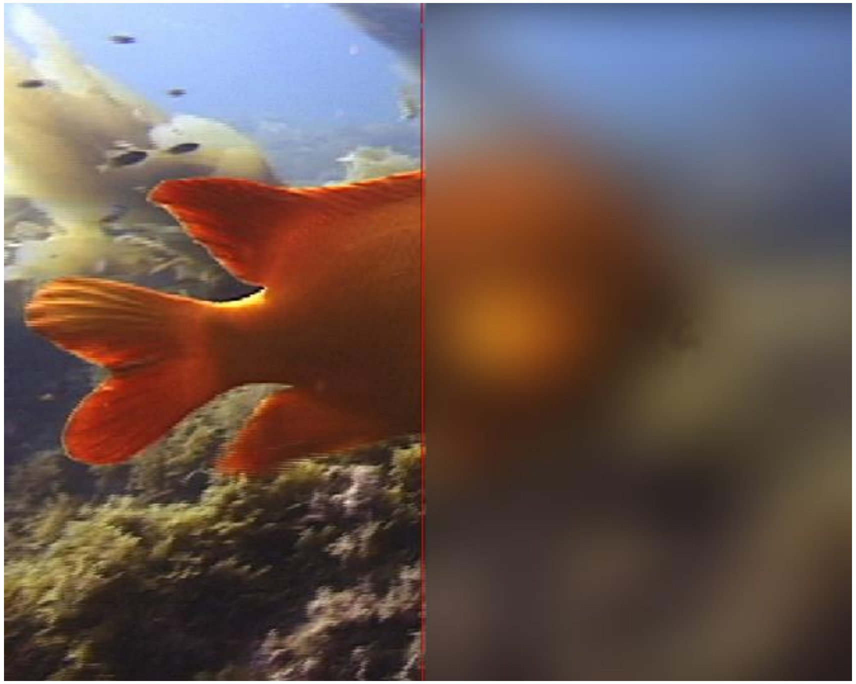
Alternating projection

Ingredients for Sinkhorn

1. Supply vector p
2. Demand vector q
3. Multiplication by K

$$K_{ij} = e^{-D_{ij}/\gamma}$$

On a Grid: Fast *K* Product

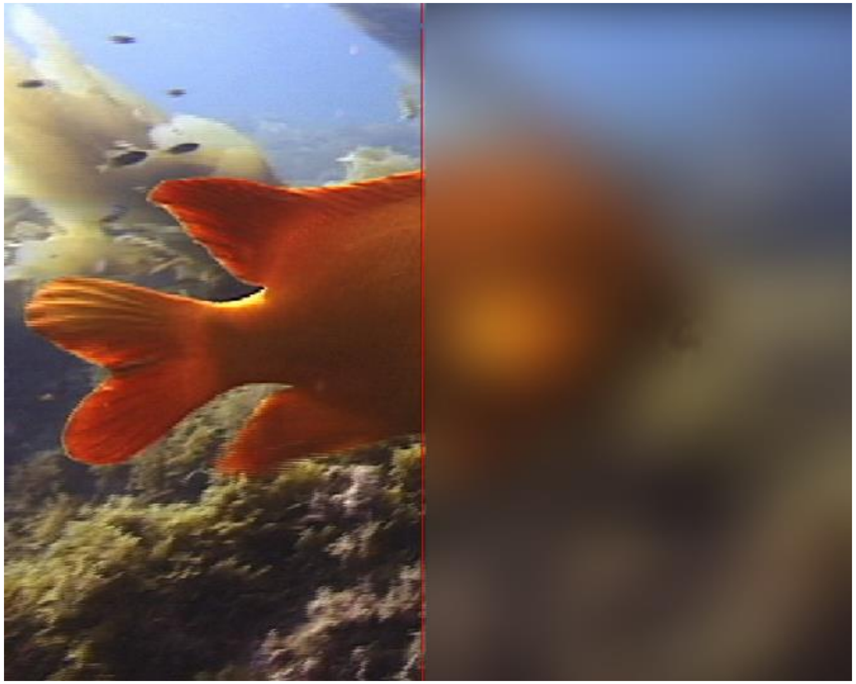


$$(Kv)_{ij} = \sum_{k\ell} g_\sigma(\|(i,j) - (k,\ell)\|_2) v_{k\ell}$$

Fish image from borisfx.com

Gaussian convolution

Sinkhorn on a Grid

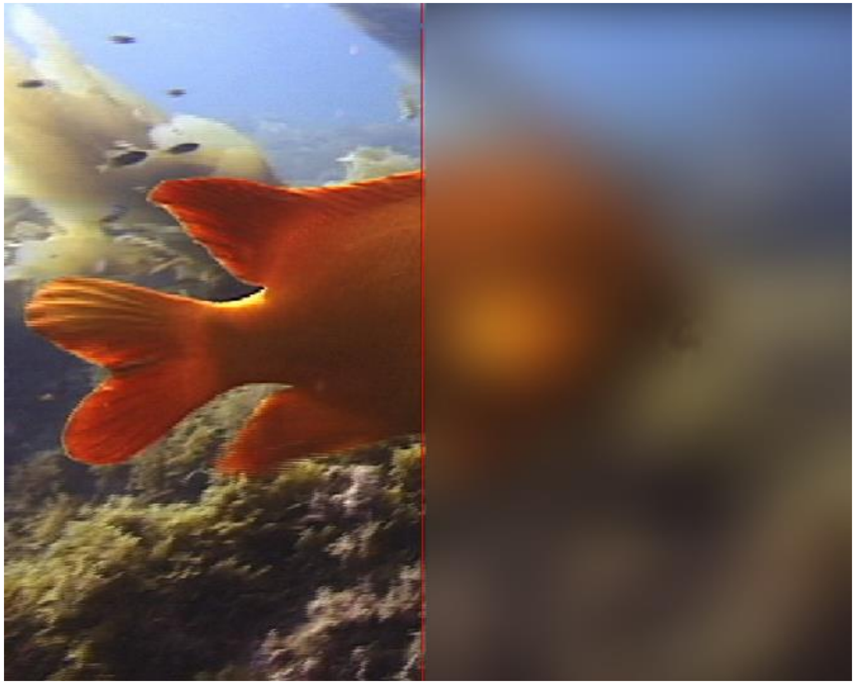


$$u \leftarrow p / K v$$

$$v \leftarrow q / K^\top u$$

No need to store K

Sinkhorn on a Grid



$$u \leftarrow p / K v$$
$$v \leftarrow q / K^\top u$$

What about surfaces?

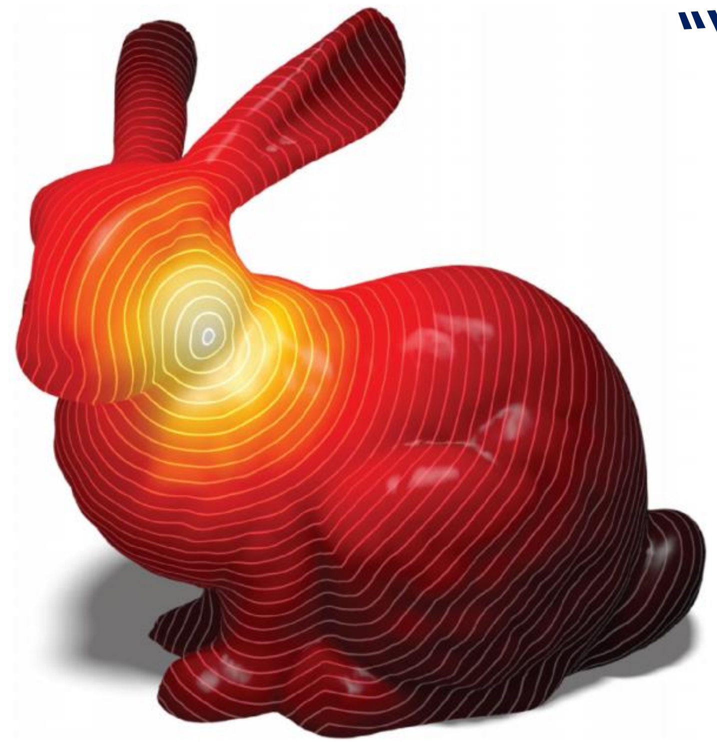
No need to store K

Recall:

Geodesic Distances

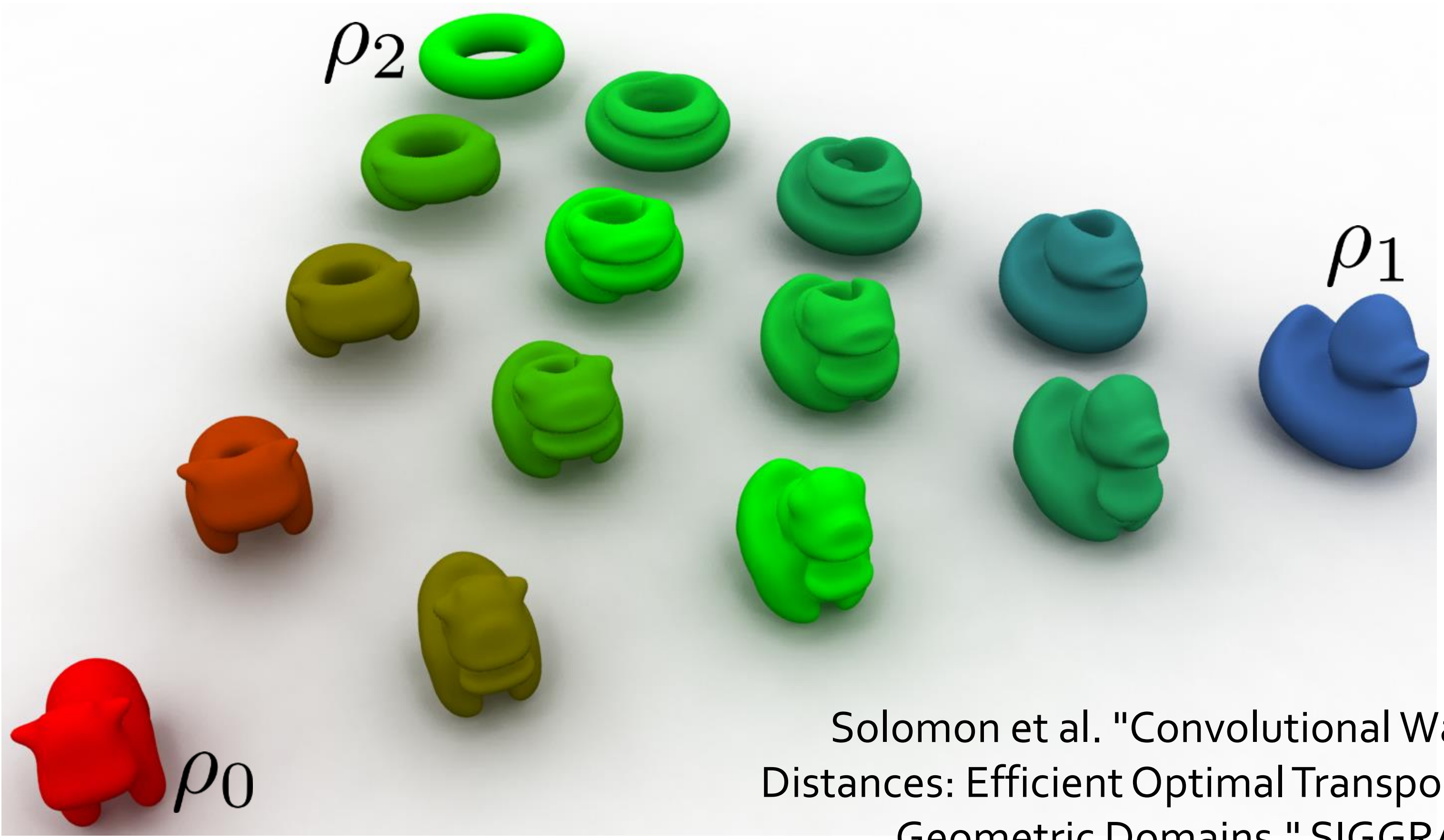
$$d_g(p,q) = \lim_{t\to 0} \sqrt{-4t \log k_{t,p}(q)}$$

"Varadhan's Theorem"



"Geodesics in heat"
Crane, Weischedel, and Wardetzky; TOG 2013

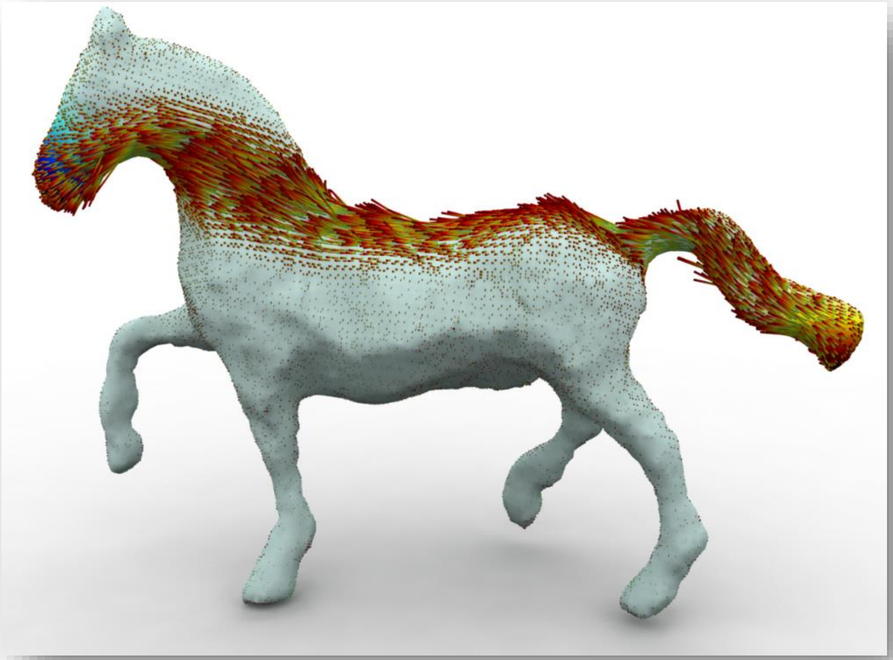
Approximate Sinkhorn



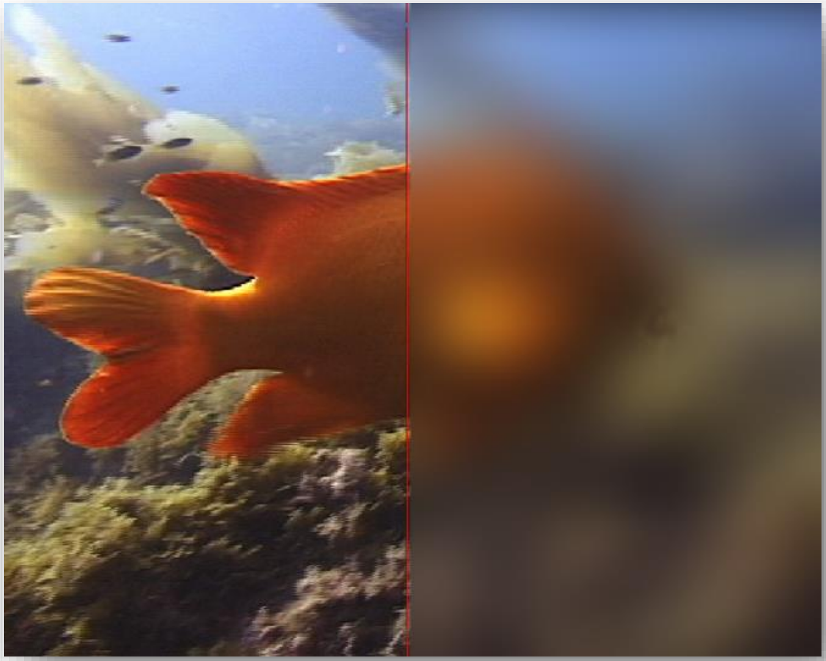
Solomon et al. "Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains." SIGGRAPH 2015.

Replace *K* with heat kernel

Curious Observation



$\mathcal{W}_1$



$\mathcal{W}_2$

Similar problems, different algorithms

Flow-Based W_2

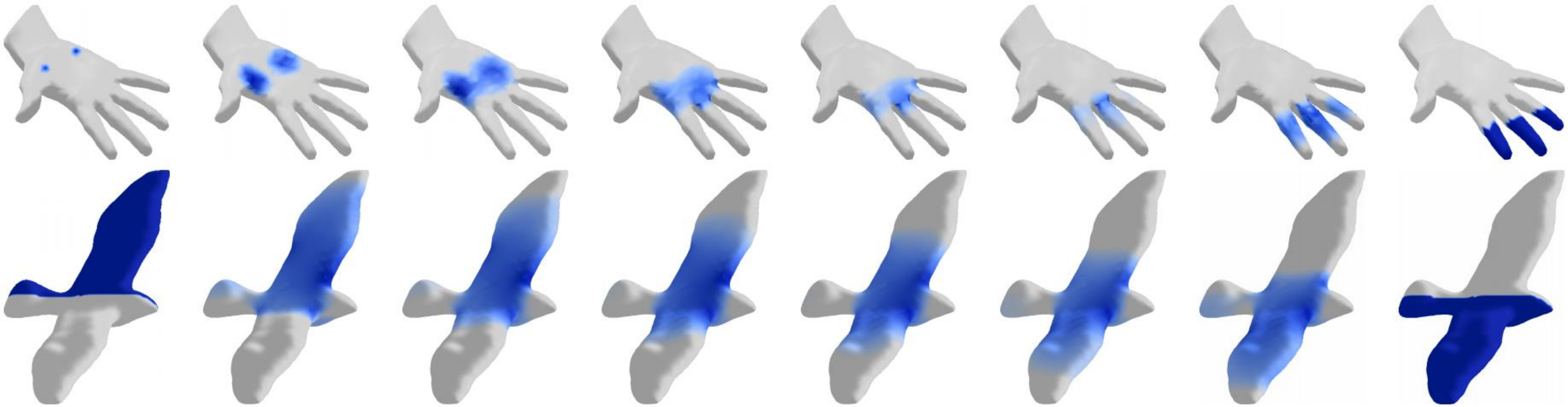
Critical theoretical idea, computationally challenging

$$\mathcal{W}_2^2(\rho_0,\rho_1)=\begin{cases}\inf_{\rho,v}\iint_{M\times[0,1]}\frac{1}{2}\rho(x,t)\|v(x,t)\|^2\,dx\,dt\\ \quad\text{s.t. }\nabla\cdot(\rho(x,t)v(x,t))=\frac{\partial\rho(x,t)}{dt}\\ \qquad v(x,t)\cdot\hat{n}(x)=0\ \forall x\in\partial M\\ \qquad \rho(x,0)=\rho_0(x)\\ \qquad \rho(x,1)=\rho_1(x)\end{cases}$$

Benamou & Brenier
"A computational fluid mechanics solution of the Monge-Kantorovich mass transfer problem"
Numer. Math. 84 (2000), pp. 375-393

Flow-Based W_2

Not impossible!

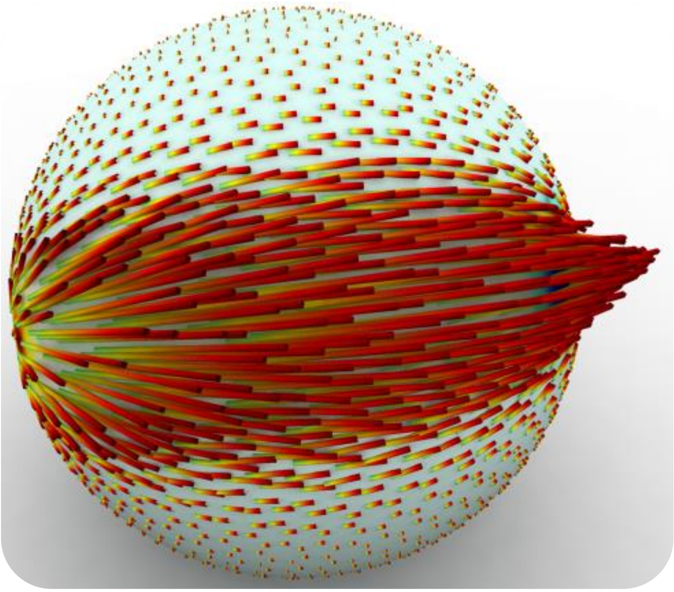


Lavenant et al.
"Dynamical Optimal Transport on Discrete Domains",
SIGGRAPH ASIA 2018

Aside:

Riemannian Structure

$$\langle V, W \rangle_\mu := \int \langle V(x), W(x) \rangle \, d\mu(x)$$



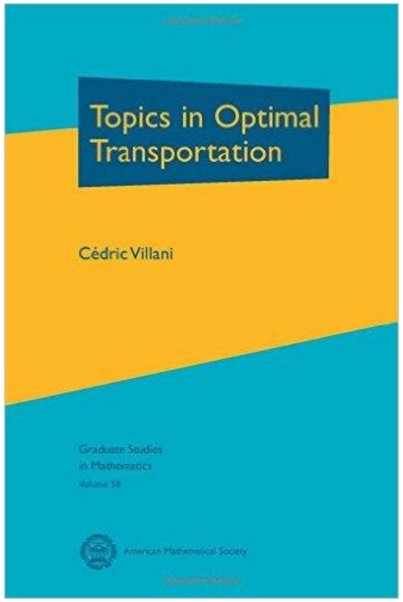
Tangent space/inner product at μ

Aside:

Parallel to Information Geometry

- Consider set of distributions as a **manifold**
- **Tangent spaces** from advection
- **Geodesics** from displacement interpolation

Only Scratching the Surface



Topics in Optimal Transportation
Villani, 2003

Giant field in modern math

Many Other Approaches

Lévy. "A numerical algorithm for L2 semi-discrete optimal transport in 3D." (2014)

Example: Semi-discrete transport

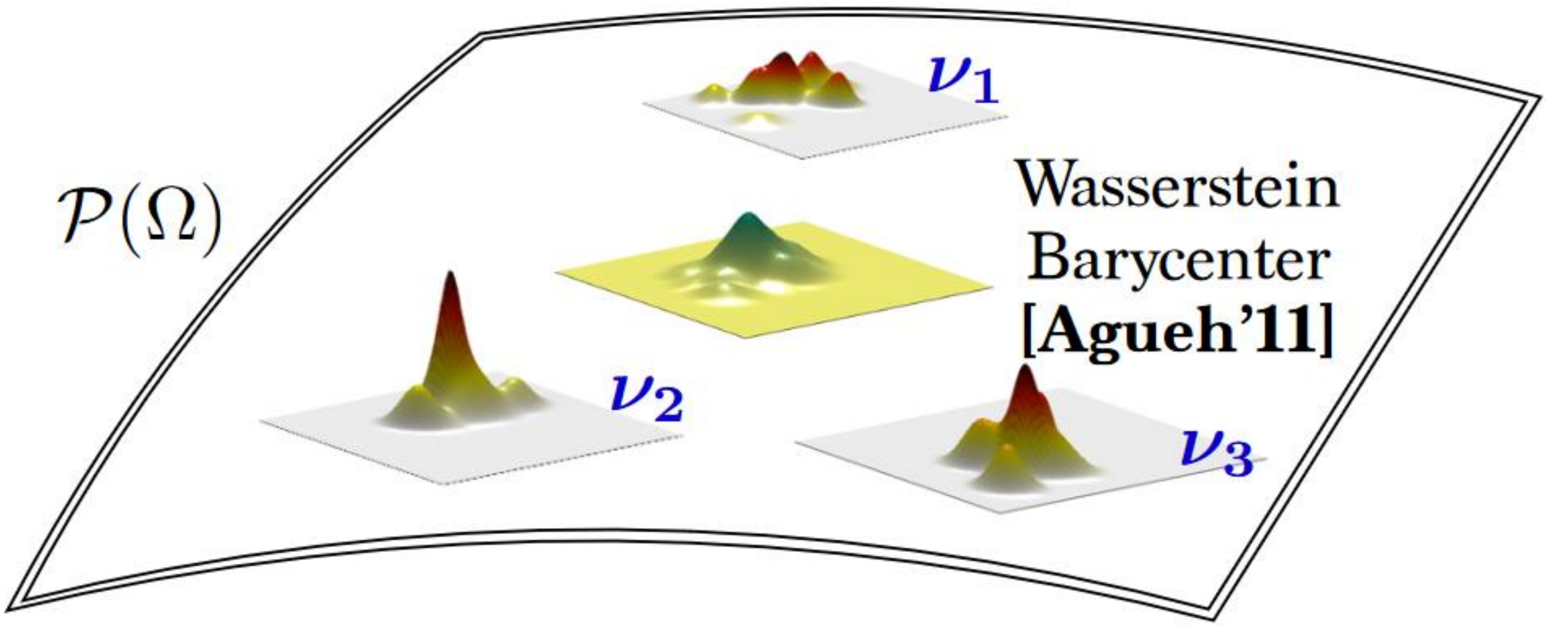
Many Other Approaches

$$\begin{aligned} \min_\phi \quad & \int c(x,\phi(x))\,d\mu(x) \\ \text{s.t.} \quad & \phi_*\mu = \nu \end{aligned}$$

Example: Monge formulation

Derived Problems

$$\min_{\mu \in \mathcal{P}(\Omega)} \sum_{i=1}^{N} \lambda_i W_p^p(\mu, \nu_i)$$



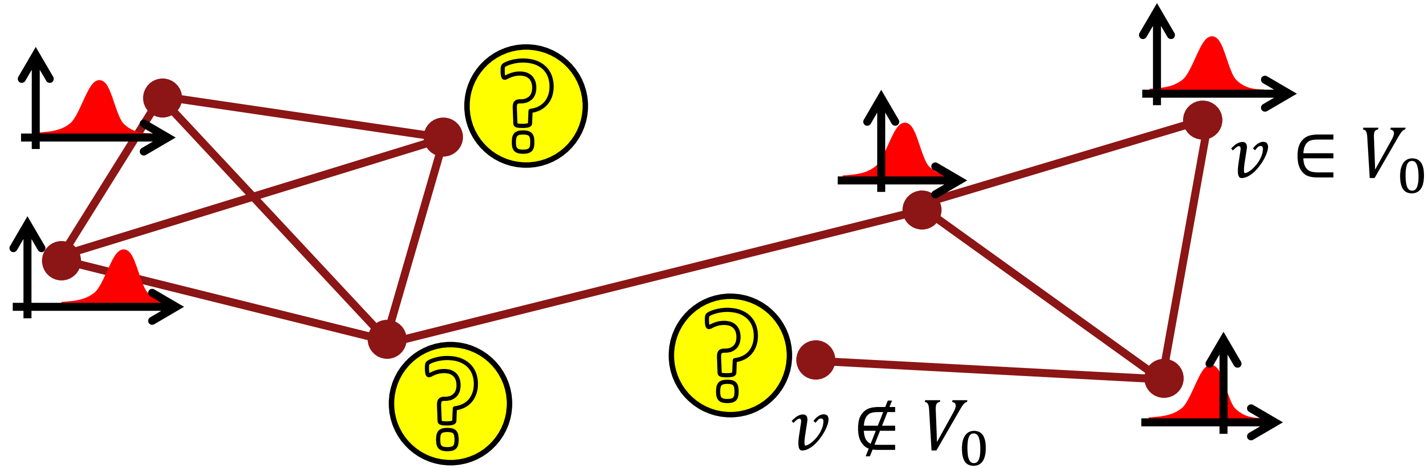
Slide courtesy M. Cuturi

Formula for Applications

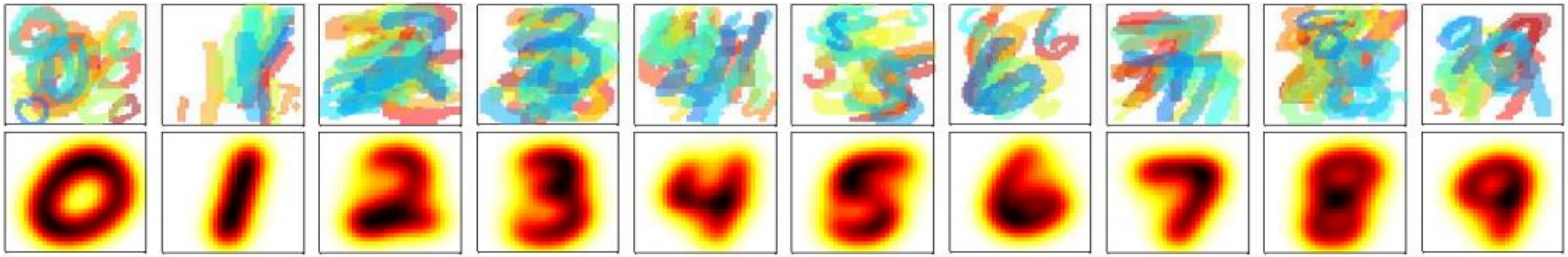
Any (ML) problem involving a **KL** or **L2** loss between (parameterized) histograms or probabilility measures can be easily *Wasserstein-ized* if we can differentiate W efficiently.

Slide courtesy M. Cuturi

Computational Applications



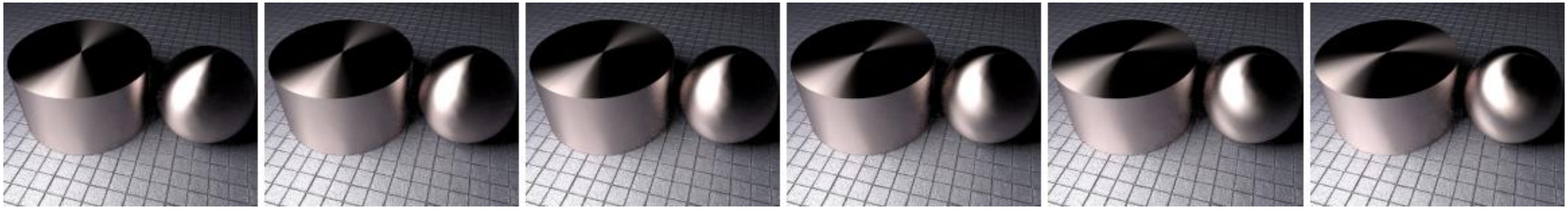
"Wasserstein Propagation for Semi-Supervised Learning" (Solomon et al.)



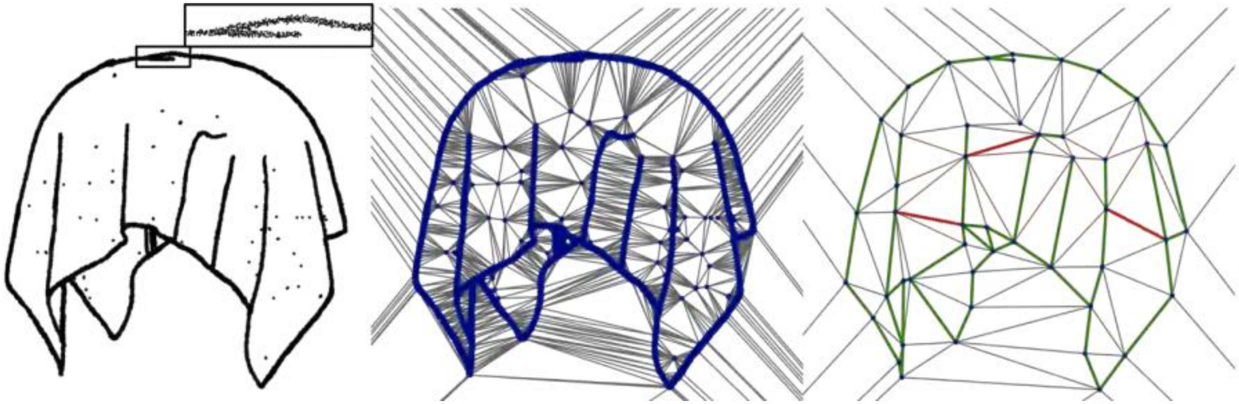
"Fast Computation of Wasserstein Barycenters" (Cuturi and Doucet)

Learning

Computational Applications



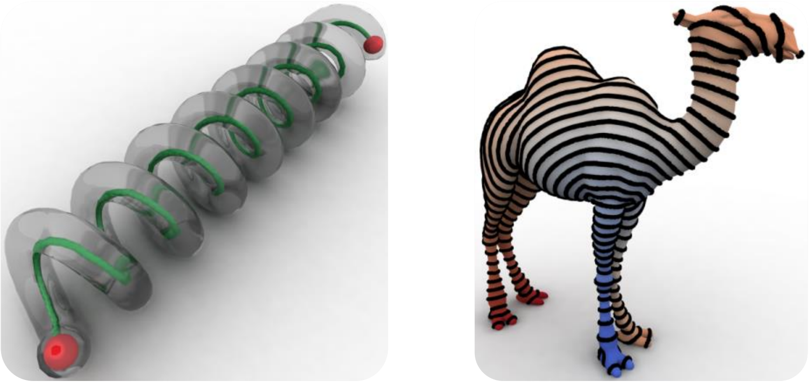
"Displacement Interpolation Using Lagrangian Mass Transport" (Bonneel et al.)



"An Optimal Transport Approach to Robust Reconstruction and Simplification of 2D Shapes" (de Goes et al.)

Morphing and registration

Computational Applications

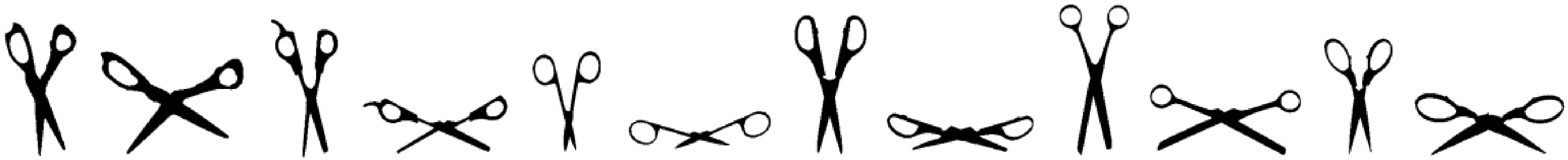


“Earth Mover’s Distances on Discrete Surfaces” (Solomon et al.)

“Blue Noise Through Optimal Transport” (de Goes et al.)

Graphics

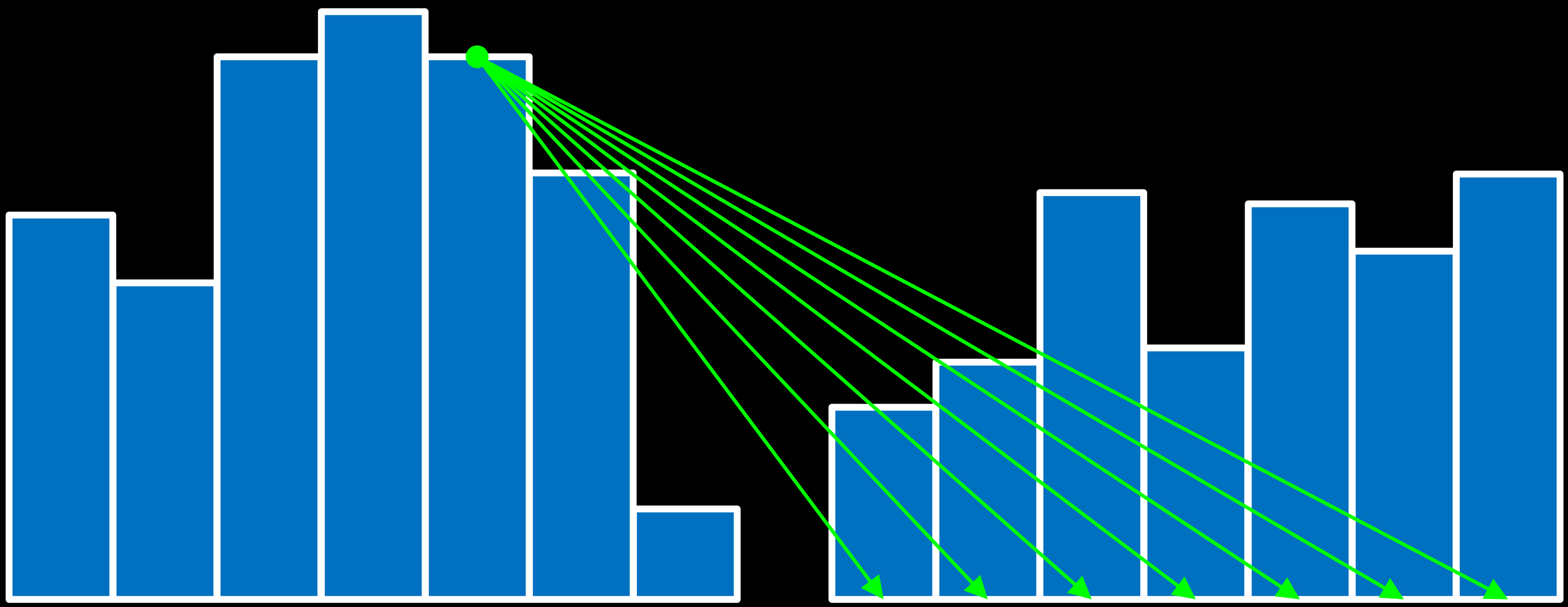
Computational Applications



"Geodesic Shape Retrieval via Optimal Mass Transport" (Rabin, Peyré, and Cohen)

"Adaptive Color Transfer with Relaxed Optimal Transport" (Rabin, Ferradans, and Papadakis)

Vision and image processing



Optimal Transport

Justin Solomon

MIT, Spring 2019