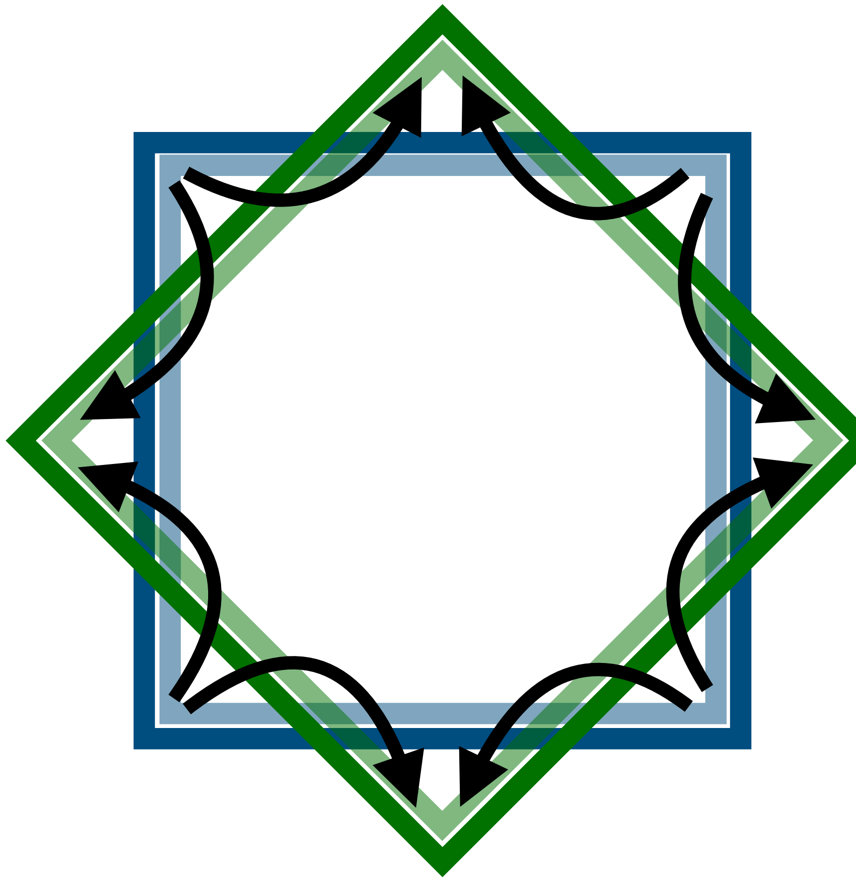


Optimal transport in high-energy physics

April 25, 2023

Tudor Manole
Carnegie Mellon University

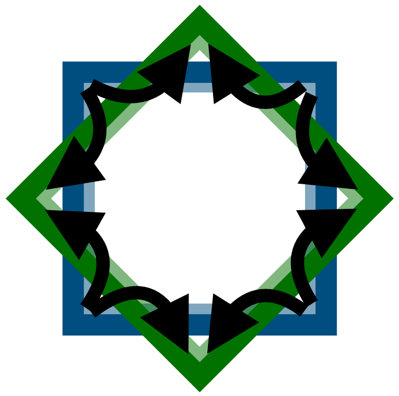
Philipp Windischhofer
University of Chicago

Carnegie Mellon University

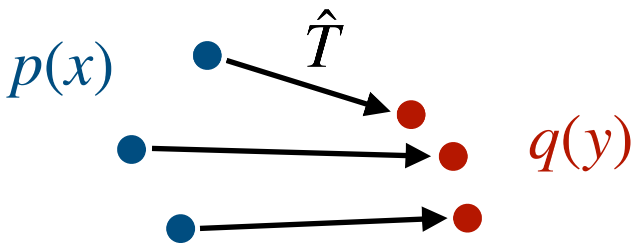
THE UNIVERSITY OF CHICAGO

What can you expect?

A *(very)* brief introduction
to the world of optimal transport

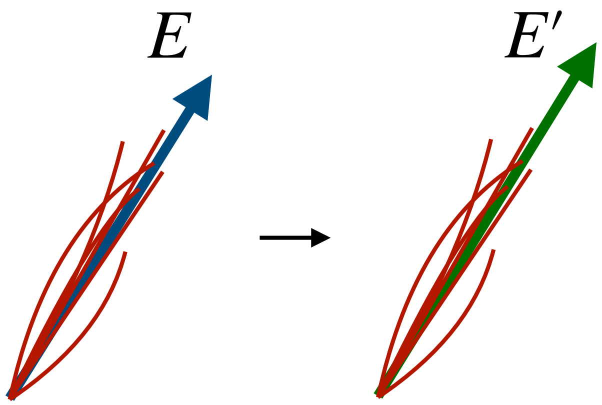


A glimpse at how to solve
optimal transport problems



(Potential) applications in
particle physics

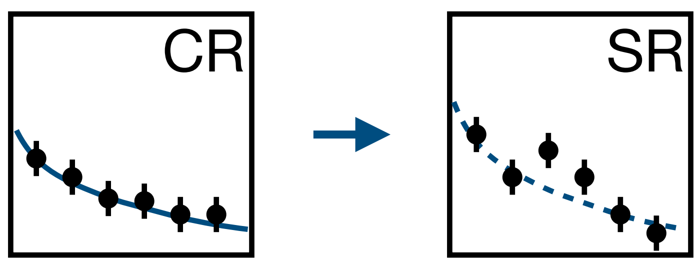
From the perspective of a
statistician *(Tudor)* and a physicist *(Philipp)*



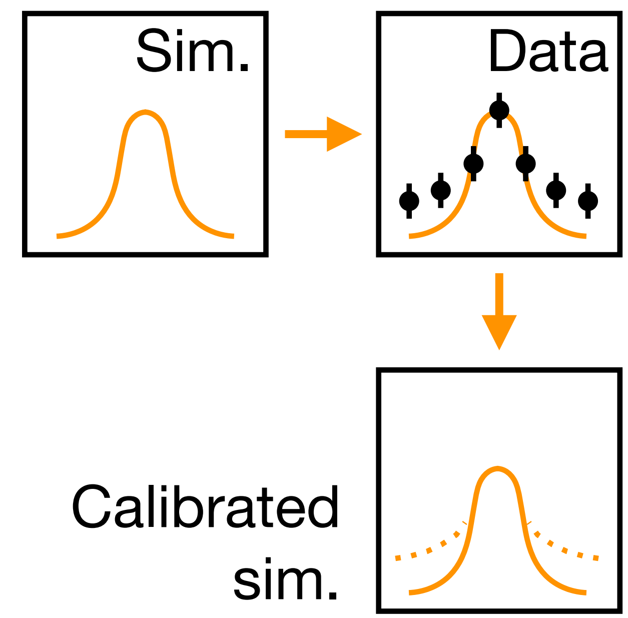
We'll be brief; let's keep the details for the discussion afterwards

Why should you care?

In particle physics, we manipulate (probability) **distributions on a daily basis …**

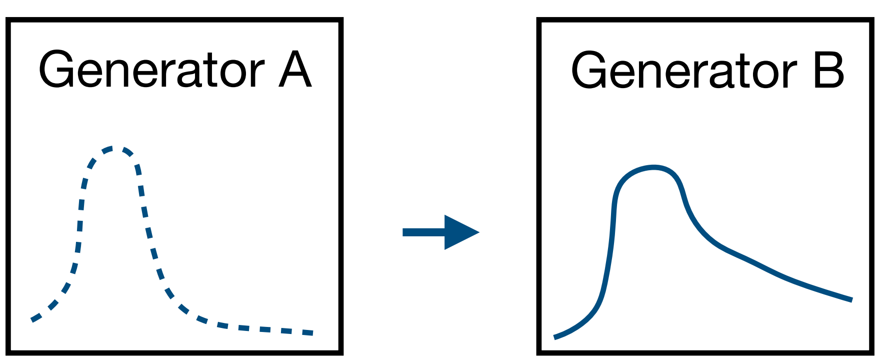


Extrapolation across phase space
(e.g. control region → signal region)



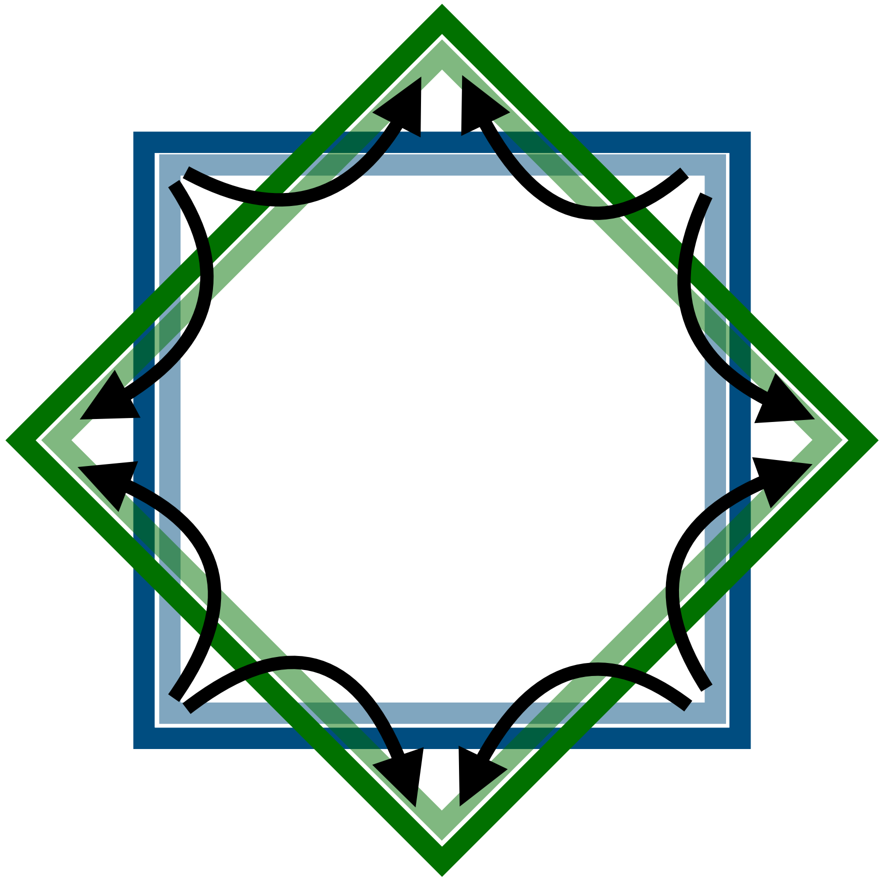
Calibrated sim.

Calibration of simulation
(e.g. Monte Carlo prediction against data side bands)



Template morphing
(e.g. 2-point systematics)

… **optimal transport** provides **useful tools** *(and a unifying perspective)* for many of these!



The theory of optimal transport

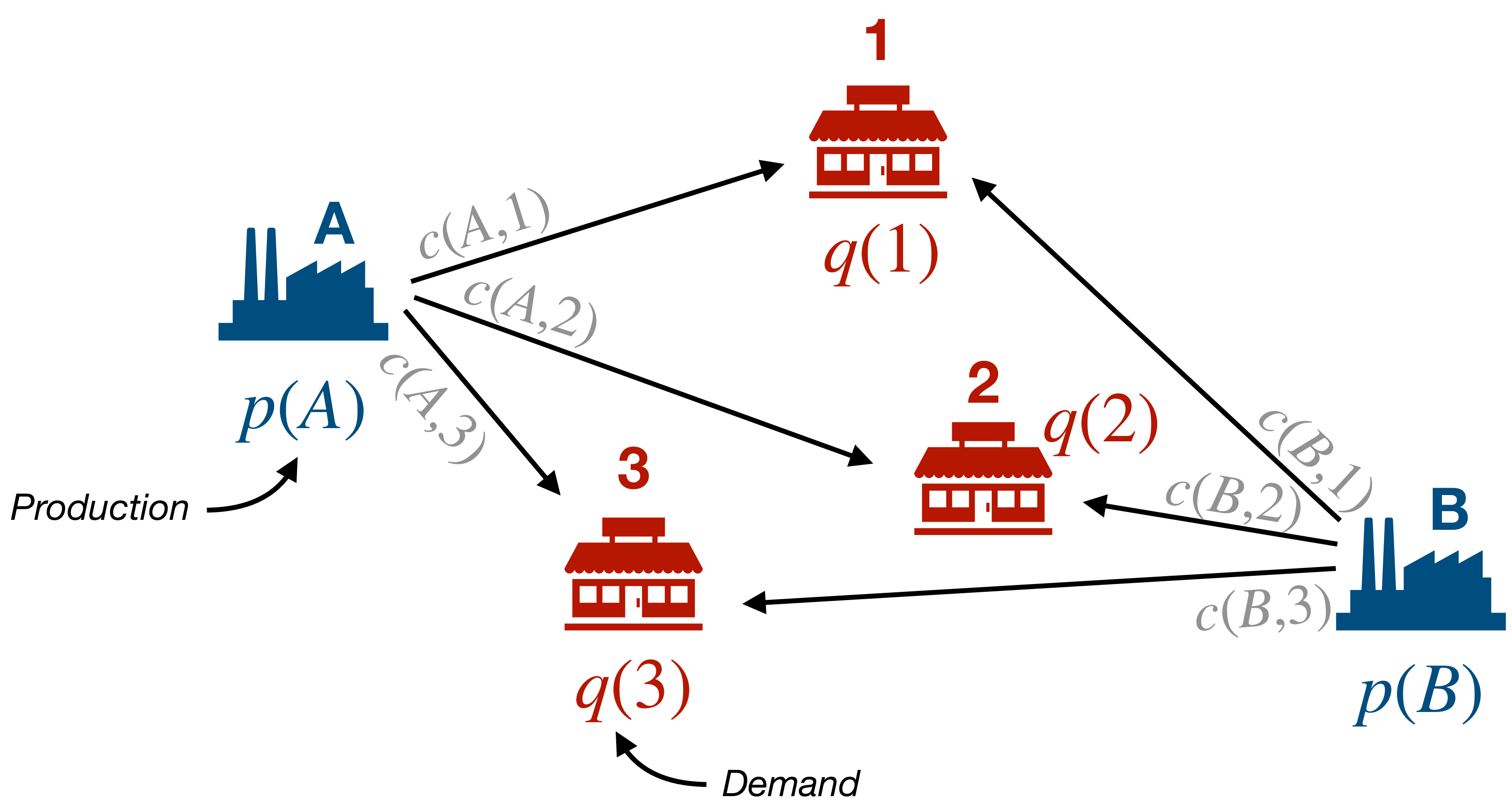
What is optimal transport?

The answer to a logistics problem!

"How to transport commodities from N ***factories*** *to M* ***stores*** *…*

… in the presence of a ***transportation cost*** *$c(a, i)$ between factory a and store i …*

… so that the total cost is minimized?



Incredibly rich mathematical problem with more than 200 years of literature

(Some of it very high-profile, Fields medal-winning work!)

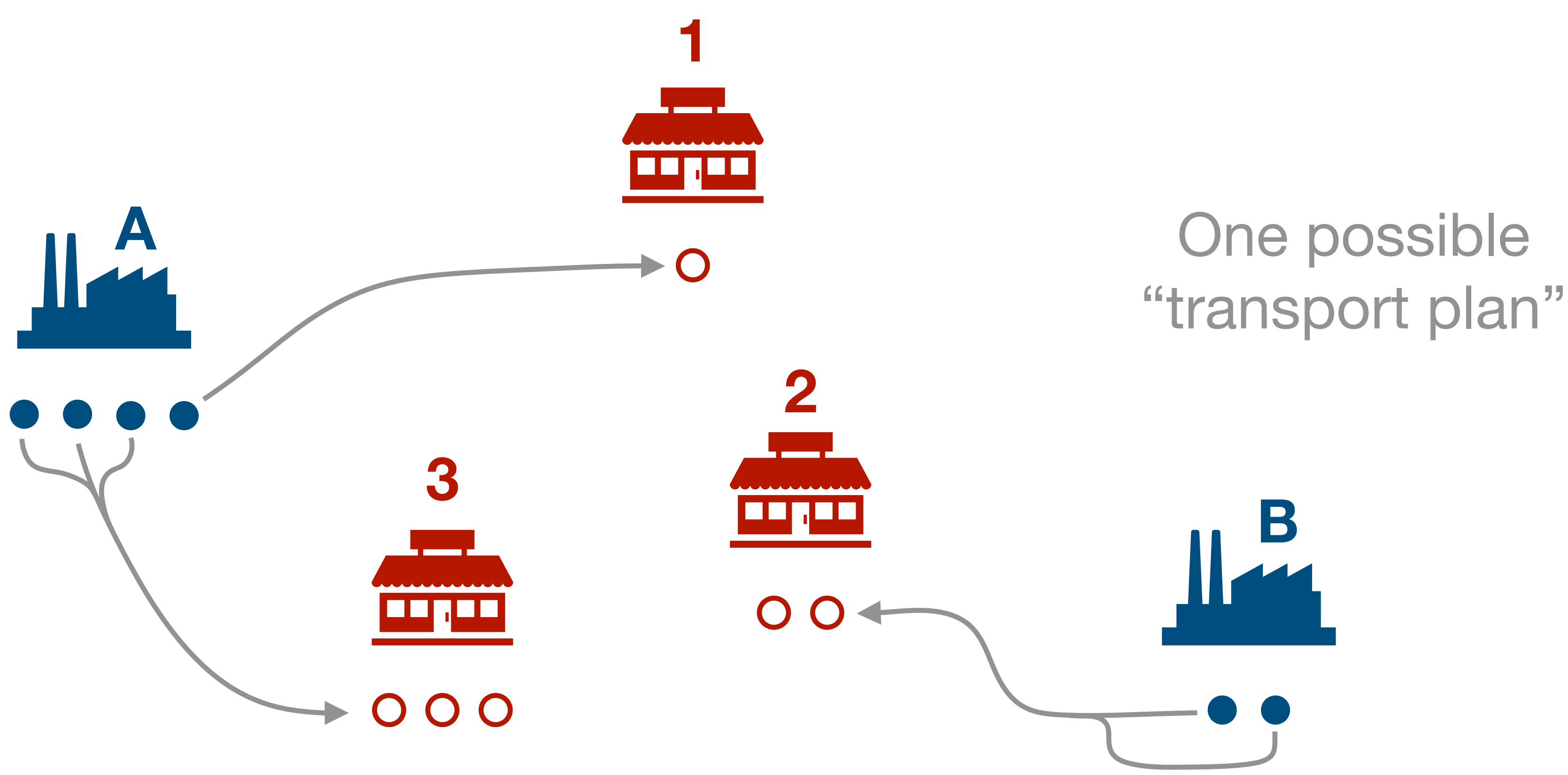
What is optimal transport?

The answer to a logistics problem!

"How to transport commodities from N ***factories*** *to M* ***stores*** *…*

… in the presence of a ***transportation cost*** *$c(a, i)$ between factory a and store i …*

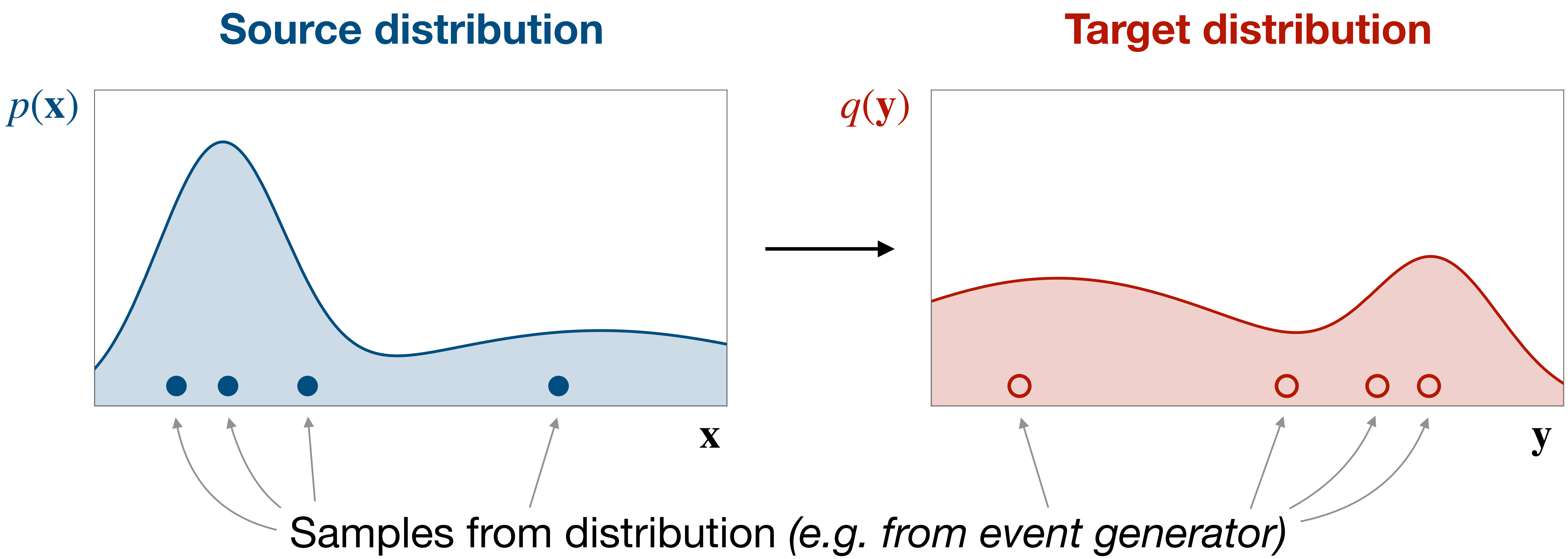
… so that the total cost is minimized?



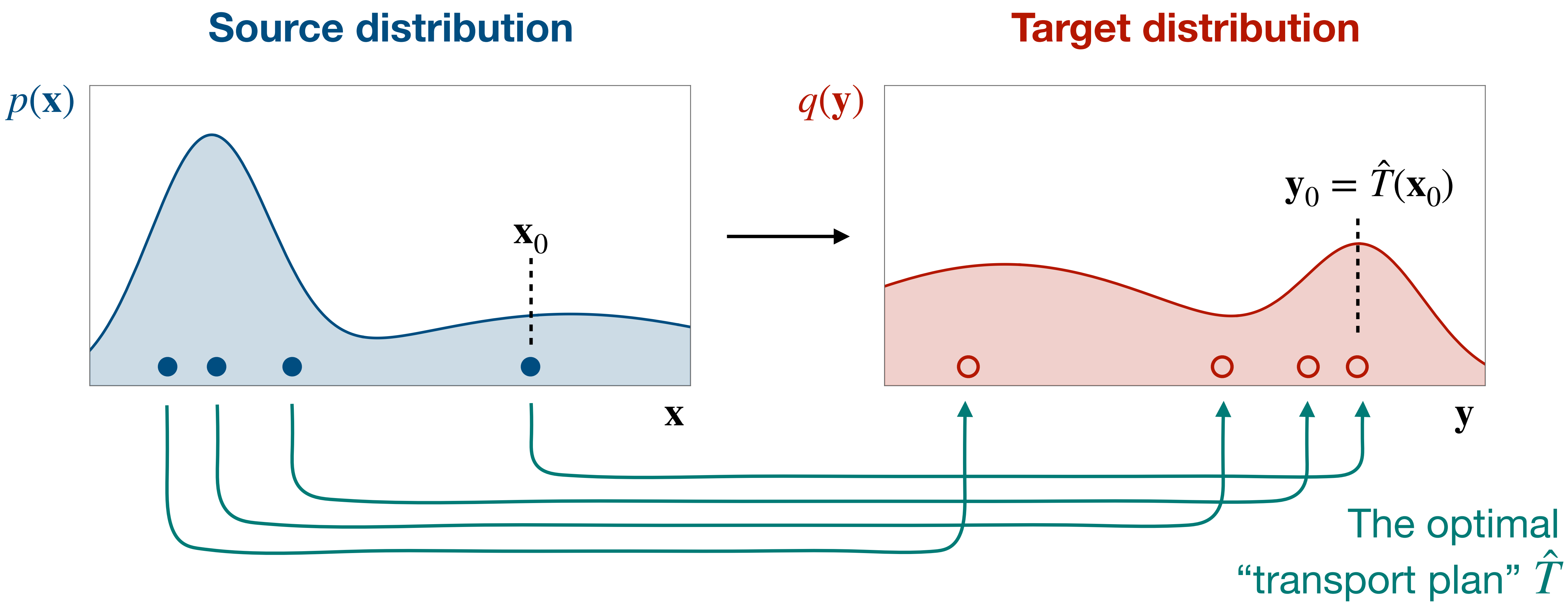
Incredibly rich mathematical problem with more than 200 years of literature

(Some of it very high-profile, Fields medal-winning work!)

Optimal transport, for a particle physicist



Optimal transport, for a particle physicist



"Monge optimal transport problem":

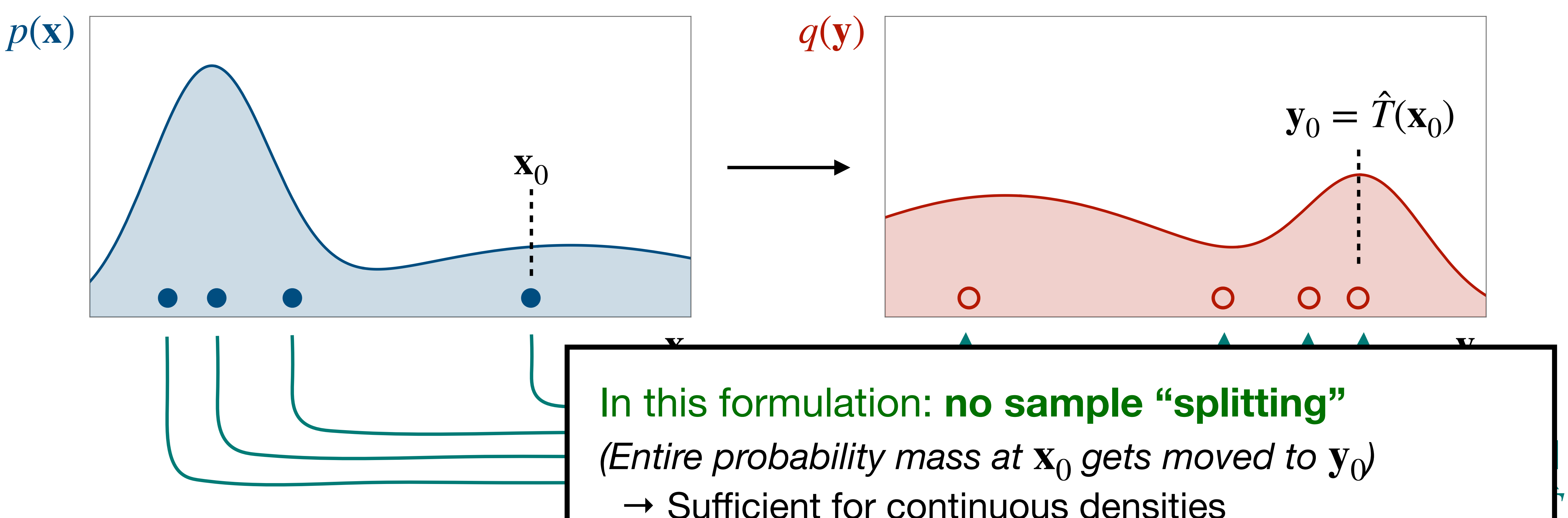
Construct a (continuous) function $\hat{T}$ that maps $p(\mathbf{x})$ into $q(\mathbf{y})$ in an optimal way by "moving" the samples:

Transport cost $c(\mathbf{x}, \mathbf{y})$ for moving sample from $\mathbf{x}$ to $\mathbf{y}$

$$\mathbf{x} \mapsto \mathbf{y} = \hat{T}(\mathbf{x})$$

Such that $q(\mathbf{y}) = p(\mathbf{x})\,(\nabla_{\mathbf{x}}\hat{T})^{-1}$ and $\hat{T} = \arg\min_T \int dx\; p(x)\; c(x, T(x))$

Optimal transport, for a particle physicist

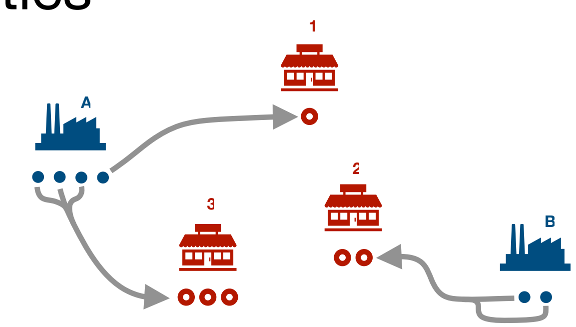


In this formulation: **no sample "splitting"**

(Entire probability mass at $\mathbf{x}_0$ gets moved to $\mathbf{y}_0$)

→ Sufficient for continuous densities

"Kantorovich problem" →



"Monge optimal transport pro

Construct a (continuous) functio

in an optimal way by "moving" t

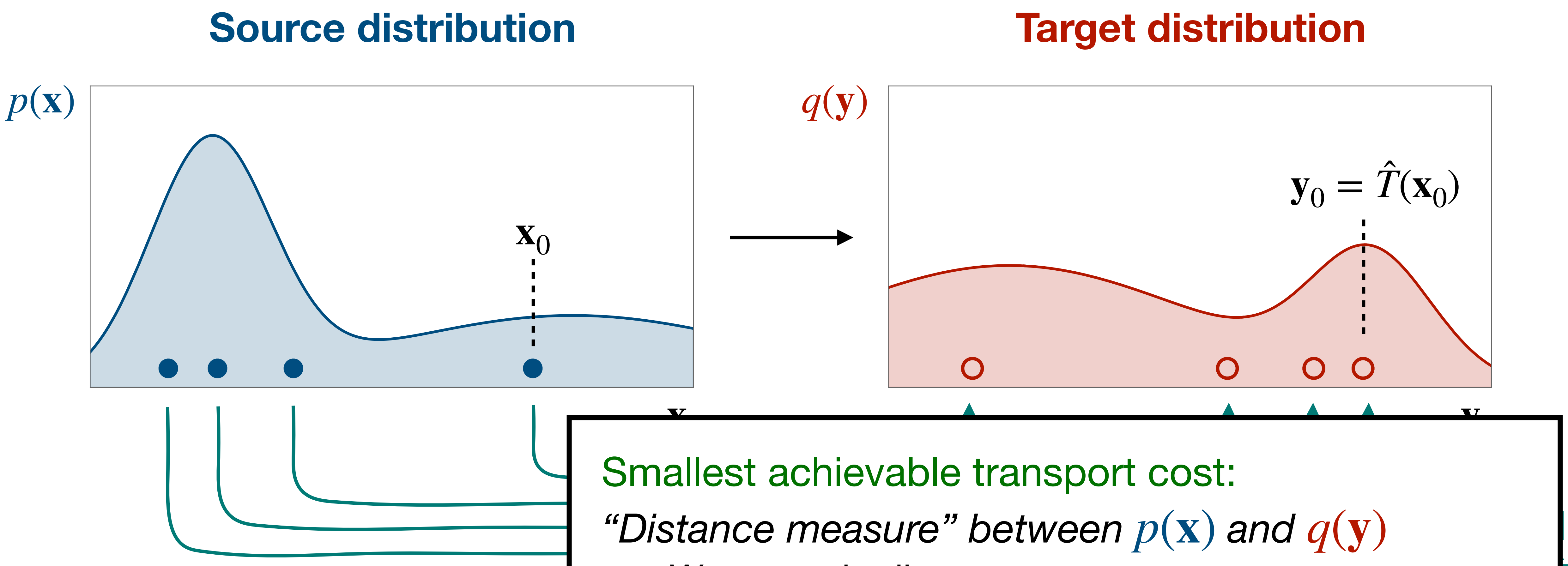
ng

sample from $\mathbf{x}$ to $\mathbf{y}$

$$\mathbf{x} \mapsto \mathbf{y} = \hat{T}(\mathbf{x})$$

Such that $q(\mathbf{y}) = p(\mathbf{x})\,(\nabla_{\mathbf{x}}\hat{T})^{-1}$ and $\hat{T} = \arg\min_T \int dx\; p(x)\; c(x, T(x))$

Optimal transport, for a particle physicist



Smallest achievable transport cost:

“Distance measure” between $p(\mathbf{x})$ *and* $q(\mathbf{y})$

→ Wasserstein distance

$$W = \min_T \int dx\, p(x)\, c(x, T(x))$$

“Monge optimal transport pro

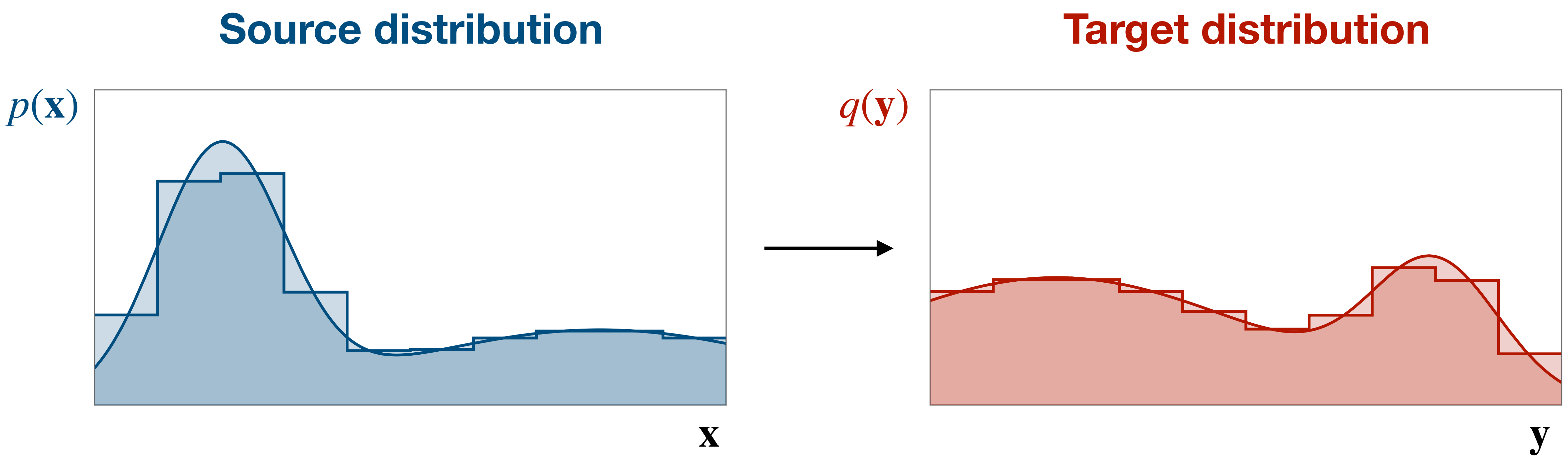
Construct a (continuous) functio
in an optimal way by “moving” t

ng
sample from $\mathbf{x}$ to $\mathbf{y}$

$$\mathbf{x} \mapsto \mathbf{y} = \hat{T}(\mathbf{x})$$

Such that $q(\mathbf{y}) = p(\mathbf{x})\,(\nabla_{\mathbf{x}} \hat{T})^{-1}$ and $\hat{T} = \arg\min_T \int dx\; p(x)\; c(x, T(x))$

Optimal transport, for a particle physicist



Operatively, this procedure gives the same results as

→ Binning $\mathbf{x}$ and $\mathbf{y}$

→ Reweighting bin contents for $\mathbf{x}$ by the density ratio $q(\mathbf{y})/p(\mathbf{x})$

… but is also **well-behaved** where the density ratio gets very large
(Empty bins when densities don't have common support)

→ *Important for applications (see later)*

How to do optimal transport?

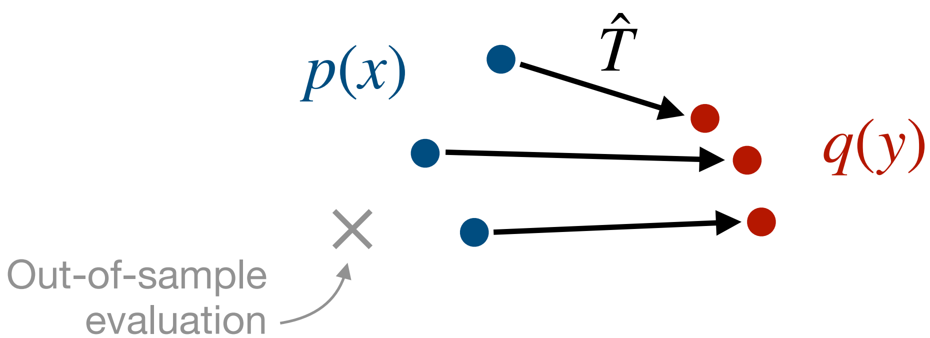
In general, the Monge problem is very difficult to solve!

$$q(\mathbf{y}) = p(\mathbf{x})\,(\nabla_{\mathbf{x}} \hat{T})^{-1}$$

(Highly nonlinear constraint!)

$$\hat{T} = \arg\min_{T} \int dx\; p(x)\; c(x, T(x))$$

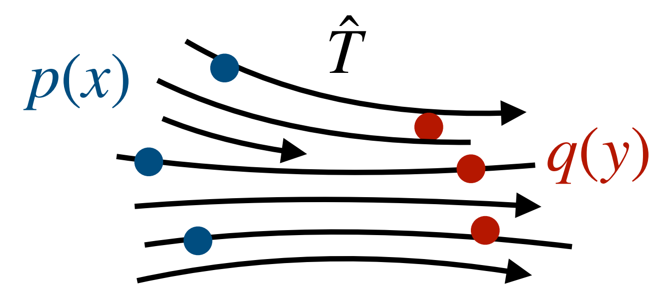
Two main classes of algorithms:



"Discrete" optimal transport

Transport empirical distributions by pairing up samples ~ $\mathcal{O}(N^2)$

Need to interpolate transport map to unseen samples



"Continuous" optimal transport

Use samples to construct continuous transport function

Need to make assumptions on underlying densities

The role of the transport cost

The character of the solution $\hat{T}$ to the Monge problem depends strongly on the cost function $c(x, y)$

Many useful cost functions are *(strictly)* convex!

E.g. $c(x, y) = |x - y|^p$ for $p > 1$

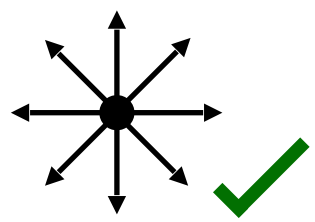
In this case: the optimal transport function is <u>unique</u> and the gradient of a potential!

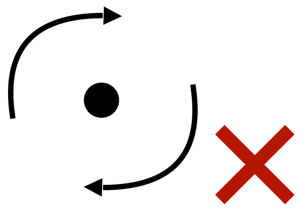
$$\hat{T}(x) = x + \nabla g(x)$$

"Transport potential"

Optimal transport ⇔ Electrostatics

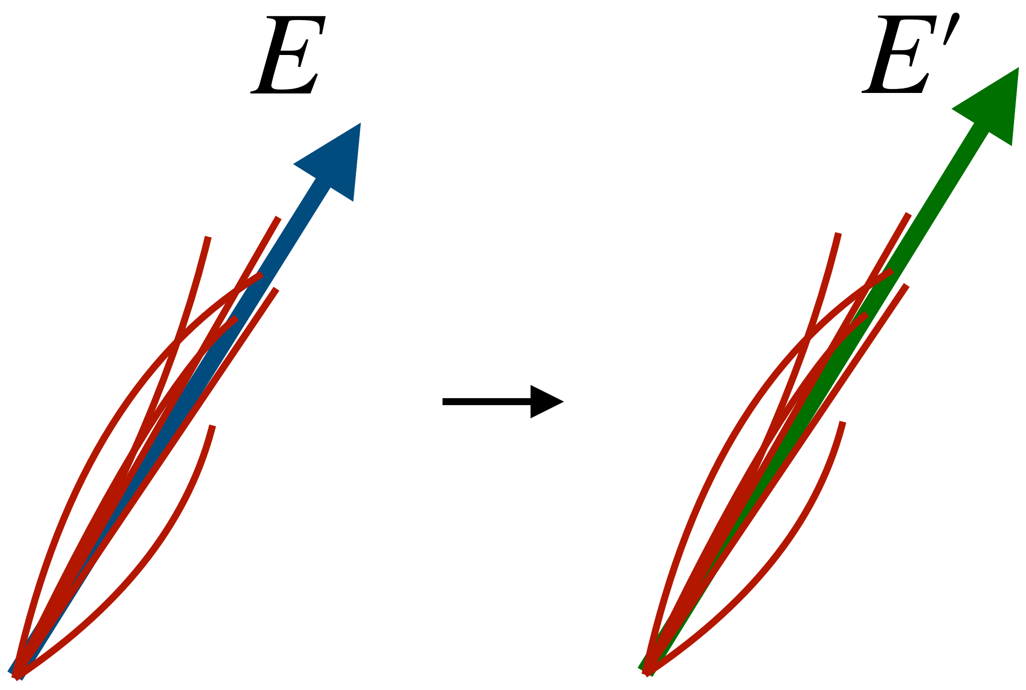
The transport vector field $\hat{T}$ has zero curl!





"Don't ship your stuff in circles."

→ More information on other cases in backup



(Potential) Applications in high-energy physics

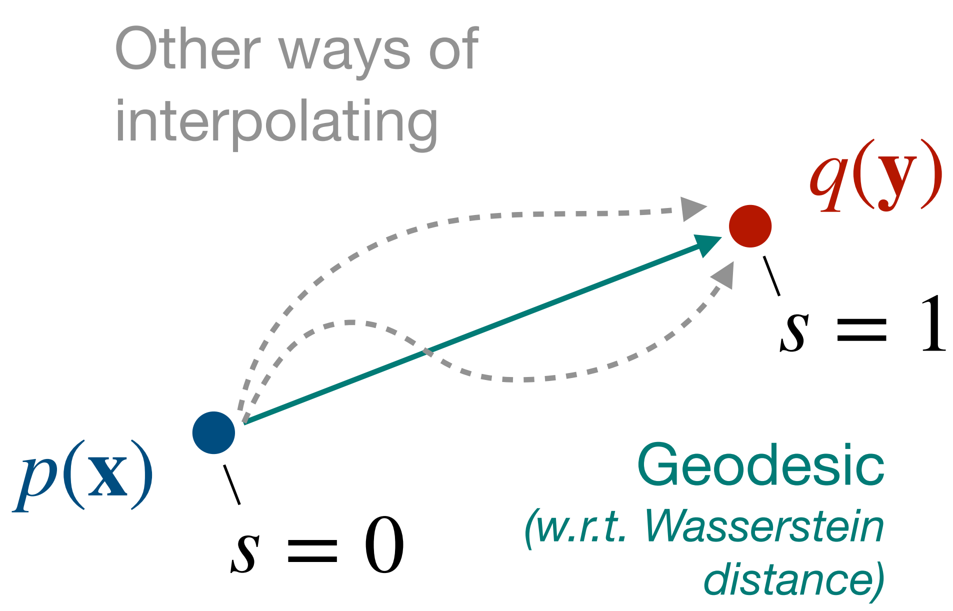
Template morphing

Optimal transport solution maps $p(\mathbf{x})$ into $q(\mathbf{y})$

$$\mathbf{x} \mapsto \mathbf{y} = \hat{T}(x) = x + \nabla g(x)$$

Can interpolate between p and q: just move each sample by a **fraction of the full gradient**

$$\hat{T}_s(x) = x + s\nabla g(x), \quad 0 \leq s \leq 1$$



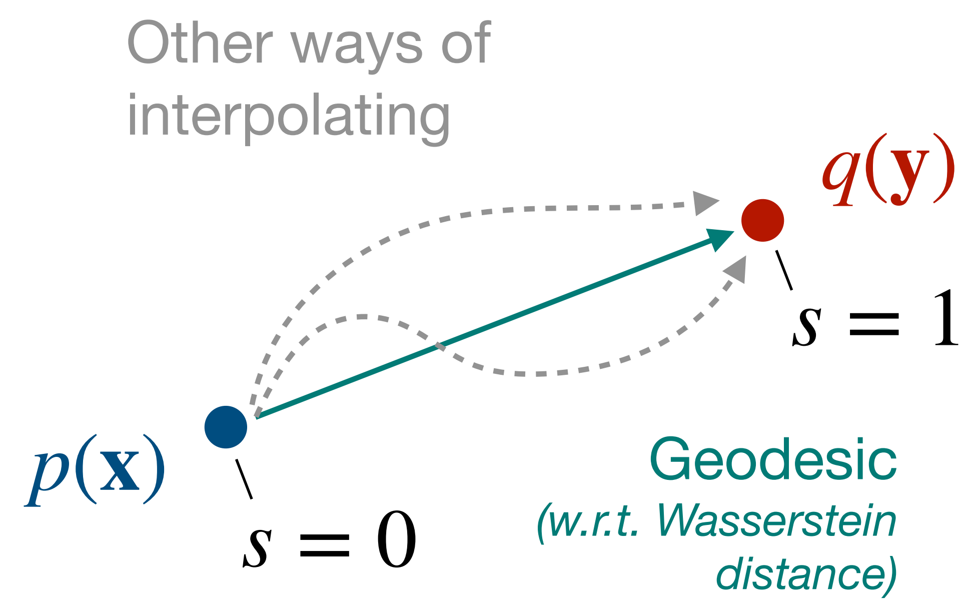
Template morphing

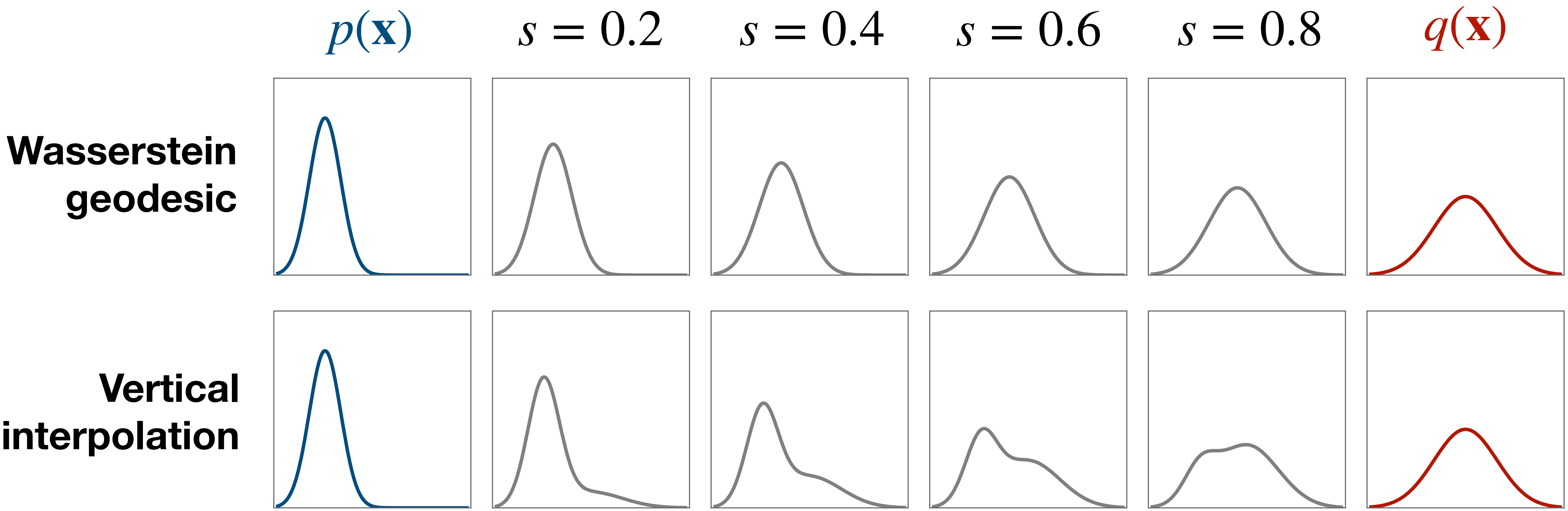
Optimal transport solution maps $p(\mathbf{x})$ into $q(\mathbf{y})$

$$\mathbf{x} \mapsto \mathbf{y} = \hat{T}(x) = x + \nabla g(x)$$

Can interpolate between p and q: just move each sample by a **fraction of the full gradient**

$$\hat{T}_s(x) = x + s\nabla g(x), \quad 0 \le s \le 1$$

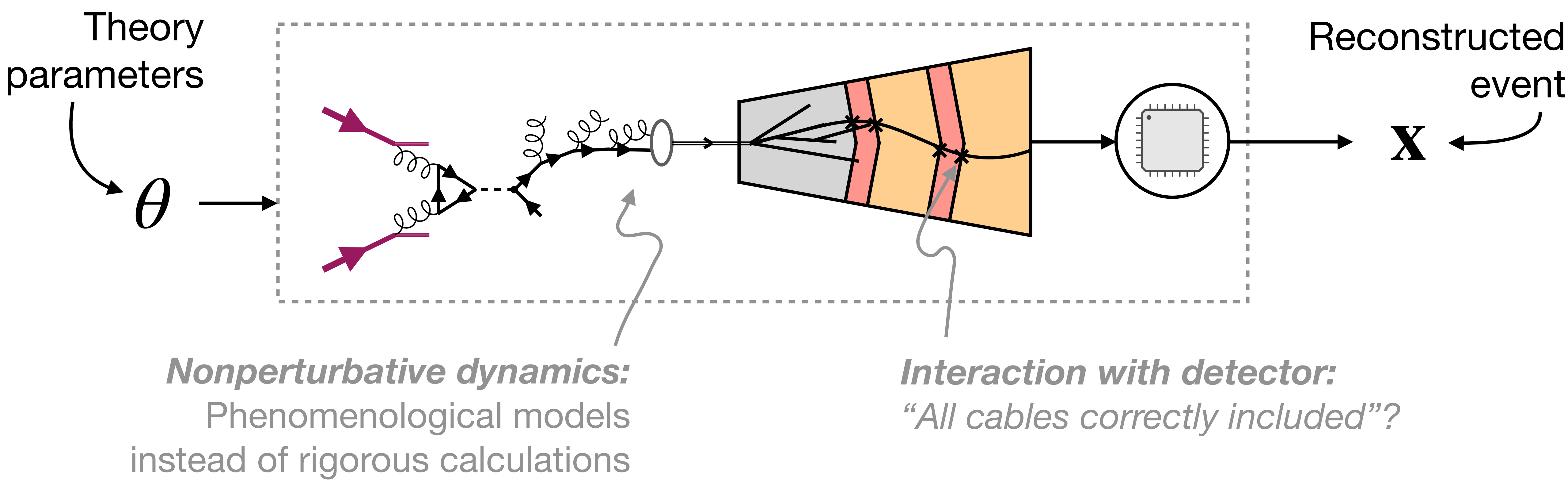




Calibrating simulations

Our field has spent several decades building extremely precise simulations …

*… they **encode** a lot of **domain knowledge**, but they are not perfect!*



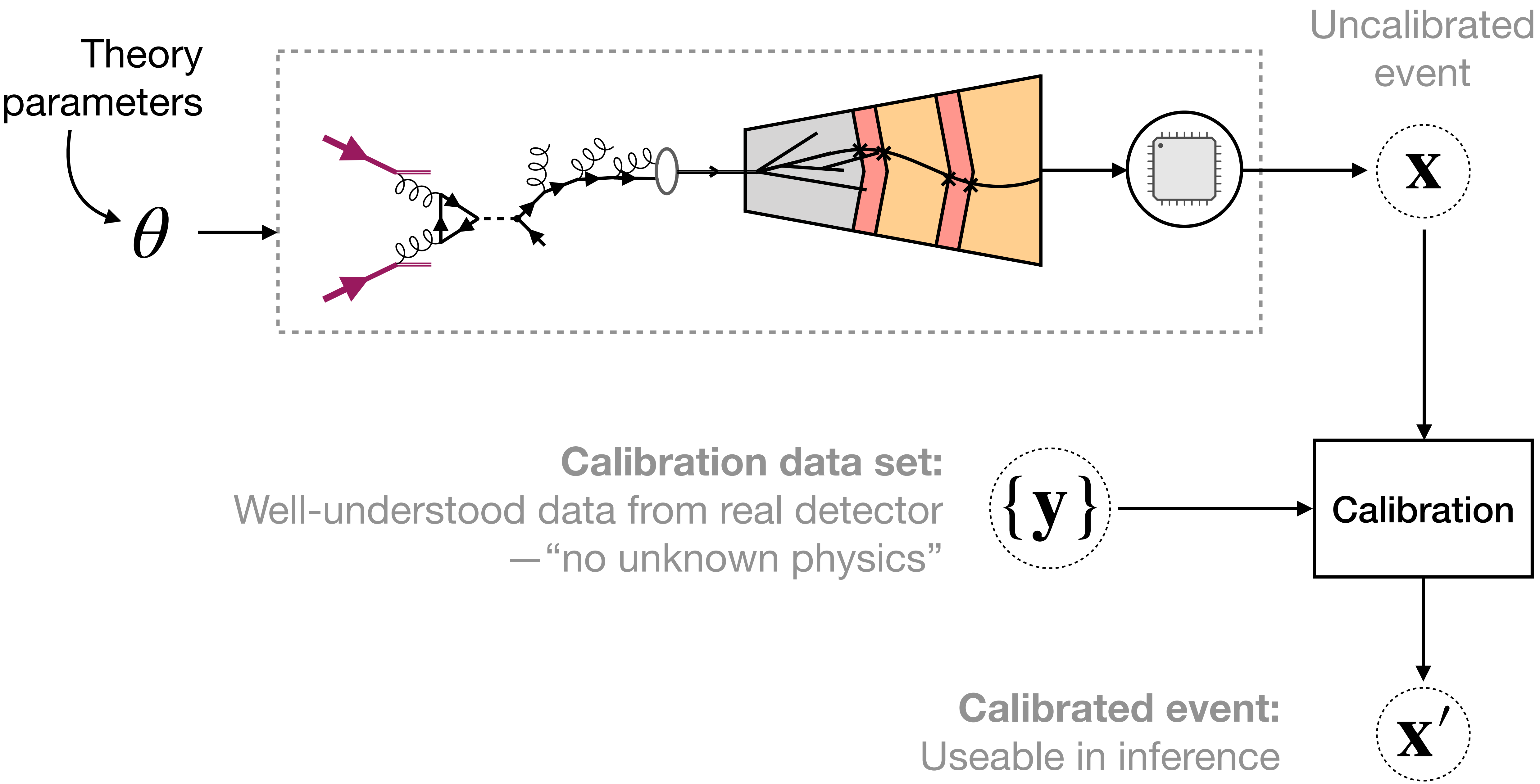
Often impossible / impractical to correct the simulation model

***Instead:** calibrate the simulator output*

Calibrating simulations

Our field has spent several decades building extremely precise simulations …

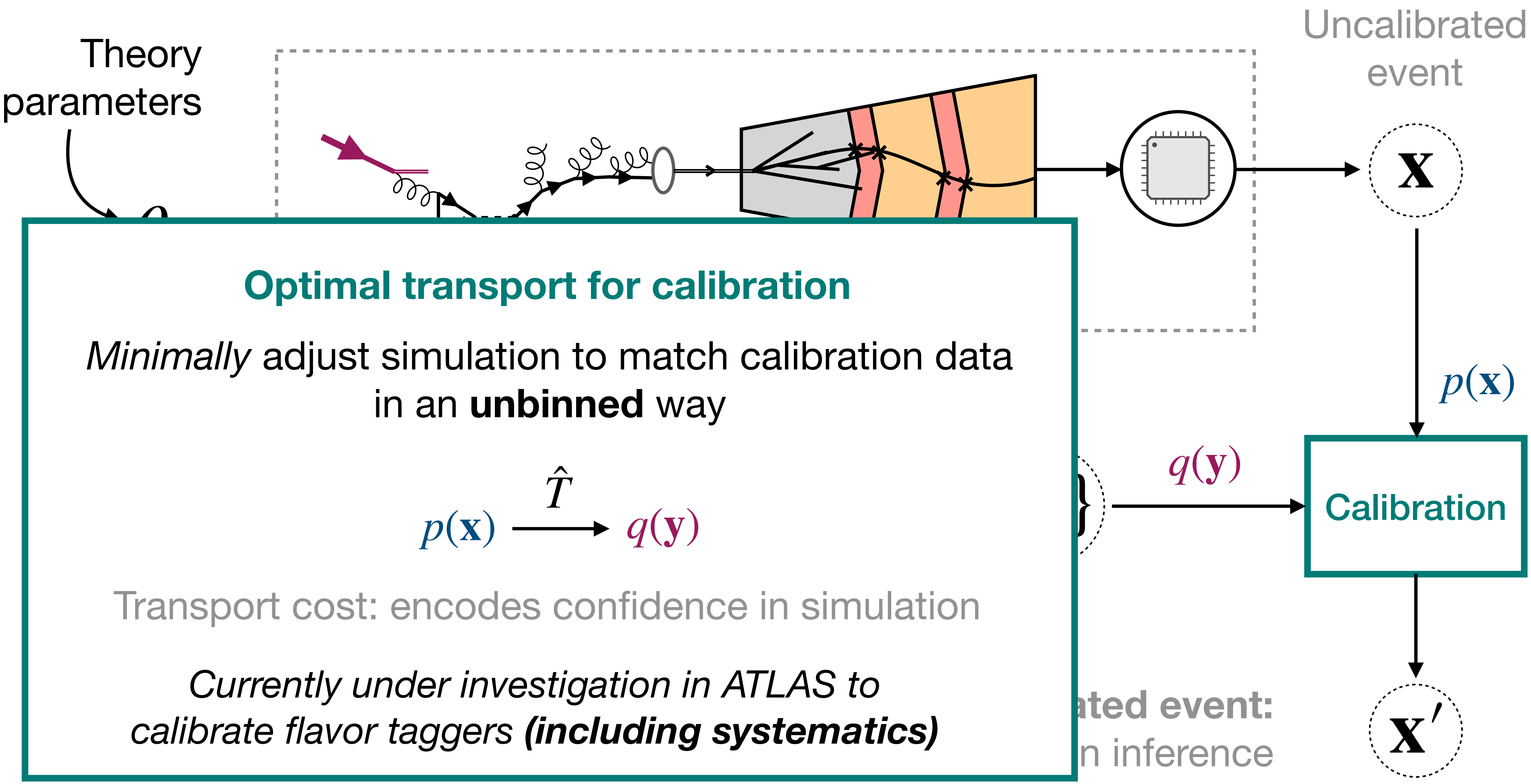
*… they **encode** a lot of **domain knowledge**, but they are not perfect!*



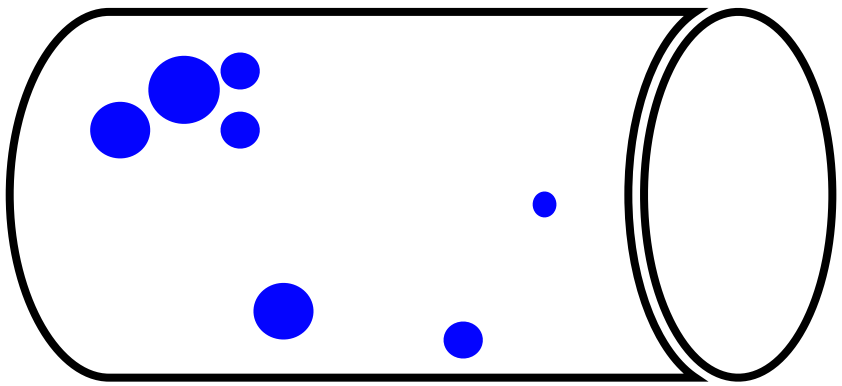
Calibrating simulations

Our field has spent several decades building extremely precise simulations …

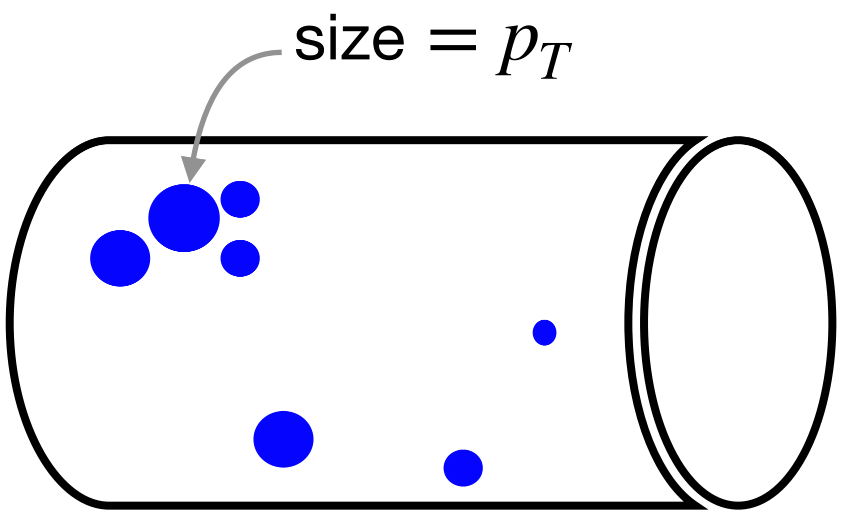
*… they **encode** a lot of **domain knowledge**, but they are not perfect!*



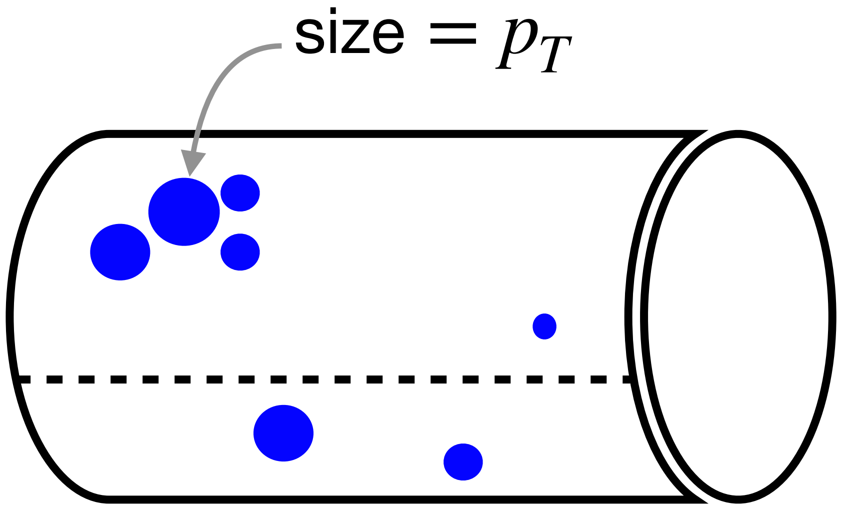
Comparing collider events (Komiske et al. 2019)



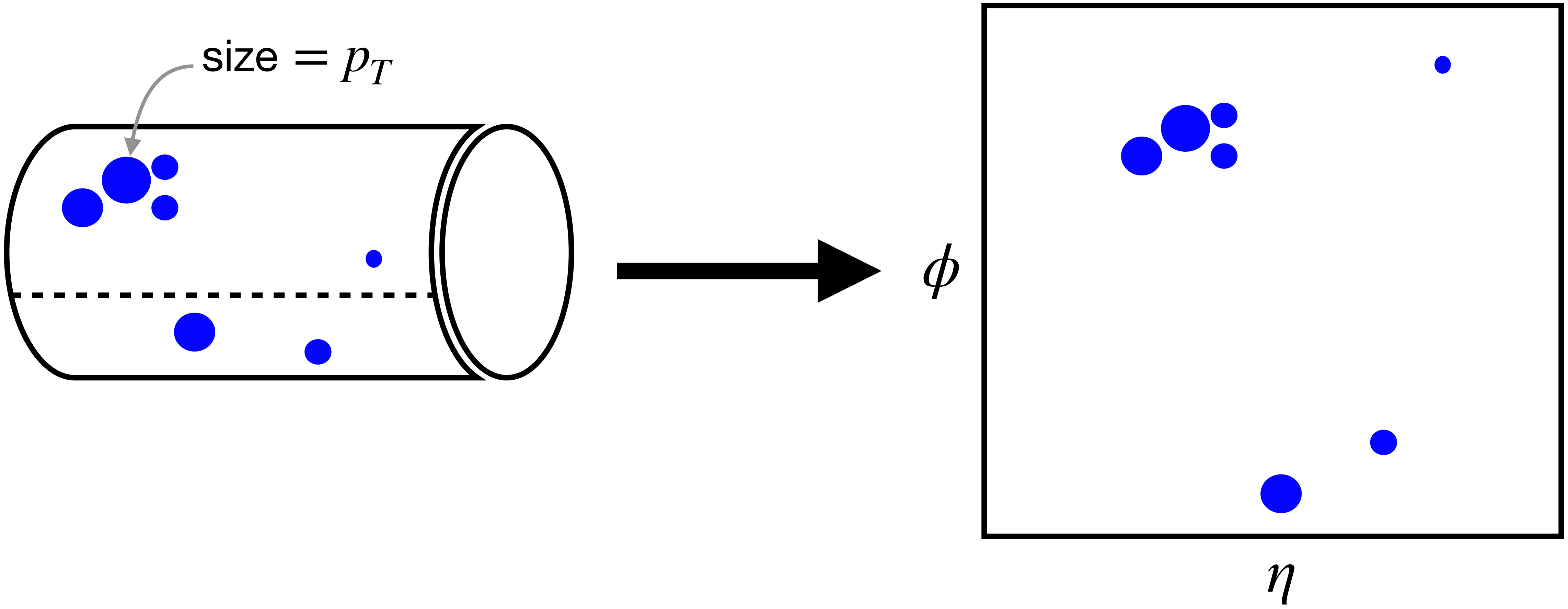
Comparing collider events (Komiske et al. 2019)



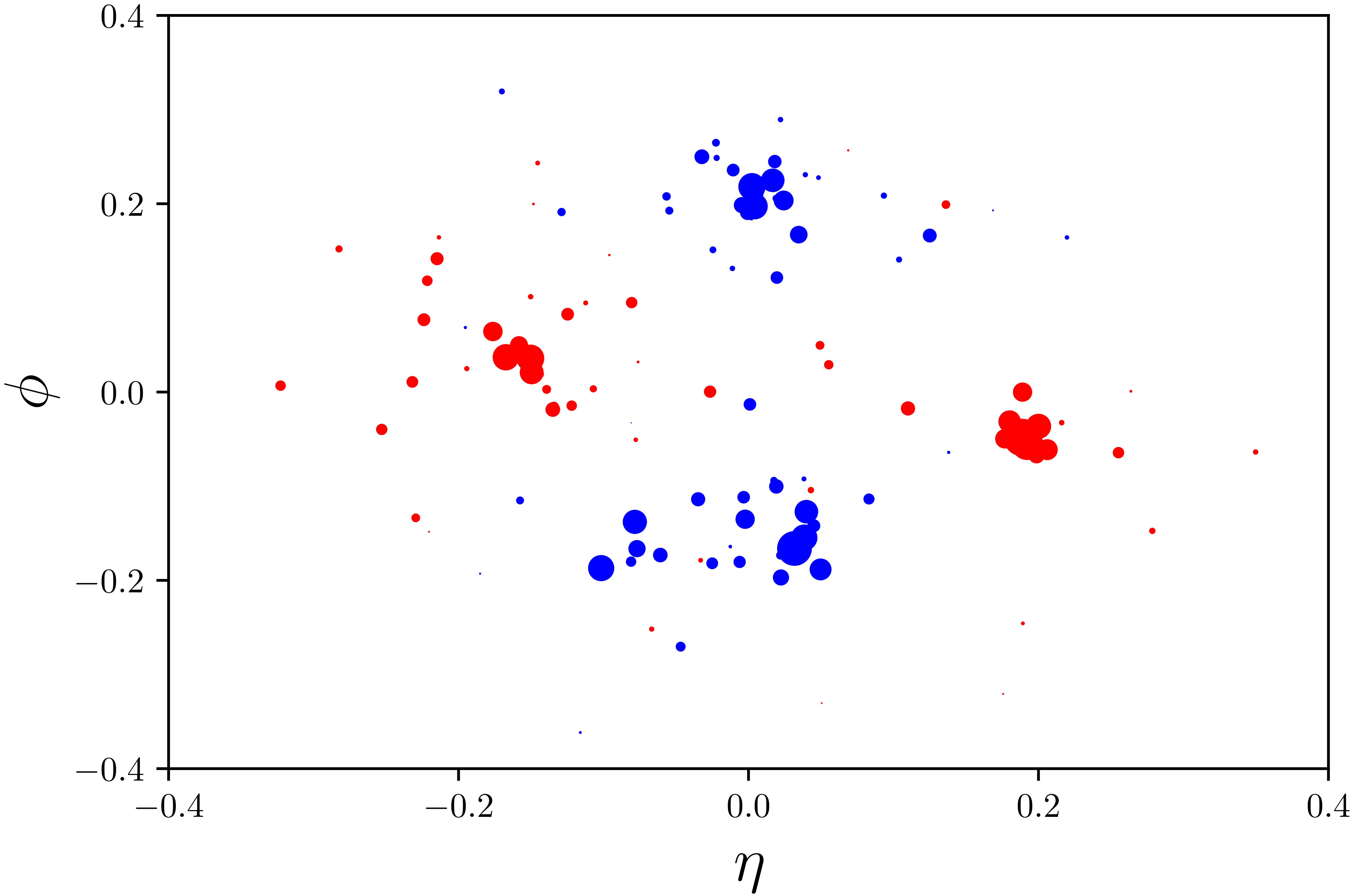
Comparing collider events (Komiske et al. 2019)



Comparing collider events (Komiske et al. 2019)

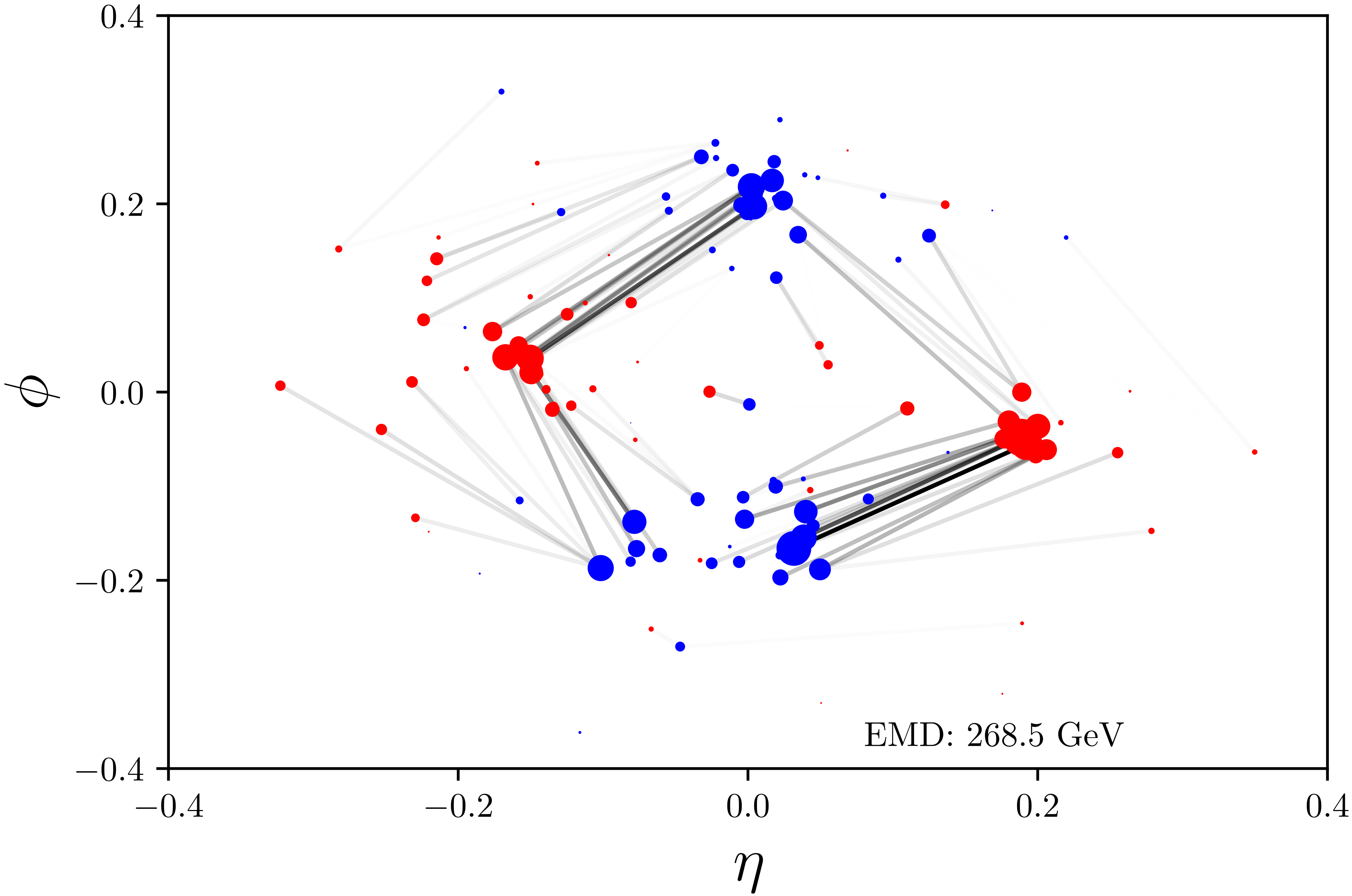


Comparing collider events (Komiske et al. 2019)



Generated with the Energyflow package based on CMS open data.

Comparing collider events (Komiske et al. 2019)



Generated with the Energyflow package based on CMS open data.

Data-driven background estimation

$$X_1, \ldots, X_n \sim f(x) = \epsilon \cdot s(x) + (1 - \epsilon) \cdot b(x)$$

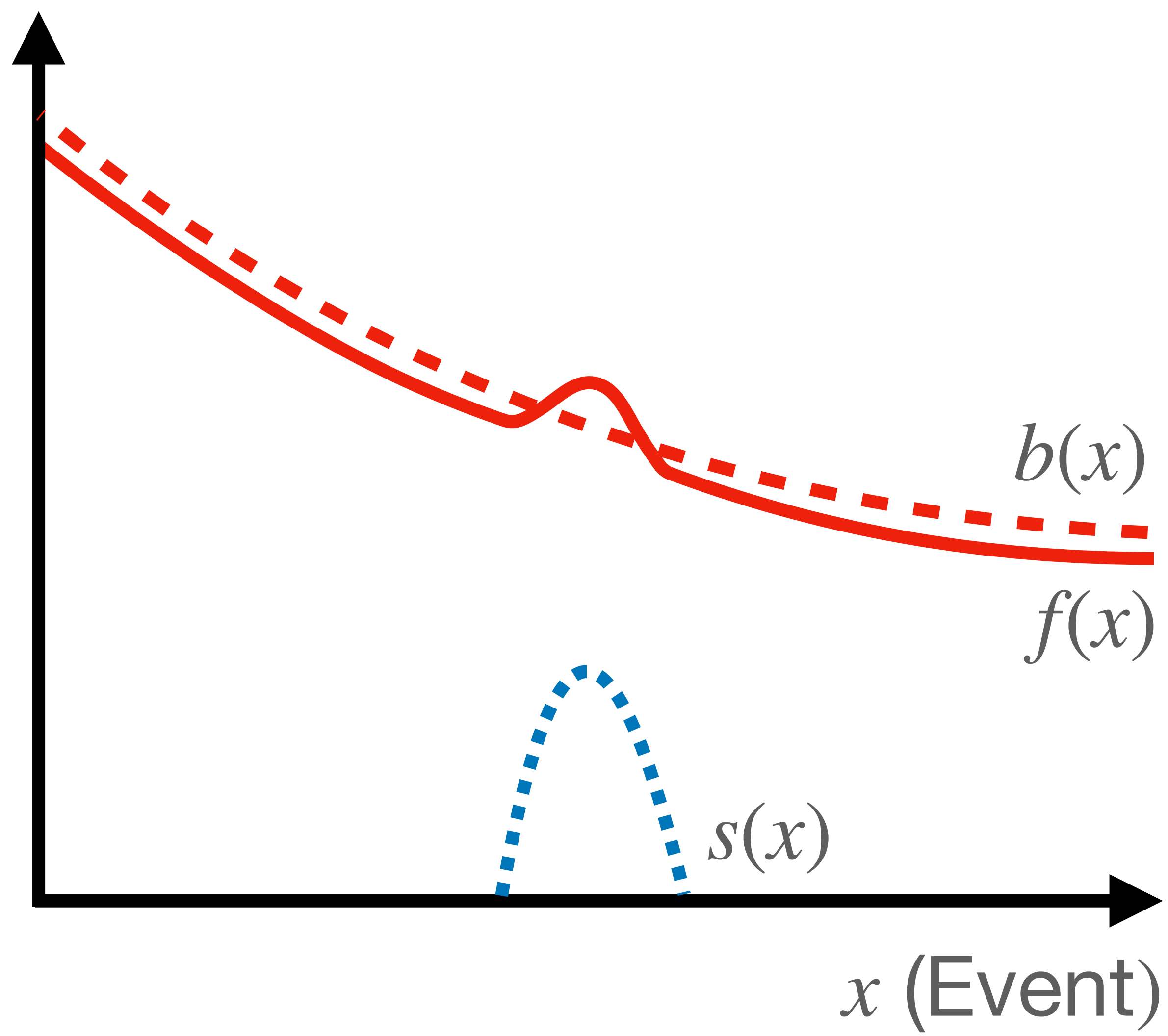
s: Known signal density

b: **Unknown** background density

ϵ: Proportion of signal

Goal: Provide a data-driven estimate of the (multidimensional) distribution b.

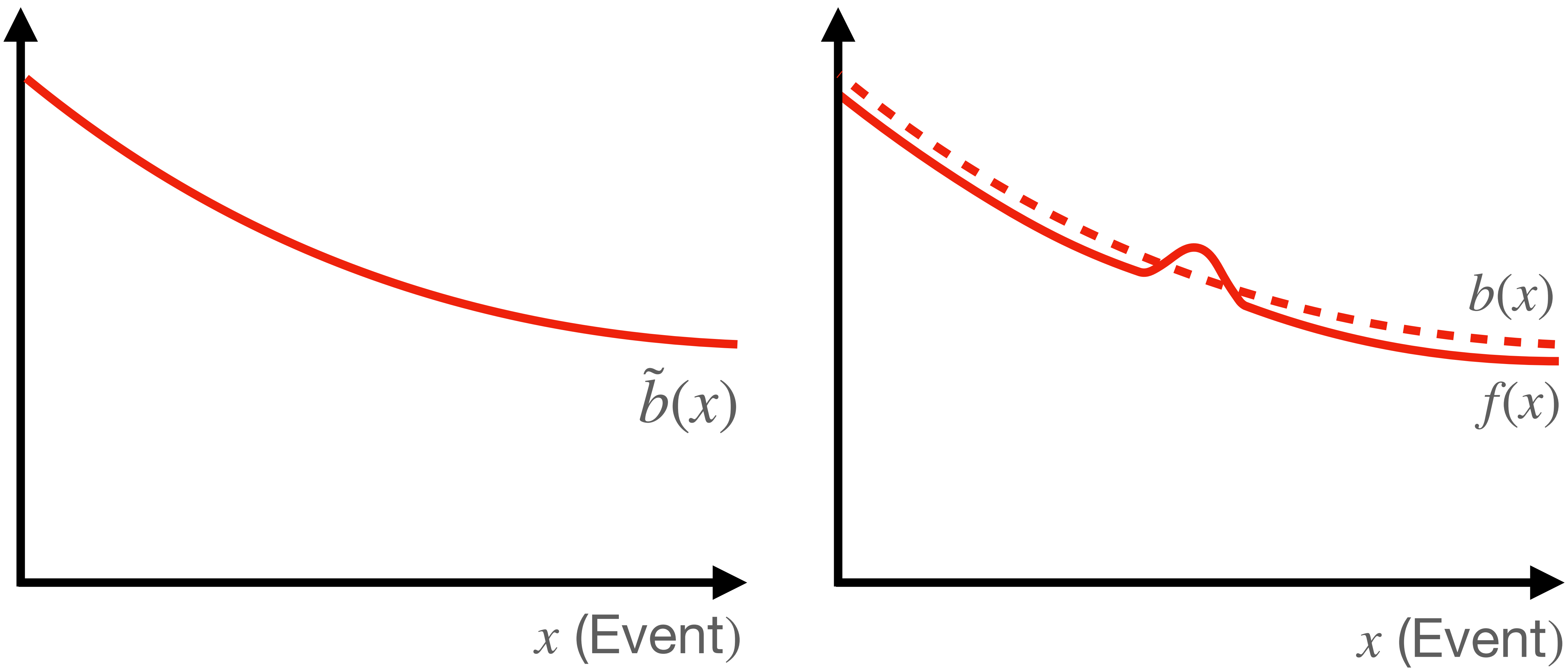
- Example: HH→4b search



Data-driven background estimation

$$X_1, \ldots, X_n \sim f(x) = \epsilon \cdot s(x) + (1 - \epsilon) \cdot b(x)$$

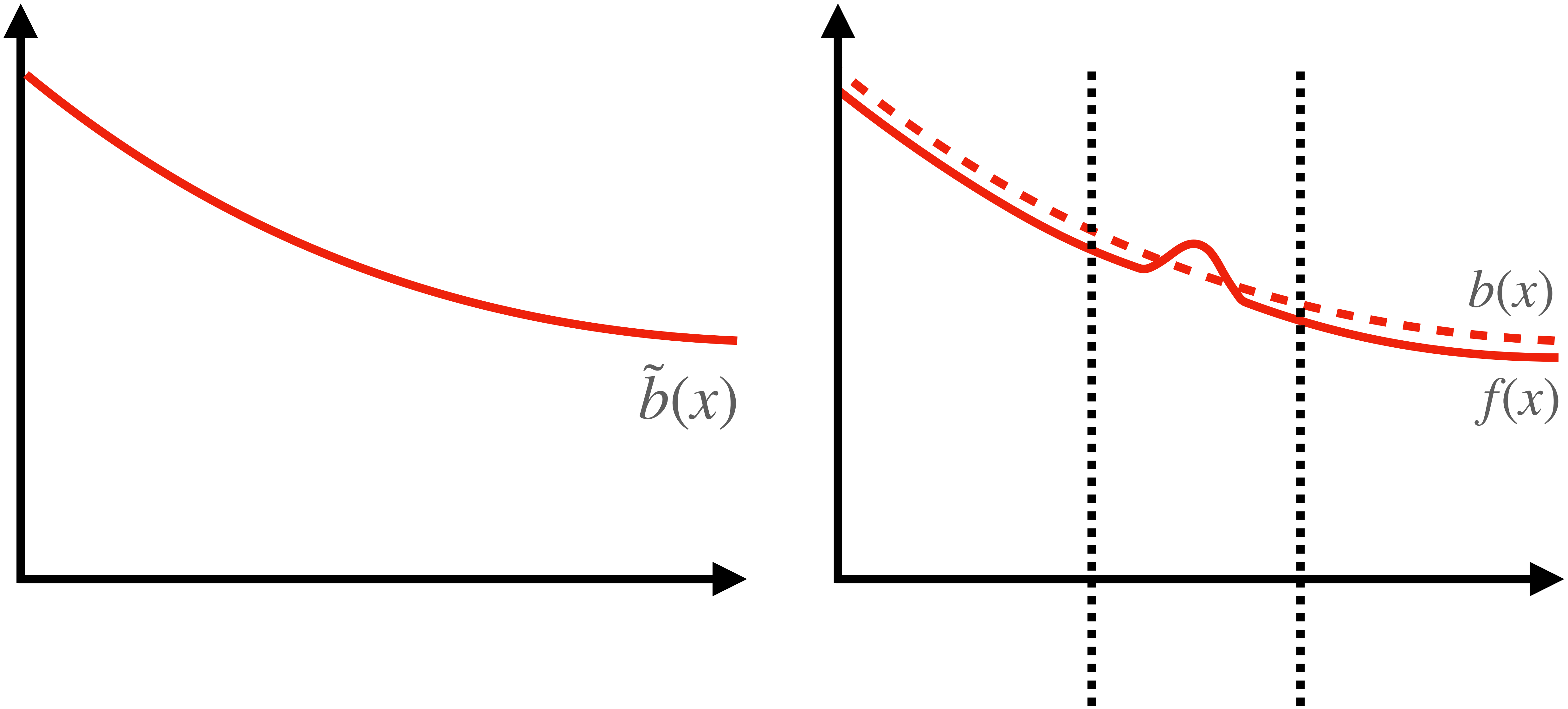
Assume we also have**:** $Y_1, \ldots, Y_m \sim \tilde{b}(x) \approx b(x)$



Data-driven background estimation

$$X_1, \ldots, X_n \sim f(x) = \epsilon \cdot s(x) + (1 - \epsilon) \cdot b(x)$$

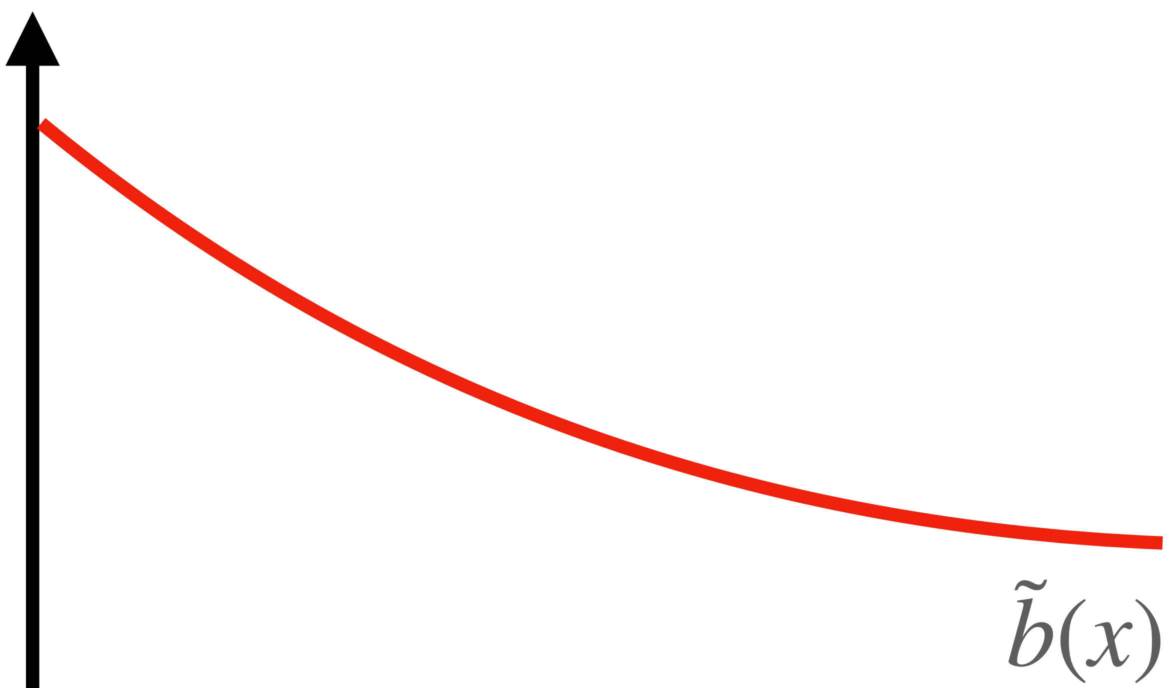
Assume we also have**:** $Y_1, \ldots, Y_m \sim \tilde{b}(x) \approx b(x)$

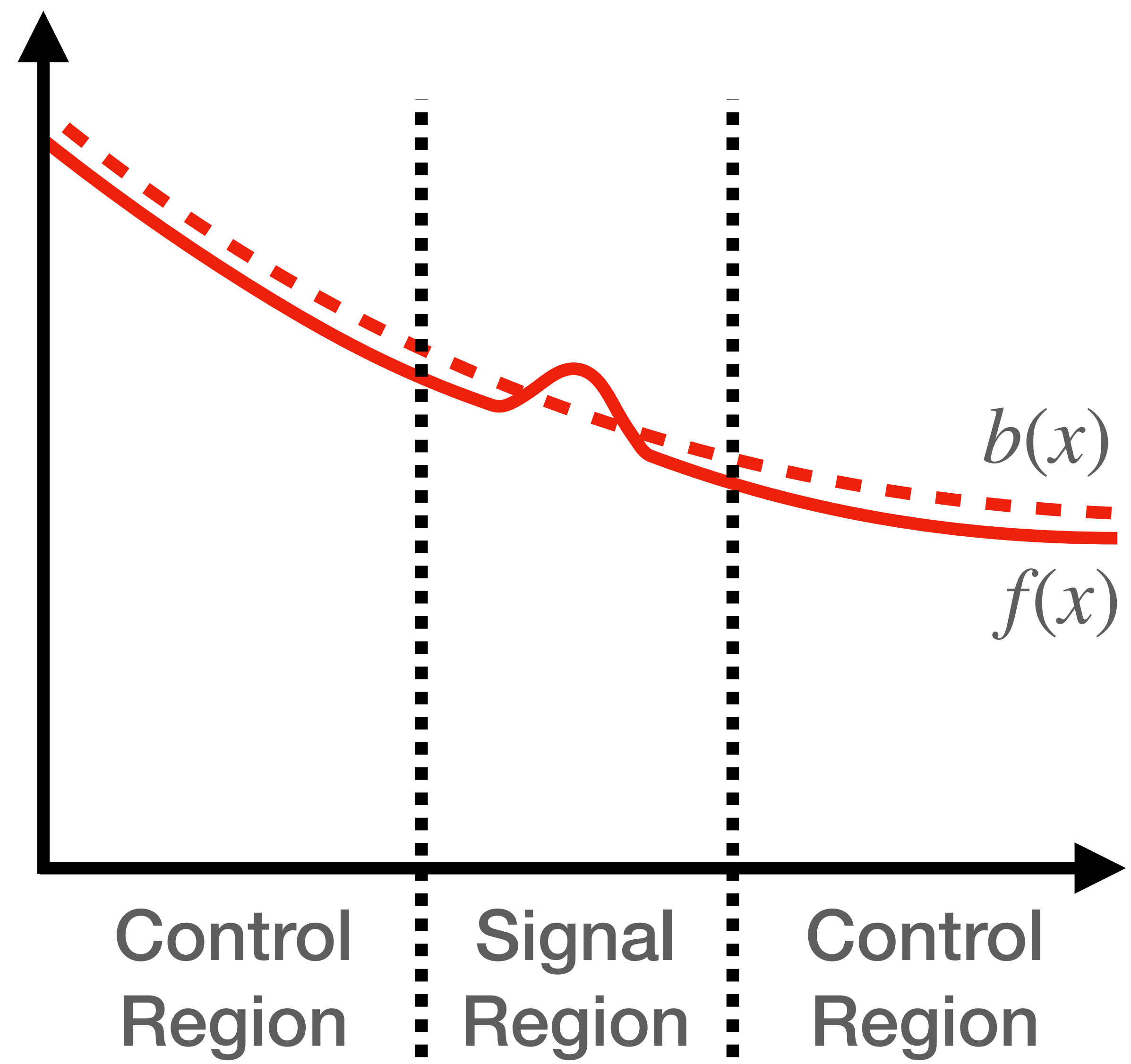


Data-driven background estimation

$$X_1, \ldots, X_n \sim f(x) = \epsilon \cdot s(x) + (1 - \epsilon) \cdot b(x)$$

Assume we also have**:** $Y_1, \ldots, Y_m \sim \tilde{b}(x) \approx b(x)$

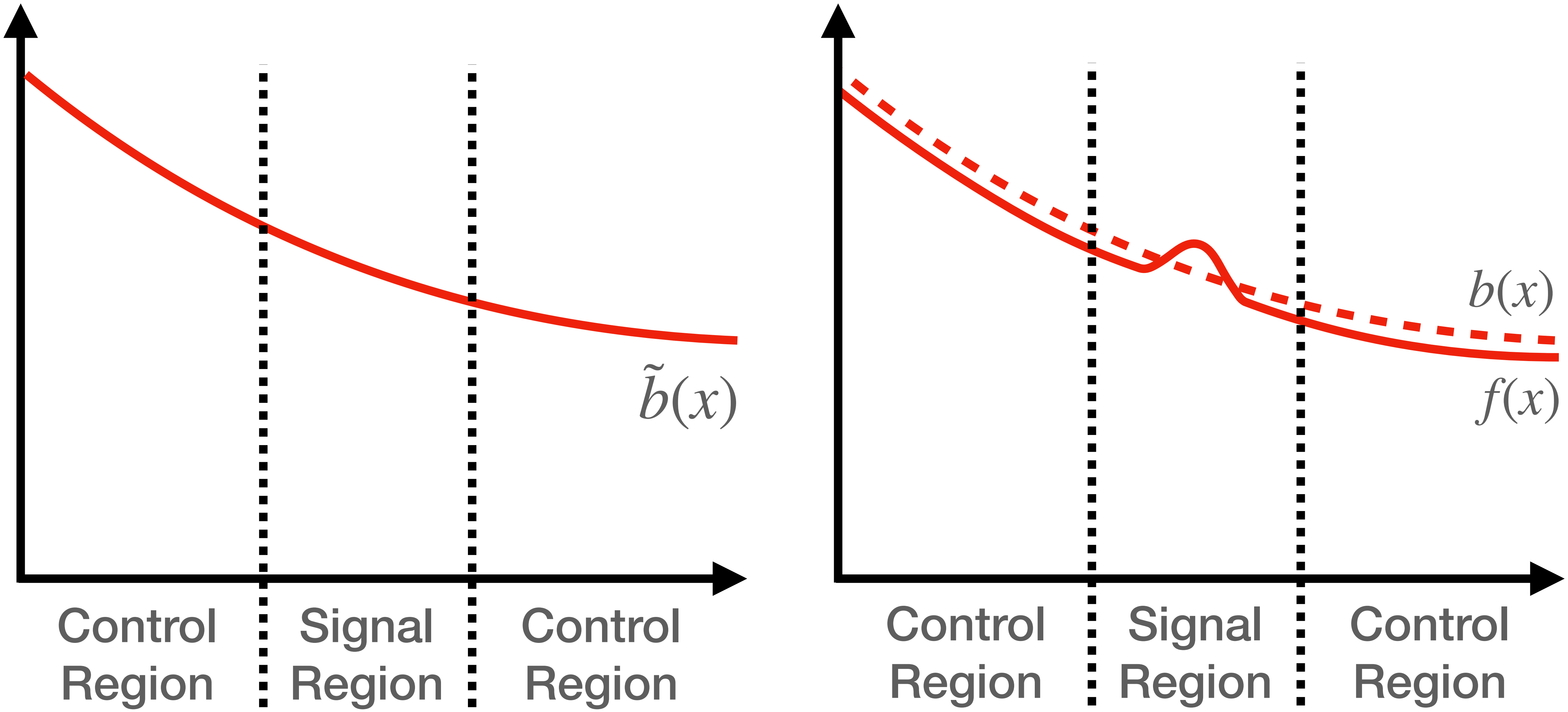




Data-driven background estimation

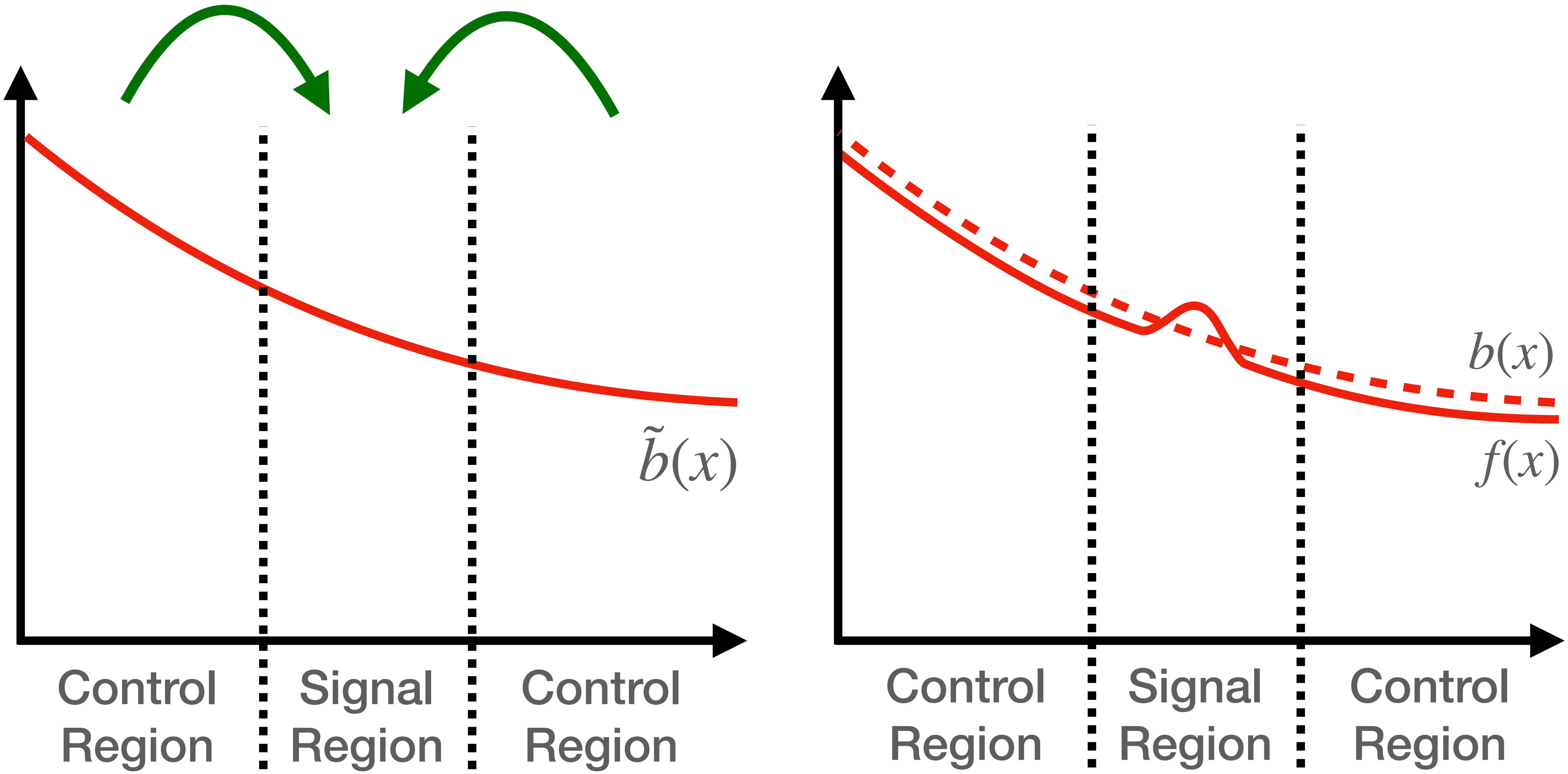
$$X_1, \ldots, X_n \sim f(x) = \epsilon \cdot s(x) + (1-\epsilon) \cdot b(x)$$

Assume we also have: $Y_1, \ldots, Y_m \sim \tilde{b}(x) \approx b(x)$



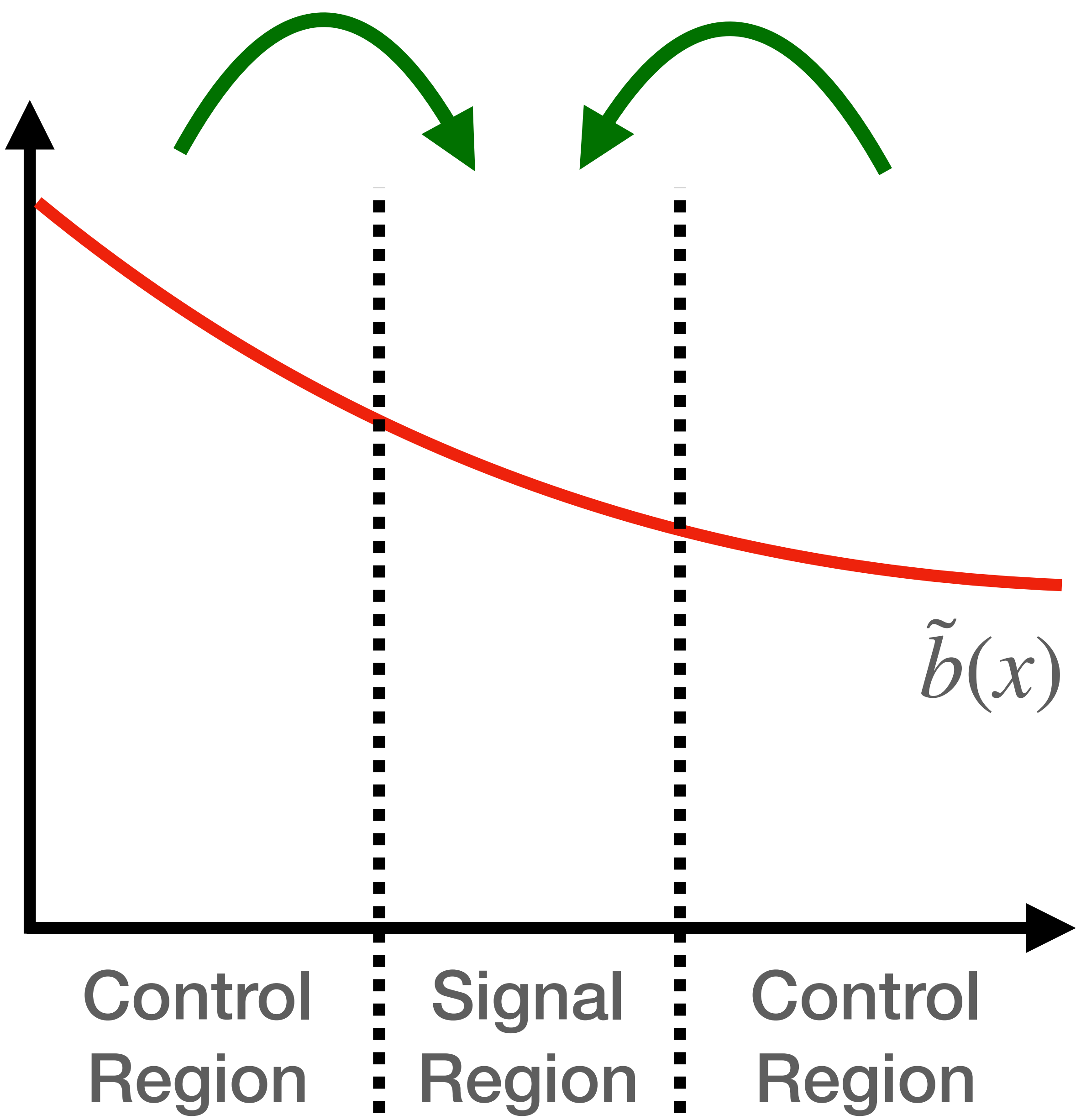
Data-driven background estimation

Step 1: Fit <u>multivariate</u> OT map
$\hat{T}$ from CR to SR of $\tilde{b}$

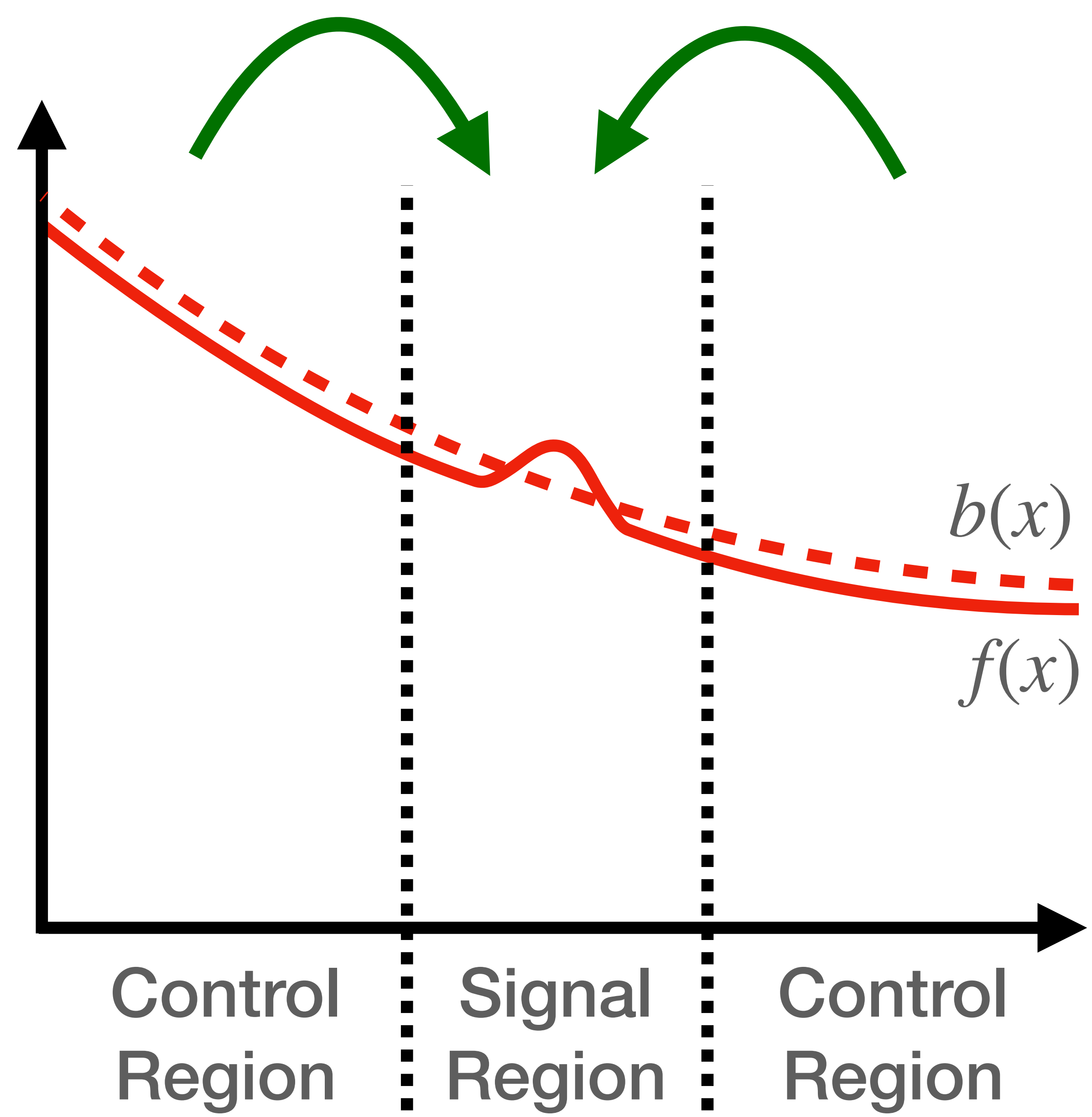


Data-driven background estimation

Step 1: Fit <u>multivariate</u> OT map $\hat{T}$ from CR to SR of $\tilde{b}$



Step 2: Evaluate on CR of b (distinct modeling assumptions from density ratio extrapolation)

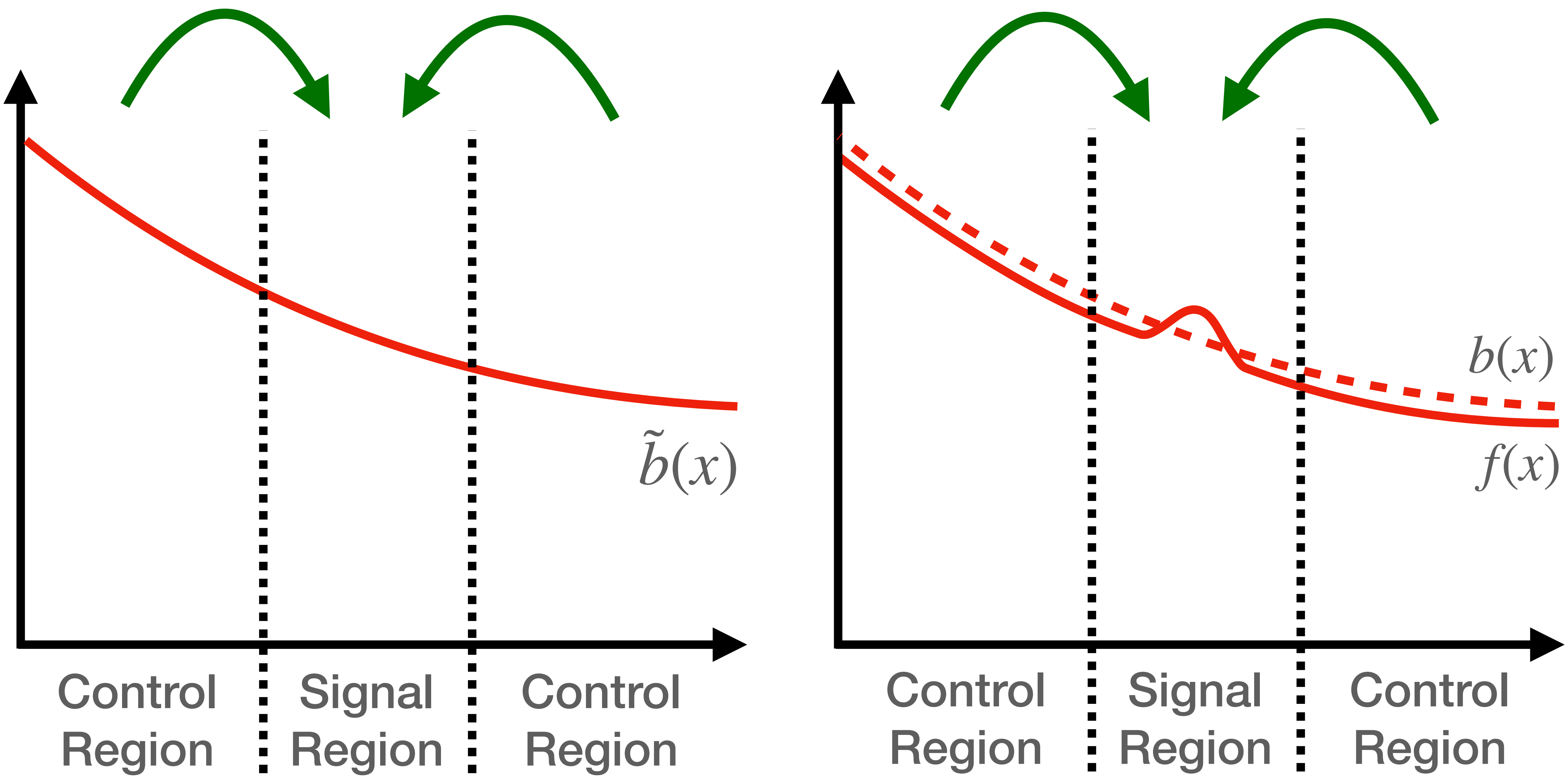


Data-driven background estimation

Hierarchical Optimal Transport:

Our ground cost is itself the OT distance of Komiske et al. (2019)!

arXiv:2208.02807



More generally: Optimal transport for transfer learning

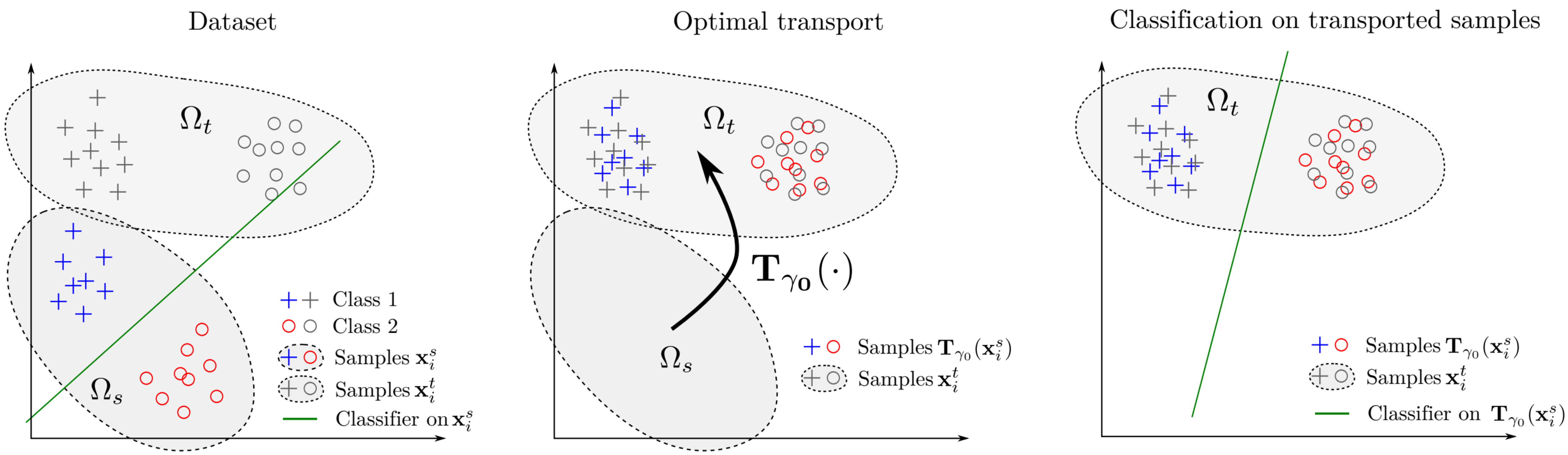
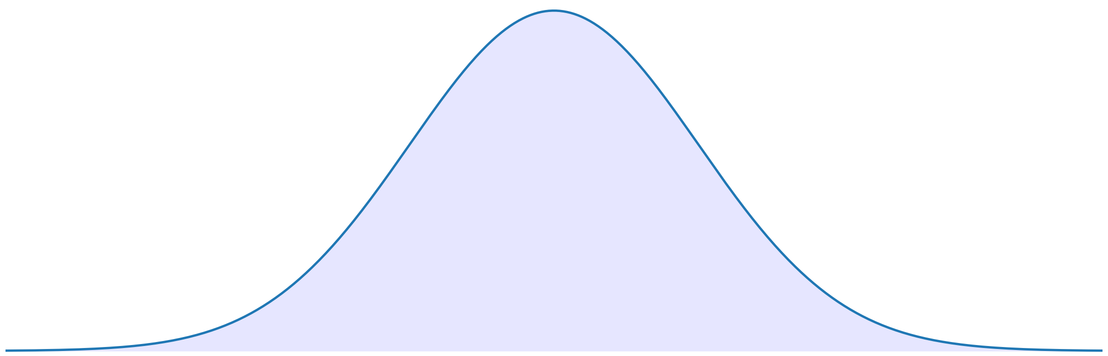


Image Credit: Courty et al (2016)

Multivariate C.D.F.s and quantiles

(Consider $c = \|\cdot\|^2$)

p

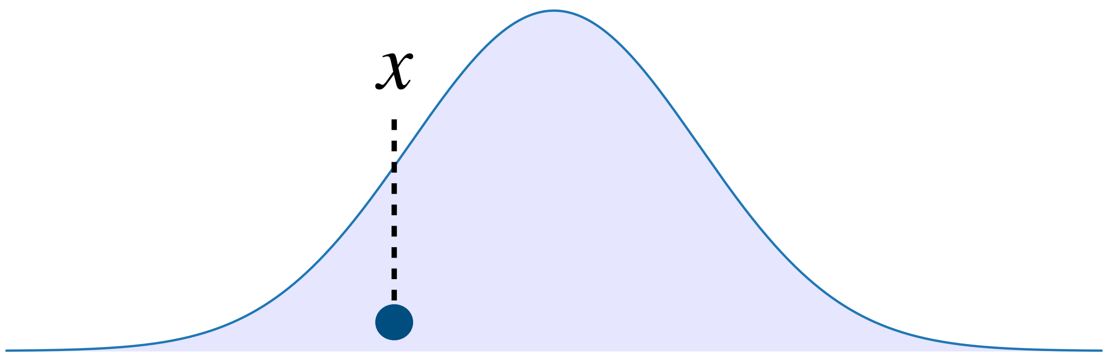


$q = \text{Unif}(0,1)$

Multivariate C.D.F.s and quantiles

(Consider $c = \|\cdot\|^2$)

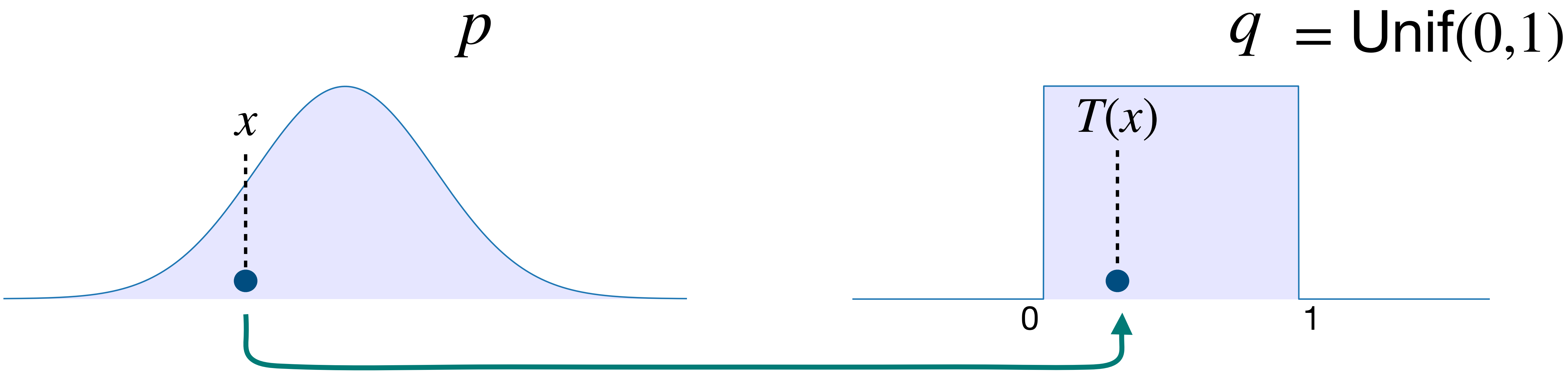
p



$q = \text{Unif}(0,1)$

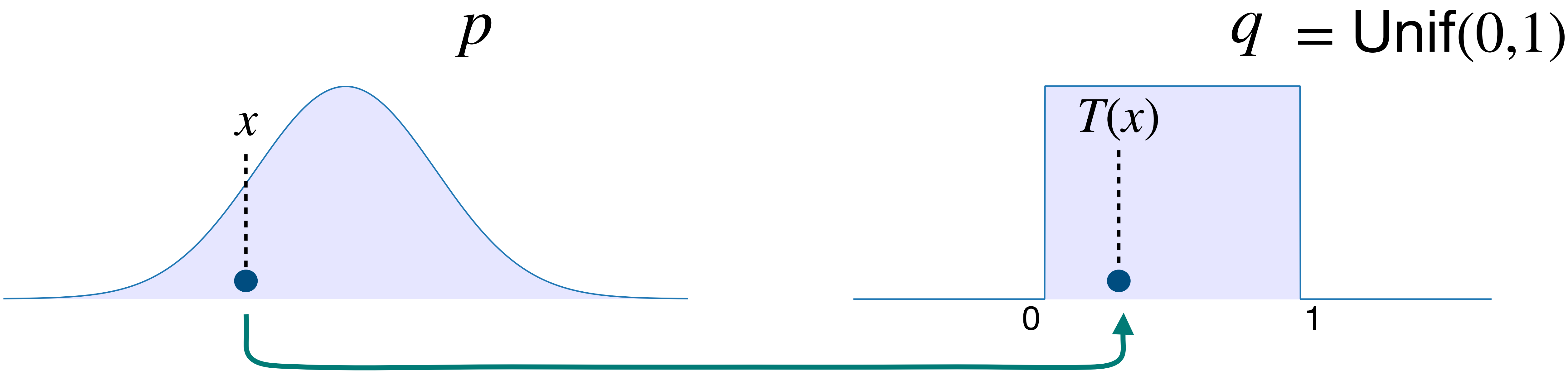
Multivariate C.D.F.s and quantiles

(Consider $c = \|\cdot\|^2$)



Multivariate C.D.F.s and quantiles

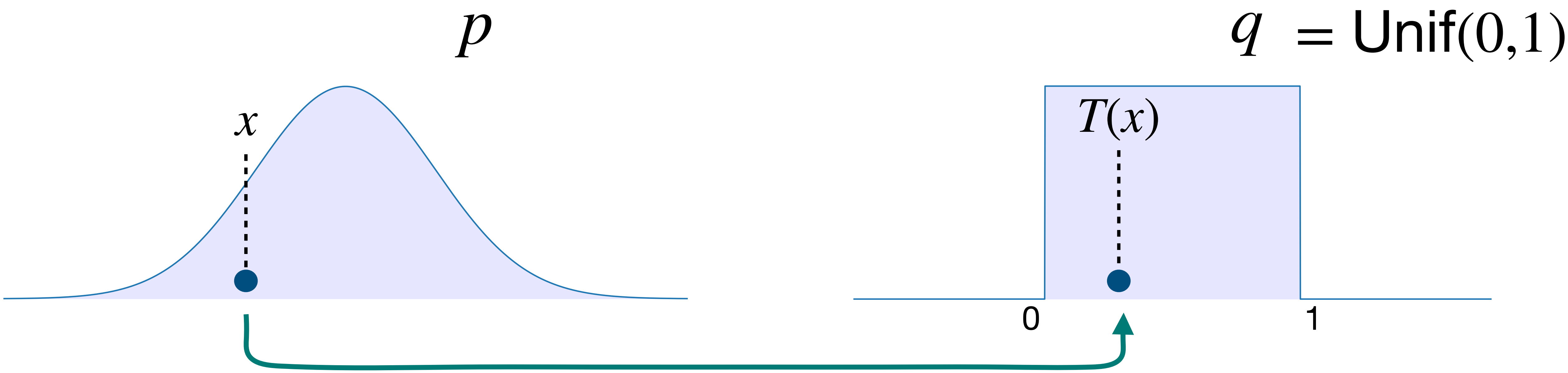
(Consider $c = \|\cdot\|^2$)



$$T(x) = \int_{\infty}^{x} p(y)dy$$

Multivariate C.D.F.s and quantiles

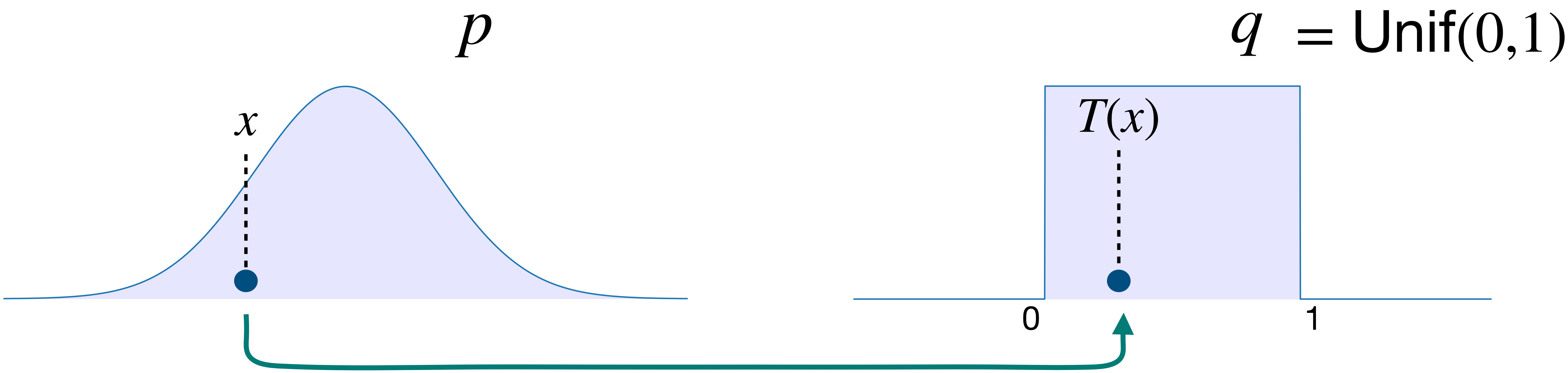
(Consider $c = \|\cdot\|^2$)



$T(x) = \int_{\infty}^{x} p(y)dy$ (T is the C.D.F. of p)!

Multivariate C.D.F.s and quantiles

(Consider $c = \|\cdot\|^2$)



$T(x) = \int_{\infty}^{x} p(y)dy$ (T is the C.D.F. of p)!

Suggests a way to define multivariate C.D.F.s and quantiles

Given a **reference density** q and a multivariate density p:

- The OT map from p to q is called the **multivariate C.D.F.** of p
- The OT map from q to p is called the **multivariate quantile** of p.

Multivariate C.D.F.s and quantiles

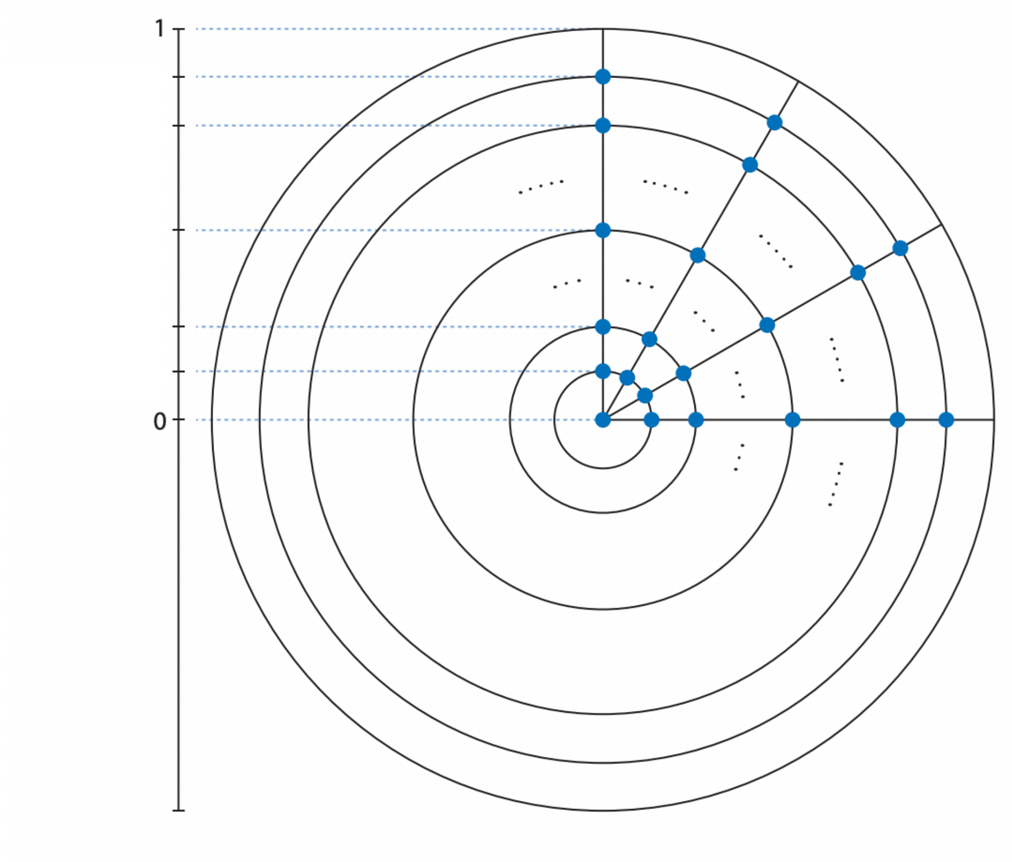


Image Credit: Hallin (2022).

Multivariate C.D.F.s and quantiles

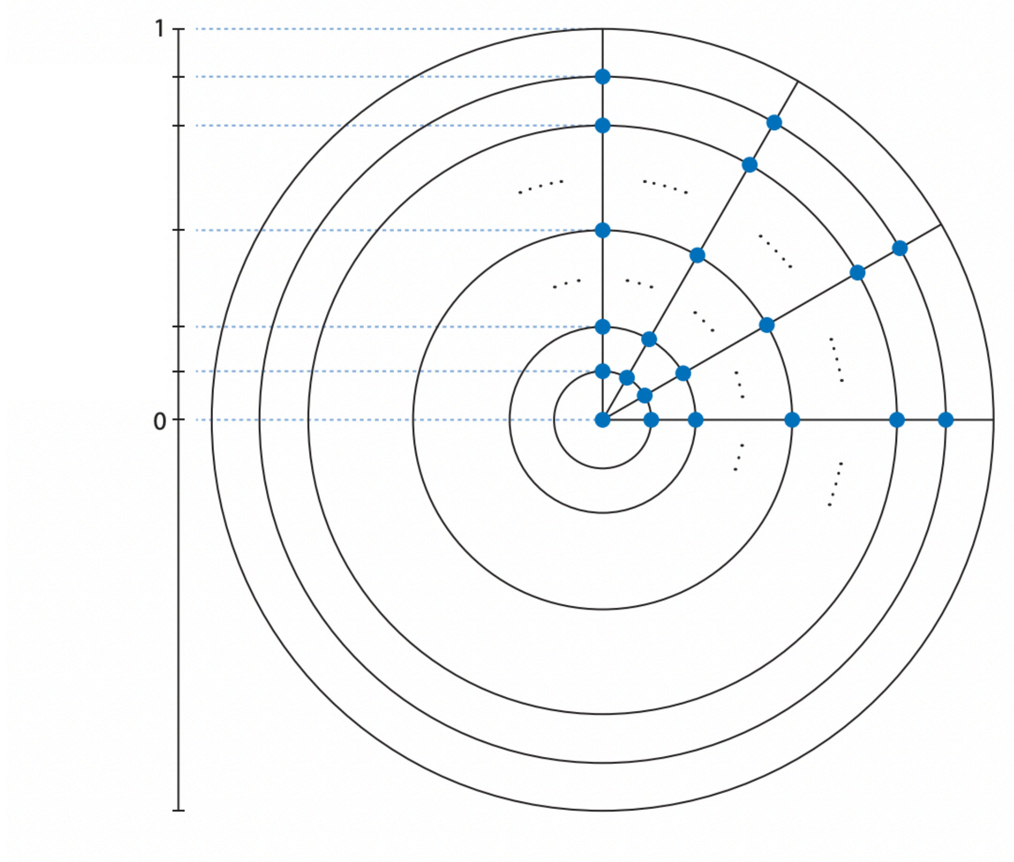
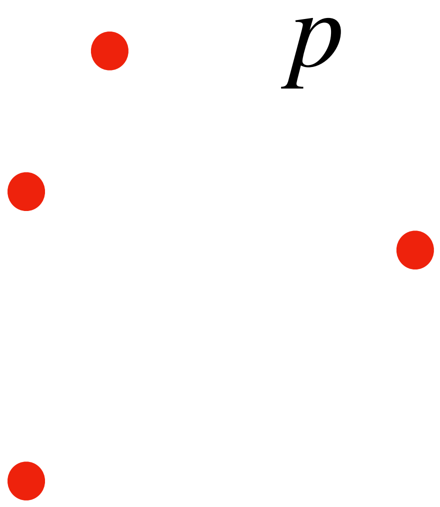


Image Credit: Hallin (2022).



Multivariate C.D.F.s and quantiles

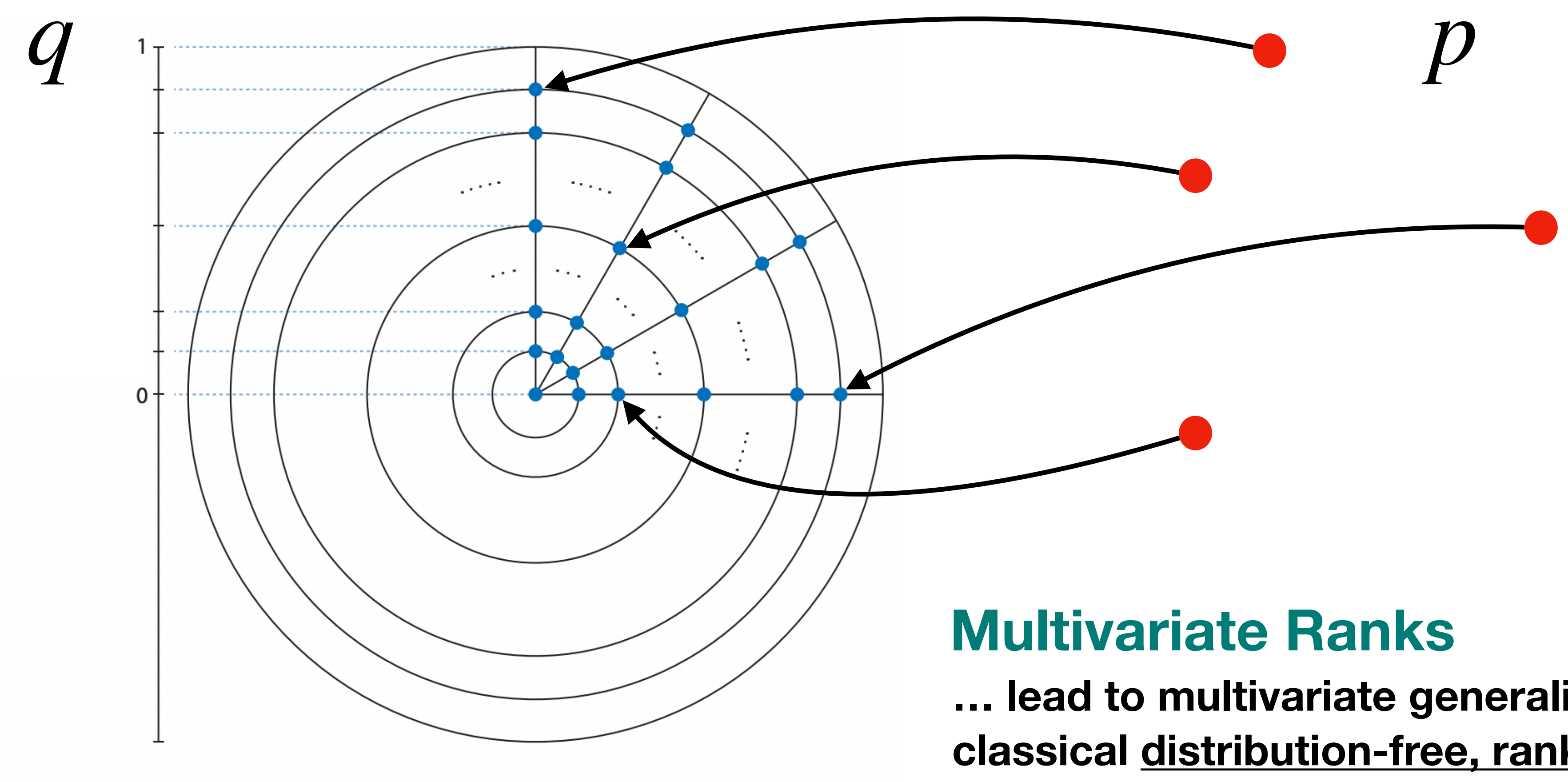


Image Credit: Hallin (2022).

Multivariate Ranks

… lead to multivariate generalizations of classical <u>distribution-free, rank-based methods</u> for Goodness-of-Fit testing, two-sample testing, independence testing, testing for symmetry, etc.

Multivariate C.D.F.s and quantiles

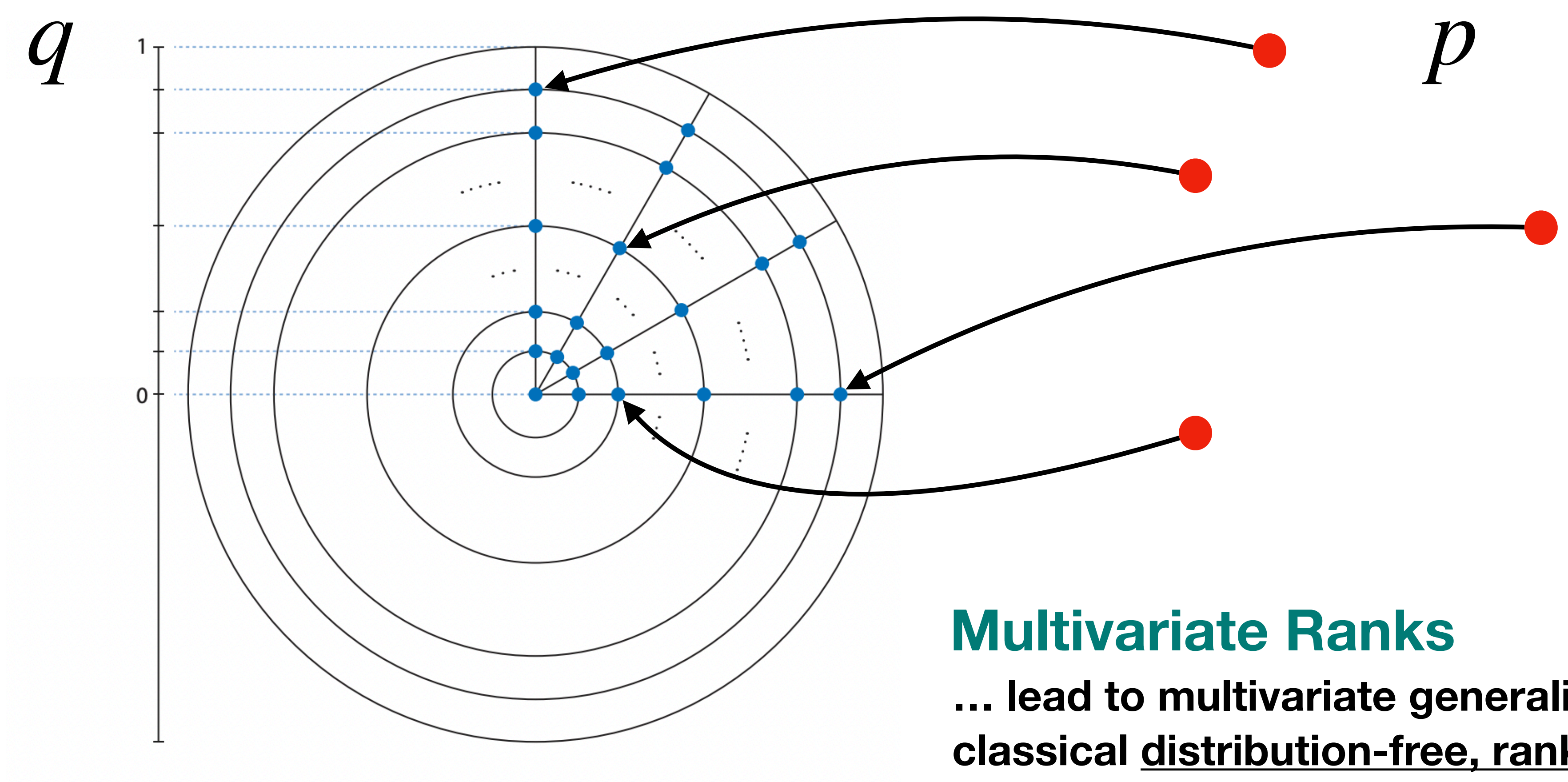


Image Credit: Hallin (2022).

Multivariate Ranks

… lead to multivariate generalizations of classical distribution-free, rank-based methods for Goodness-of-Fit testing, two-sample testing, independence testing, testing for symmetry, etc.

Suggests a way to define multivariate C.D.F.s and quantiles

Given a **reference density** q and a multivariate density p:

- The OT map from p to q is called the **multivariate C.D.F.** of p
- The OT map from q to p is called the **multivariate quantile** of p.

Outlook and Open Problems

Outlook and Open Problems

Optimal transport has become popular in statistics/HEP-ex because it:

- Provides a canonical way to transport probability distributions
- Stays faithful to the underlying geometry of the space (via the choice of c).
- Yields a metric between distributions for which smoothing is not needed.
- Generalizes traditional statistical notions related to monotonicity (quantiles, CDFs, etc.).
- …

Outlook and Open Problems

Optimal transport has become popular in statistics/HEP-ex because it:

- Provides a canonical way to transport probability distributions
- Stays faithful to the underlying geometry of the space (via the choice of c).
- Yields a metric between distributions for which smoothing is not needed.
- Generalizes traditional statistical notions related to monotonicity (quantiles, CDFs, etc.).
- …

Many open problems remain!

Outlook and Open Problems

Optimal transport has become popular in statistics/HEP-ex because it:

- Provides a canonical way to transport probability distributions
- Stays faithful to the underlying geometry of the space (via the choice of c).
- Yields a metric between distributions for which smoothing is not needed.
- Generalizes traditional statistical notions related to monotonicity (quantiles, CDFs, etc.).
- …

Many open problems remain!

- *Computationally and statistically efficient estimators of OT maps?*

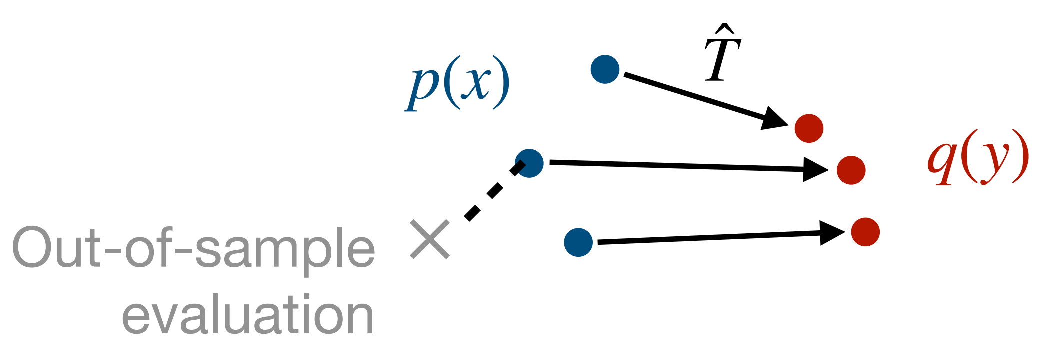
Outlook and Open Problems

Optimal transport has become popular in statistics/HEP-ex because it:

- Provides a canonical way to transport probability distributions
- Stays faithful to the underlying geometry of the space (via the choice of c).
- Yields a metric between distributions for which smoothing is not needed.
- Generalizes traditional statistical notions related to monotonicity (quantiles, CDFs, etc.).
- …

Many open problems remain!

- *Computationally and statistically efficient estimators of OT maps?*
  - "Map-then-smooth estimators"



"Map-then-smooth"

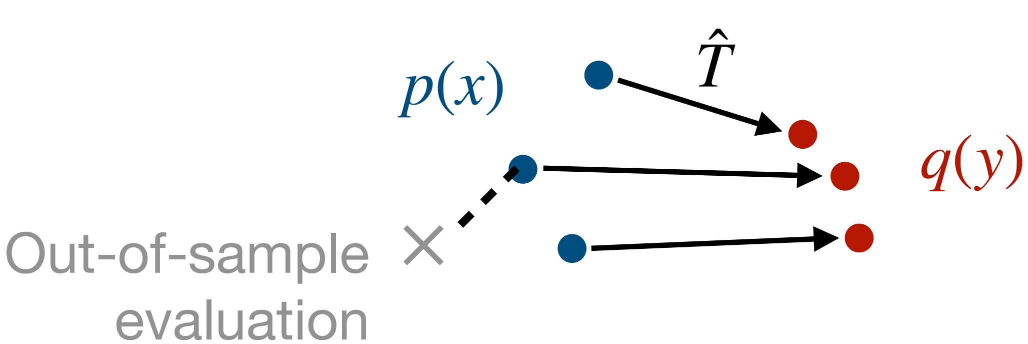
Outlook and Open Problems

Optimal transport has become popular in statistics/HEP-ex because it:

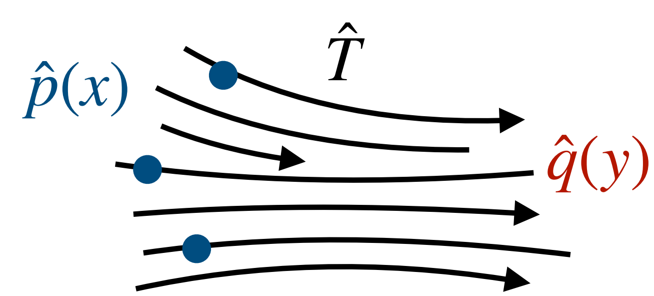
- Provides a canonical way to transport probability distributions
- Stays faithful to the underlying geometry of the space (via the choice of c).
- Yields a metric between distributions for which smoothing is not needed.
- Generalizes traditional statistical notions related to monotonicity (quantiles, CDFs, etc.).
- …

Many open problems remain!

- *Computationally and statistically efficient estimators of OT maps?*
  - "Map-then-smooth estimators"
  - "Smooth-then-map estimators"



"Map-then-smooth"



"Smooth-then-map"

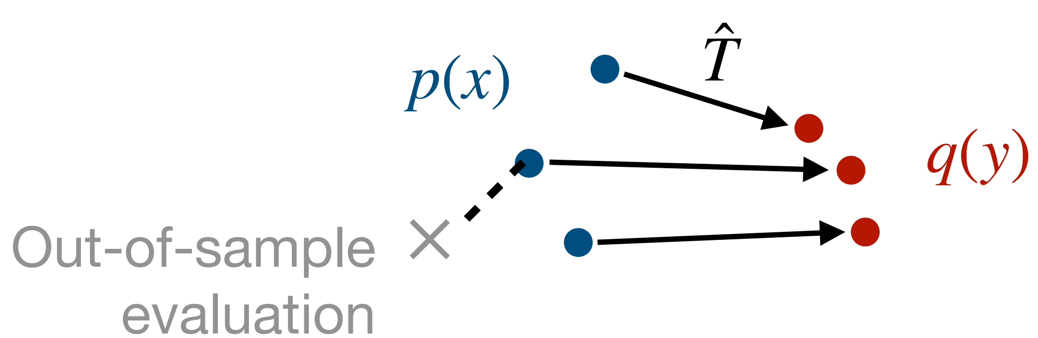
Outlook and Open Problems

Optimal transport has become popular in statistics/HEP-ex because it:

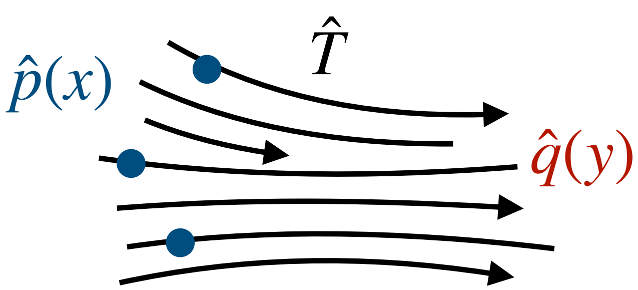
- Provides a canonical way to transport probability distributions
- Stays faithful to the underlying geometry of the space (via the choice of c).
- Yields a metric between distributions for which smoothing is not needed.
- Generalizes traditional statistical notions related to monotonicity (quantiles, CDFs, etc.).
- …

Many open problems remain!

- <u>*Computationally and statistically efficient estimators of OT maps?*</u>
  - "Map-then-smooth estimators"
  - "Smooth-then-map estimators"
  - Other heuristics: input convex neural networks, etc.



"Map-then-smooth"



"Smooth-then-map"

Outlook and Open Problems

Optimal transport has become popular in statistics/HEP-ex because it:

- Provides a canonical way to transport probability distributions
- Stays faithful to the underlying geometry of the space (via the choice of c).
- Yields a metric between distributions for which smoothing is not needed.
- Generalizes traditional statistical notions related to monotonicity (quantiles, CDFs, etc.).
- …

Many open problems remain!

- *Computationally and statistically efficient estimators of OT maps?*
  - "Map-then-smooth estimators"
  - "Smooth-then-map estimators"
  - Other heuristics: input convex neural networks, etc.

- *Quantifying statistical uncertainty for OT maps?*
  - For smooth-then-map estimators, we recently showed that (under regularity conditions!), for some $\Sigma_n(x)$,

$$\Sigma_n(x)\left(\hat{T}_n(x) - T(x)\right) \rightsquigarrow N(0, I_d)\,.$$

  - Does this hold for more practical estimators?
  - Is the bootstrap valid?

References

1. Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society. Series B, 81*.

2. Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, *8*.

3. Chernozhukov, V., Galichon, A., Hallin, M., & Henry, M. (2017). Monge-Kantorovich depth, quantiles, ranks and signs. *Annals of Statistics*, *45*(1), 223-256.

4. Deb, N., Bhattacharya, B. B., & Sen, B. (2021). Efficiency lower bounds for distribution-free Hotelling-type two-sample tests based on optimal transport. *arXiv preprint arXiv:2104.01986*.

5. Flamary, R., Courty, N., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, *1*.

6. Hallin, M., Gilles M., and Johan S. Multivariate goodness-of-fit tests based on Wasserstein distance. (2021) *Electronic Journal of Statistics* 15.

7. Hallin, M., Del Barrio, E., Cuesta-Albertos, J., & Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics, 49.*

8. Komiske, P. T., Metodiev, E. M., & Thaler, J. (2019). Metric space of collider events. *Physical Review Letters*, *123*.

9. Makkuva, A., Taghvaei, A., Oh, S., & Lee, J. (2020). Optimal transport mapping via input convex neural networks. *International Conference on Machine Learning 37*.

10. Manole, T., Bryant, P., Alison, J., Kuusela, M., & Wasserman, L. (2022). Background Modeling for Double Higgs Boson Production: Density Ratios and Optimal Transport. *arXiv preprint arXiv:2208.02807*.

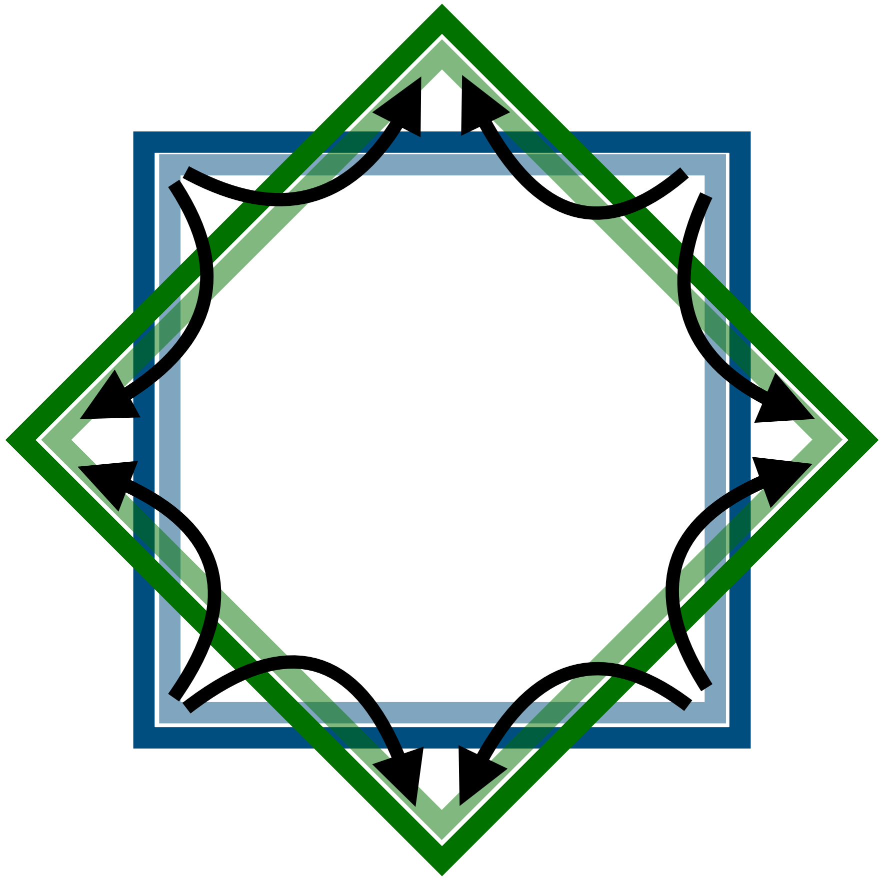
11. Manole, T., Balakrishnan, S., Niles-Weed, J., & Wasserman, L. (2021). Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*.

12. Peyré, G., & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, *11*.

13. Pollard, C., & Windischhofer, P. (2022). Transport away your problems: Calibrating stochastic simulations with optimal transport. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *1027*.

14. Read, A. L. (1999). Linear interpolation of histograms. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 425(1-2), 357-360.

15. Sommerfeld, M., & Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society. Series B*, *80*.



Backup

What is optimal transportation?

The answer to a logistics problem!

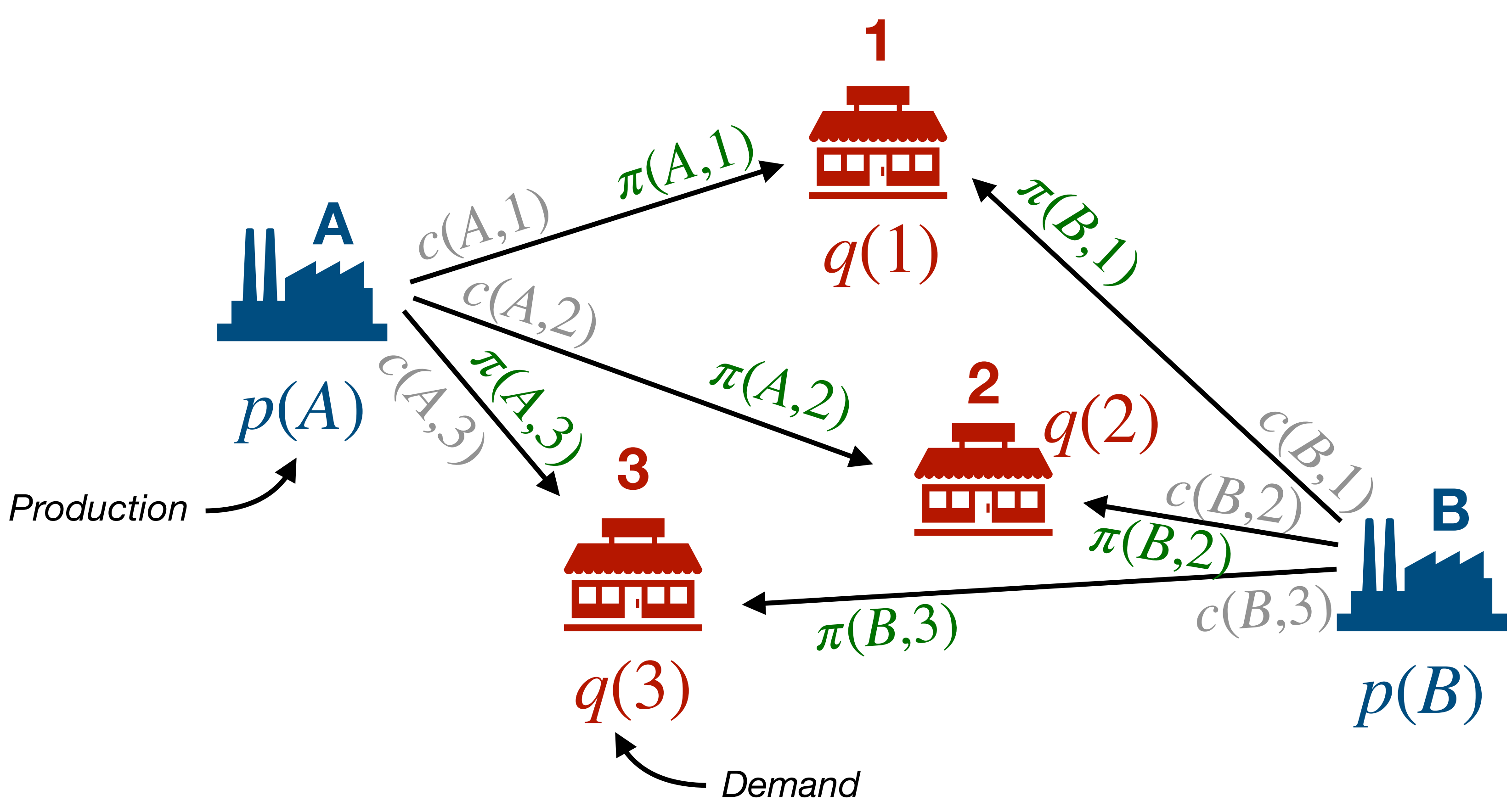
$$\hat{\pi} = \arg\min_{\pi} \sum_{a} \sum_{i} \pi(a, i)\, c(a, i)$$

Optimal transportation plan → $\hat{\pi}$

Optimization over all possible transportation plans

Transportation cost *(per unit mass)*

Mass transported from factory a to store i *("transportation plan")*



Assume total production $p(A) + p(B)$ *equals total demand* $q(1) + q(2) + q(3)$

What is optimal transportation?

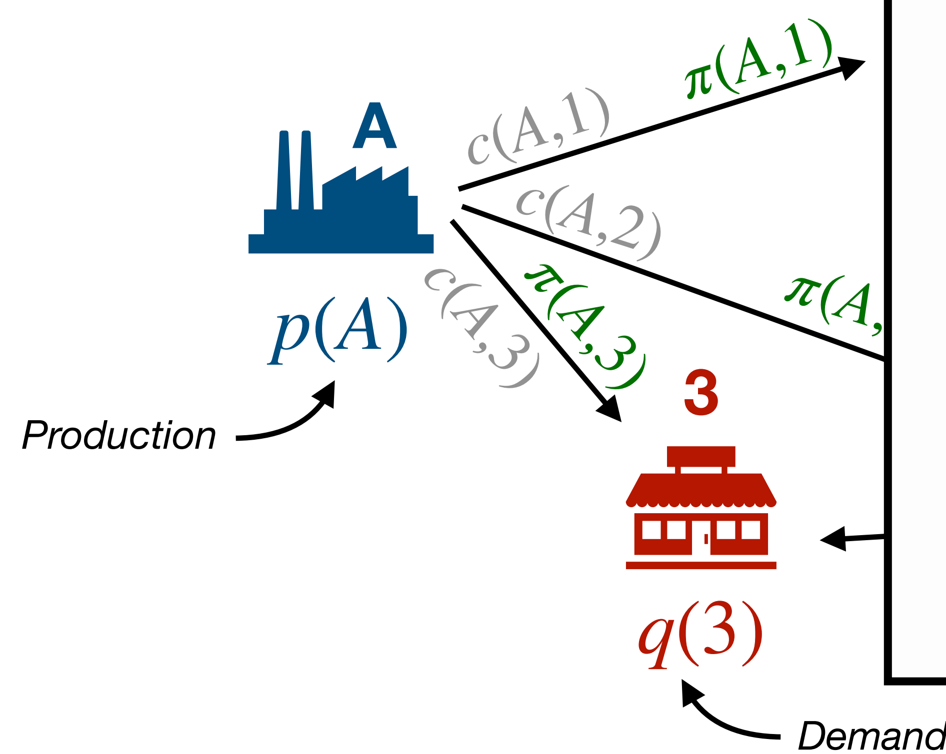
The answer to a logistics problem!

Optimal transportation plan → $\hat{\pi} = \arg\min_{\pi} \sum_a \sum_i \pi(a,i)\, c(a,i)$

Transportation cost *(per unit mass)*

Optimization over all possible transportation plans

Mass transported from factory a to store i *("transportation plan")*



"All possible" transportation plans?

Must satisfy two conditions:

$$\sum_i \pi(a,i) = p(a) \quad \forall a$$

"Entire production is shipped"

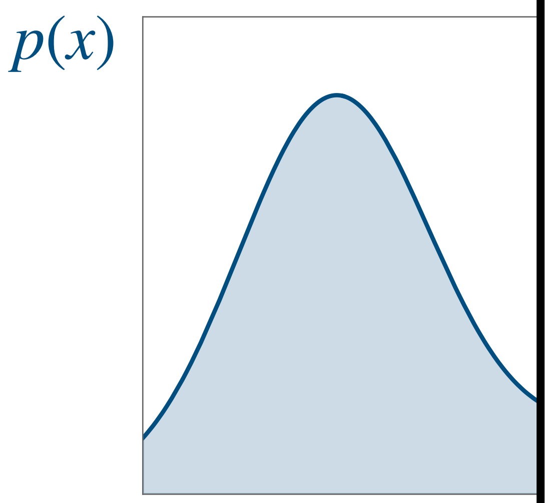
$$\sum_a \pi(a,i) = q(i) \quad \forall i$$

"Entire demand is satisfied"

Assume total production $p(A) + p(B)$ *equals total demand* $q(1) + q(2) + q(3)$

Optimal transport, now continuous

How about a continuous **distribution of production** $p(x)$ and a **continuous distribution of demand** $q(y)$?



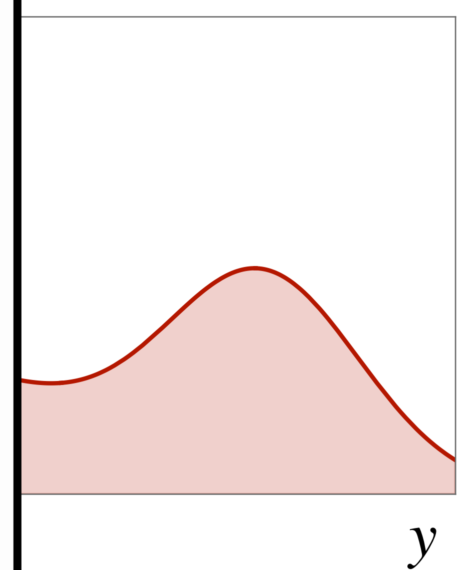
Remember: the marginals of any admissible transport plan must give the source and target distributions:

$$\int dy \; \pi(x, y) = p(x)$$

"Entire mass picked up"

$$\int dx \; \pi(x, y) = q(y)$$

"Entire mass delivered"



Cost to transport one unit of mass from x to y: $c(x, y)$

Transport plan: move an amount $\pi(x, y)$ from x to y

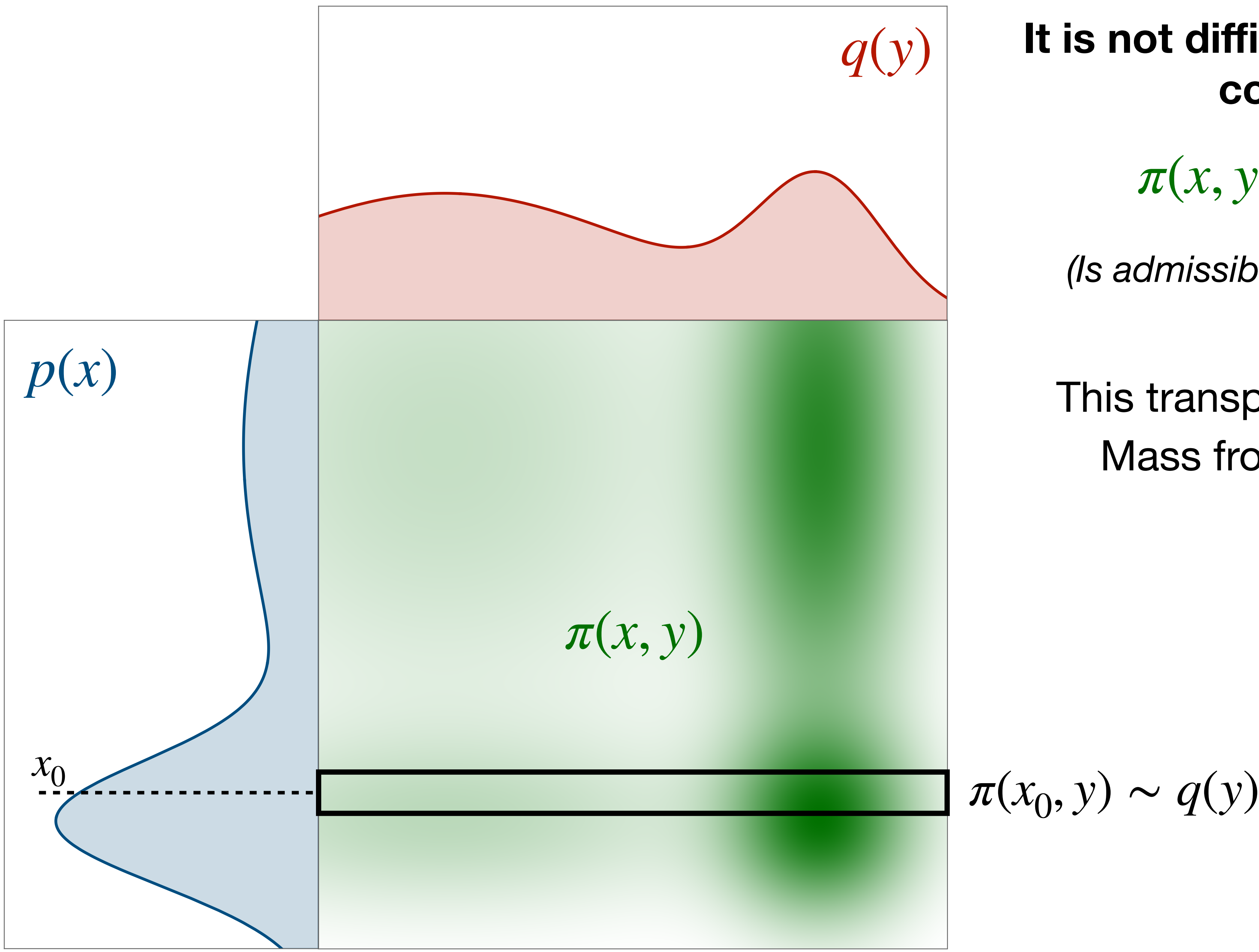
Transport plan with minimal cost:

$$\hat{\pi} = \arg\min_{\pi} \int dx\, dy \; \pi(x, y) \; c(x, y)$$

"Kantorovich optimal transport problem"

Optimal transport, now continuous

How about a continuous **distribution of production** $p(x)$ and a **continuous distribution of demand** $q(y)$?



It is not difficult to satisfy these constraints!

$$\pi(x, y) = p(x)\, q(y)$$

(Is admissible, but rarely minimal)

This transport plan distributes Mass from x_0 across all y

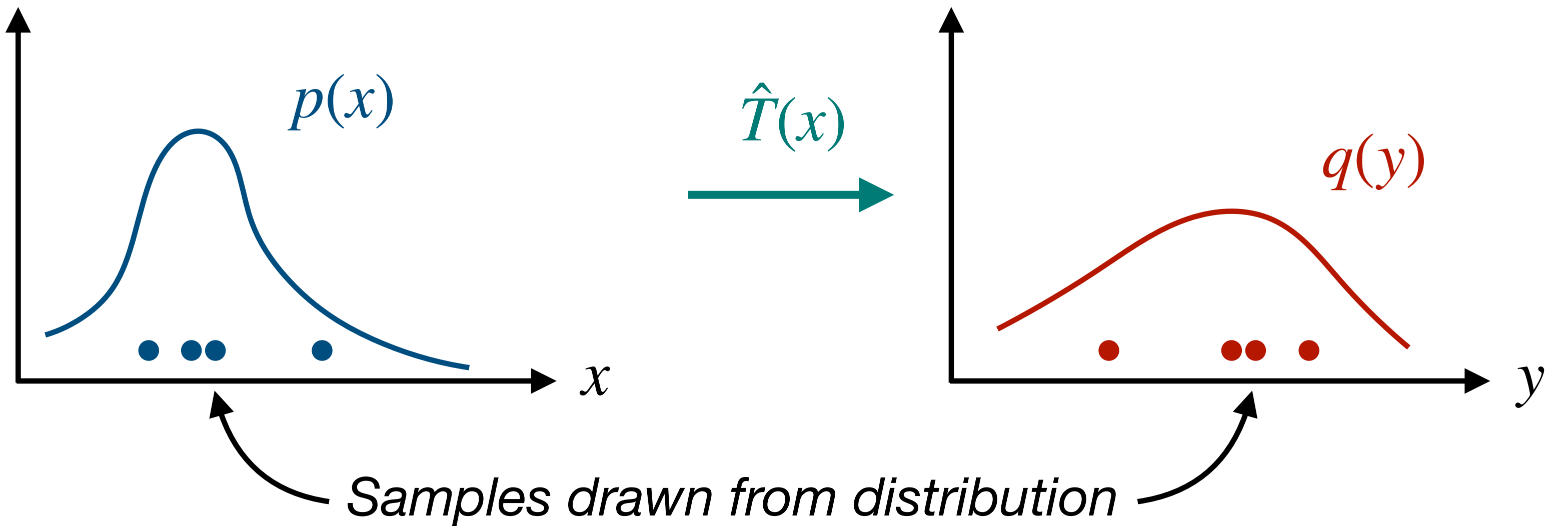
Constraints:

$$\int dy \;\; \pi(x, y) = p(x)$$

$$\int dx \;\; \pi(x, y) = q(y)$$

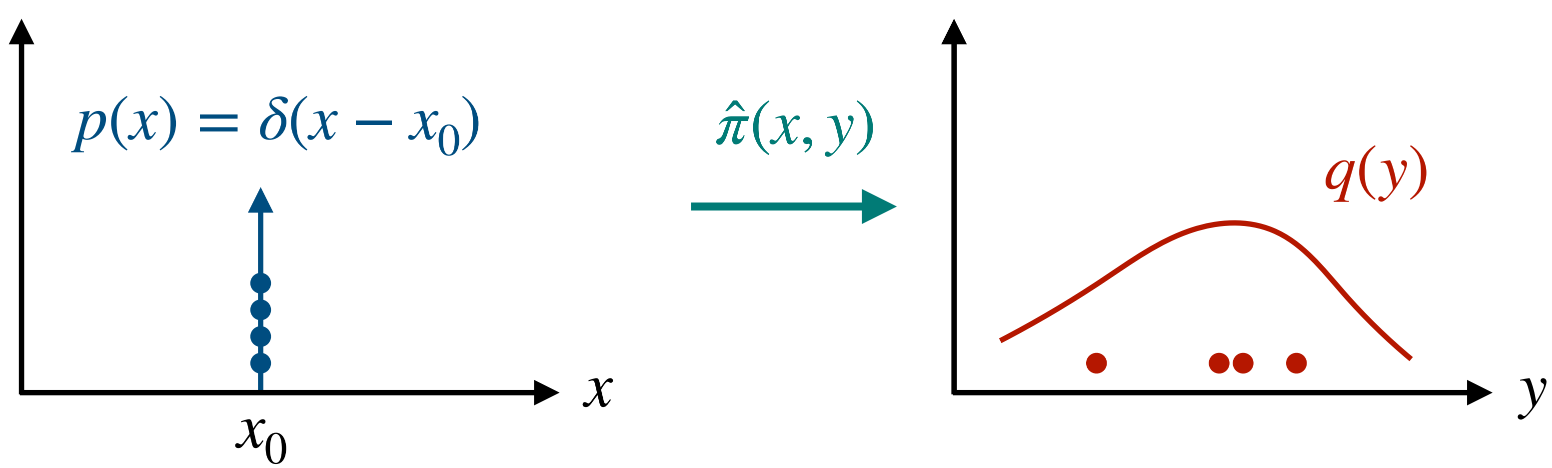
Monge vs. Kantorovich

Transport between two smooth distributions:



Deterministic transport
("reordering of samples") sufficient
→ ***Monge problem***

Transport between non-smooth and smooth distribution:



Need stochastic transport
("random smearing of samples")
→ ***Kantorovich problem***

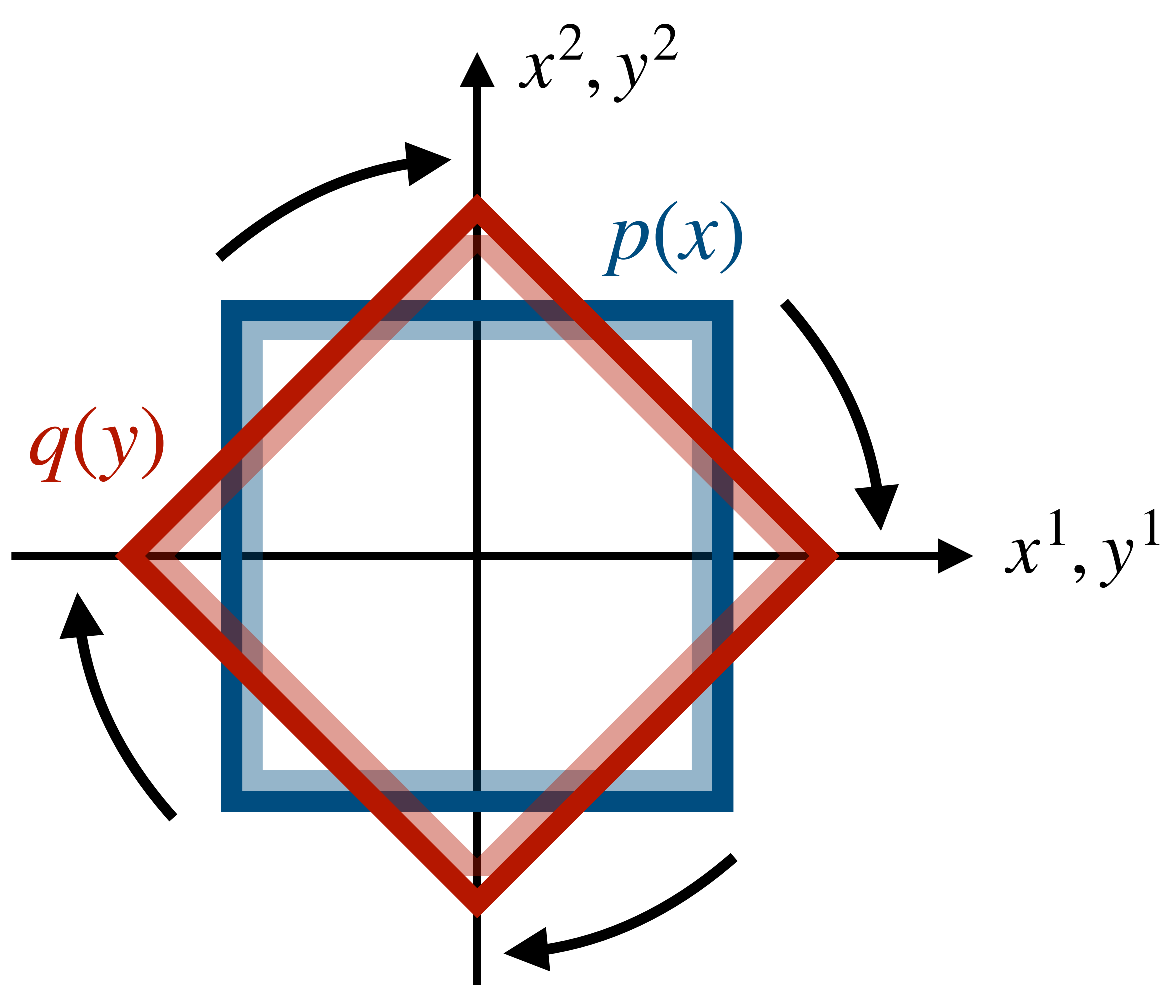
The choice of cost function

Many useful cost functions are convex!

E.g. $c(x, y) = |x - y|^p$ for $p > 1$

… let's look at a few examples!

$p = 2$, i.e. $c(x, y) = |x - y|^2$



Example:

Source distribution $p(x)$ populates inside of axis-aligned square

Target distribution $q(y)$ populates "rotated" square

But: rotation is not a gradient vector field!

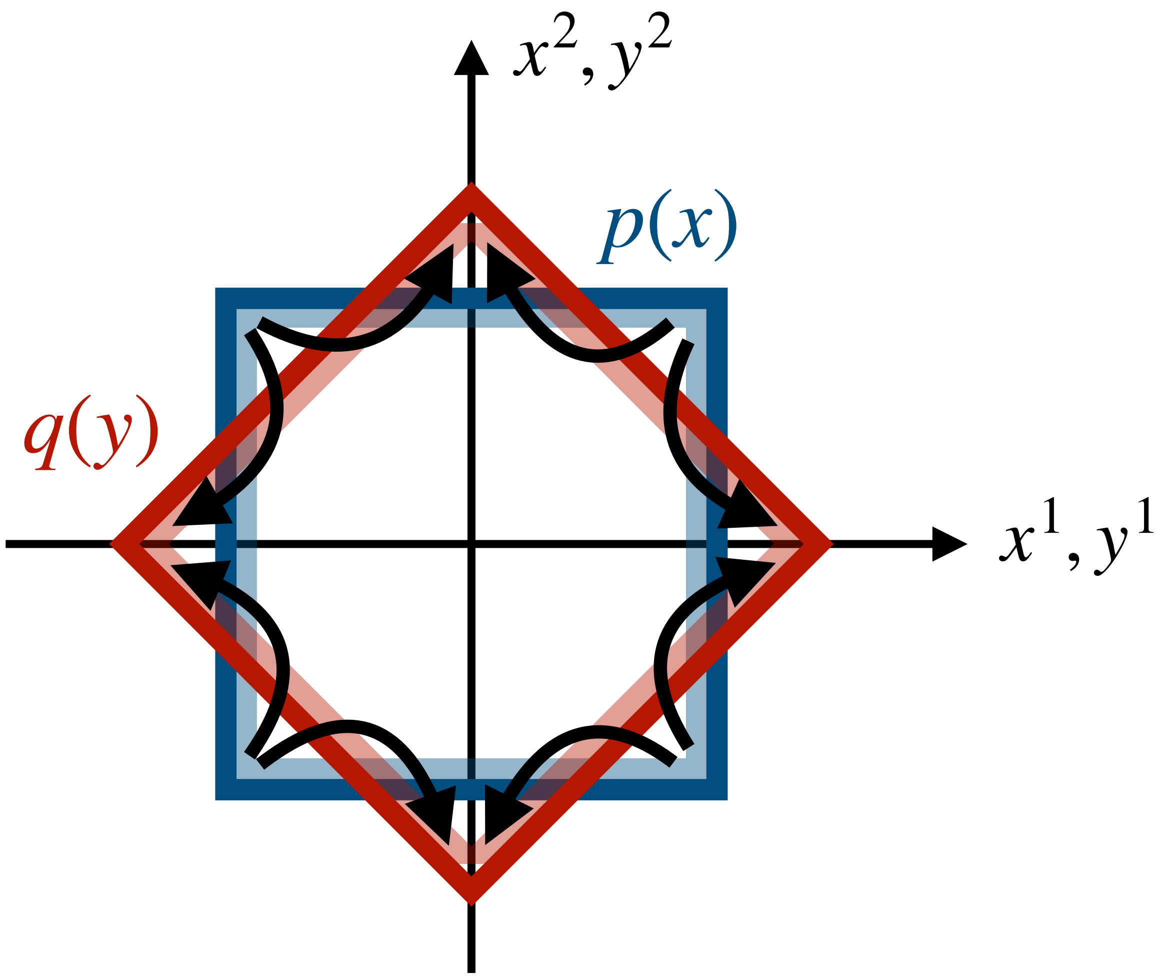
The choice of cost function

Many useful cost functions are convex!

E.g. $c(x, y) = |x - y|^p$ for $p > 1$

… let's look at a few examples!

$p = 2$, i.e. $c(x, y) = |x - y|^2$



Example:

Source distribution $p(x)$ populates inside of axis-aligned square

Target distribution $q(y)$ populates "rotated" square

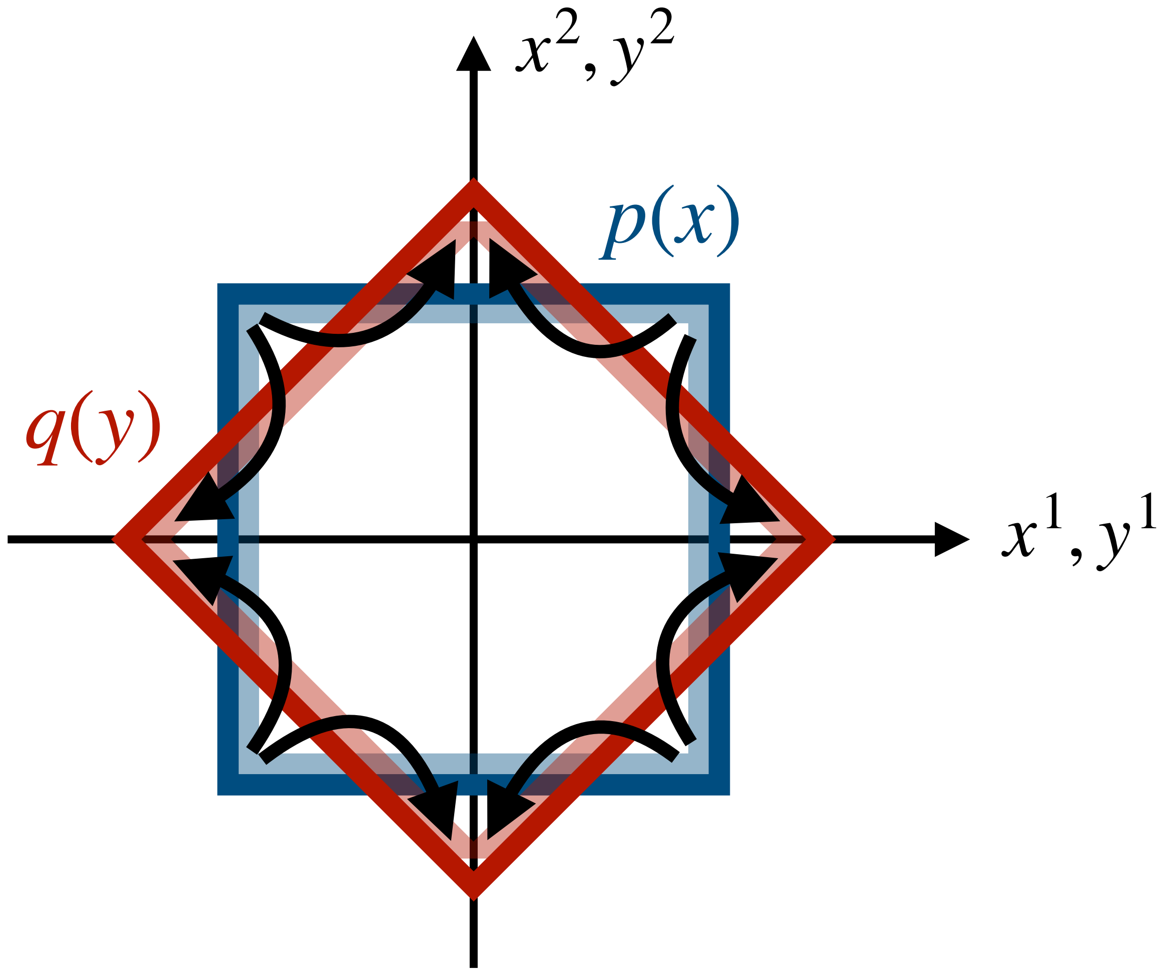
But: rotation is not a gradient vector field!

The optimal transport solution looks like this

Calibrating simulations: the right cost function

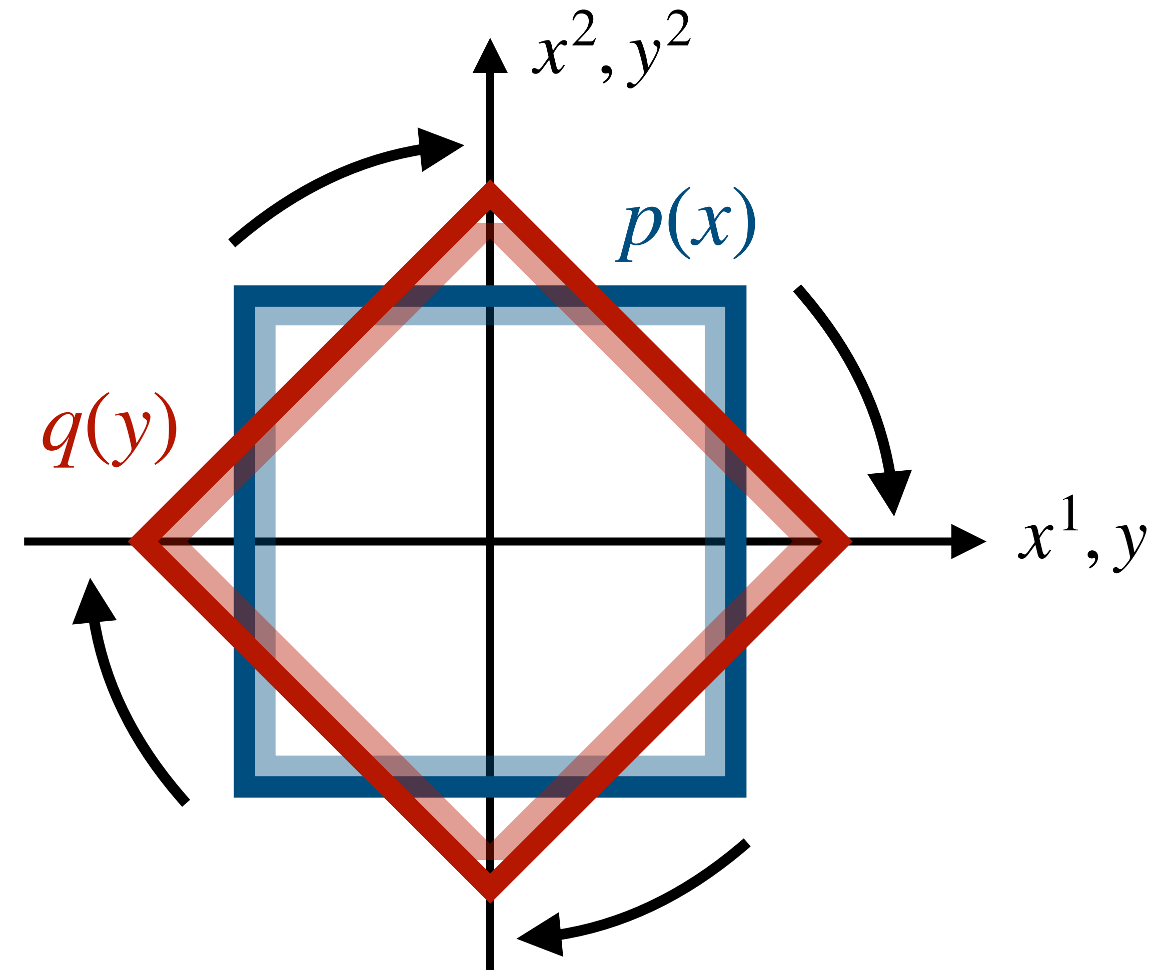
Example from before: simulation of a square, but rotation angle incorrectly modeled

Uncalibrated simulation **Calibration data**



Optimal in Euclidean plane

$$ds^2 = dr^2 + r^2 d\phi^2$$



Optimal on a cone manifold

$$ds^2 = \alpha^2 dr^2 + r^2 d\phi^2, \alpha > 1$$

Use this if rotational degree of freedom is known to be poorly modeled

The choice of cost function

Many useful cost functions are convex!

E.g. $c(x, y) = |x - y|^p$ for $p > 1$

… let's look at a few examples!

$p = 2$, i.e. $c(x, y) = |x - y|^2$

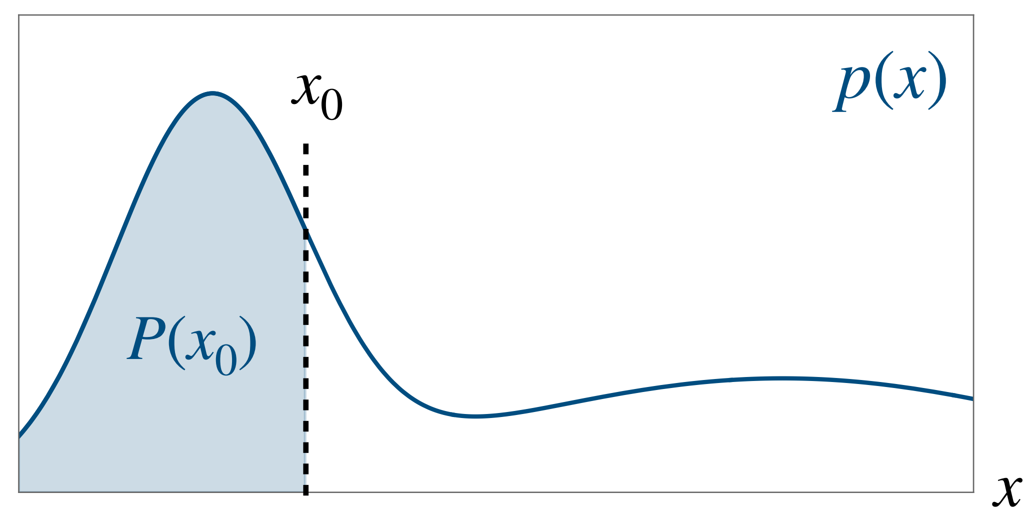
For 1-dimensional distributions:

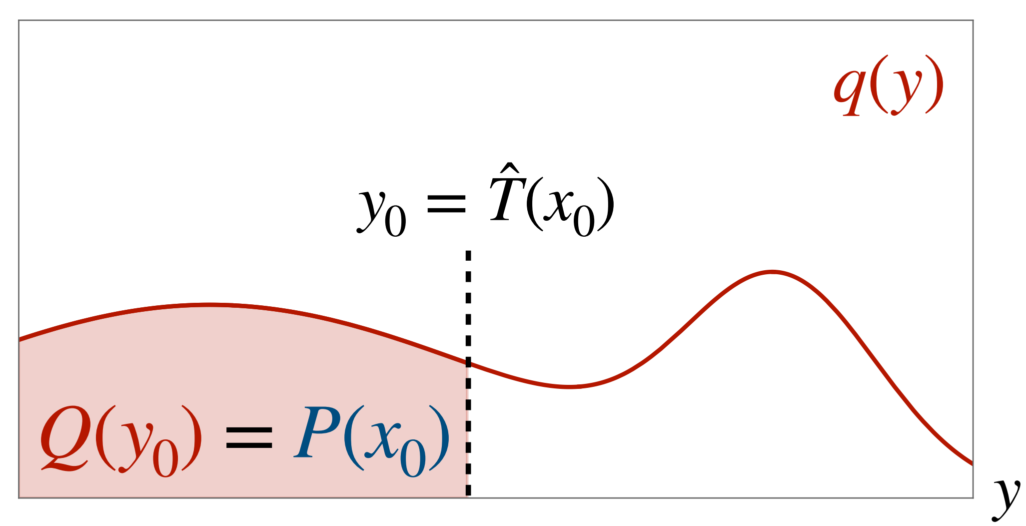
The optimal transport solution performs quantile-matching *(works for all convex cost functions!)*

$$\hat{T}(x) = Q^{-1}(P(x))$$

Cumulative distributions of $p(x)$, $q(y)$:

Generically: $F(x) = \int_0^x dx' f(x')$





The choice of cost function

Many useful cost functions are convex!

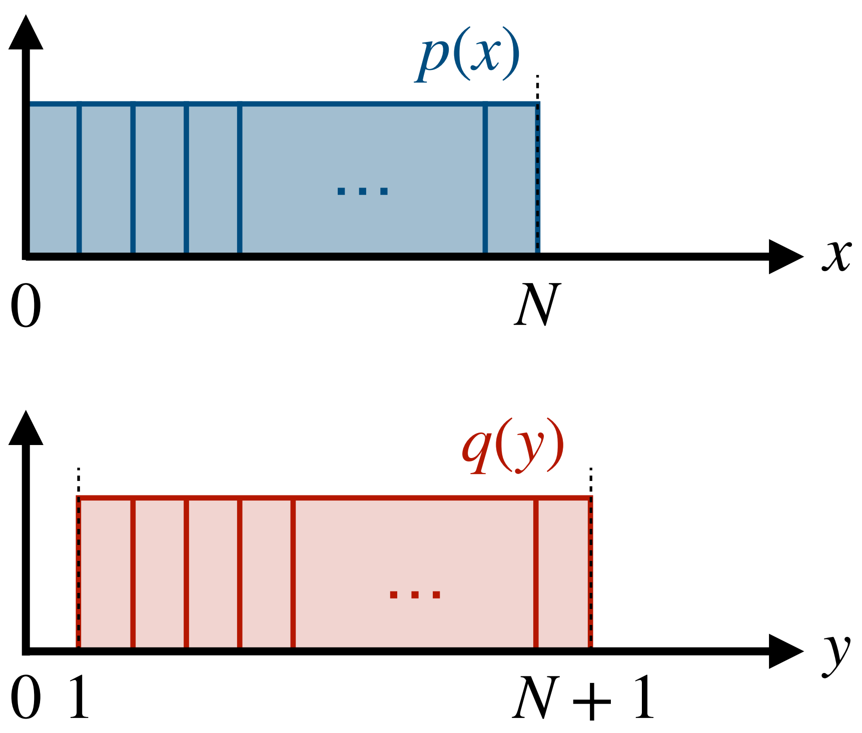
E.g. $c(x, y) = |x - y|^p$ for $p > 1$

… let's look at a few examples!

$p = 1$, i.e. $c(x, y) = |x - y|$ *(Monge's original problem)*

This is a much more complicated case!

Solutions exist for smooth distributions, but no longer unique!



Example:

Uniform source and target distributions
(e.g. rows of N books, shifted by one)

The choice of cost function

Many useful cost functions are convex!

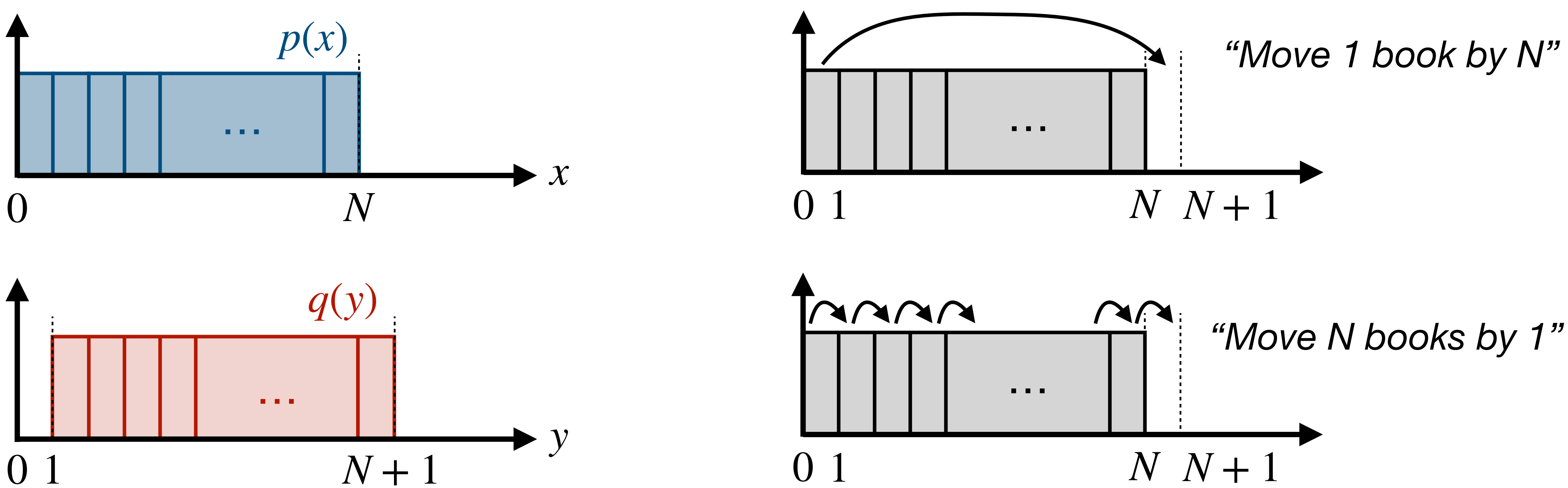
E.g. $c(x, y) = |x - y|^p$ for $p > 1$

… let's look at a few examples!

$p = 1$, i.e. $c(x, y) = |x - y|$ *(Monge's original problem)*

This is a much more complicated case!

Solutions exist for smooth distributions, but no longer unique!



A solution sketch

$\hat{T} = \nabla \hat{g}$

Monge problem

Nonlinear constraint

Equivalence for smooth distributions

Kantorovich problem

Linear constraints!

Kantorovich-Rubinstein duality

Dual Kantorovich problem

Convex constraints
→ manageable!

$$\hat{T} = \arg\min_{T} \int dx \; p(x) \; c(x, T(x))$$

$$\pi(x,y) = p(x)\, \delta[y - T(x)] \qquad q(y) = p(x) \left(\frac{dT}{dx} \right)^{-1}$$

$$\hat{\pi} = \arg\min_{\pi} \int dx\, dy \; \pi(x,y) \; c(x,y)$$

$$\int dy \; \pi(x,y) = p(x) \qquad \int dx \; \pi(x,y) = q(y)$$

$$\hat{f}, \hat{g} = \arg\max_{f,g} \int dy\, q(y)\, f(y) + \int dx\, p(x) g(x)$$

$$g(x) + f(y) \leq c(x,y)$$

The Kantorovich-Rubinstein duality

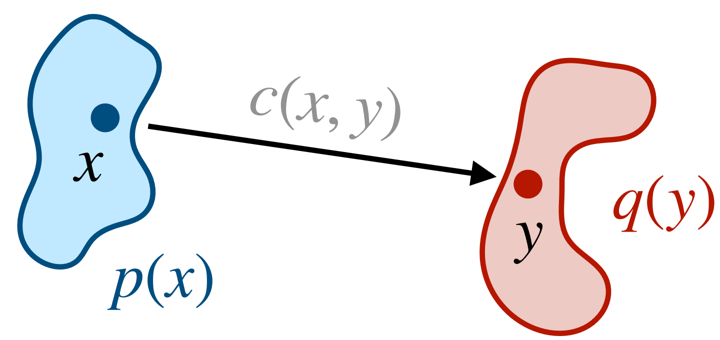
Primal problem:

$$\hat{\pi} = \arg\min_{\pi} \int dx\, dy\ \pi(x,y)\, c(x,y)$$

$$\int dy\ \pi(x,y) = p(x) \qquad \int dx\ \pi(x,y) = q(y)$$

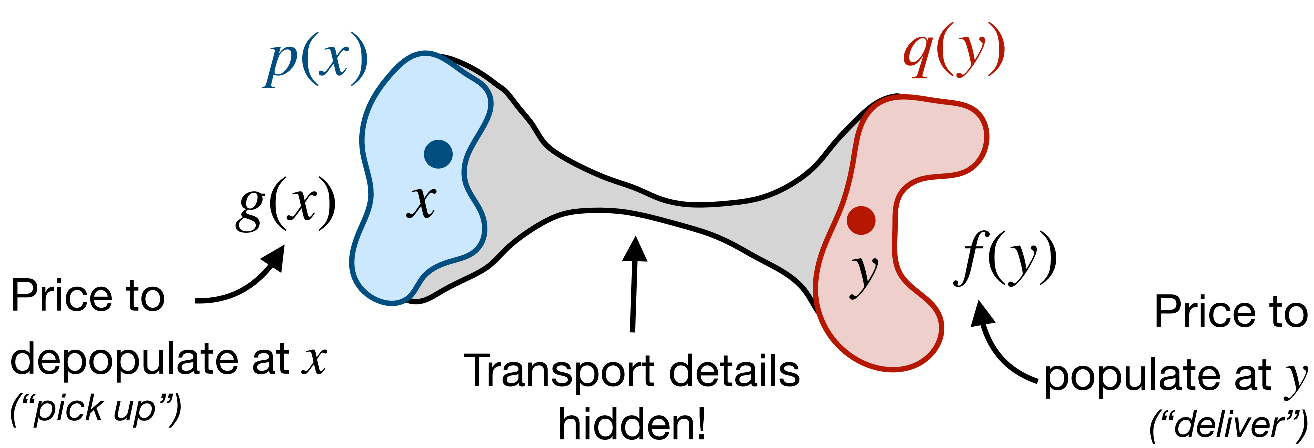
"Operative perspective":

Optimise transportation plan based on point-to-point cost $c(x, y)$



"Black-box perspective":

Optimize prices $g(x)$ and $f(y)$: maximize revenue while underbidding point-to-point transport



Dual problem:

$$\hat{f}, \hat{g} = \arg\max_{f,g} \int dy\, q(y)\, f(y) + \int dx\, p(x) g(x)$$

$$g(x) + f(y) \le c(x, y)$$

The dual problem

The dual problem is (much) easier to solve numerically:

$$\hat{f}, \hat{g} = \arg\max_{f,g} \int dy\, q(y)\, f(y) + \int dx\, p(x) g(x)$$

$$g(x) + f(y) \leq c(x, y)$$

For $c(x,y) = |x-y|^2$, $\hat{f}$ and $\hat{g}$ are Legendre-conjugates!

Legendre transform in classical mechanics:

$$H(p) + L(\dot{q}) = p\dot{q}$$

Hamiltonian *Lagrangian*

$$\hat{g} = \arg\max_{g \in \text{cvx}} \int dy\, q(y)\, g^*(y) + \int dx\, p(x) g(x)$$

Legendre transform: $g^*(y) = \max_x \left[x \cdot y - g(x) \right]$

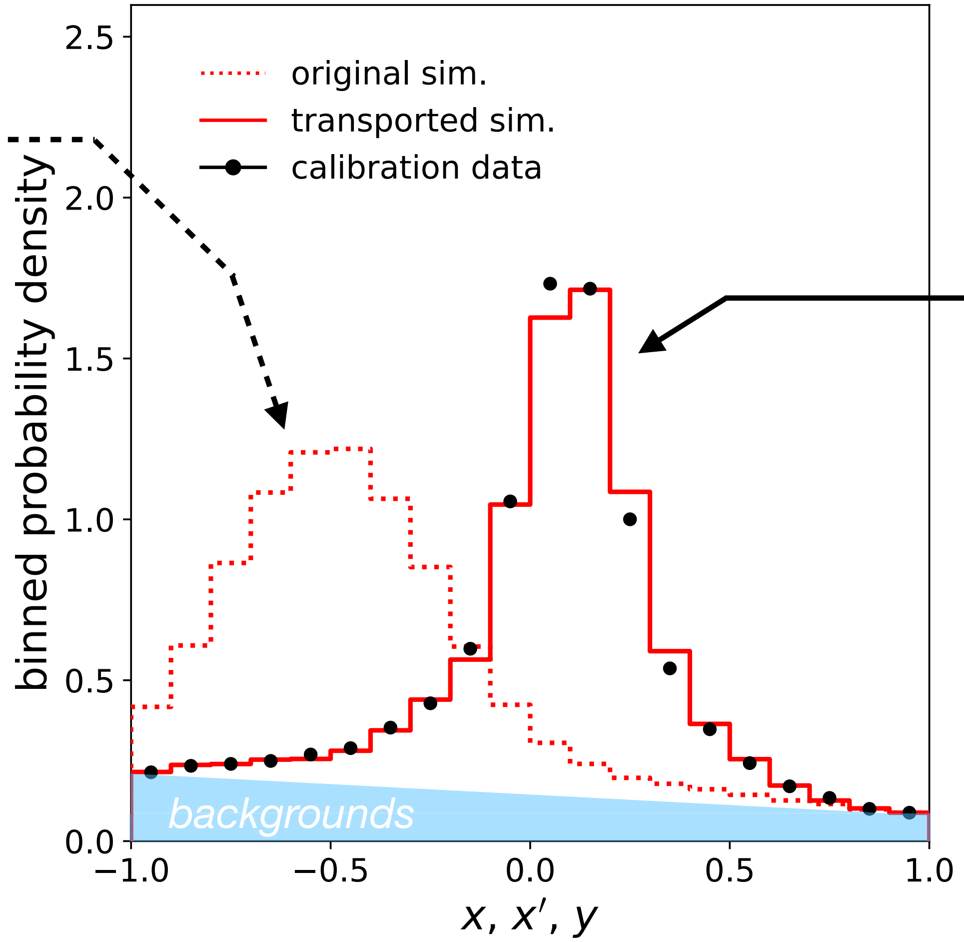
Maximise this "loss function" over all convex functions $g(x)$

Recover optimal transport function $\hat{T} = \nabla \hat{g}$

Optimal transport at colliders

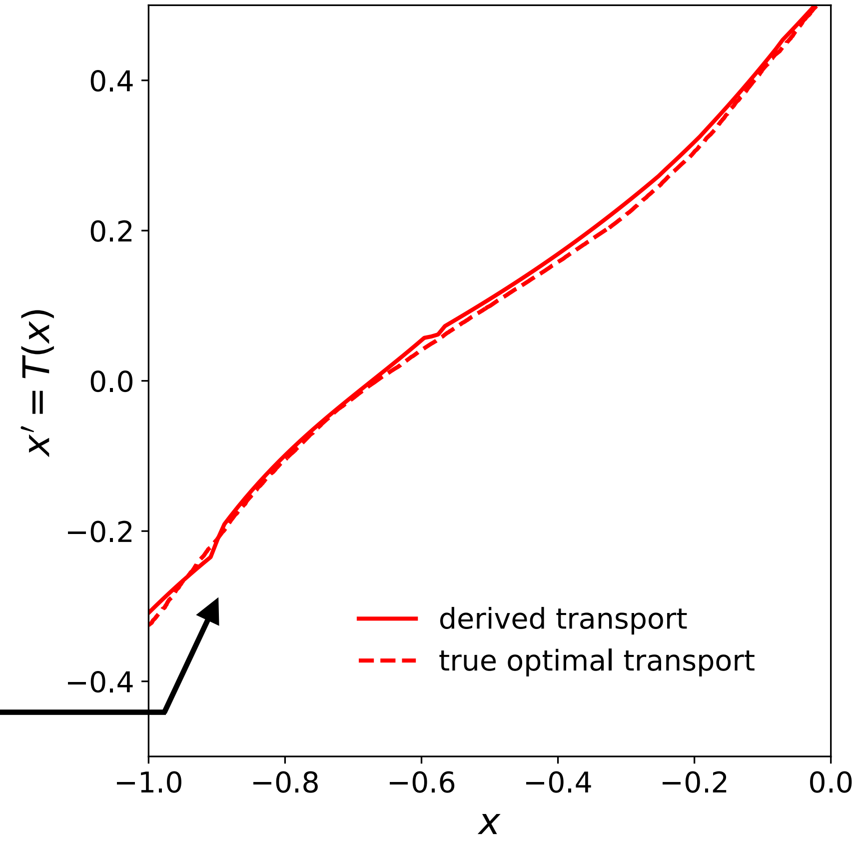
Common situation: measurement of meta-stable particle as "**resonance bump**" on top of **smooth background**

Uncalibrated simulator



Calibrated simulator

Transport function $\hat{T}$

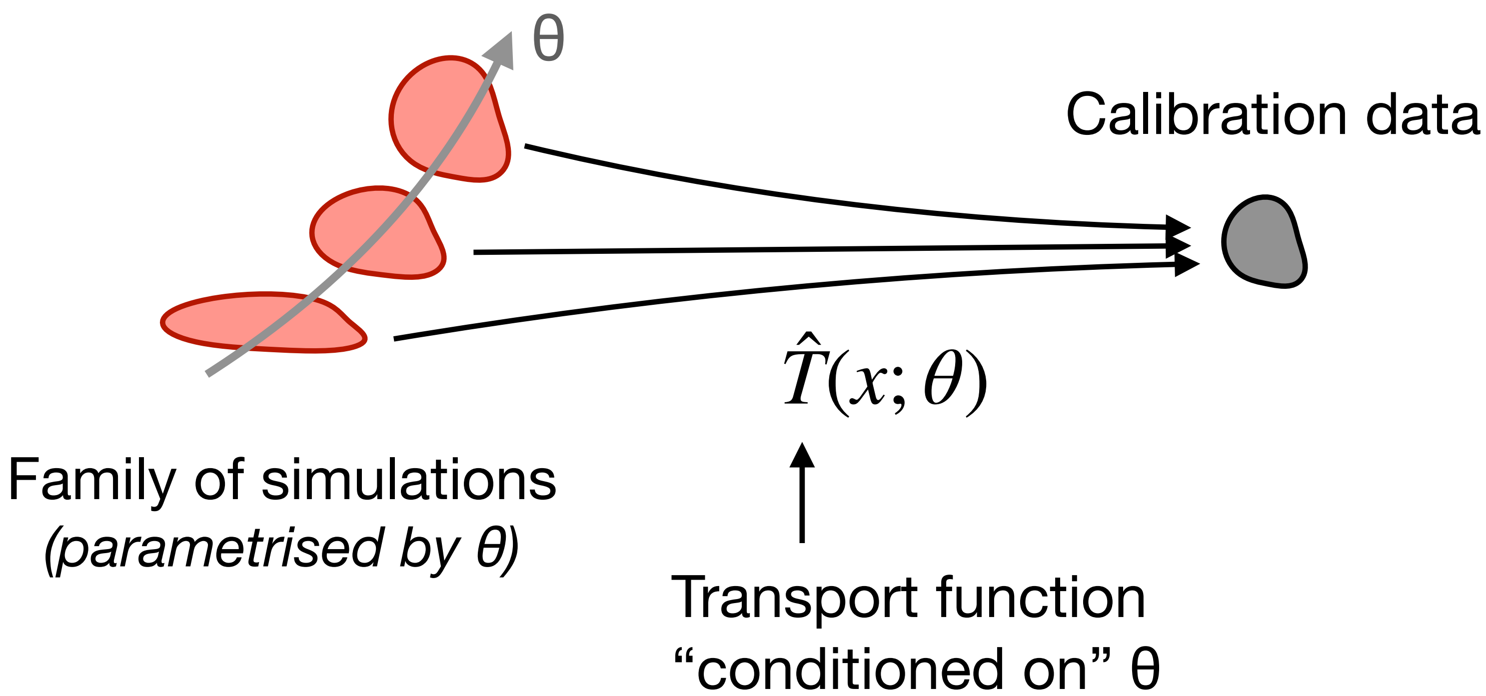


Simple **toy example** with **analytic solution**, more complex **applications in development at collider experiments**

Chris Pollard, PW, Nucl. Instrum. Meth. A, 1027 (2022) 166119 [link]

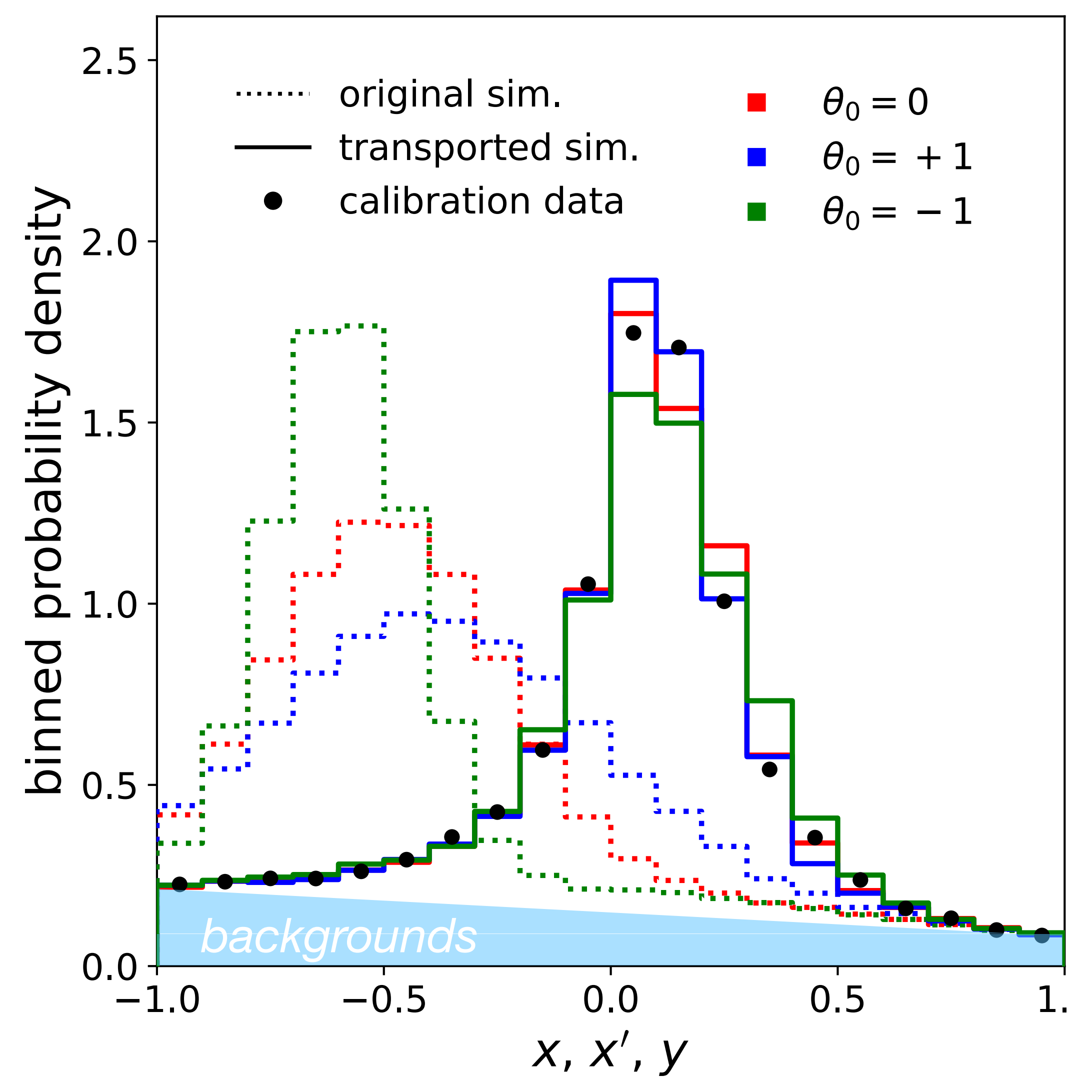
Systematic uncertainties

Simulations typically have adjustable “nuisance parameters”



Integrates naturally into input-convex neural networks

(Nuisance parameters as additional inputs without convexity requirements)



Chris Pollard, PW, Nucl. Instrum. Meth. A, 1027 (2022) 166119 [link]

Some statistical applications of Wasserstein distances

Some statistical applications of Wasserstein distances

- **Goodness-of-fit Testing:** Given $X_1, \ldots, X_n \sim p$ and known q, one can test

$$H_0 : p = q, \quad H_1 : p \neq q$$

using the test statistic $W_p(P_n, q)$, where P_n is the empirical distribution.

  - Similar ideas apply to **two-sample testing**.

dMinimum-distance Estimation: Given a parametric model $(p_\theta)_{\theta \in \Theta}$ and $X_1, \ldots, X_n \sim p_{\theta_0}$, construct the following estimator for θ_0:

Some statistical applications of Wasserstein distances

- **Goodness-of-fit Testing:** Given $X_1, \ldots, X_n \sim p$ and known q, one can test

$$H_0 : p = q, \quad H_1 : p \neq q$$

using the test statistic $W_p(P_n, q)$, where P_n is the empirical distribution.

  - Similar ideas apply to **two-sample testing.**

dMinimum-distance Estimation: Given a parametric model $(p_\theta)_{\theta \in \Theta}$ and $X_1, \ldots, X_n \sim p_{\theta_0}$, construct the following estimator for θ_0:

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} W_p(P_n, p_\theta)\,.$$

Some statistical applications of Wasserstein distances

- **Goodness-of-fit Testing:** Given $X_1, \ldots, X_n \sim p$ and known q, one can test

$$H_0 : p = q, \quad H_1 : p \neq q$$

using the test statistic $W_p(P_n, q)$, where P_n is the empirical distribution.

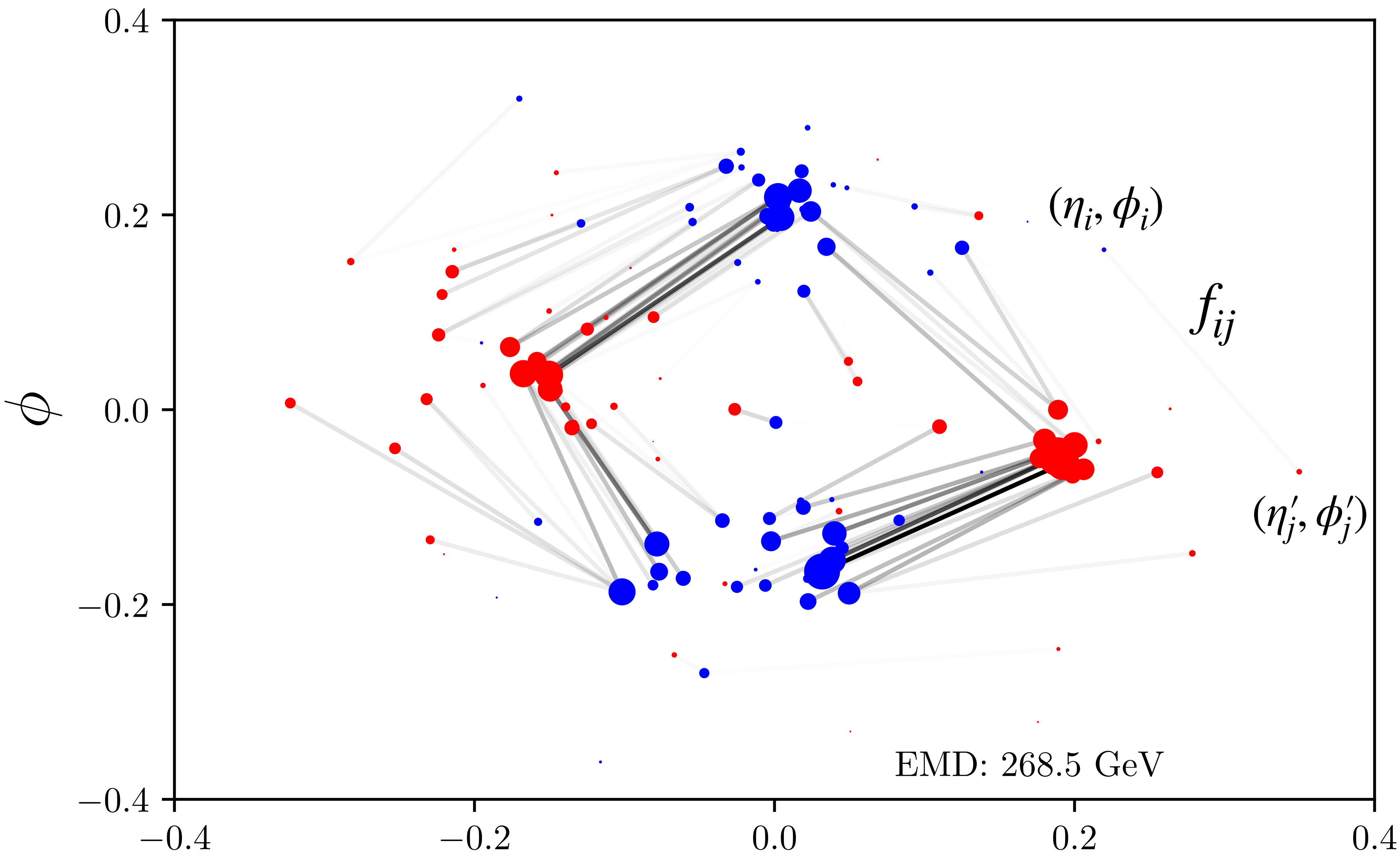
  - Similar ideas apply to **two-sample testing**.

dMinimum-distance Estimation: Given a parametric model $(p_\theta)_{\theta \in \Theta}$ and $X_1, \ldots, X_n \sim p_{\theta_0}$, construct the following estimator for θ_0:

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} W_p(P_n, p_\theta)\,.$$

Broad message: Unlike many classical metrics, the Wasserstein distance is well-defined for empirical measures, and provides a useful <u>data analytic tool</u>.

Comparing collider events (Komiske et al. 2019)



$$\mathbf{EMD}(\mathcal{E}, \mathcal{E}') = \frac{1}{R} \sum_{i,j} f_{ij} \| (\eta_i, \phi_i) - (\eta'_j, \phi'_j) \| + | s_T - s'_T |$$

The Earth Mover's Distance a.k.a. Partial OT)

$$\mathrm{EMD}(\mathcal{E}, \mathcal{E}') = \min_{\{f_{ij} \geq 0\}} \sum_{ij} f_{ij} \frac{\theta_{ij}}{R} + \left| \sum_i E_i - \sum_j E'_j \right|,$$

$$\sum_j f_{ij} \leq E_i, \qquad \sum_i f_{ij} \leq E'_j, \qquad \sum_{ij} f_{ij} = E_{\min},$$

See Komiske et al., 2019.