Optimal Transport and the Geometry of Collider Data

Katy and Nathaniel Craig
UC Santa Barbara
UCSB

Based in part on

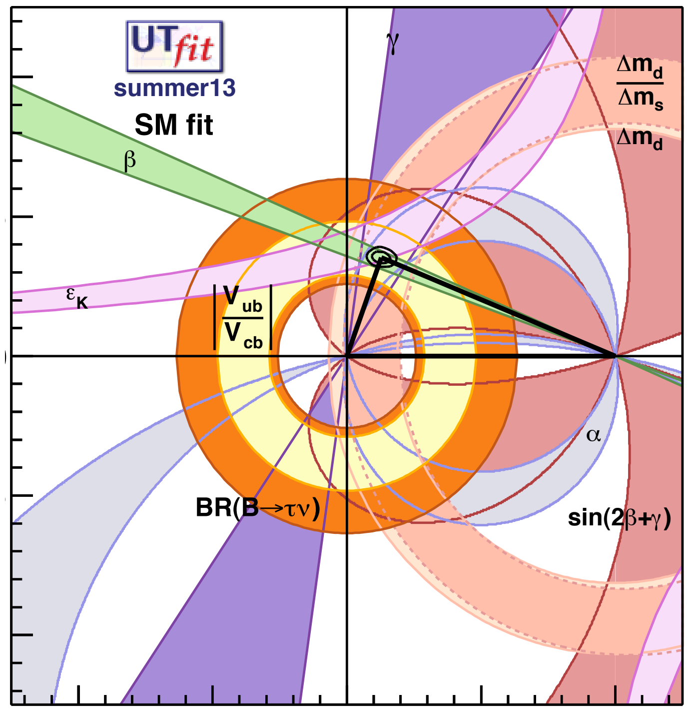
“Linearized Optimal Transport for Collider Events” [arXiv:2008.08604]
w/ **Tianji Cai & Junyi Cheng**

“The Linearized Hellinger-Kantorovich Distance” [arXiv: 2102.08807]
by **T. Cai, J. Cheng**, B. Schmitzer, M. Thorpe

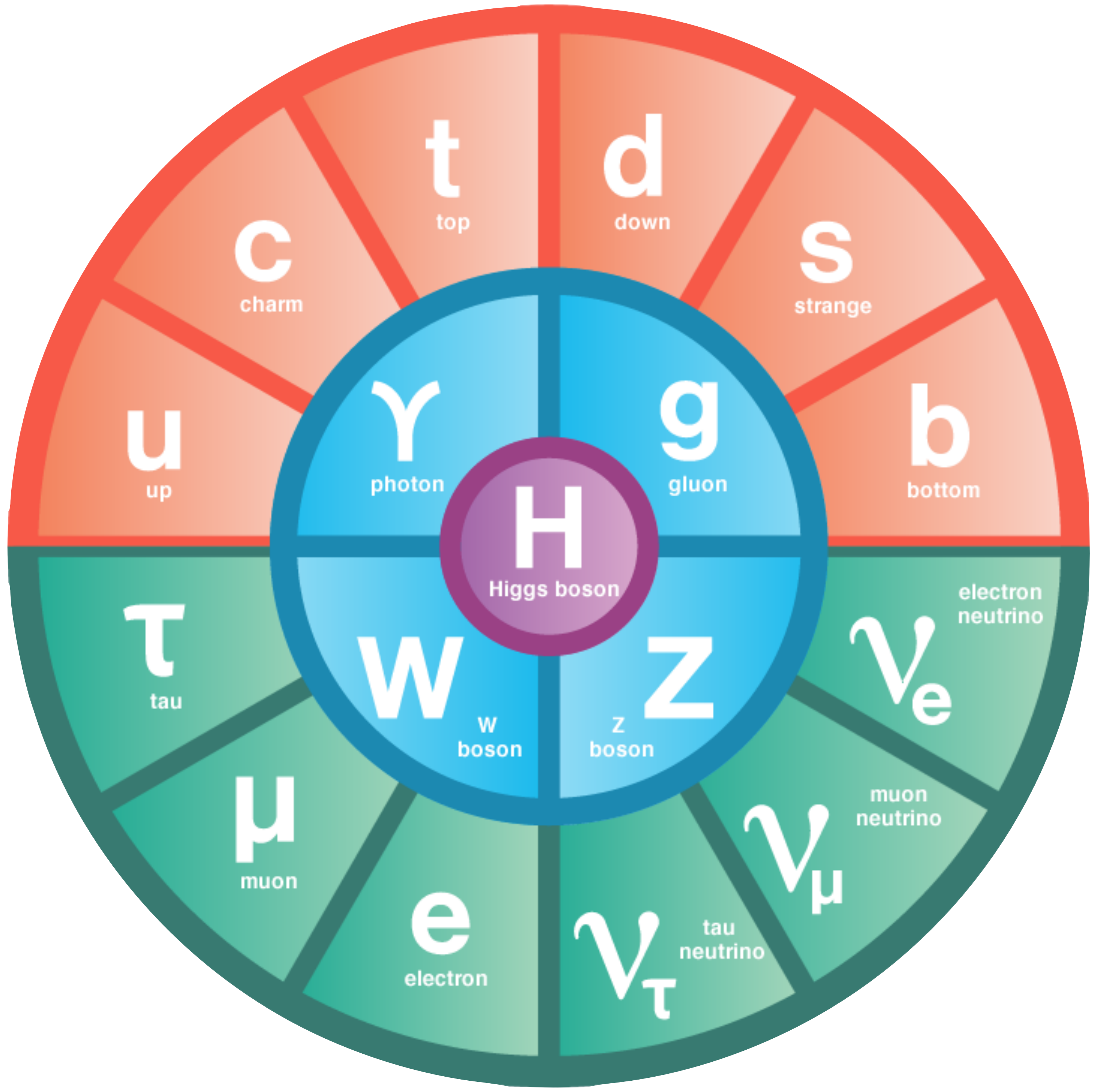
“Which Metric on the Space of Collider Events?” [arXiv: 2111.03670]
w/ **Tianji Cai & Junyi Cheng**

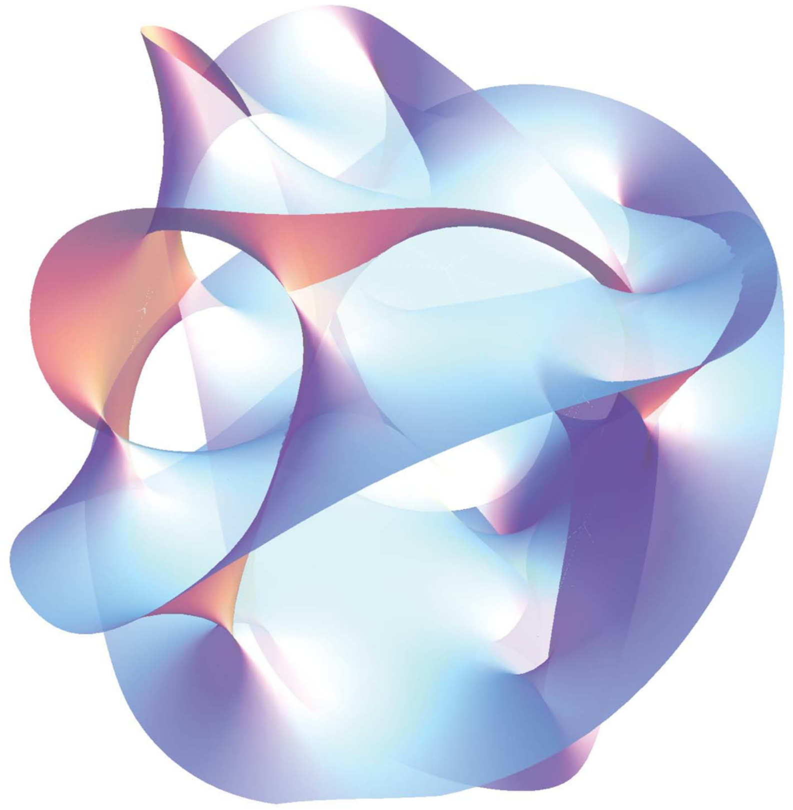

Berkeley Machine Learning and Science Forum, December 6th, 2021

The Standard Model & Beyond

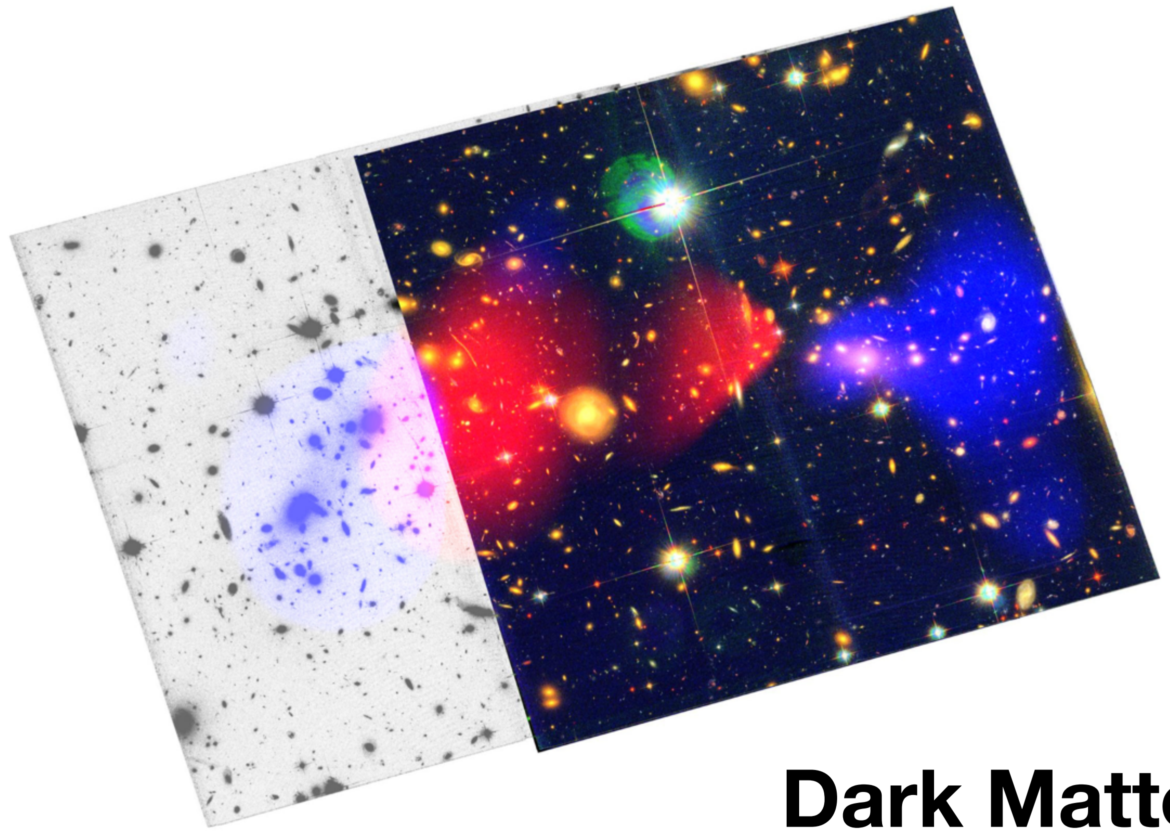


Flavor





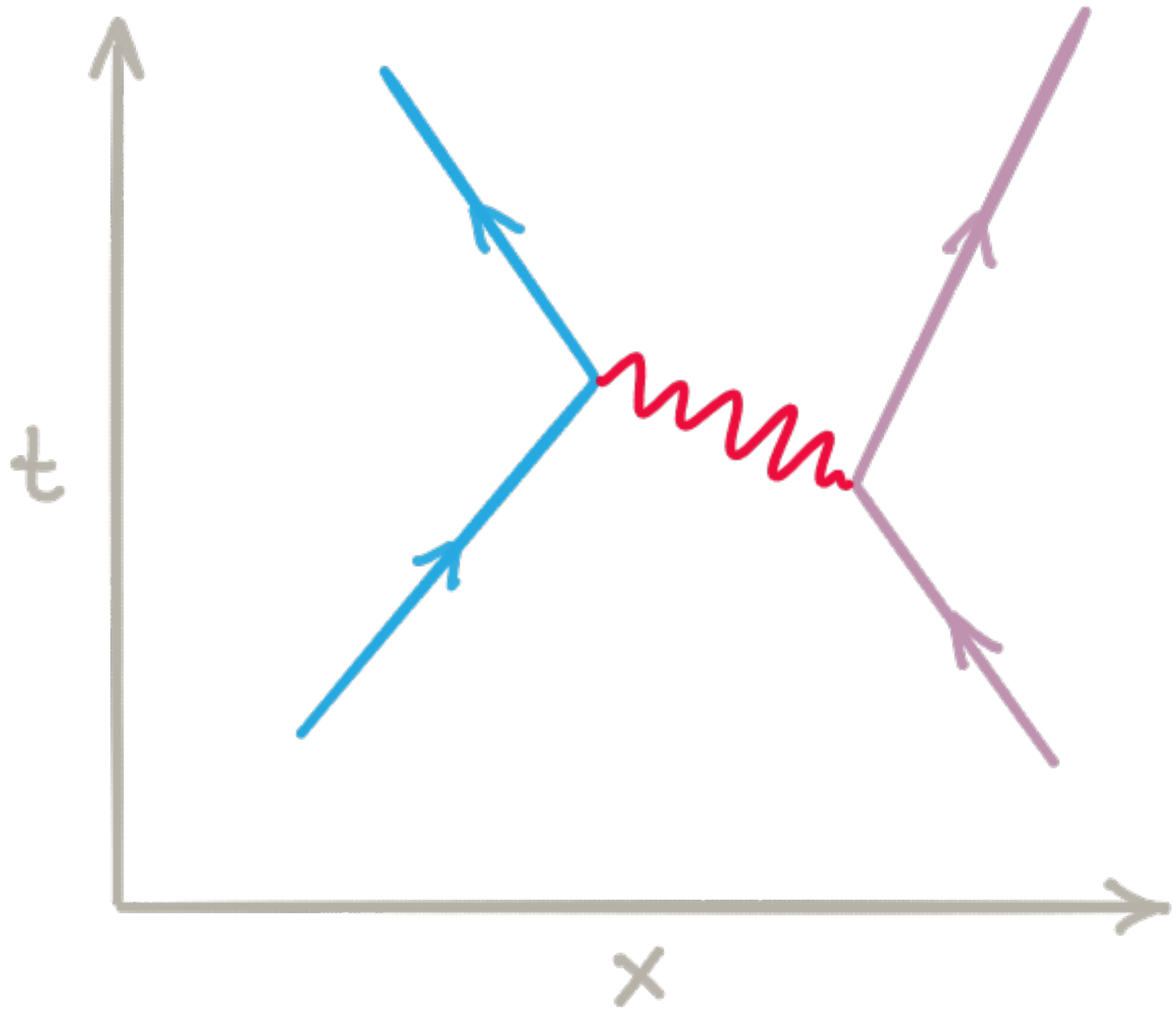
Extra dimensions

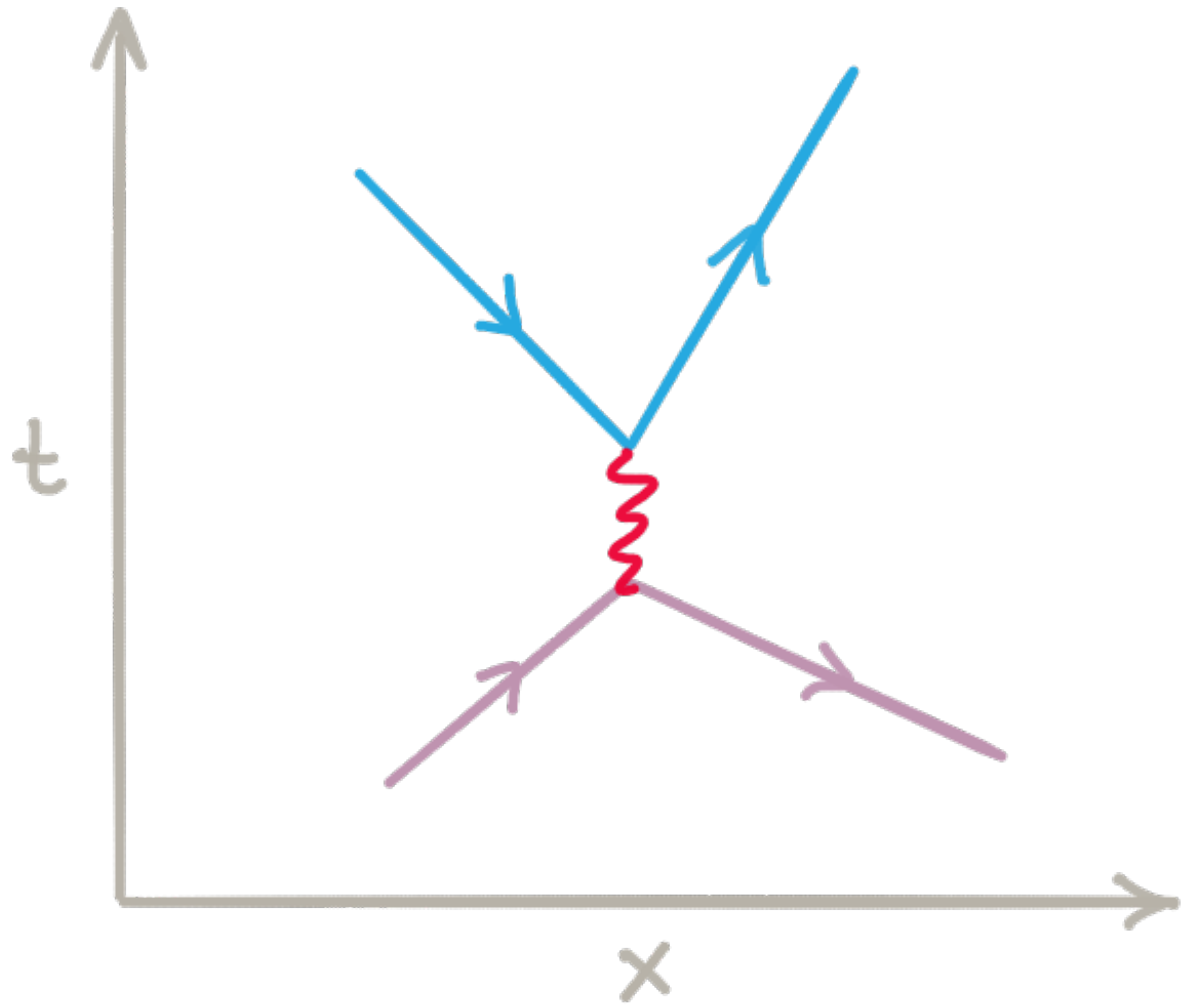


Dark Matter

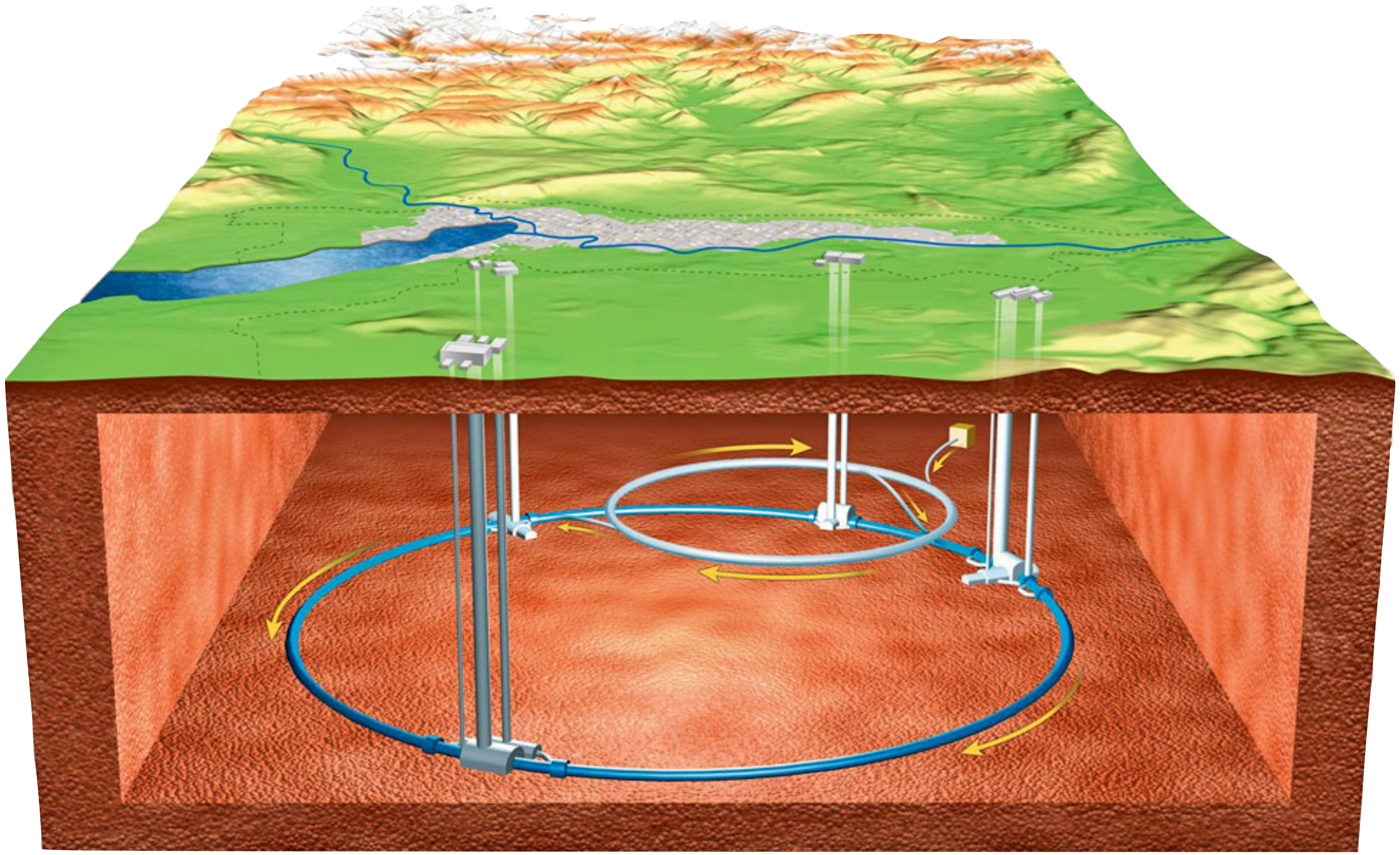
Supersymmetry

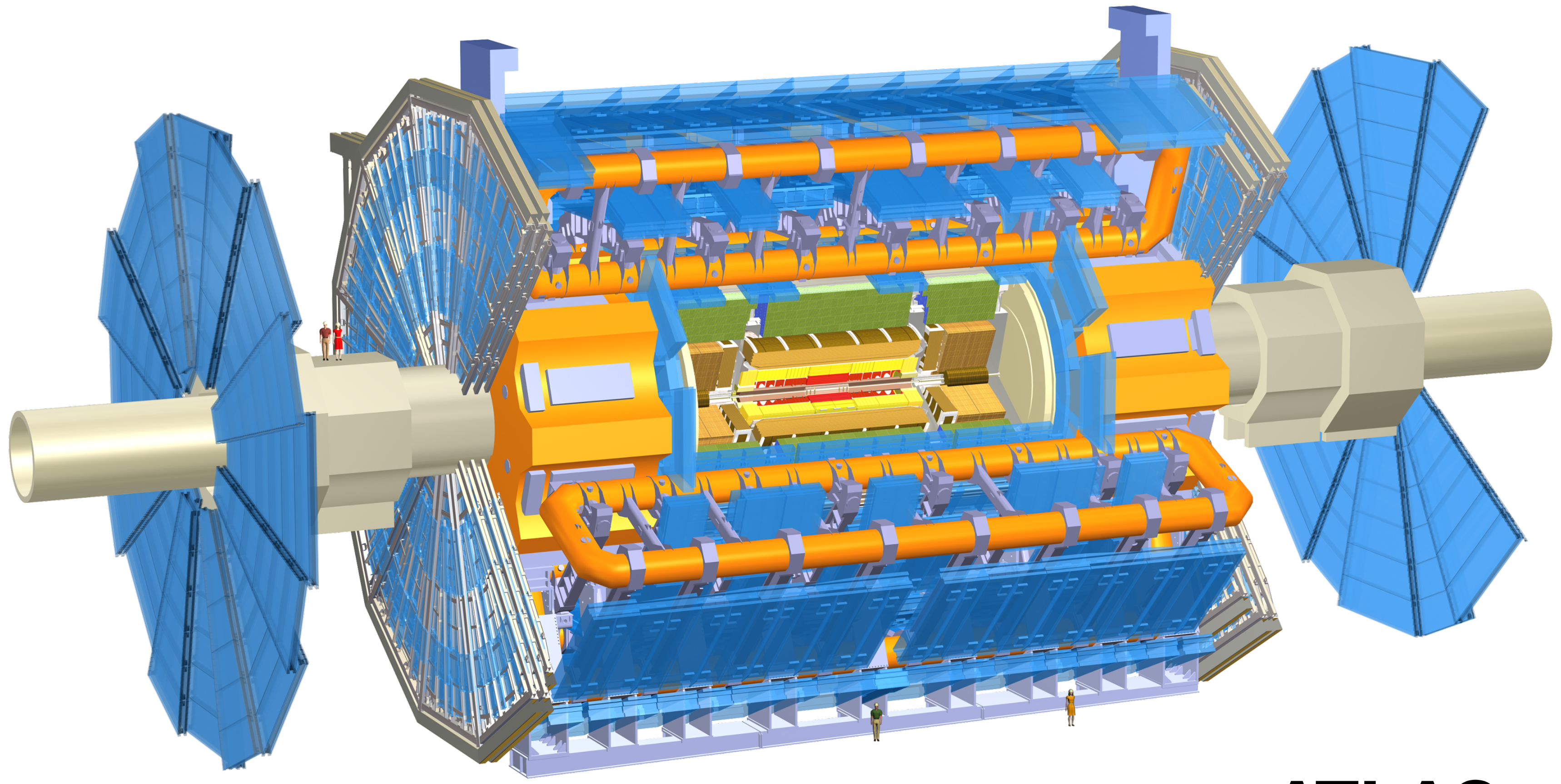
Collider Physics as Alchemy



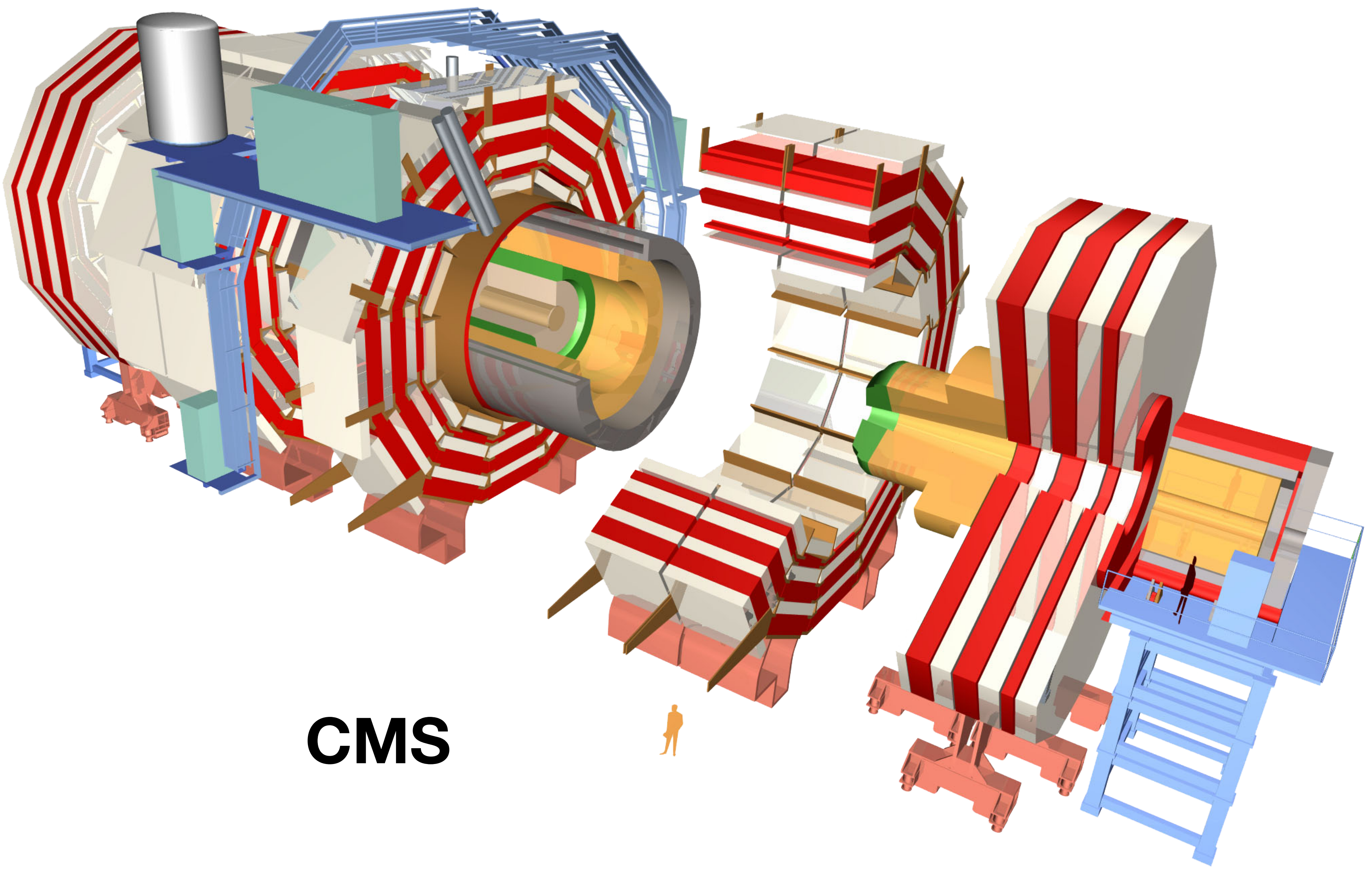


The Large Hadron Collider





ATLAS

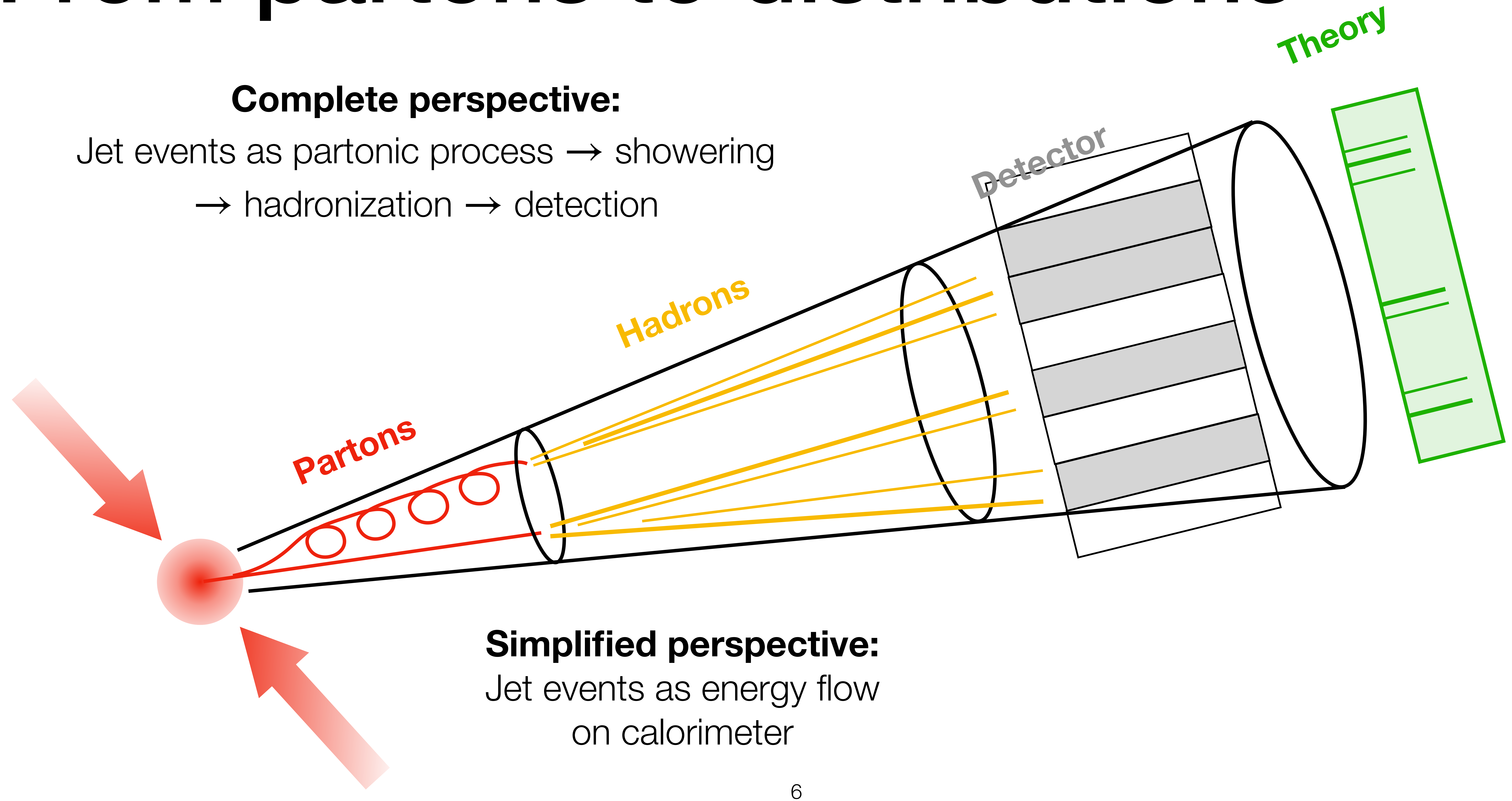


CMS


ATLAS
EXPERIMENT

From partons to distributions

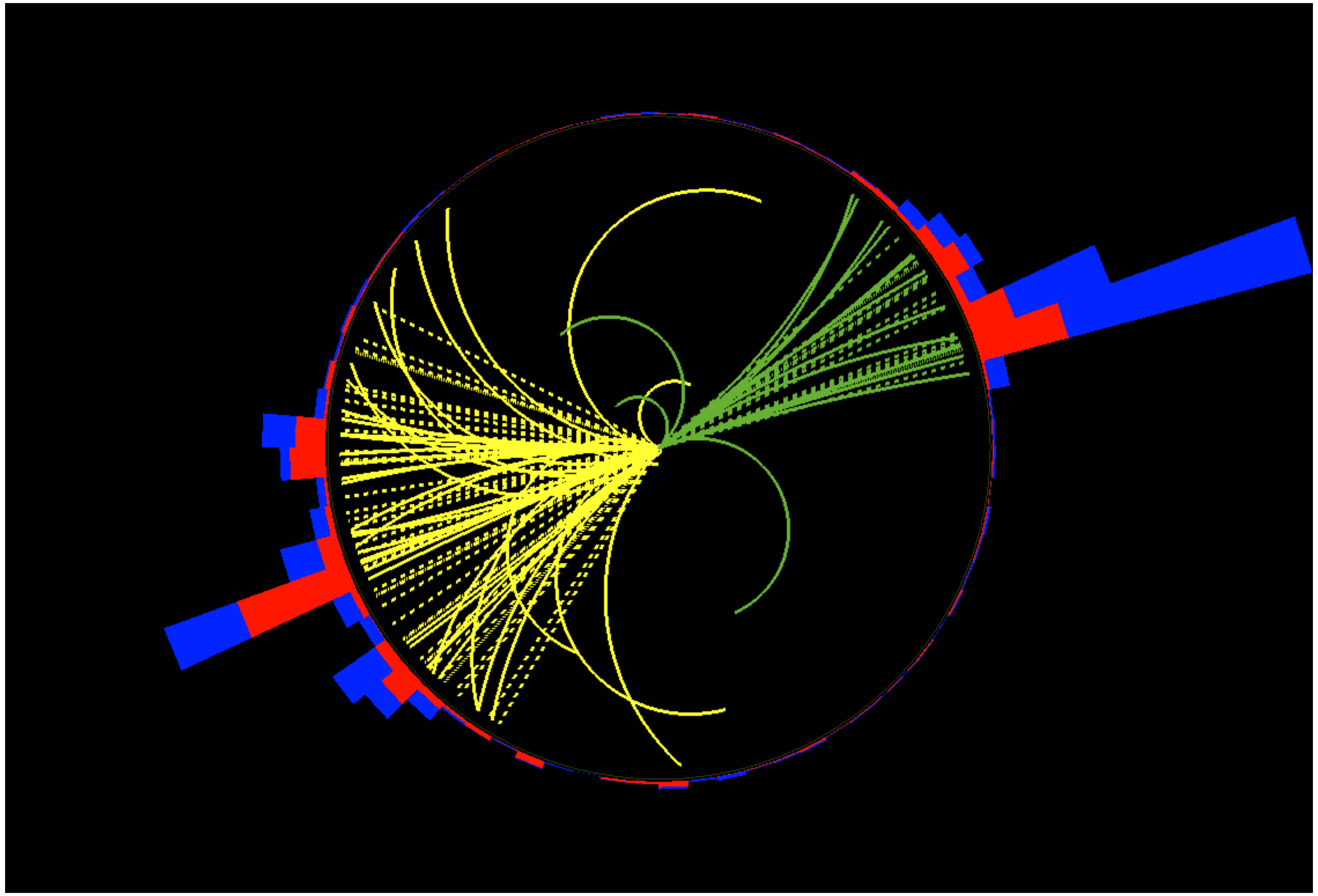
Complete perspective:
Jet events as partonic process → showering → hadronization → detection

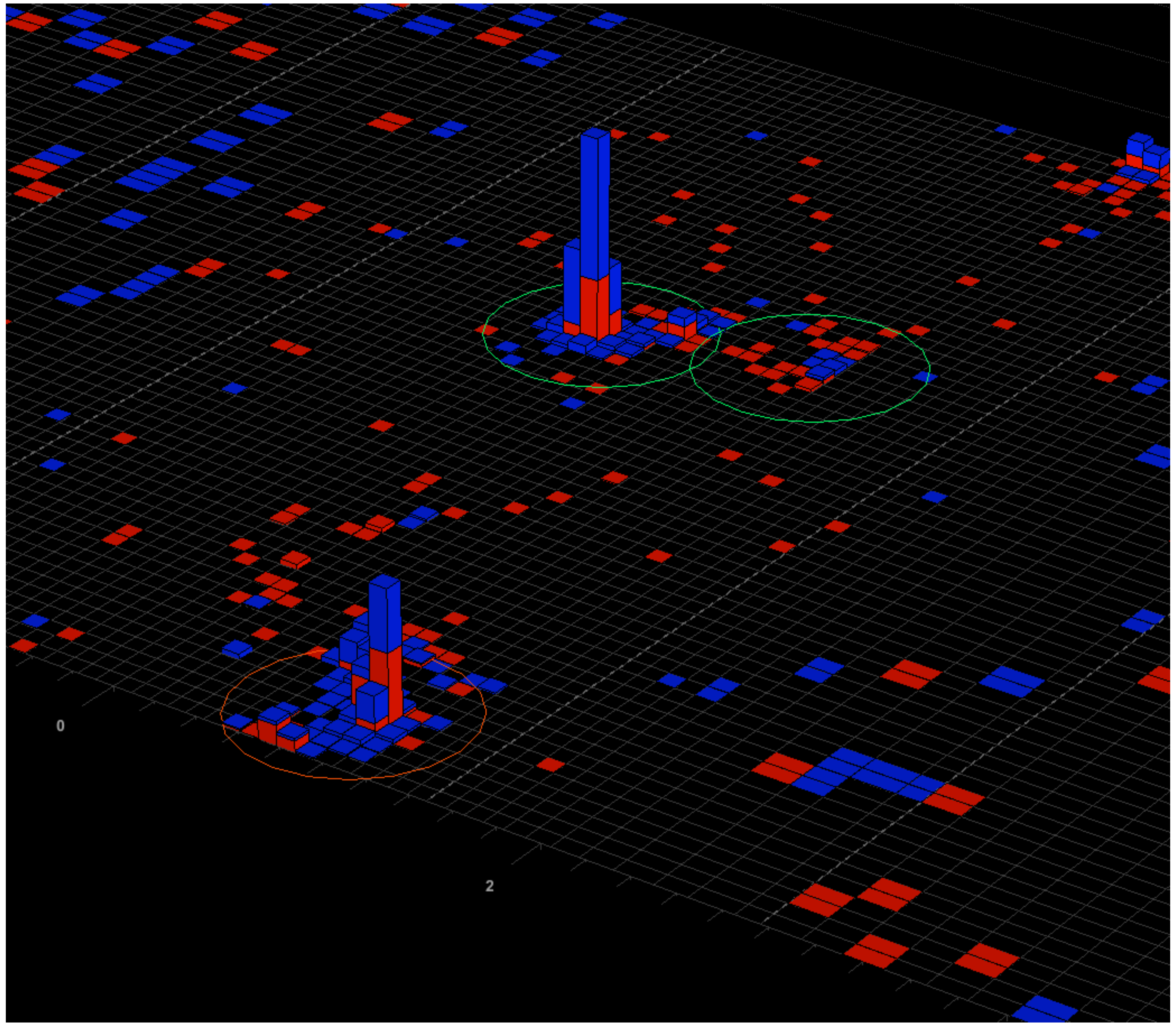


Simplified perspective:
Jet events as energy flow on calorimeter

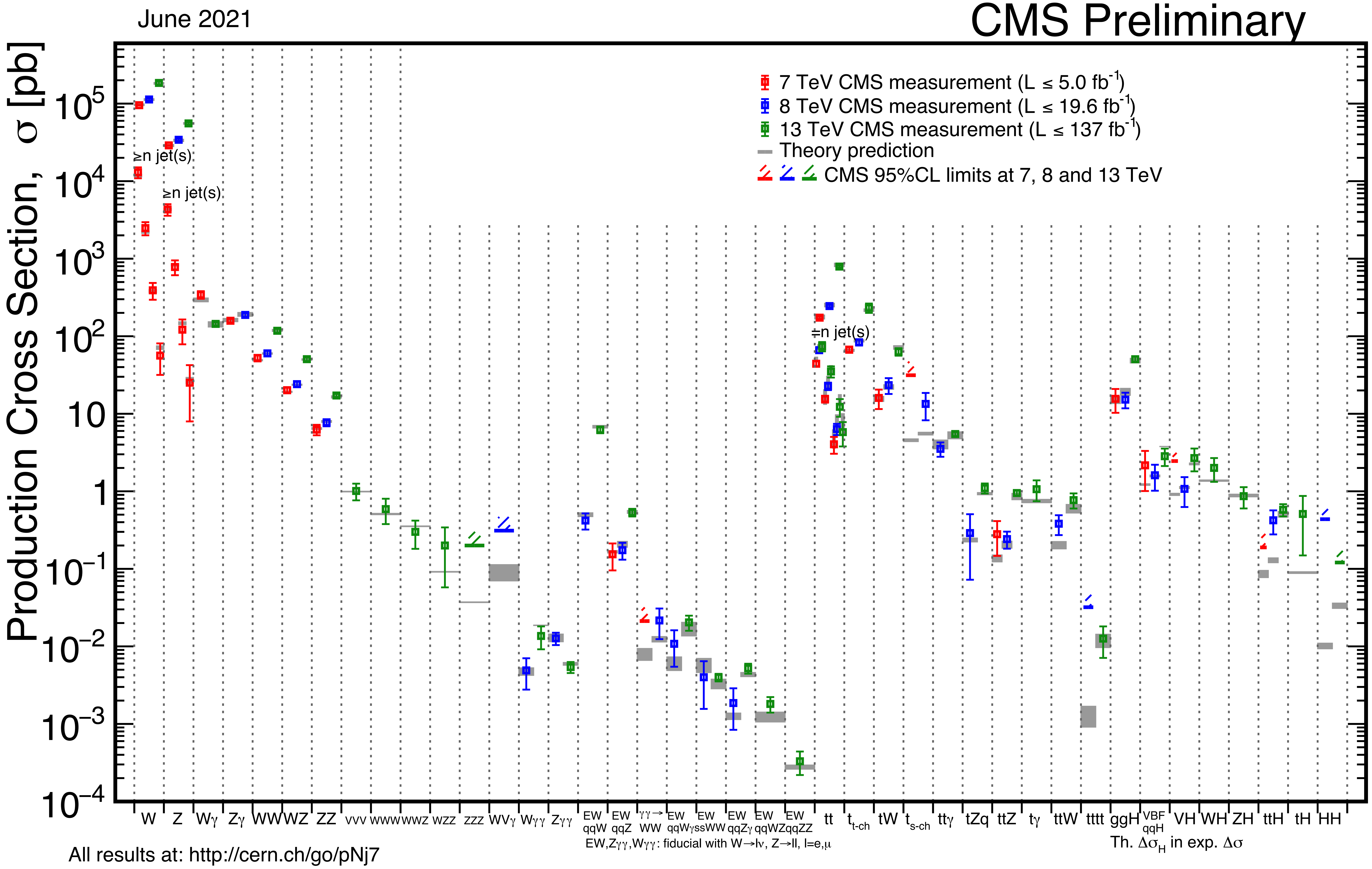
From partons to distributions

Treat jets as energy flow distributions on a 2d domain $\mathcal{E}(\hat{n}) = \sum_{i \in J} E_i \, \delta(\hat{n} - \hat{n}_i)$





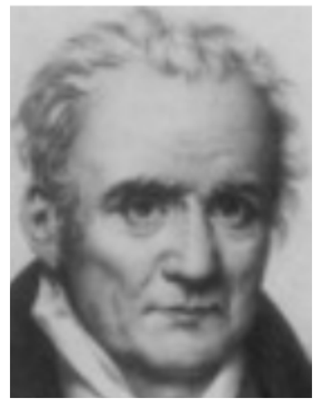
Signals and Backgrounds

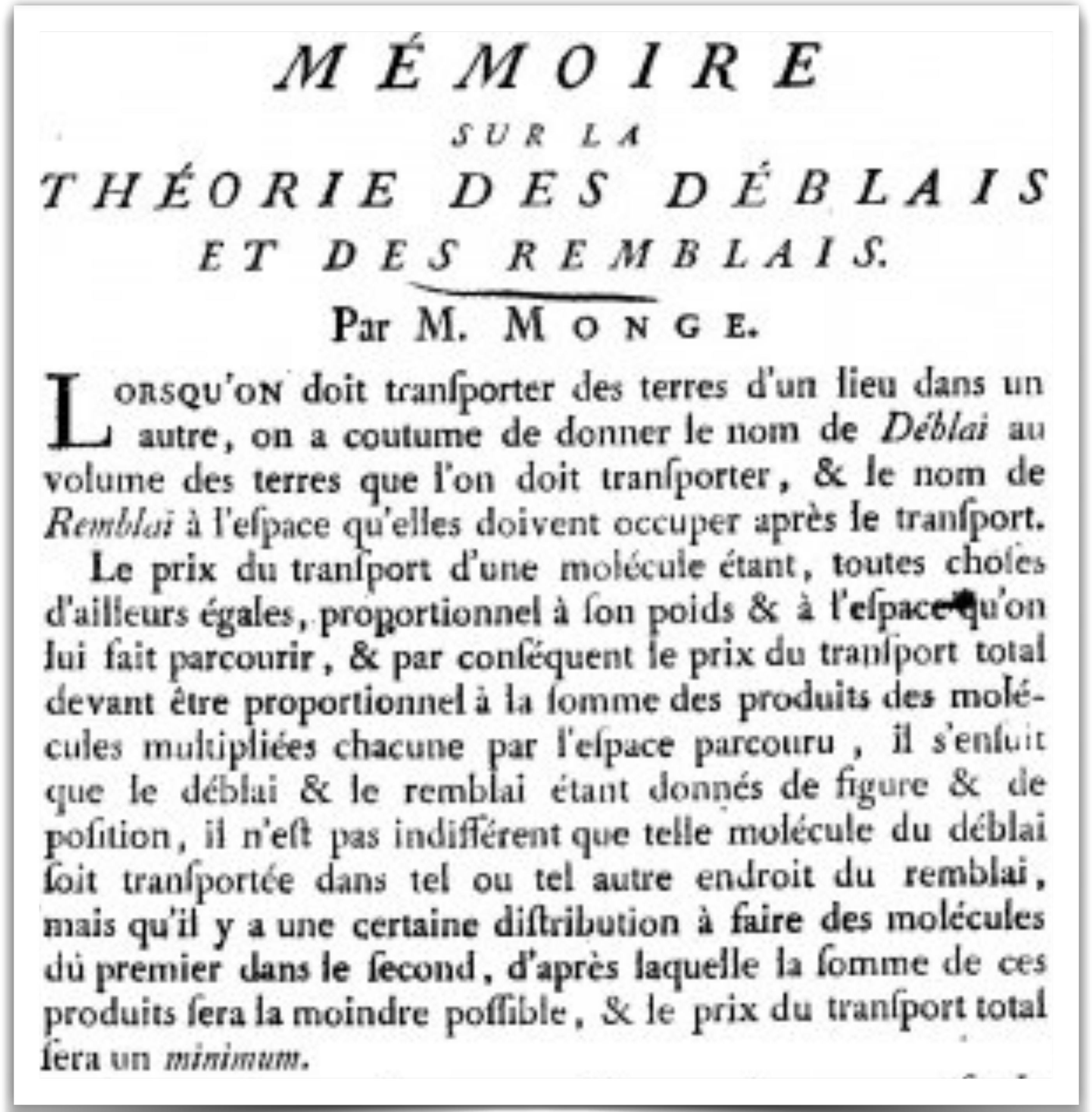


Typical size of beyond-the-Standard Model signals

Optimal Transport

Gaspard Monge
1781





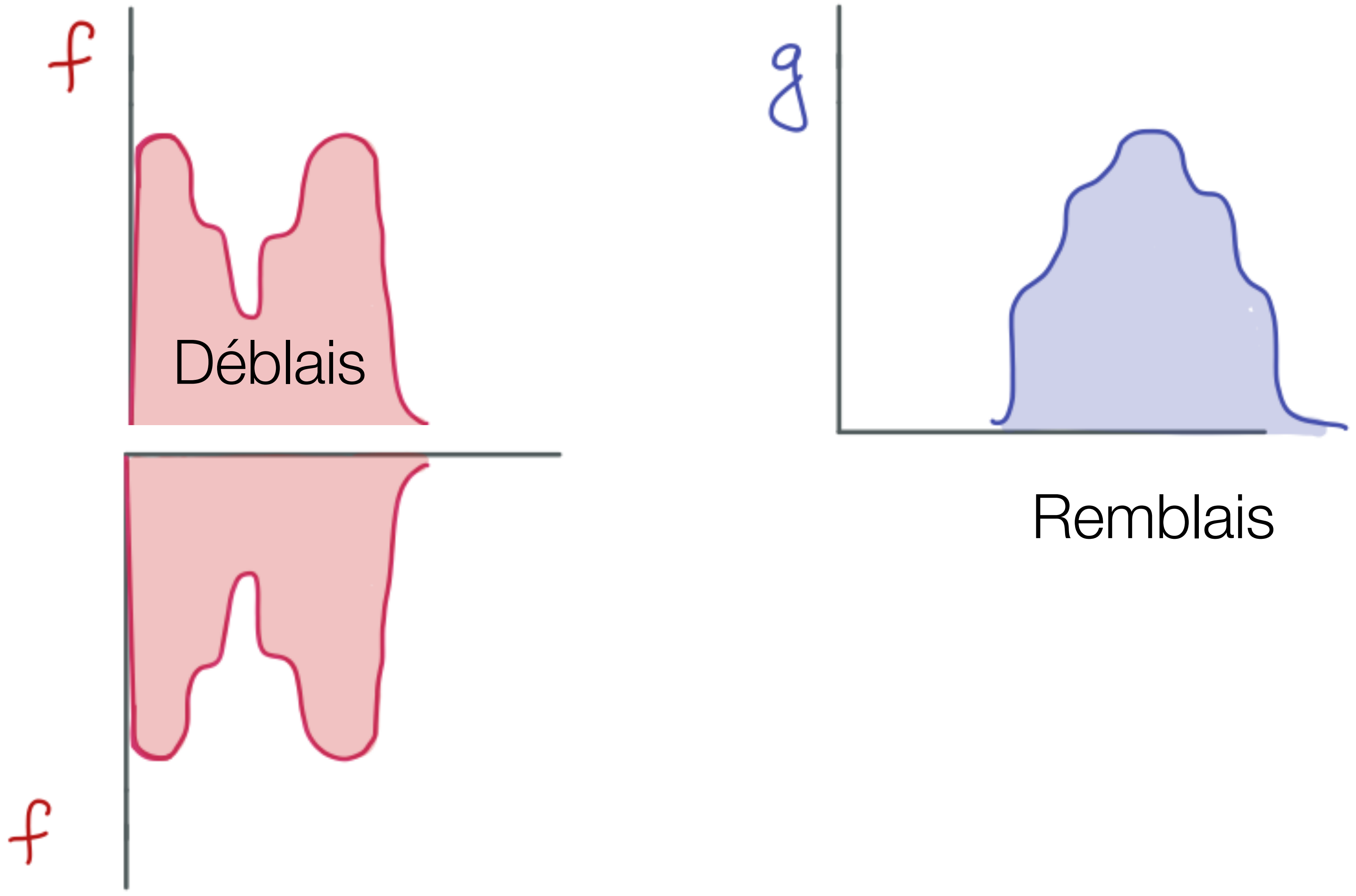
MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.
Par M. MONGE.

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.

Le prix du tranſport d'une molécule étant, toutes choſes d'ailleurs égales, proportionnel à ſon poids & à l'eſpace qu'on lui fait parcourir, & par conſéquent le prix du tranſport total devant être proportionnel à la ſomme des produits des molécules multipliées chacune par l'eſpace parcouru, il s'enſuit que le déblai & le remblai étant donnés de figure & de poſition, il n'eſt pas indifférent que telle molécule du déblai ſoit tranſportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine diſtribution à faire des molécules du premier dans le ſecond, d'après laquelle la ſomme de ces produits ſera la moindre poſſible, & le prix du tranſport total ſera un *minimum*.

Fundamental problem of **optimal transport:**

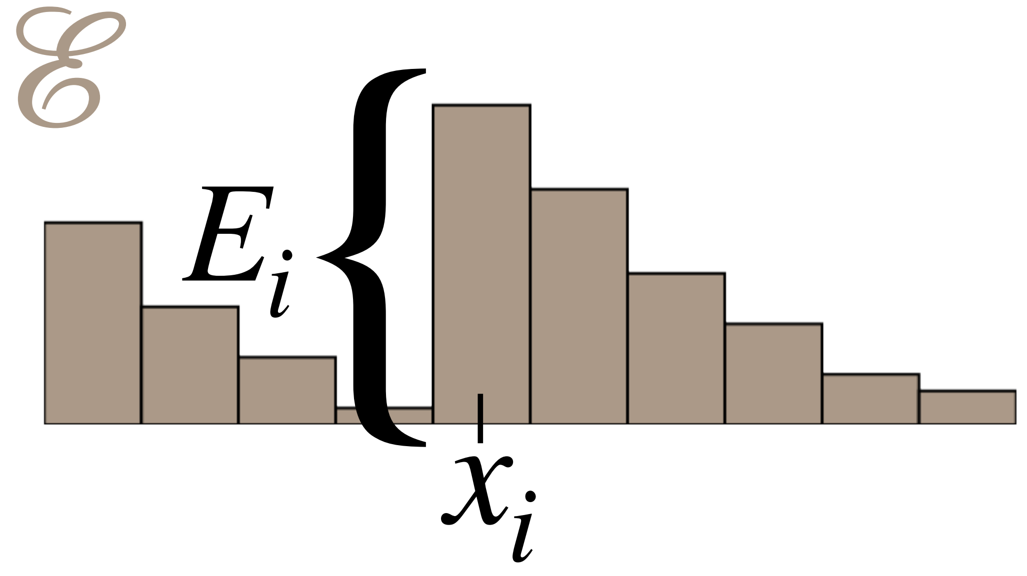
How to rearrange **f** to look like **g**
with the **least amount of "work"?**

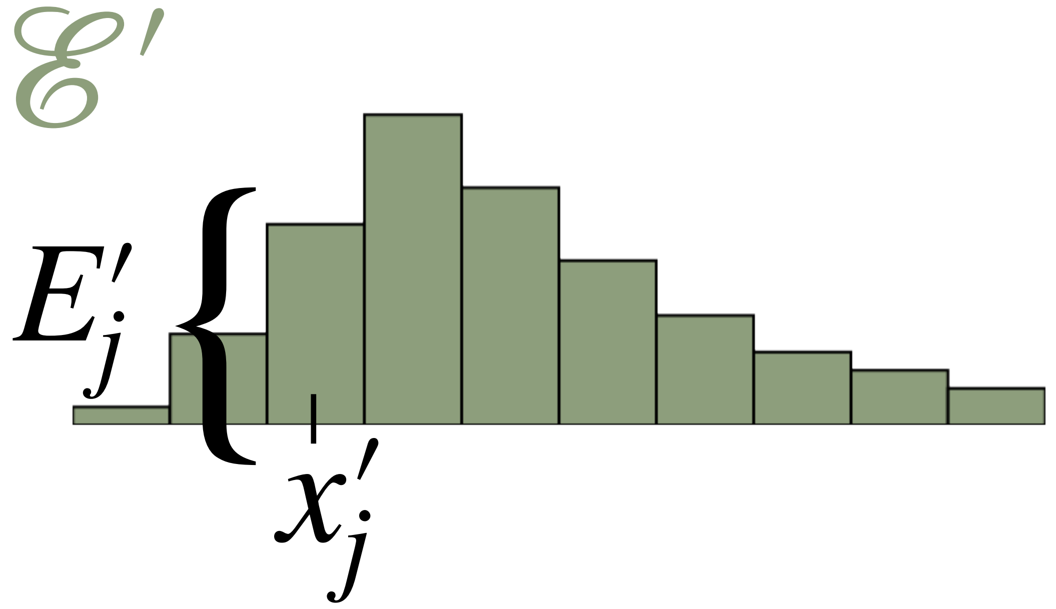


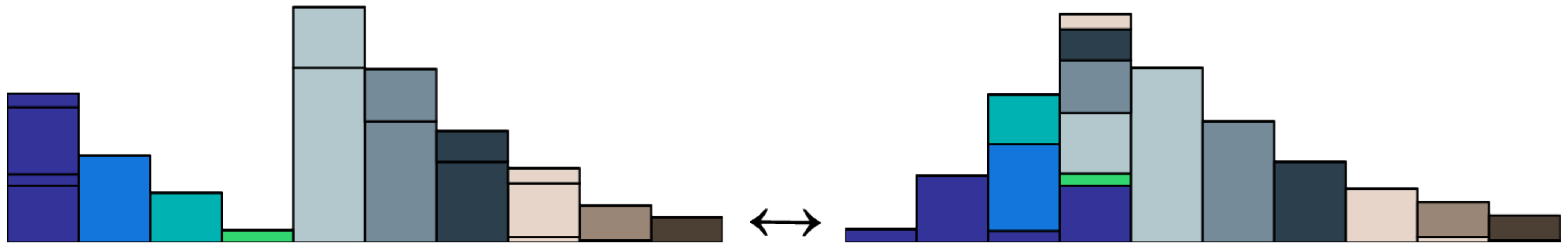
In other words, how can we optimally transport *f* to *g*?

Figures adapted from V. Herrmann, https://vincentherrmann.github.io/blog/wasserstein/

Optimal Transport





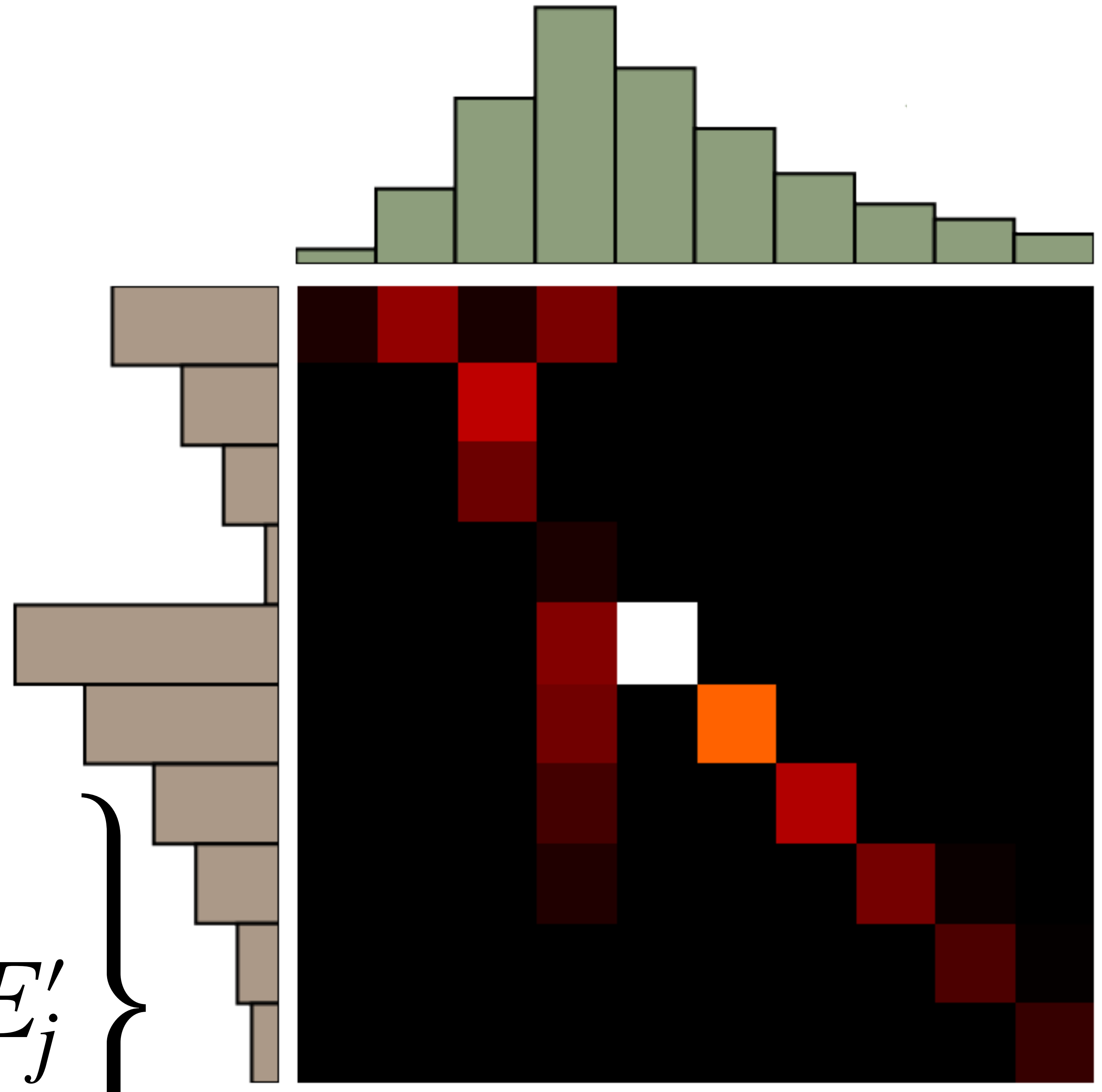


p-Wasserstein Distance, $p \geq 2$

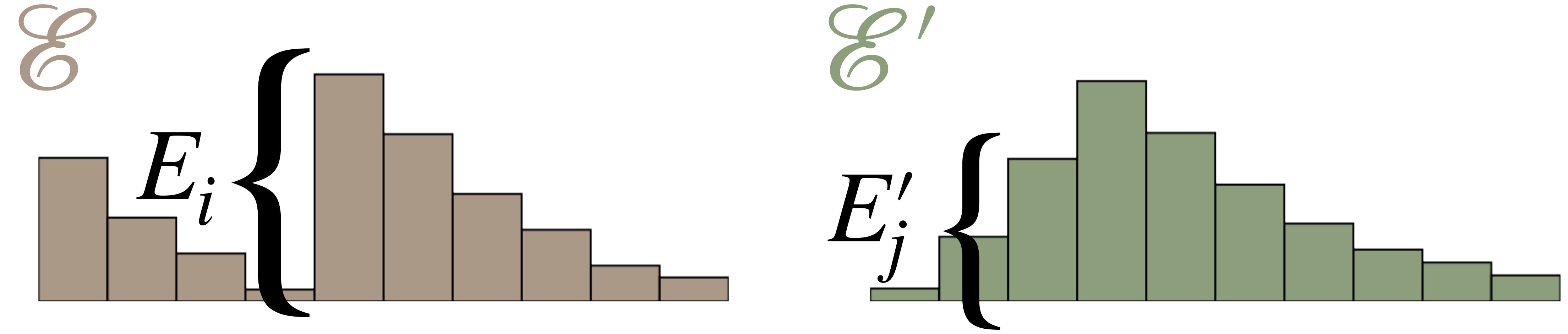
$$W_p(\mathscr{E}, \mathscr{E}') = \min_{\gamma_{ij} \in \Gamma_{(\mathscr{E}, \mathscr{E}')}} \left(\sum_{i,j} \|x_i - x_j'\|^p \gamma_{ij} \right)^{1/p}$$

Transport Plans from $\mathscr{E}$ to $\mathscr{E}'$

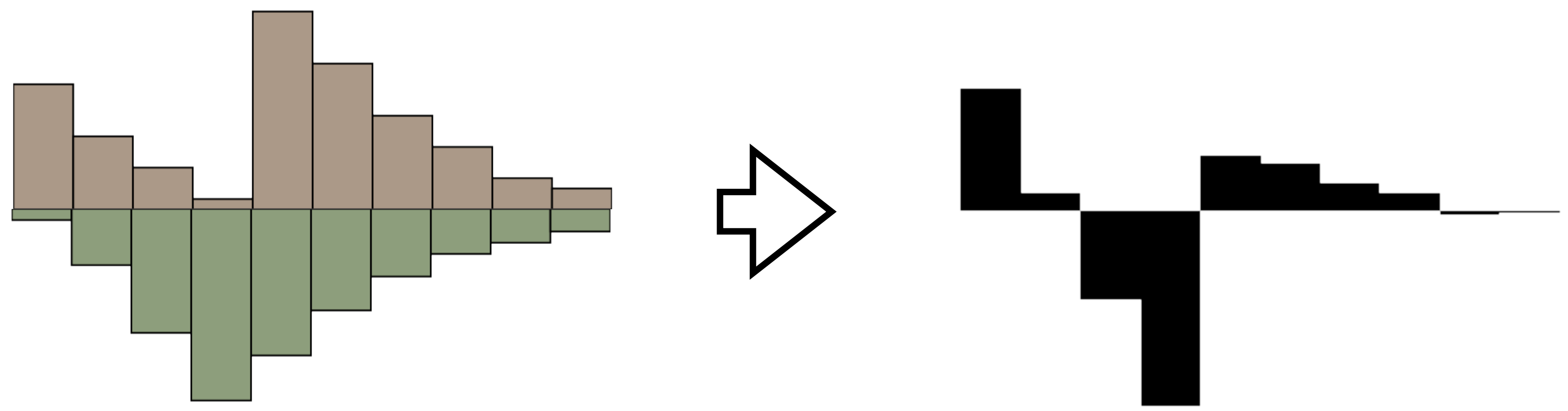
$$\Gamma_{(\mathscr{E}, \mathscr{E}')} = \left\{ \gamma_{ij} \in M_{n \times n}(\mathbb{R}) : \gamma_{ij} \geq 0, \sum_j \gamma_{ij} = E_i, \sum_i \gamma_{ij} = E_j' \right\}$$



OT respects underlying geometry

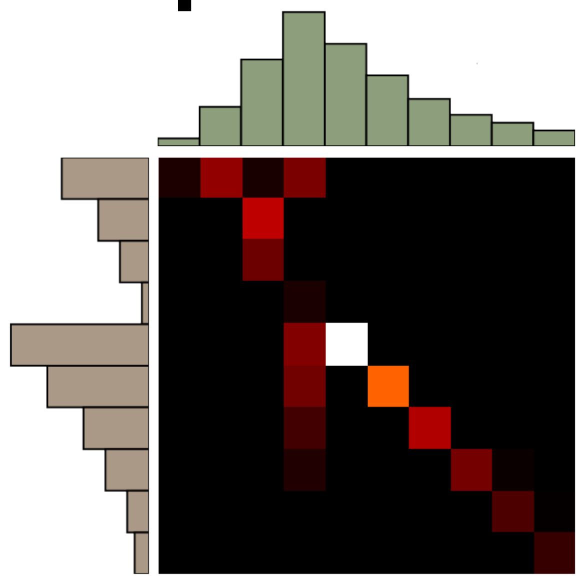


L² distance



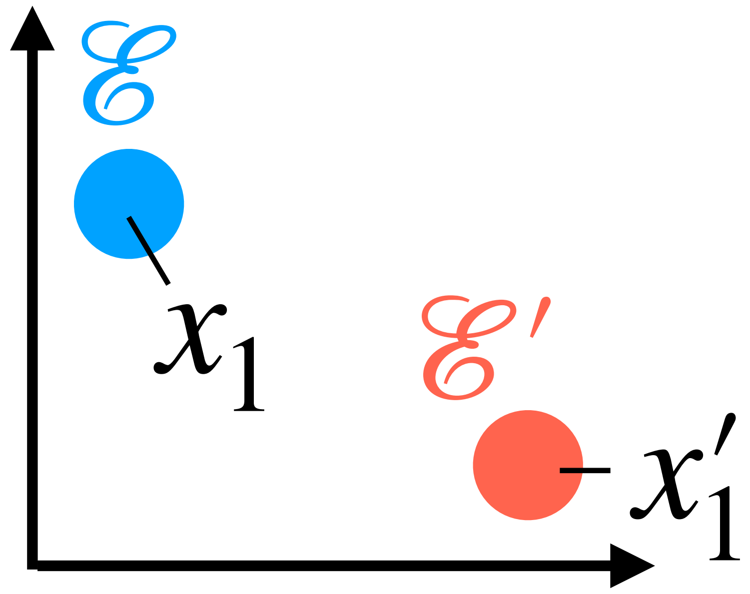
$$\left(\sum_{i=1}^{N} |E_i - E'_i|^2 \right)^{1/2} = 0.30$$

W



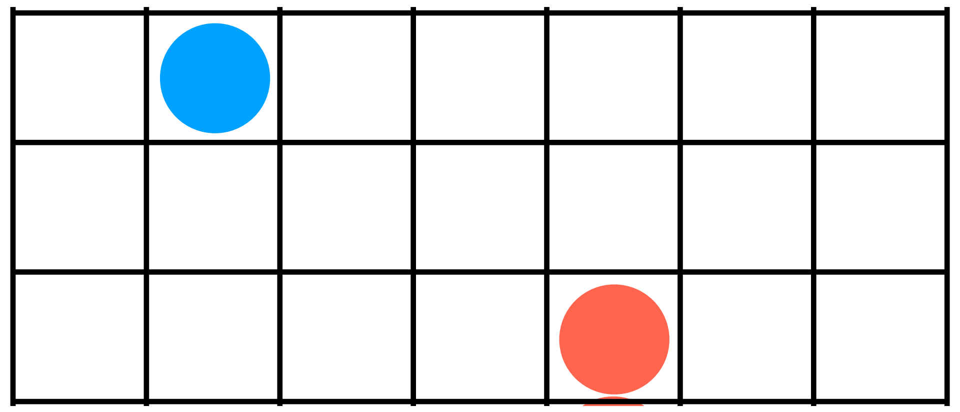
$$W_1(\mathscr{E}, \mathscr{E}') = 0.71$$

OT respects underlying geometry



$$E_1 = E_1' = 1$$

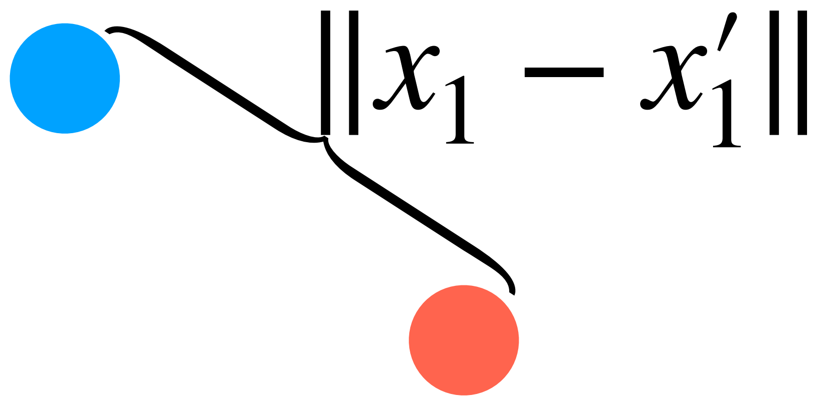
L^2 distance



$$\left(\sum_{i=1}^{N} |E_i - E_i'|^2\right)^{1/2} = \begin{cases} \sqrt{2} \\ 0 \end{cases}$$

disregards spatial information

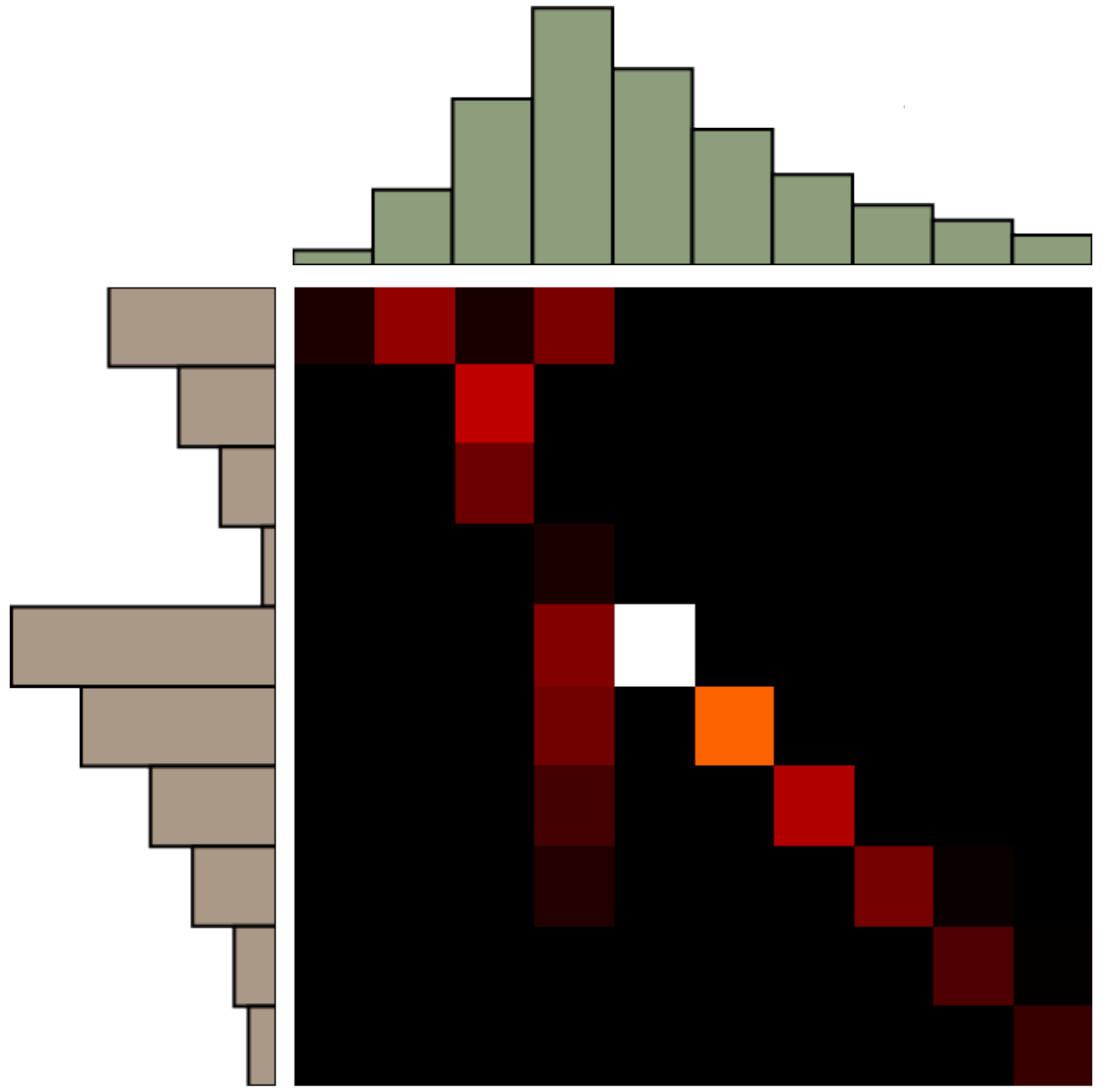
W_p distance



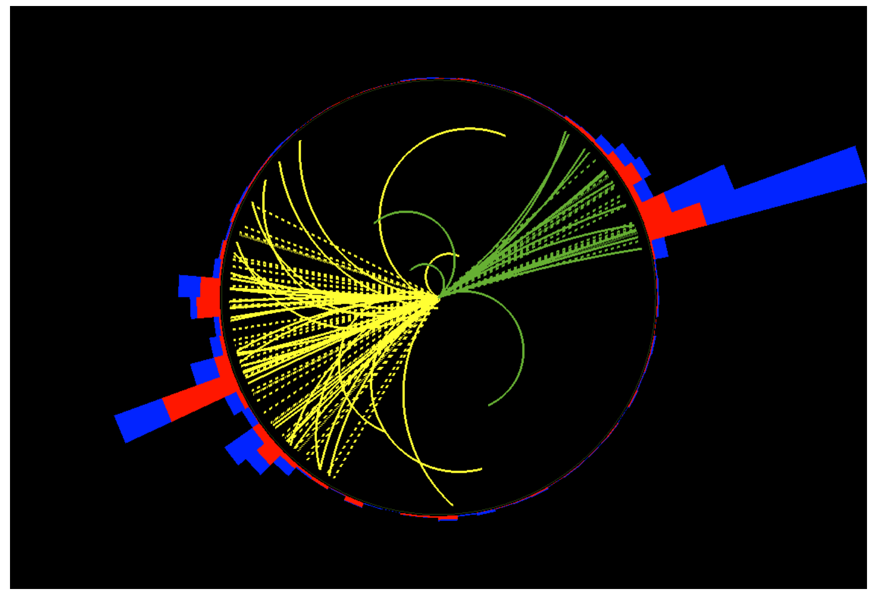
$$W_p(\mathscr{E}, \mathscr{E}') = \|x_1 - x_1'\|$$

lifts geometry of underlying space to space of distributions

OT for Particle Physics



⊕



“Particle physics is better with OT”

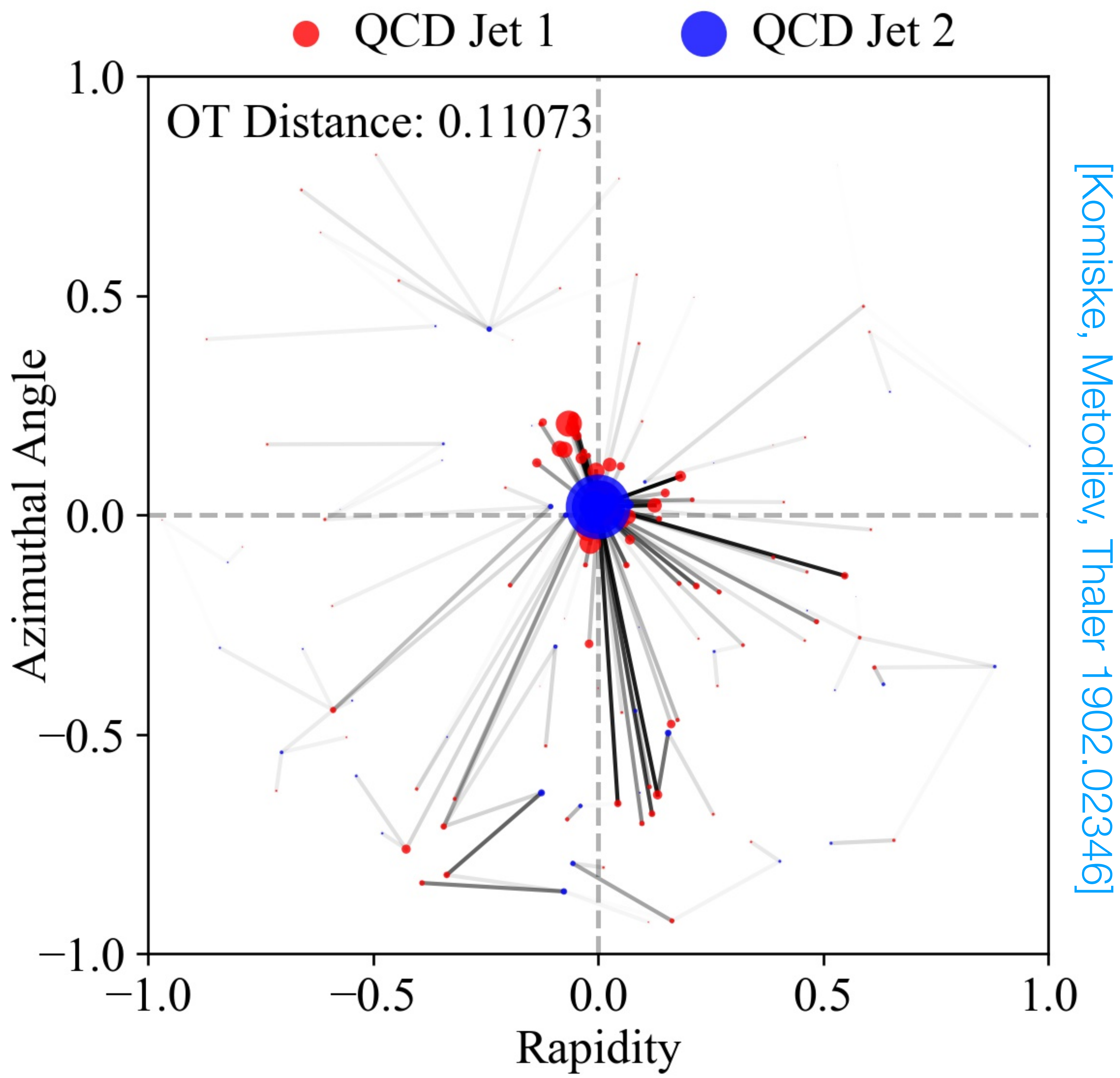
- *Metric Space of Collider Events* (Komiske, Metodiev, Thaler 2019)
- *The Hidden Geometry of Particle Collisions* (Komiske, Metodiev, Thaler, 2020)
- *A Robust Measure of Event Isotropy at Colliders* (Cesarotti, Thaler 2020)
- …   ⇒ this talk

“Machine Learning is better with OT”

- Wasserstein GAN (Arjovsky et al. ‘17)
- Wasserstein AE (Tolstikhin et al. ‘17)
- Sliced Iterative Normalizing Flows (Dai, Seljak ’20)
- …

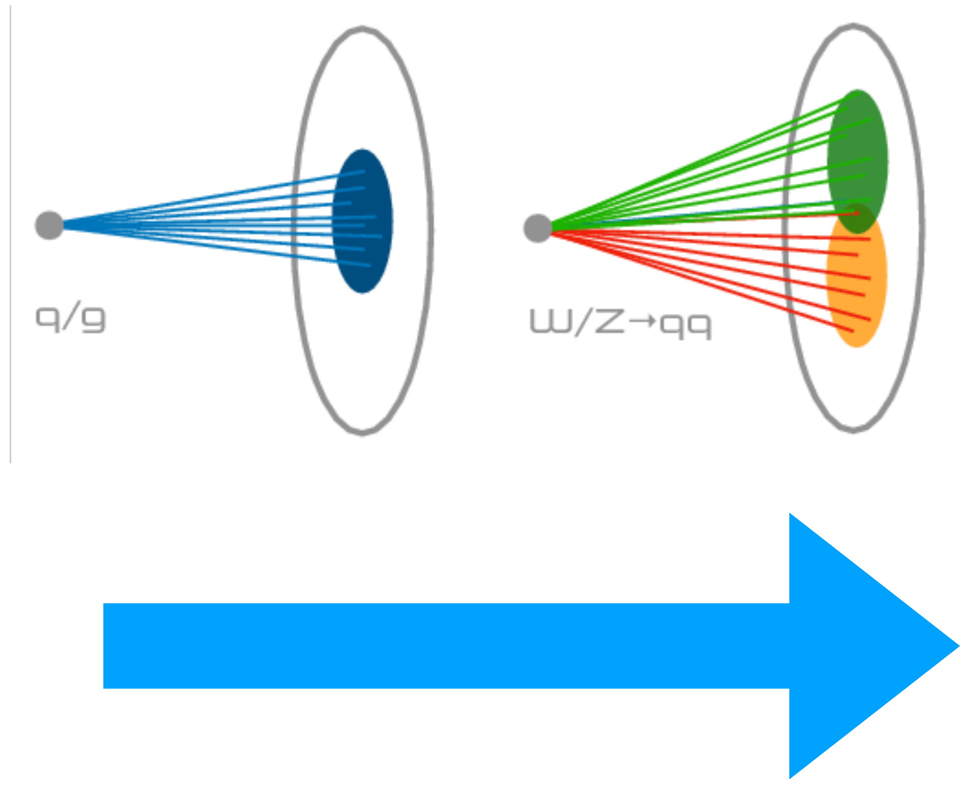
OT for Particle Physics

Komiske, Metodiev, Thaler 1902.02346: OT (EMD) is useful for collider physics

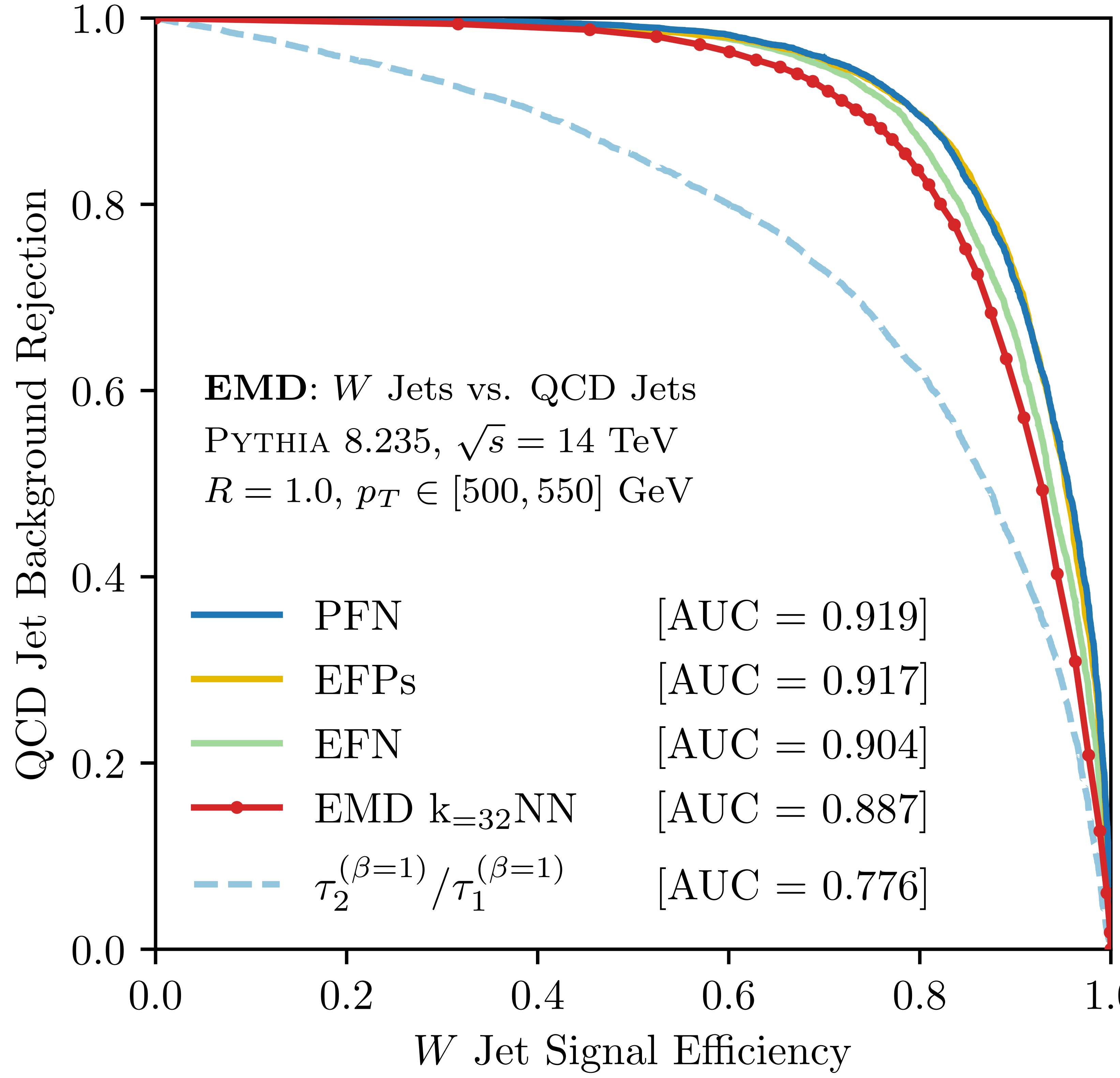


[Komiske, Metodiev, Thaler 1902.02346]

(1) Compute pairwise EMD distances between all events in signal & background samples



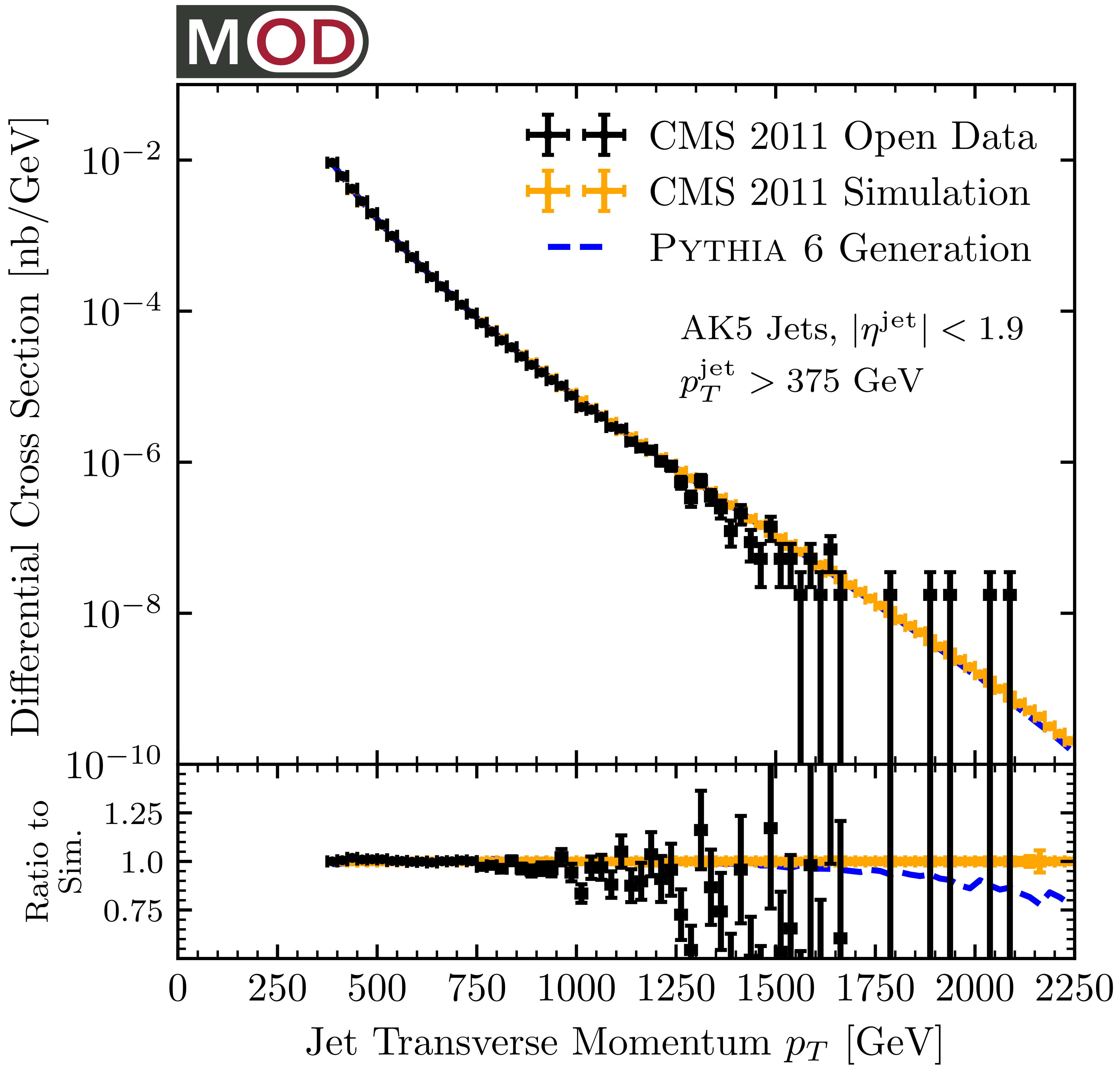
(2) Use EMD distances as input to "off-the-shelf" ML algorithms (e.g. kNN) for event classification

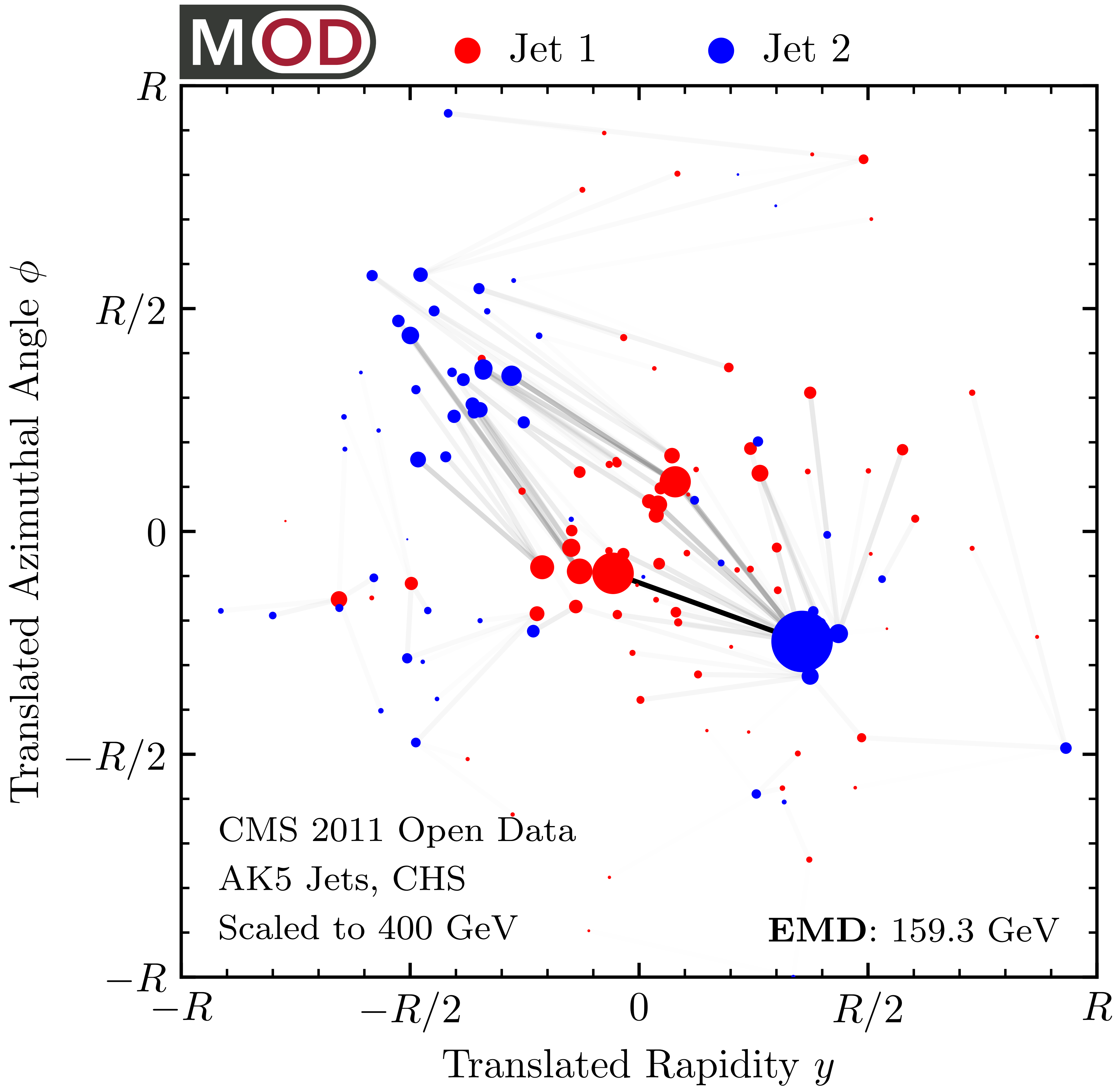


[Komiske, Metodiev, Thaler 1902.02346]

OT on Open Data

Komiske, Mastandrea, Metodiev, Naik, Thaler [1908.08542]





Not so fast (literally)

- Computing OT distances between N_{evt} events requires $\mathcal{O}(N_{\text{evt}}^2)$ evaluations.

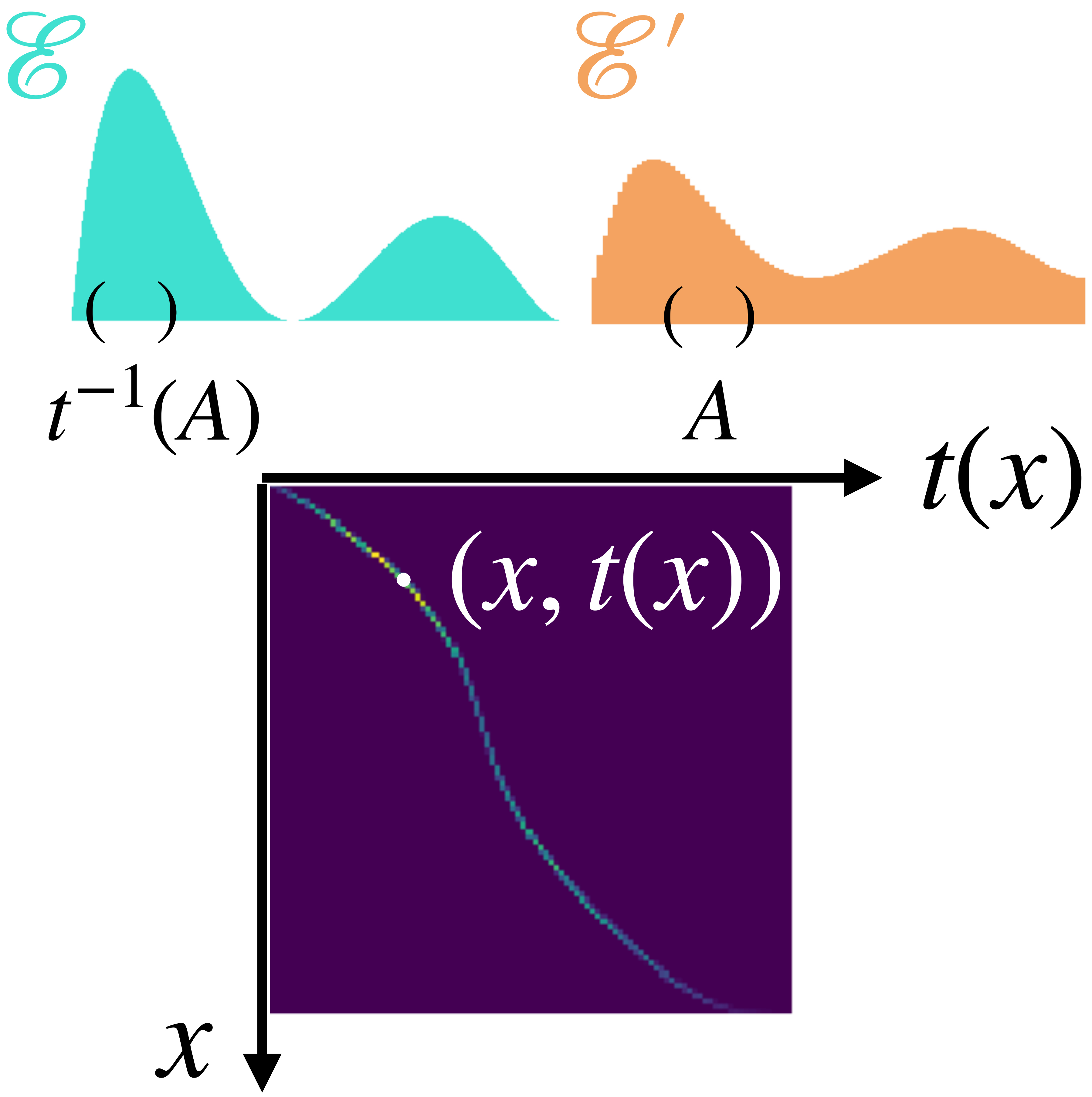
- Evaluating one OT distance between two typical events takes ~0.1s on a desktop using e.g. Python Optimal Transport library.

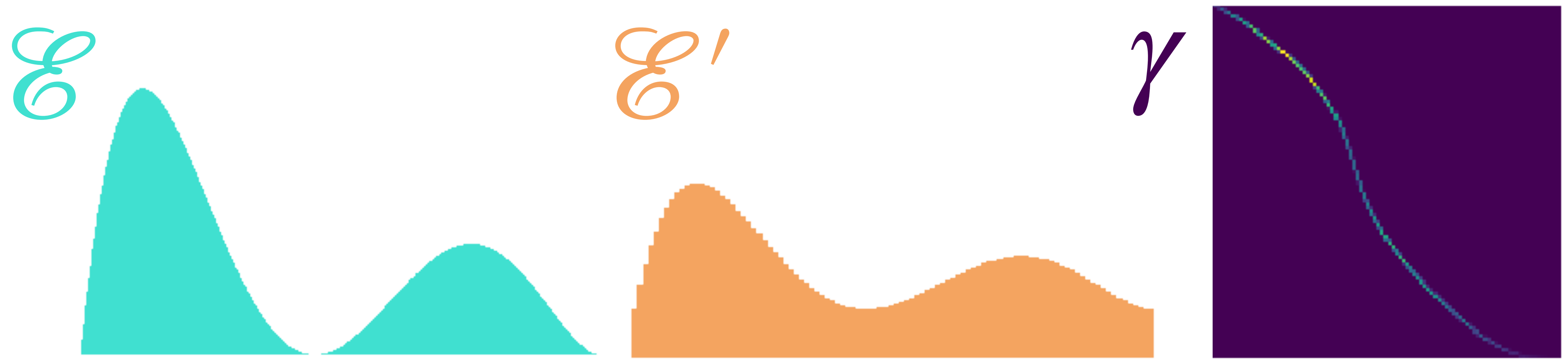
- $\Rightarrow$ A naive construction of pairwise distance matrix for 100k events would take ~16 years on a desktop.

- Various speedups, parallelization possible, but clearly cumbersome

if computational burden is comparable to NNs, what is gained?

From Kantorovich to Monge





Continuum Kantorovich

$$W_2(\mathscr{E}, \mathscr{E}') = \min_{\gamma \in \Gamma_{(\mathscr{E},\mathscr{E}')}} \left(\iint \|x - y\|^2 d\Gamma(x, y) \right)^{1/2}$$

$$\Gamma_{(\mathscr{E},\mathscr{E}')} = \left\{ \gamma \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d) : \pi_1 \# \gamma = \mathscr{E}, \pi_2 \# \gamma = \mathscr{E}' \right\}$$

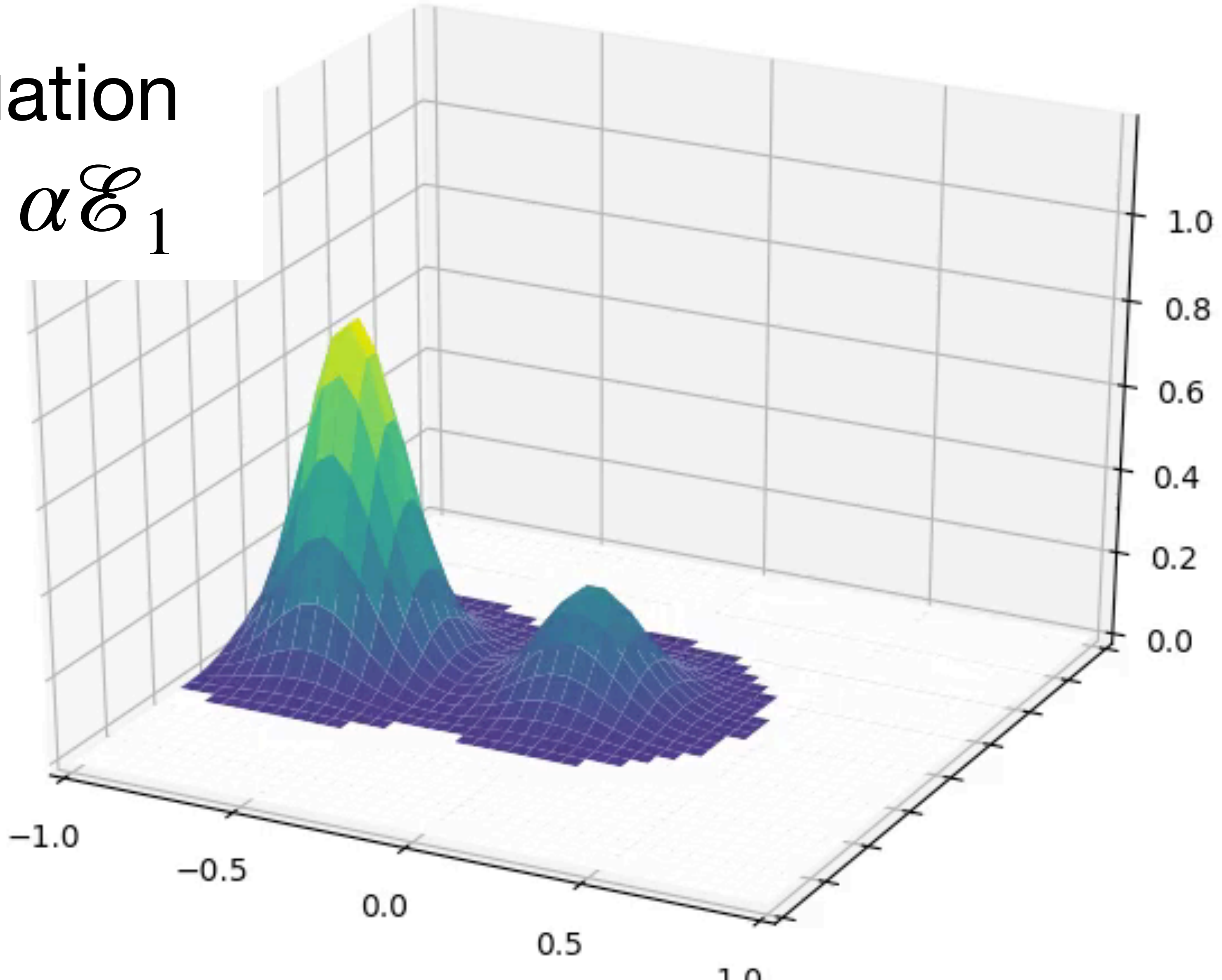
Thm (Brenier '91): If $\mathscr{E}$ is a continuum distribution, ∃! optimal transport map.

Geodesics

[Peyré, Papadakis, Oudet 2013]

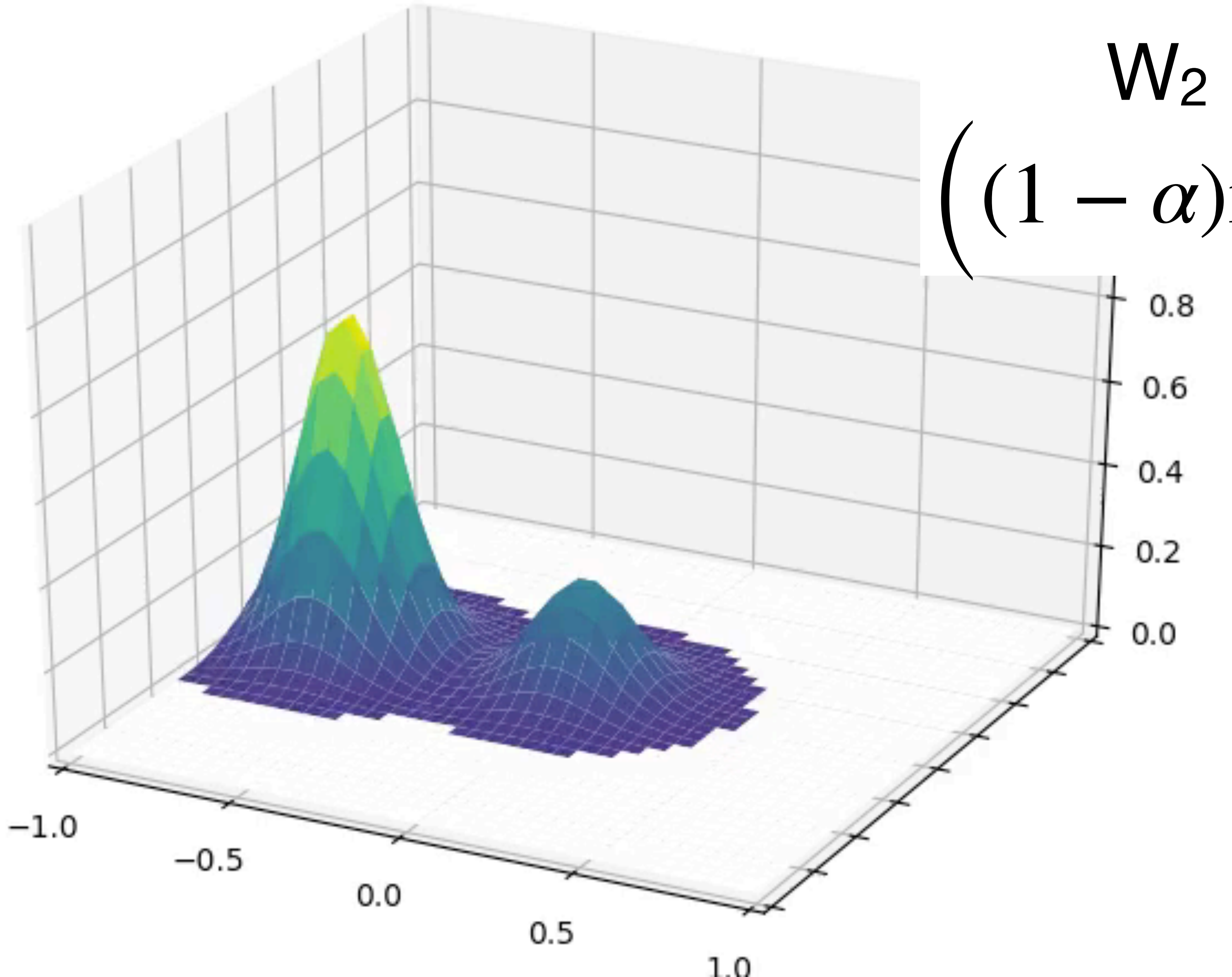
linear interpolation

$(1-\alpha)\mathscr{E}_0 + \alpha\mathscr{E}_1$



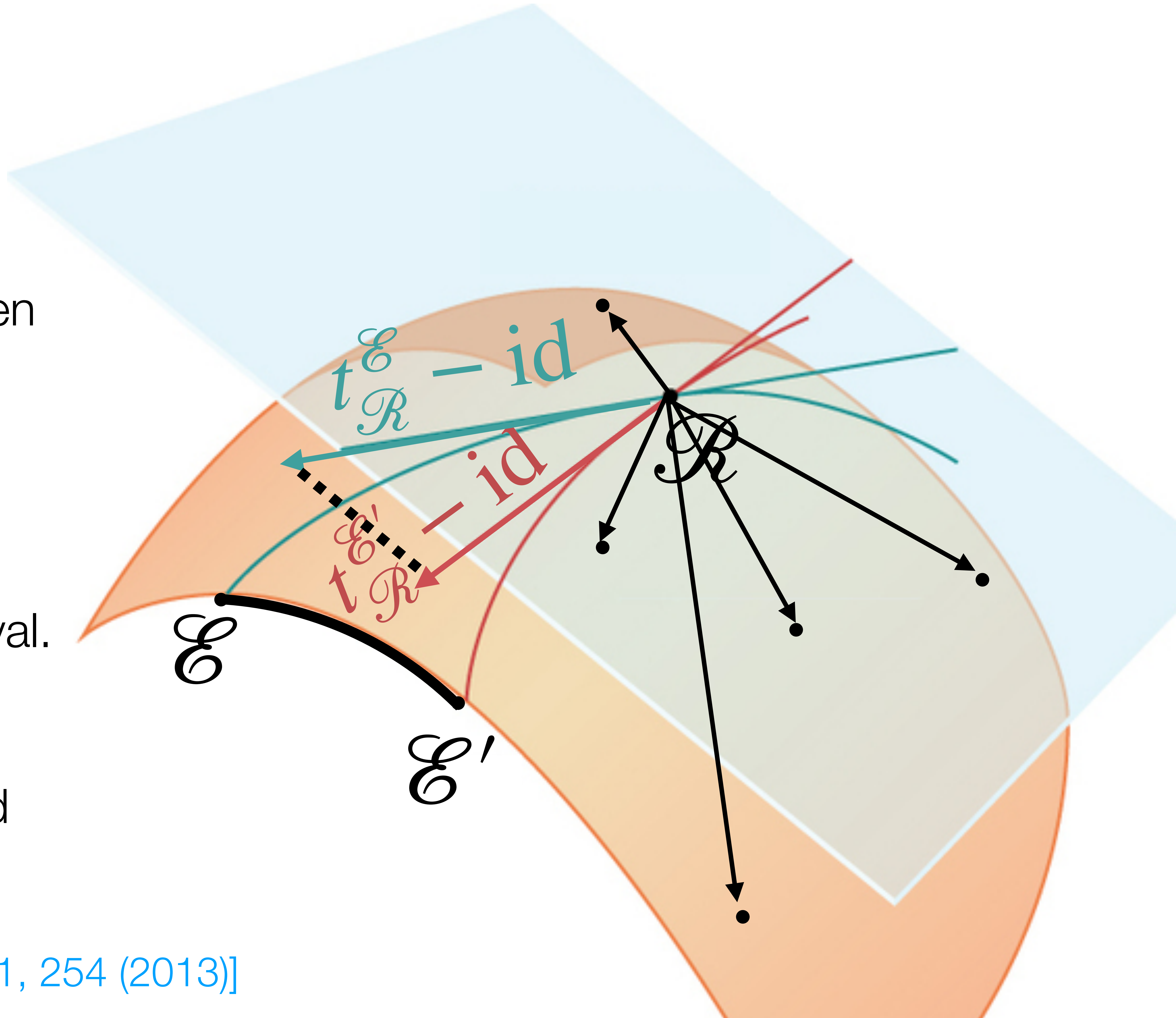
W_2 geodesic

$\left((1-\alpha)\mathrm{id} + \alpha t_{\mathscr{E}_0}^{\mathscr{E}_1}\right)\#\mathscr{E}_0$



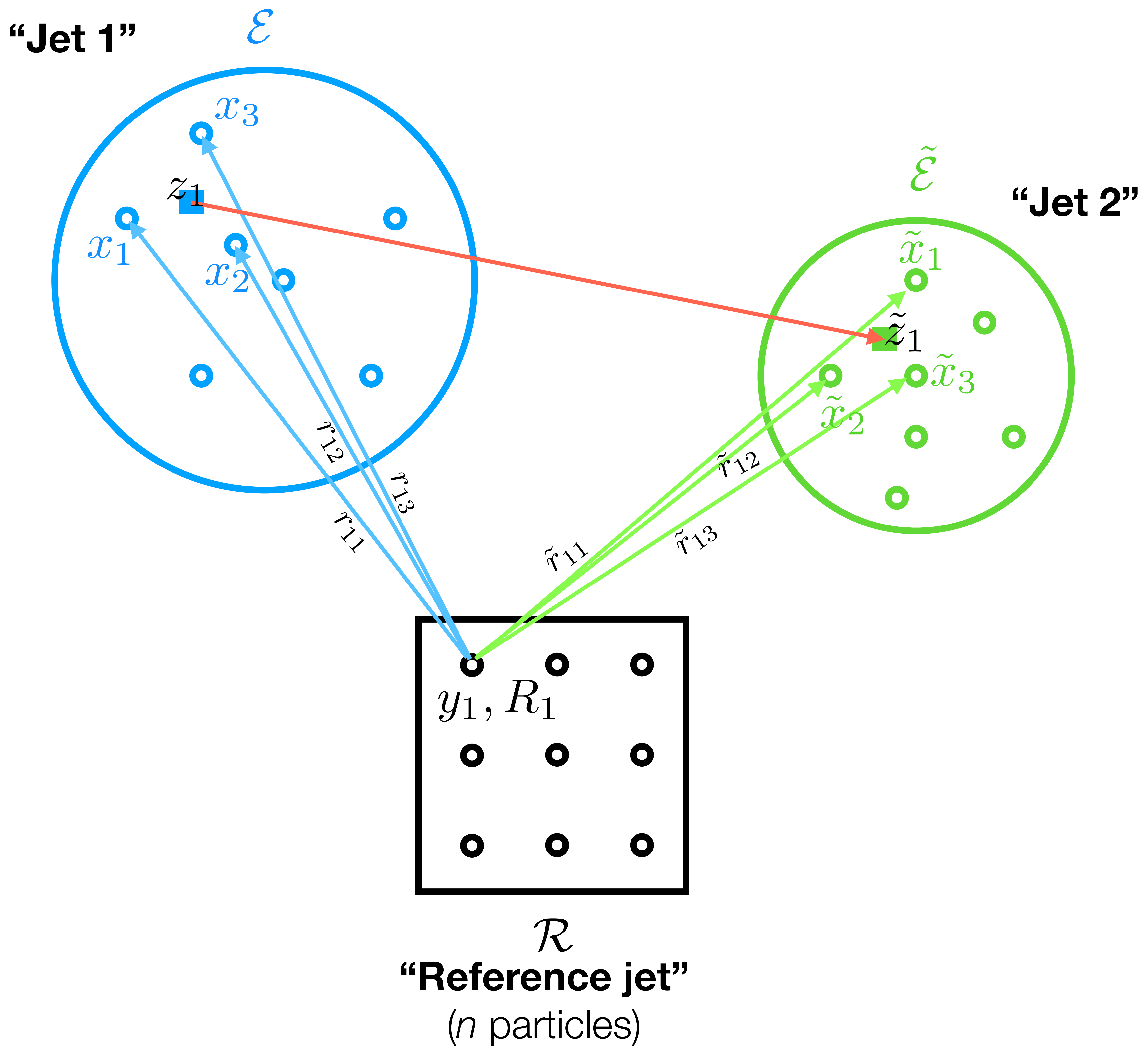
Riemannian Structure → Linearize

- The W_2 metric between continuum distributions has Riemannian structure.
- Basic idea of linearization:
  1. Project onto tangent plane at a chosen reference event $\mathscr{R}$.
  2. Compute Euclidean distances.
- Enormous computational benefit:
  - $\mathscr{O}(N_{\text{evt}})$ W_2 eval., $\mathscr{O}(N_{\text{evt}}^2)$ Euclidean eval.
  - Euclidean embedding
- Sound mathematical footing: the linearized distance $W_{2,\mathscr{R}}$ is also a metric.



[Wang et al., International Journal of Computer Vision 101, 254 (2013)]

Linearized OT in Practice



- distance from (discrete) reference to event:

$$W_2(\mathcal{R}, \mathcal{E}) = \min_{\gamma_{ij} \in \Gamma(\mathcal{E}, \mathcal{E}')} \left(\sum_{ij} \|x_i - x_j'\|^2 \gamma_{ij} \right)^{1/2}$$

- γ_{ij} : **optimal transport plan** (minimizer of W_2)

- z_i : **barycenter** (avg. of locations to which *i*th particle is sent, weighted by transport plan)

$$z_i = \frac{1}{R_i} \sum_j \gamma_{ij} x_j$$

← Map from $\mathcal{E}$ to a vector in $\mathbb{R}^{2n}$; approximates OT map $t(x_i)$

- LOT approximation of $W_2(\mathcal{E}, \mathcal{E}')$:

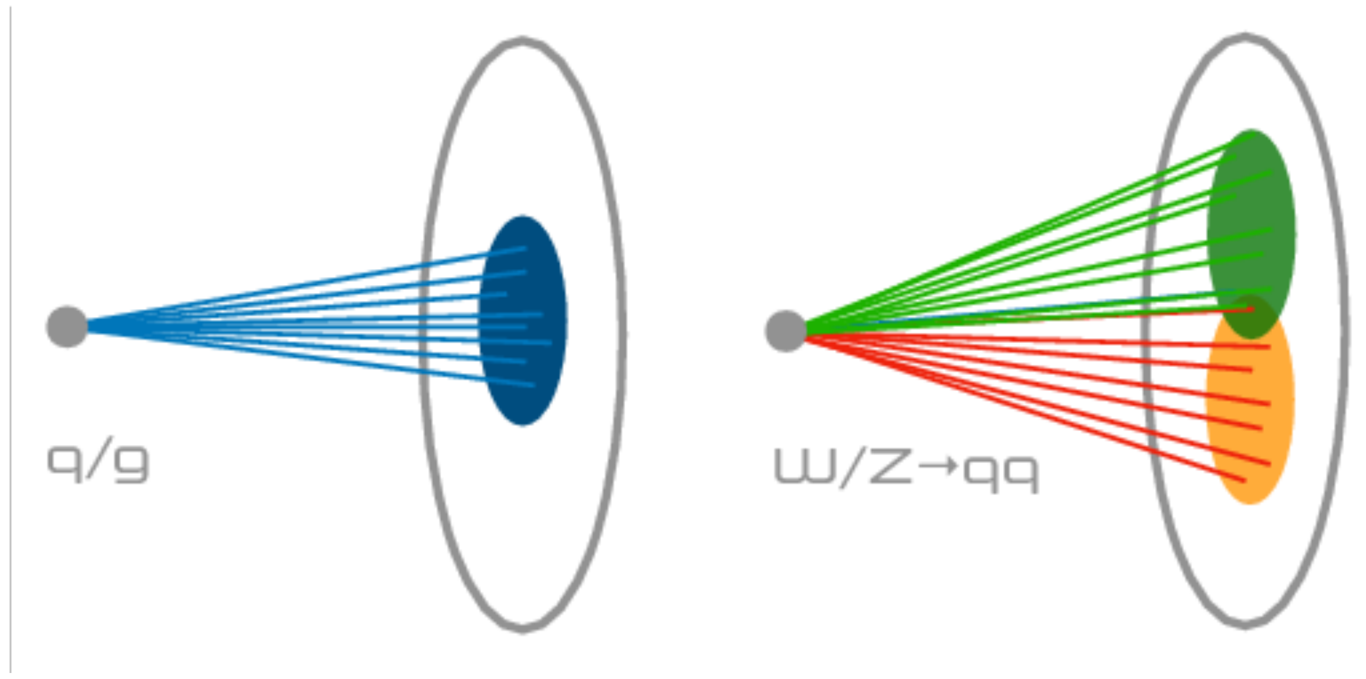
$$LOT_{r,r'}(\mathcal{E}, \mathcal{E}') = \left(\sum_i \|z_i - z_i'\|^2 R_i \right)^{1/2}$$

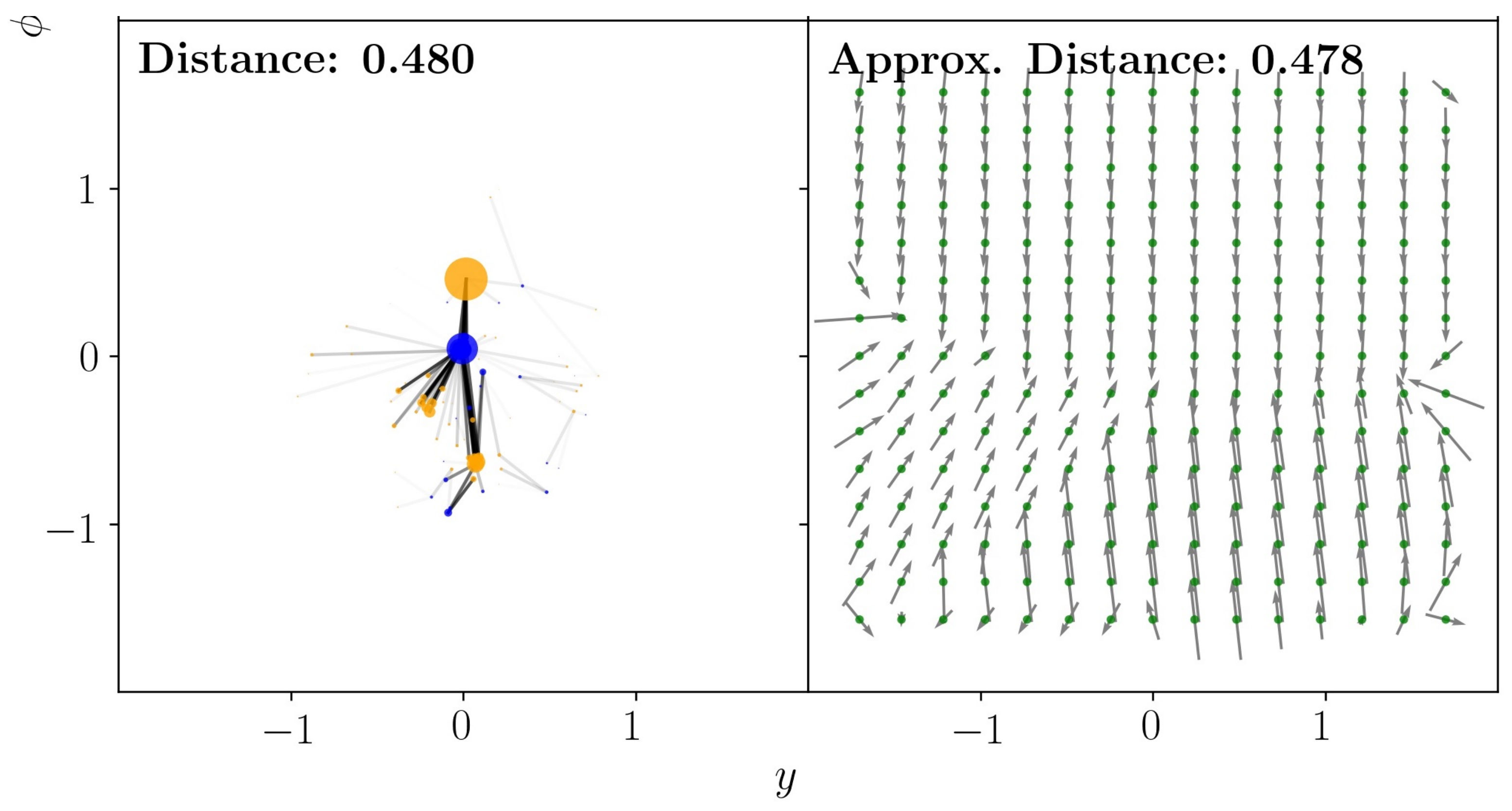
[Wang et al., International Journal of Computer Vision 101, 254 (2013)]

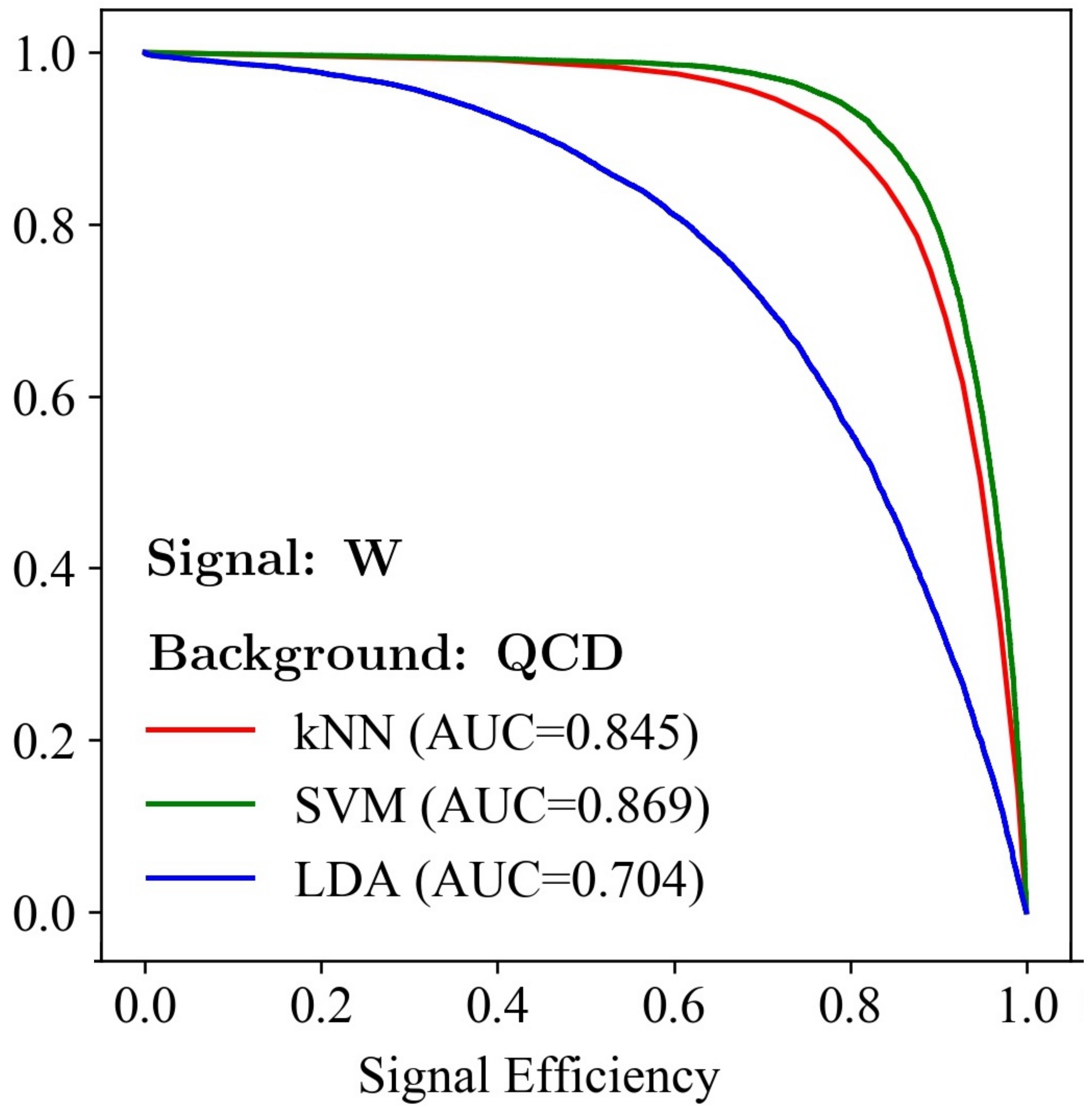
Supervised ML w/ LOT

- **k-Nearest Neighbor (kNN):** classification via majority vote of closest k neighbors in training set
- **Support Vector Machine (SVM):** lift inputs into high-dim space, find optimal hyperplane separating data
- **Linear Discriminate Analysis (LDA):** projects input data onto most discriminating linear combination

W/QCD jet classification

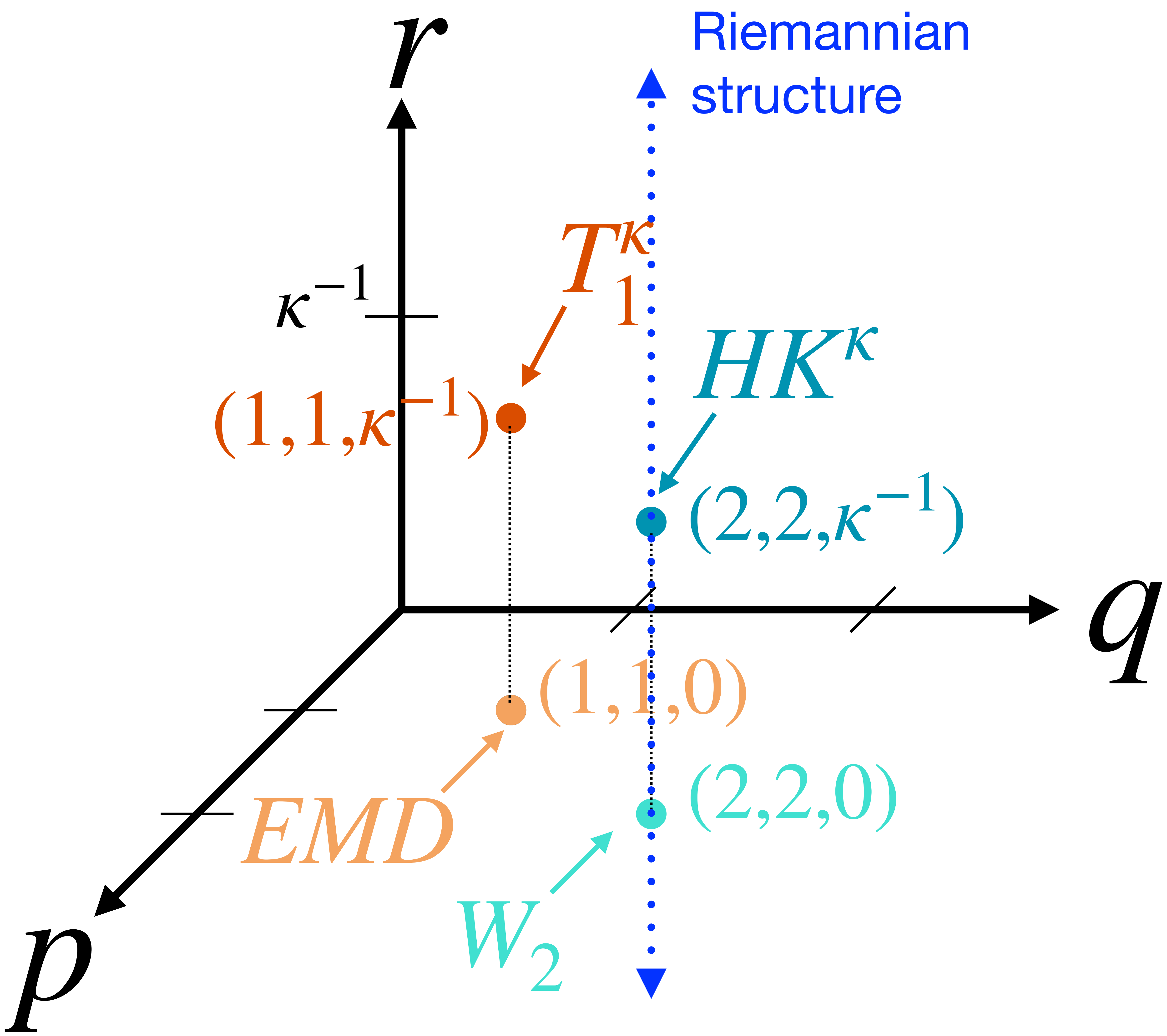






Datasets	Model	AUC
Our work	$k_{=20}$NN-LOT	0.845
	SVM-LOT	0.869
	LDA-LOT	0.704
Komiske, Metodiev, Thaler 1902.02346	$k_{=32}$NN-EMD	0.887
	$\tau_2^{\beta=1} / \tau_1^{\beta=1}$	0.776
	PFN	0.919
	EFPs	0.917
	EFN	0.904

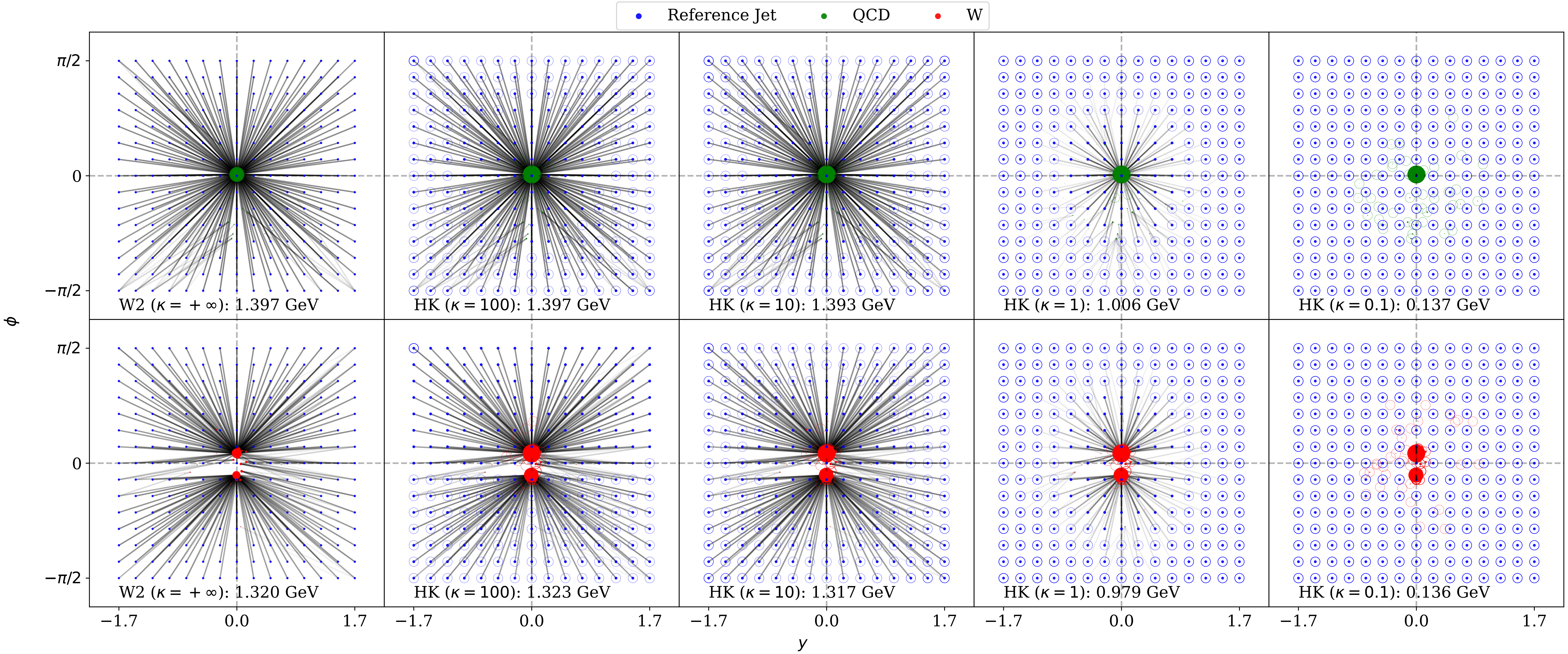
Which Metric on the Space of Collider Events?



- Optimal transport metrics are a natural choice to compare collider events since they preserve spatial information.
- A key limitation of W_p metrics is that they require distributions to be normalized.
- The **partial optimal transport** metric T_1^κ generalizes **EMD** to allow for creation and destruction of mass [Piccoli, Rossi, '14, '16], [Komiske, et. al. '19].
- The **Hellinger-Kantorovich** metric HK^κ generalizes $\mathbf{W_2}$ to allow for creation and destruction of mass [Liero, et. al. '16, '18], [Chizat, et. al. '18], [Kondratyev et. al. '16]; mass never moves more than distance κ.

Linearized Unbalanced Optimal Transport

“Linearized Hellinger-Kantorovich Distance” (Cai, Cheng, Schmitzer, Thorpe [2102.08807, math.OC])



$\kappa = \infty$: original W_2 distance ⟵ κ ⟶ **$\kappa = 0$: rescaled Euclidean distance**

Particle Linearized Unbalanced Optimal Transport (PLUOT)

"Which Metric on the Space of Collider Events?" [arXiv: 2111.03670] w/ **Tianji Cai & Junyi Cheng**

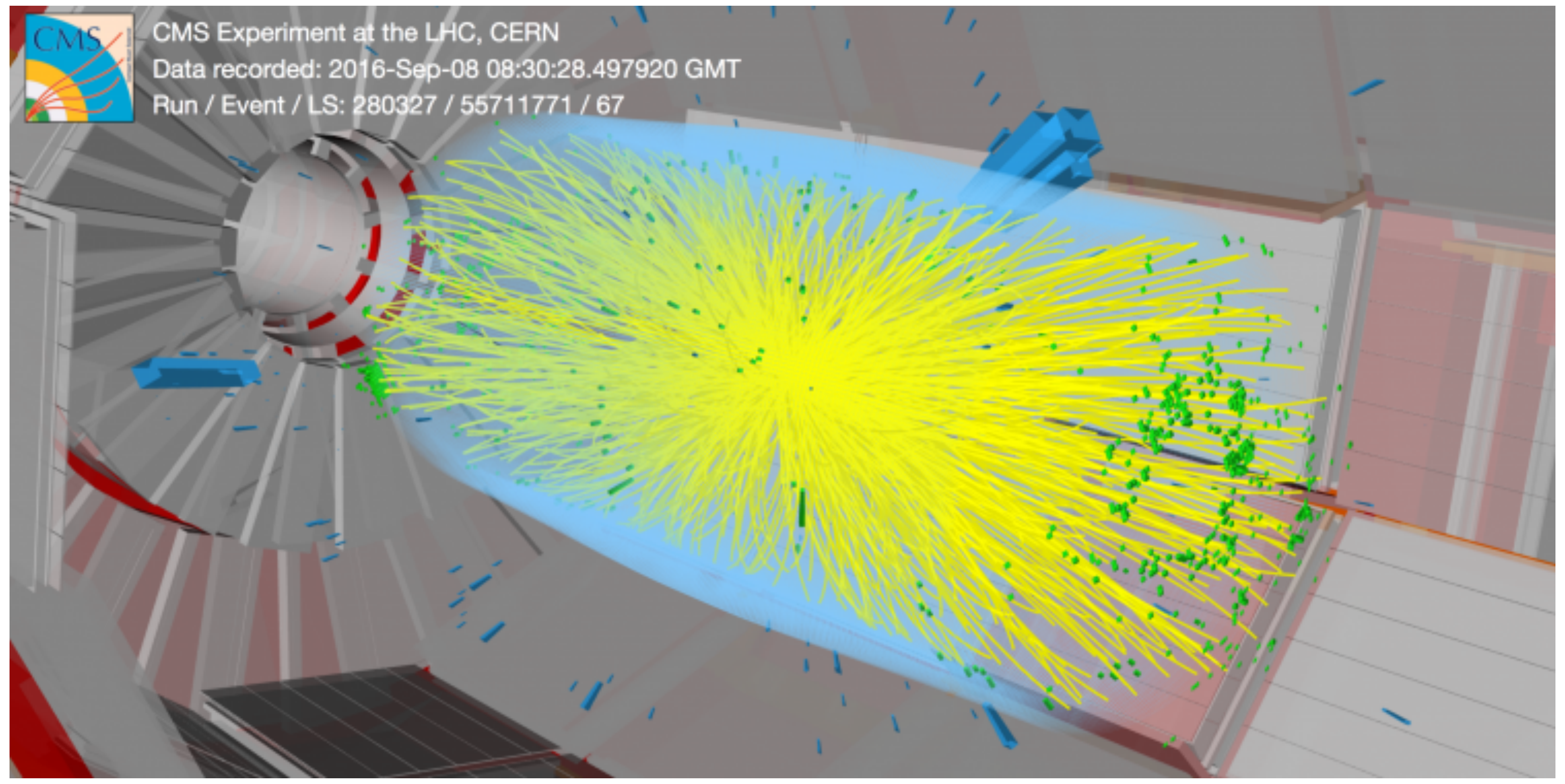
Low-κ falloff set by scale of reference measure

Anything EMD can do, linearized HK can do faster

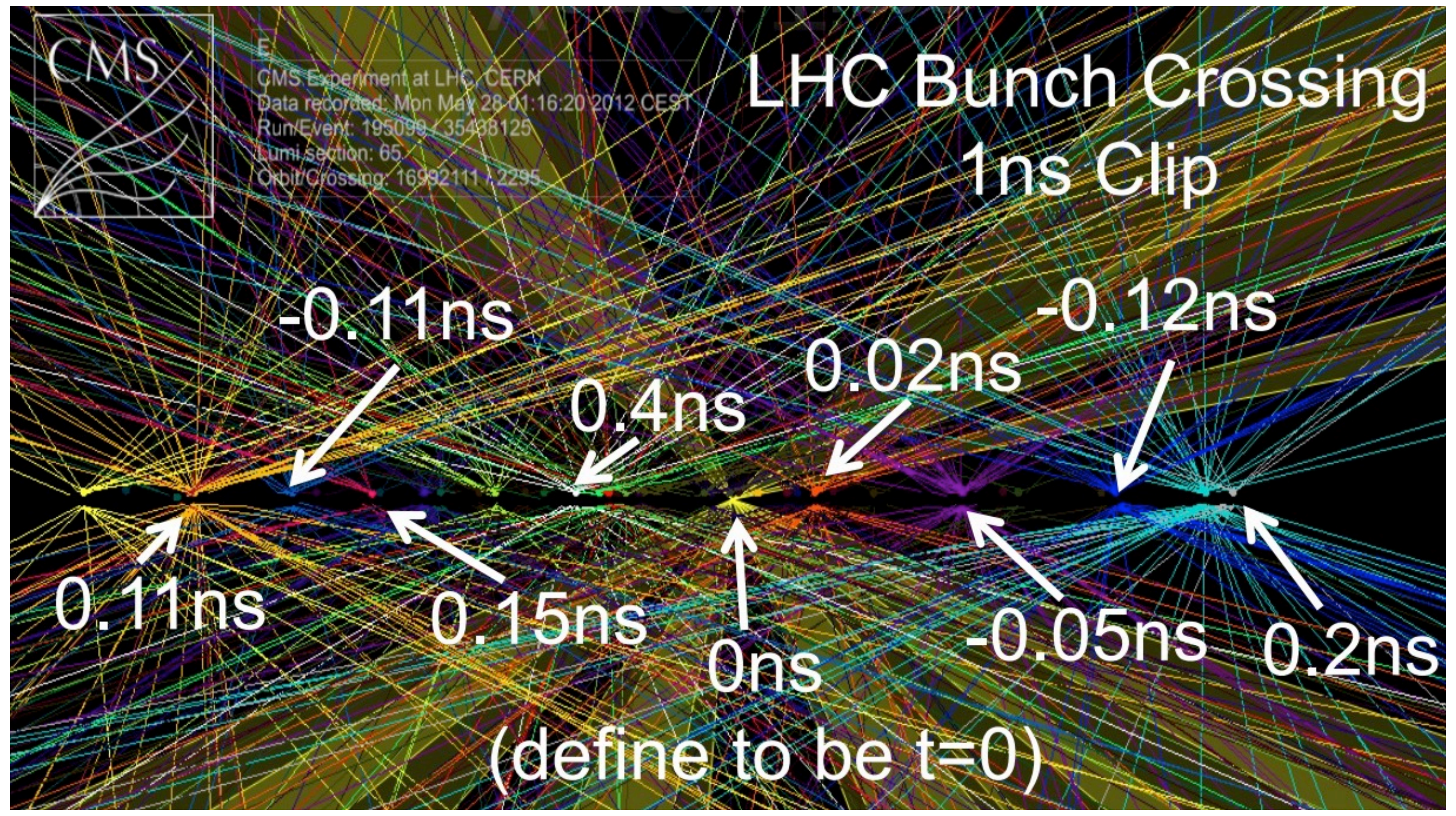
High-κ falloff for non-normalized events because total energy differences have increasing cost

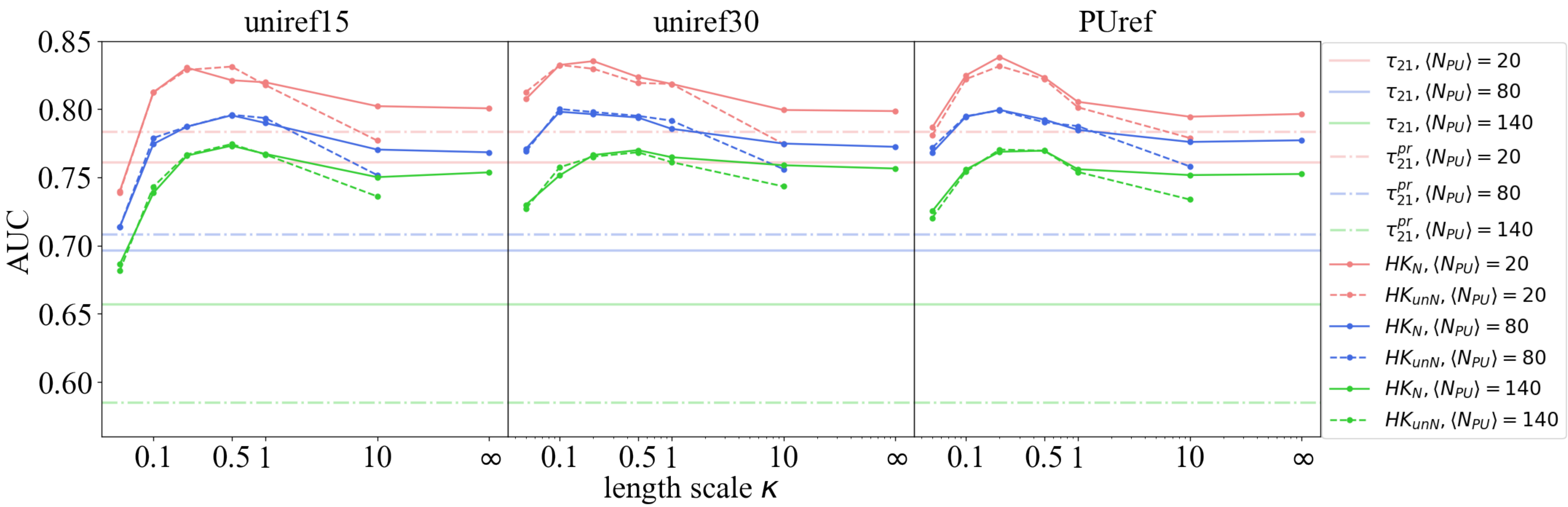
***Linearized HK distance** has something particularly compelling to offer when pT range of events grows large.*

Pileup (Noise) Robustness



[N. Cartiglia, INFN, Torino - EPS Venice 07/07/17]





Some things to do…

- Explore different choices of ground measure, perhaps reflecting underlying symmetries (see e.g. [Larkoski & Melia 2008.06508])
- Explore the space of Kantorovich dual inequalities and their consequences
- Treat full events in OT (including various object categories) with multi-species optimal transport
- Use OT distances as input to a wider array of ML-based analyses strategies for particle physics (see e.g. [Fraser et al. 2110.06948])
- Connect the geometry of collider events and pQCD calculations
- Use collider events (simulated or Open Data) as a benchmark for other ML tools

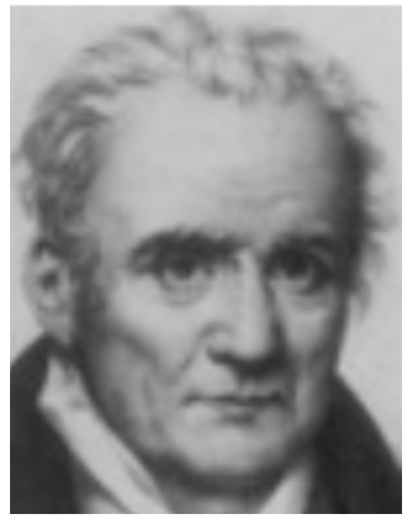
Conclusions

- Optimal transport increasingly important in PDEs, geometry, statistics, economics, image processing, and machine learning, but party is just getting started in particle physics.
- Provides a natural metric on the space of collider events with ideal properties.
- Useful for geometrization of LHC data, event classification, nonperturbative bounds, unifying description of collider observables…
- …and now easily computable on your laptop using Linearized Optimal Transport.
- More broadly, there may be considerable advantages for exploring balanced and unbalanced OT with Riemannian structure in the context of particle physics.
- *Much to explore!*

Thank you!

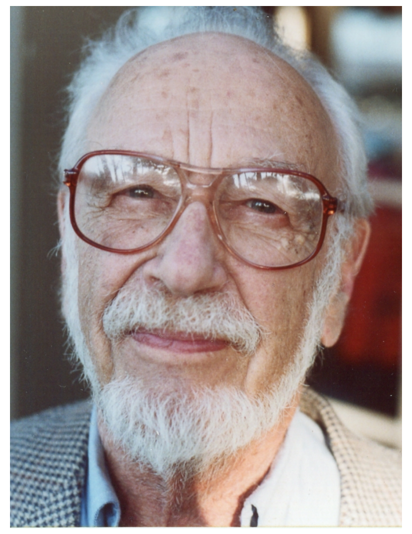
Optimal Transport

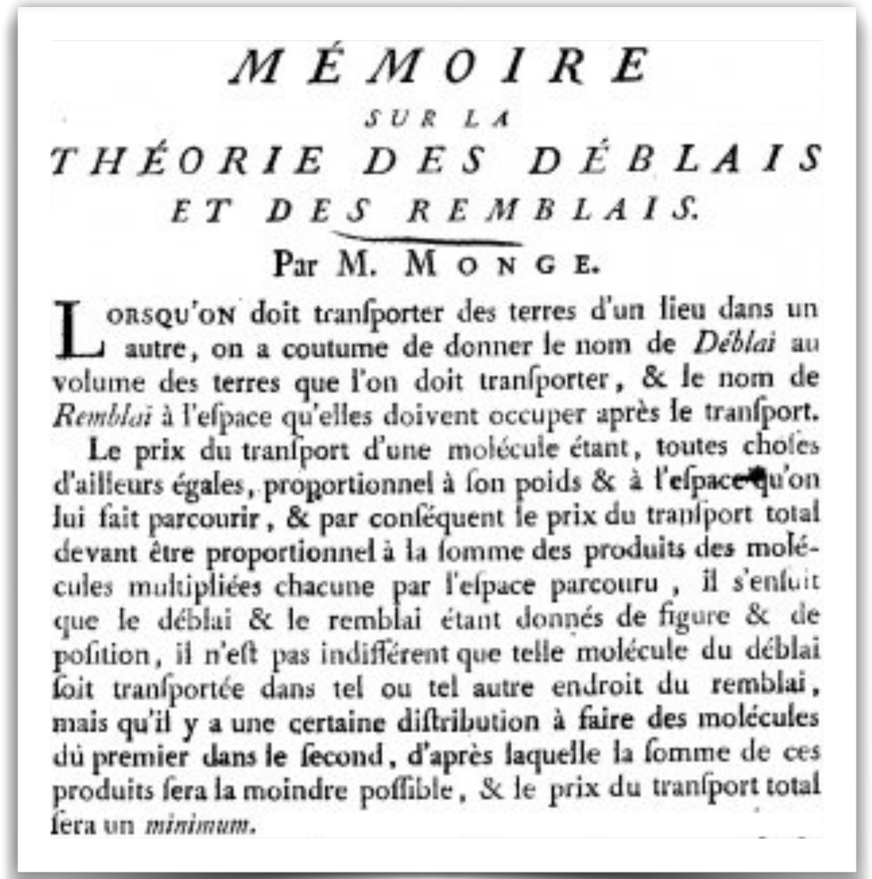
- 1781: Gaspard Monge,
  Mémoire sur la théorie des deblais et des remblais
- 1942: Leonid Kantorovich,
  On the translocation of masses
- 2000: Felix Otto, Cedric Villani,
  Generalization of an inequality by Talagrand, as a consequence of the logarithmic Sobolev inequality
- 2010: Cedric Villani wins Fields medal
- 2012: Shapely and Roth win Nobel Prize in economics
- 2018: Alessio Figalli wins Fields medal









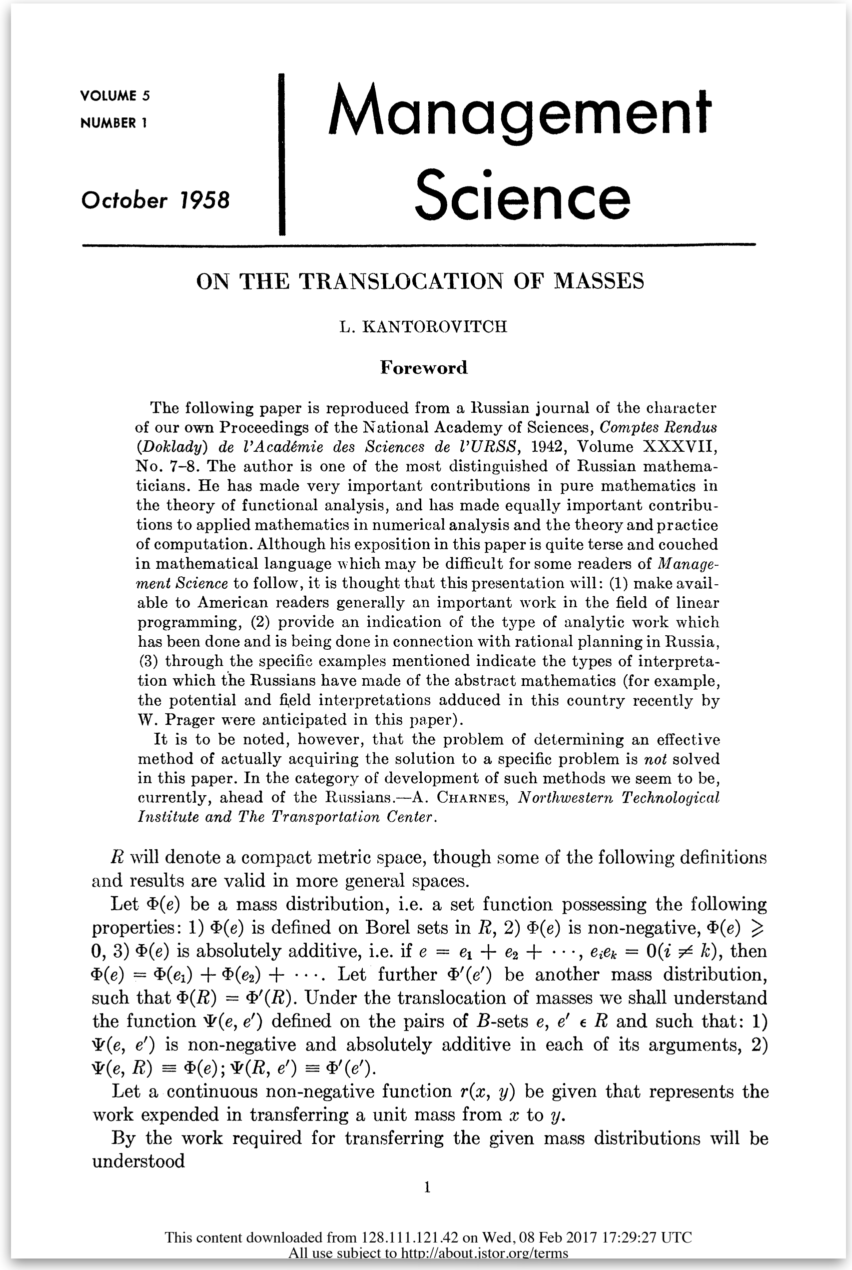


MÉMOIRE SUR LA THÉORIE DES DÉBLAIS ET DES REMBLAIS.

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.

Le prix du tranſport d'une molécule étant, toutes choſes d'ailleurs égales, proportionnel à ſon poids & à l'eſpace qu'on lui fait parcourir, & par conſéquent le prix du tranſport total devant être proportionnel à la ſomme des produits des molécules multipliées chacune par l'eſpace parcouru, il s'enſuit que le déblai & le remblai étant donnés de figure & de poſition, il n'eſt pas indifférent que telle molécule du déblai ſoit tranſportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine diſtribution à faire des molécules du premier dans le ſecond, d'après laquelle la ſomme de ces produits ſera la moindre poſſible, & le prix du tranſport total ſera un *minimum*.



VOLUME 5
NUMBER 1
October 1958
Management Science

ON THE TRANSLOCATION OF MASSES

L. KANTOROVITCH

Foreword

The following paper is reproduced from a Russian journal of the character of our own Proceedings of the National Academy of Sciences, *Comptes Rendus (Doklady) de l'Académie des Sciences de l'URSS*, 1942, Volume XXXVII, No. 7–8. The author is one of the most distinguished of Russian mathematicians. He has made very important contributions in pure mathematics in the theory of functional analysis, and has made equally important contributions to applied mathematics in numerical analysis and the theory and practice of computation. Although his exposition in this paper is quite terse and couched in mathematical language which may be difficult for some readers of *Management Science* to follow, it is thought that this presentation will: (1) make available to American readers generally an important work in the field of linear programming, (2) provide an indication of the type of analytic work which has been done and is being done in connection with rational planning in Russia, (3) through the specific examples mentioned indicate the types of interpretation which the Russians have made of the abstract mathematics (for example, the potential and field interpretations adduced in this country recently by W. Prager were anticipated in this paper).

It is to be noted, however, that the problem of determining an effective method of actually acquiring the solution to a specific problem is *not* solved in this paper. In the category of development of such methods we seem to be, currently, ahead of the Russians.—A. CHARNES, *Northwestern Technological Institute and The Transportation Center*.

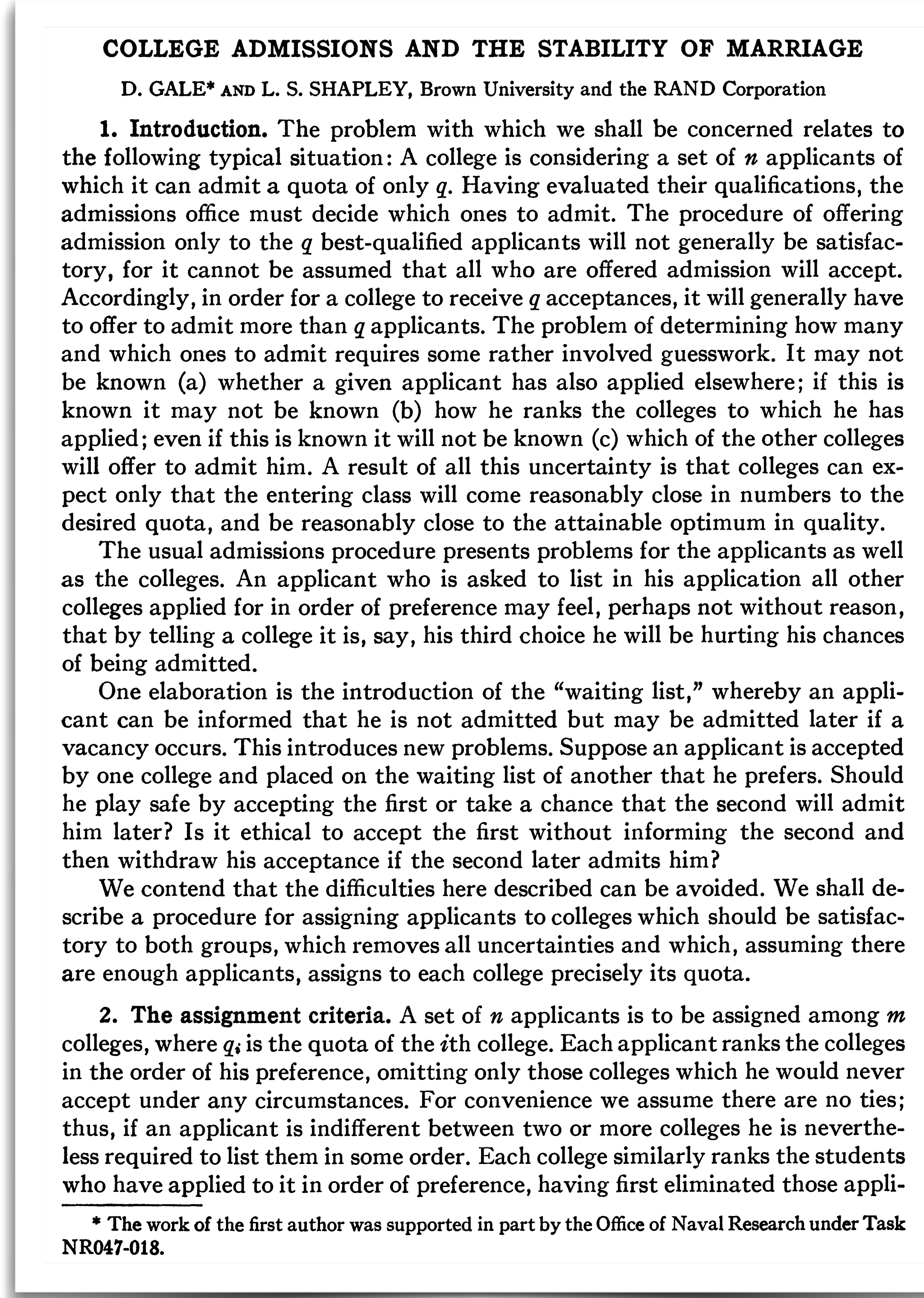
R will denote a compact metric space, though some of the following definitions and results are valid in more general spaces.

Let $\Phi(e)$ be a mass distribution, i.e. a set function possessing the following properties: 1) $\Phi(e)$ is defined on Borel sets in R, 2) $\Phi(e)$ is non-negative, $\Phi(e) \geqslant 0$, 3) $\Phi(e)$ is absolutely additive, i.e. if $e = e_1 + e_2 + \cdots$, $e_i e_k = 0 (i \neq k)$, then $\Phi(e) = \Phi(e_1) + \Phi(e_2) + \cdots$. Let further $\Phi'(e')$ be another mass distribution, such that $\Phi(R) = \Phi'(R)$. Under the translocation of masses we shall understand the function $\Psi(e, e')$ defined on the pairs of B-sets $e, e' \in R$ and such that: 1) $\Psi(e, e')$ is non-negative and absolutely additive in each of its arguments, 2) $\Psi(e, R) \equiv \Phi(e)$; $\Psi(R, e') \equiv \Phi'(e')$.

Let a continuous non-negative function $r(x, y)$ be given that represents the work expended in transferring a unit mass from x to y.

By the work required for transferring the given mass distributions will be understood

1



COLLEGE ADMISSIONS AND THE STABILITY OF MARRIAGE

D. GALE* AND L. S. SHAPLEY, Brown University and the RAND Corporation

1. Introduction. The problem with which we shall be concerned relates to the following typical situation: A college is considering a set of n applicants of which it can admit a quota of only q. Having evaluated their qualifications, the admissions office must decide which ones to admit. The procedure of offering admission only to the q best-qualified applicants will not generally be satisfactory, for it cannot be assumed that all who are offered admission will accept. Accordingly, in order for a college to receive q acceptances, it will generally have to offer to admit more than q applicants. The problem of determining how many and which ones to admit requires some rather involved guesswork. It may not be known (a) whether a given applicant has also applied elsewhere; if this is known it may not be known (b) how he ranks the colleges to which he has applied; even if this is known it will not be known (c) which of the other colleges will offer to admit him. A result of all this uncertainty is that colleges can expect only that the entering class will come reasonably close in numbers to the desired quota, and be reasonably close to the attainable optimum in quality.

The usual admissions procedure presents problems for the applicants as well as the colleges. An applicant who is asked to list in his application all other colleges applied for in order of preference may feel, perhaps not without reason, that by telling a college it is, say, his third choice he will be hurting his chances of being admitted.

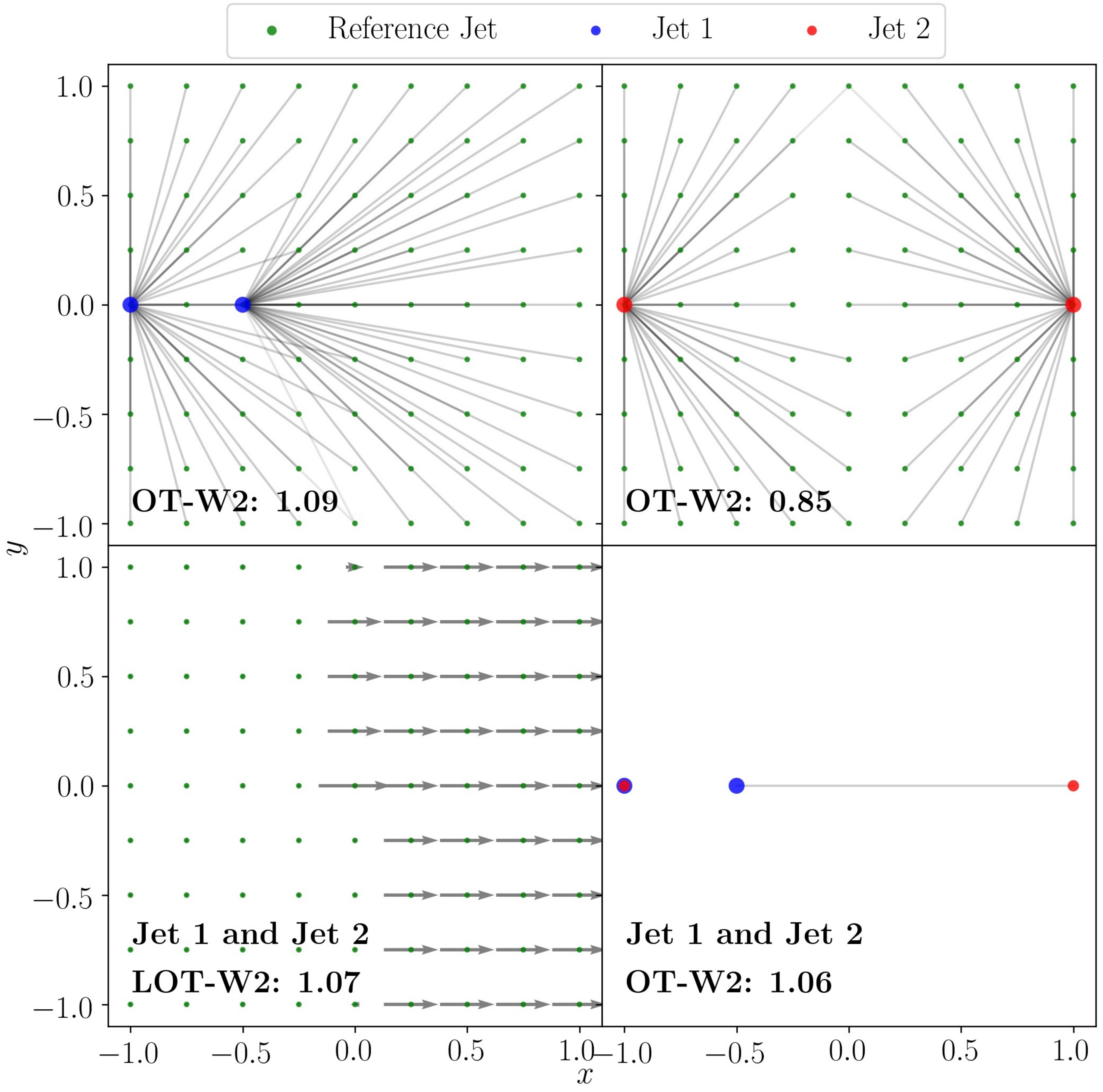
One elaboration is the introduction of the "waiting list," whereby an applicant can be informed that he is not admitted but may be admitted later if a vacancy occurs. This introduces new problems. Suppose an applicant is accepted by one college and placed on the waiting list of another that he prefers. Should he play safe by accepting the first or take a chance that the second will admit him later? Is it ethical to accept the first without informing the second and then withdraw his acceptance if the second later admits him?

We contend that the difficulties here described can be avoided. We shall describe a procedure for assigning applicants to colleges which should be satisfactory to both groups, which removes all uncertainties and which, assuming there are enough applicants, assigns to each college precisely its quota.

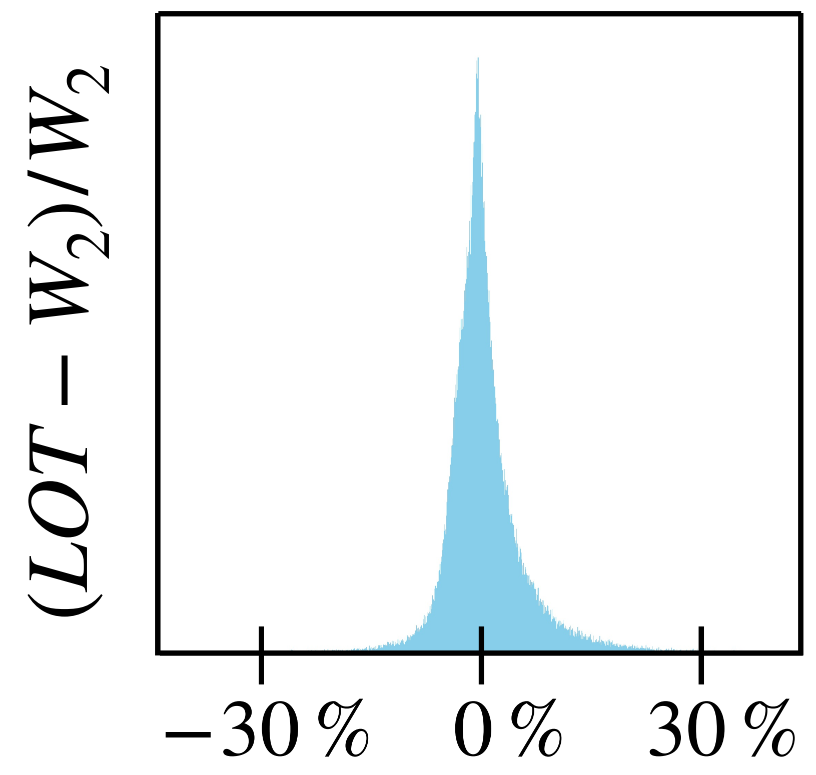
2. The assignment criteria. A set of n applicants is to be assigned among m colleges, where q_i is the quota of the ith college. Each applicant ranks the colleges in the order of his preference, omitting only those colleges which he would never accept under any circumstances. For convenience we assume there are no ties; thus, if an applicant is indifferent between two or more colleges he is nevertheless required to list them in some order. Each college similarly ranks the students who have applied to it in order of preference, having first eliminated those appli-

* The work of the first author was supported in part by the Office of Naval Research under Task NR047-018.

Linearized OT in Practice



- We choose uniform reference on 15 x 15 grid.
- **Thm** (Cai, Cheng, C., C., '20): Fully discrete $LOT_{r,r'}(\mathcal{E}, \mathcal{E}')$ converges to continuum $W_{2,\mathcal{R}}(\mathcal{E}, \mathcal{E}')$, as discrete reference converges to a continuum reference.
- **Thm** (Delalande, Merigot '21): If $\mathcal{R}$ is the uniform distribution on the domain $\mathbf{\Omega}$, $W_2(\mathcal{E}, \mathcal{E}') \leq W_{2,\mathcal{R}}(\mathcal{E}, \mathcal{E}') \leq C_\Omega W_2(\mathcal{E}, \mathcal{E}')^{1/6}$



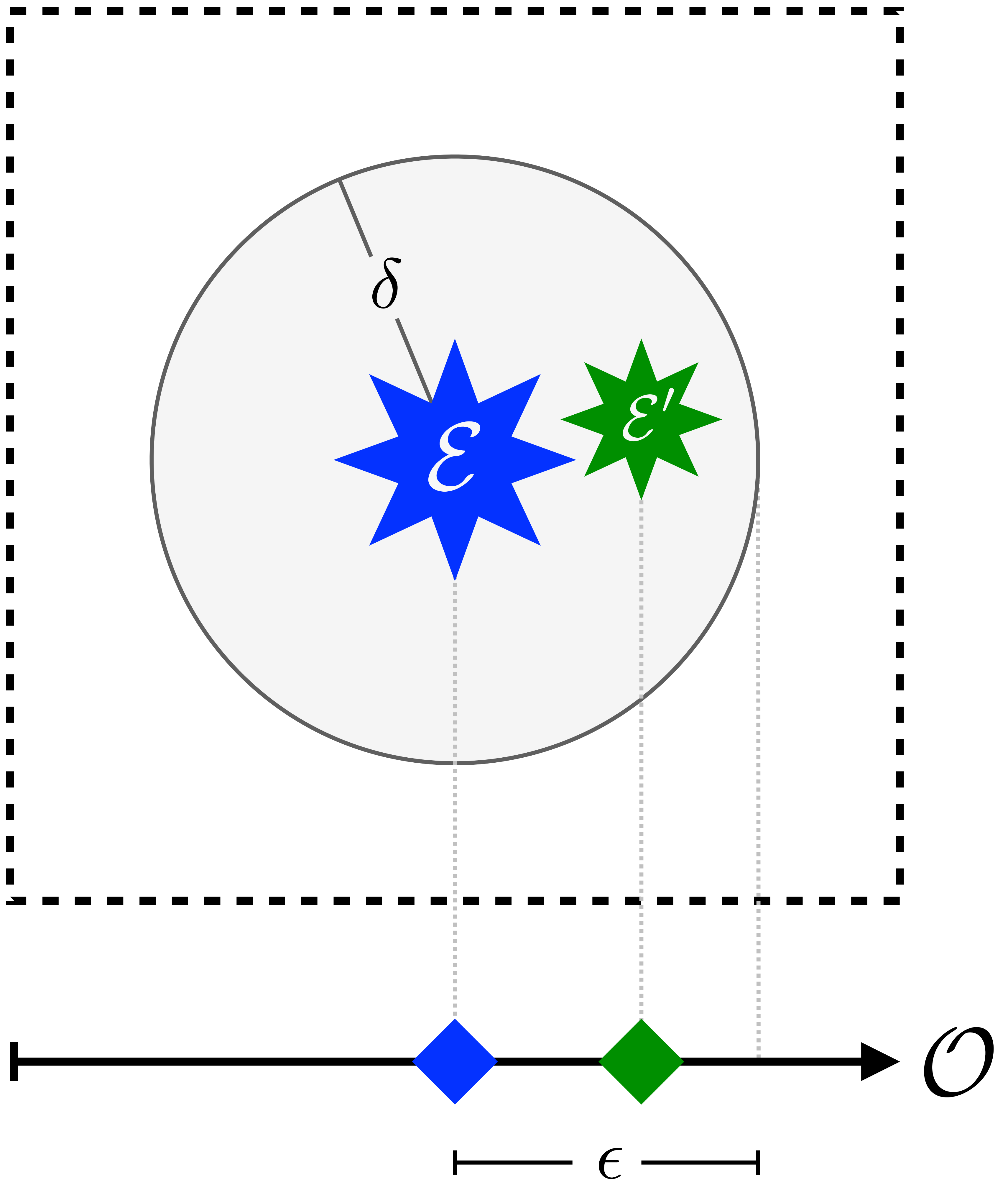
avg = 0.67%

std = 5.82%

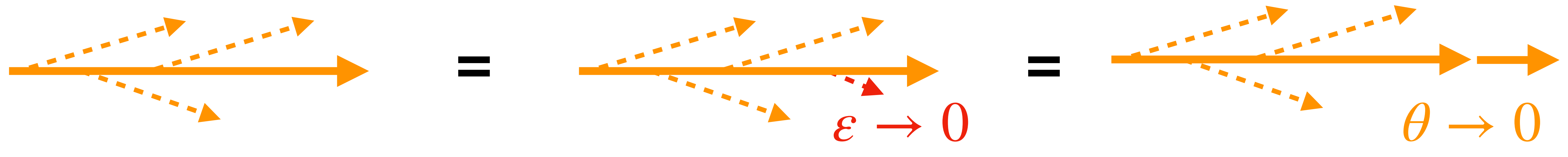
Tianji Cai, **Junyi Cheng**, **KC**, **NC** [arXiv:2008.08604]

OT Infrared/Collinear Safety

Komiske, Metodiev, Thaler [2004.04159]



OT distances insensitive to infinitesimal soft/collinear emission



OT continuity of observable $\mathcal{O}$ for event $\mathscr{E}$
(KMT [2004.04159]): for any $\epsilon > 0$ there exists δ s.t.

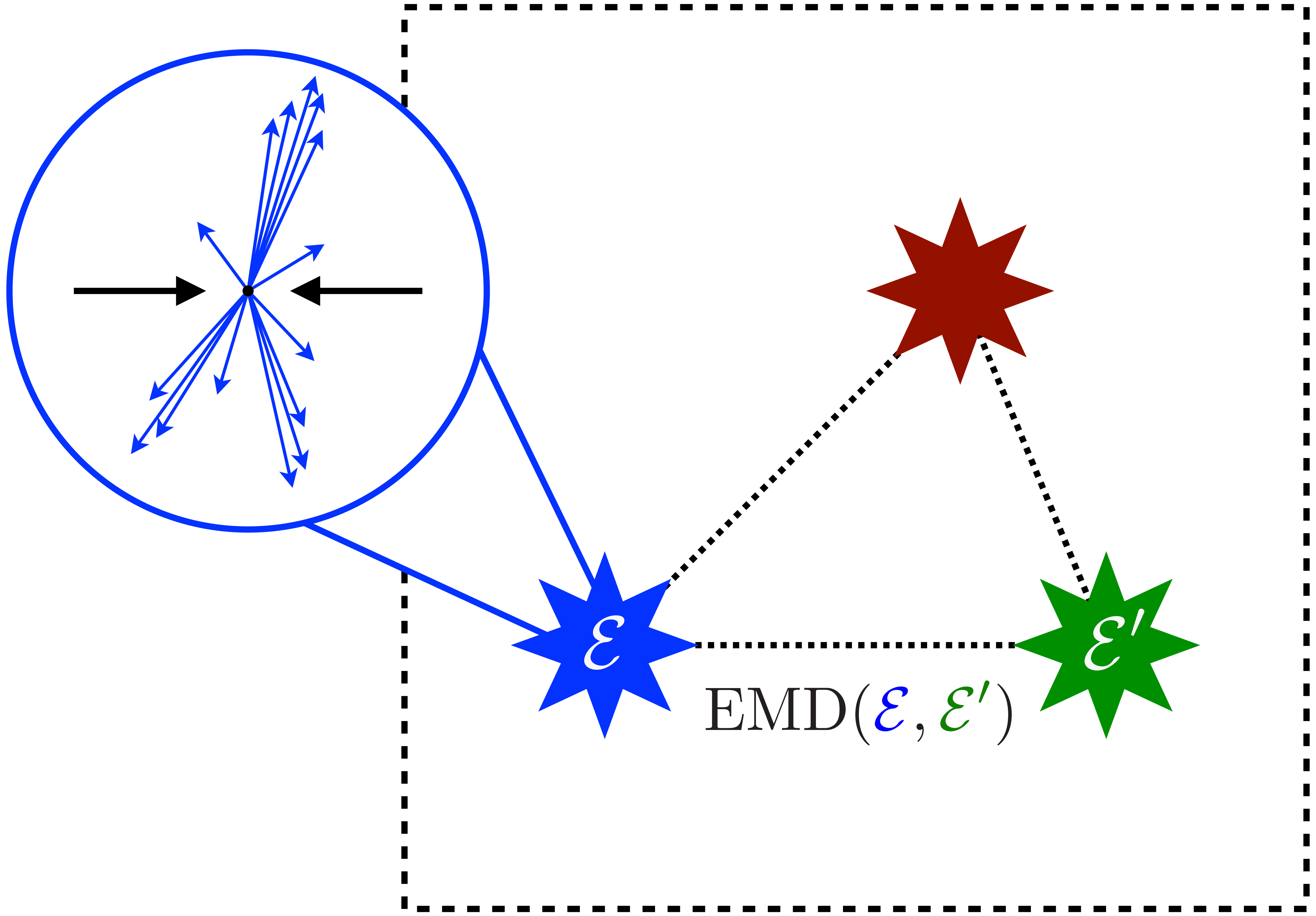
$$W_1(\mathcal{E}, \tilde{\mathcal{E}}) < \delta \to |\mathcal{O}(\mathcal{E}) - \mathcal{O}(\tilde{\mathcal{E}})| < \epsilon$$

Robust notion of IRC safety: An observable is IRC safe if it is OT continuous for all energy flows, except potentially on a negligible set of events.

OT for Particle Physics

Komiske, Metodiev, Thaler 1902.02346: OT (EMD) is useful for collider physics

[Komiske, Metodiev, Thaler 2004.04159]



Defined **“Energy Mover’s Distance”**

Akin to **“Earth Mover’s Distance”**

Accounting for **energy difference** (non-unique)

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

θ_{ij} angular distance in y-ϕ plane, ***R*** weights terms

$$f_{ij} \geq 0, \sum_j f_{ij} \leq E_i, \sum_i f_{ij} \leq \tilde{E}_j, \sum_{ij} f_{ij} = \min(\sum_i E_i, \sum_j \tilde{E}_j)$$