Topological Structure of Shot Selection in Basketball

Ricardo Reyes and Evan Ginsburg

November 2025

Abstract

Basketball analytics typically summarize shot selection with coarse counts and averages (shot charts, zones, expected value). In this project I use tools from Topological Data Analysis (TDA) to study the *geometry* of shot selection in large spatial point clouds of made and missed shots from the NBA and NCAA men's college basketball. I model shot locations as point clouds in $\mathbb{R}^2$, build Vietoris–Rips filtrations, and compute persistent homology in dimensions H_0 and H_1 using Ripser. Bottleneck and Wasserstein distances between persistence diagrams provide a metric on offensive "shapes."

I apply this framework to three related tasks. First, I compare the global offensive geometry of three NBA eras (2012–14, 2015–18, 2019–24) and quantify how far early and mid-era offenses are from the modern "pace and space" era. Second, I condition on shot-clock phase (transition, half-court set, late-clock) and show that desperation possessions exhibit materially different topological signatures than transition and set offense. Third, I construct a TDA-based distance matrix between high-volume NBA players and embed it into two dimensions to obtain a continuous "player archetype map" based purely on shot geometry. I also compare the topology of aggregated NCAA MBB shot selection to the modern NBA.

Overall, the results suggest that (i) the modern NBA's shot geometry was already partially present in earlier eras, (ii) late-clock offense is topologically distinct from the rest of the possession, and (iii) persistent homology yields a compact, model-free representation of player and league-level shot selection that could be used for similarity search and pre-draft role projection.

1 Introduction

Basketball offenses are often described informally in terms of "pace and space," "rim pressure," or "five-out" systems, but most quantitative work still reduces shot selection to coarse summaries: the share of attempts at the rim, in the midrange, or from three, along with expected points per shot. Hex–binned shot charts and location–based expected value models add spatial detail, yet they still treat locations largely independently and do not provide a global notion of how an offense *is arranged* in space. In particular, it is difficult to formalize questions such as whether two offenses "have the same shape" or how far the modern NBA is from earlier eras in a way that respects the geometry of the court.

Topological Data Analysis (TDA) offers a model–free way to summarize the geometry of large point clouds. Given a finite set of points in $\mathbb{R}^2$, one can build a Vietoris–Rips filtration by connecting points whose pairwise distance is below a scale parameter ε and then tracking how connected components and loops appear and disappear as ε grows. Persistent homology encodes this evolution in a persistence diagram, a multiset of points (b, d) recording the birth and death scales of topological features such as connected components (H_0) and loops (H_1). Distances such as the bottleneck and p-Wasserstein metrics then give a principled way to compare diagrams, and hence to compare the underlying point clouds, without specifying a parametric model for shot selection.

In this project I apply persistent homology to large-scale shot location data from the NBA and NCAA men's college basketball. I treat made and missed shot locations as point clouds in $\mathbb{R}^2$ and use Ripser to compute H_0 and H_1 persistence diagrams on Vietoris–Rips filtrations. Distances between the resulting diagrams are interpreted as distances between offensive "shapes." This viewpoint allows me to study how offensive geometry changes across time, across phases of the shot clock, and across different players and leagues.

More concretely, I address three questions:

(i) **Era comparison.** How different is the global shot geometry of the modern NBA from earlier eras? I aggregate shots by era, compute persistence diagrams for each era, and use bottleneck and Wasserstein distances to quantify how far early (2012–2014) and mid (2015–2018) eras are from the recent 2019–2024 "pace and space" era.

(ii) **Shot-clock phases.** Does the geometry of shot selection depend on where a possession is in the shot clock? I partition NBA shots into transition (17–24 seconds), set offense (7–17 seconds), and late-clock "desperation" (0–7 seconds), then compare their H_1 persistence diagrams to test whether late-clock possessions are topologically distinct from the rest of the offense.

(iii) **Player and league archetypes.** Can persistent homology provide a low-dimensional map of player archetypes and relate NCAA shot selection to the NBA? For high-volume NBA players I compute pairwise bottleneck distances between individual shot clouds and embed the resulting distance matrix into two dimensions to obtain a continuous "player archetype map." I also compare an aggregated NCAA MBB shot cloud to the modern NBA to see how closely college offenses mirror professional shot geometry.

Taken together, these experiments illustrate that persistent homology can capture meaningful differences in offensive geometry at the level of eras, shot-clock phases, and individual players. The resulting topology-based distances provide a compact summary of shot selection that complements more traditional location-based efficiency models and suggests a pathway toward similarity search and role-based player comparison grounded directly in spatial shot patterns.

2 Data and Methods

2.1 Shot location data

I work with a cleaned shot–level data set of NBA and NCAA men's basketball games from the 2012–2024 seasons. Each row corresponds to a single field–goal attempt and includes the game and team identifiers, period and clock information, the shot result, and continuous (x, y) court coordinates in feet. Throughout, I treat the provided coordinates as points in $\mathbb{R}^2$, so that an entire offense in a given season, era, or context is represented as a large planar point cloud of made and missed shots.

For the NBA, I use all regular–season and playoff shots from 2012–2024. After dropping rows with missing coordinates, this yields roughly 3.6 million shot locations in total. For the NCAA men's data (MBB), I aggregate all available Division I games from 2012–2024; after removing missing coordinates this produces on the order of 1.4 million shots. In both leagues I ignore shot outcome and other contextual variables and focus purely on the geometry of where shots are taken.

2.2 Experimental designs

The three questions in the introduction correspond to three ways of slicing these point clouds.

NBA eras. For the era comparison I partition NBA seasons into three eras:

- Era 1 (2012–2014), pre–"pace and space";
- Era 2 (2015–2018), early spread pick–and–roll and increasing three–point volume;
- Era 3 (2019–2024), the modern high–spacing era.

Within each era I pool all shots into a single point cloud $X_{\text{era}} \subset \mathbb{R}^2$. Because persistent homology on millions of points is computationally expensive, I draw a uniform random subsample of $n = 5000$ shots from each era without replacement. I then run persistent homology on these subsampled point clouds, yielding one H_0 and one H_1 persistence diagram per era.

Shot–clock phases. To study how geometry varies over the shot clock, I construct a shot–clock proxy $s \in [0, 24]$ for each NBA attempt and discretize it into three phases:

$$\text{Transition: } 17 \leq s \leq 24,$$
$$\text{Set offense: } 7 \leq s < 17,$$
$$\text{Desperation: } 0 \leq s < 7.$$

The first bucket is intended to capture early–offense and fast–break shots, the second captures half–court actions run with a comfortable amount of time, and the third captures late–clock heaves and bailout attempts. Again I form three large point clouds of shot locations, one for each phase, and randomly subsample $n = 8000$ points from each before computing persistent homology.

Player and league archetypes. For the player archetype map I restrict to the 2024 NBA season and group shots by shooter. I keep only "high–volume" players with at least 300 attempts, which yields 290 players. For each player p I collect their shot locations into a point cloud $X_p \subset \mathbb{R}^2$ and, to control runtime, draw a subsample of $n = 1000$ points before computing H_1 persistence. This produces one H_1 diagram per player. To compare the college and professional geometries, I also build a single aggregated NCAA shot cloud by pooling all MBB seasons, subsampling $n = 5000$ points, and computing its persistent homology.

2.3 Persistent homology pipeline

Given any point cloud $X \subset \mathbb{R}^2$ from the constructions above, I build a Vietoris–Rips filtration using the Euclidean metric. For a scale parameter $\varepsilon > 0$ I connect any pair of points within distance ε and fill in higher–dimensional simplices whenever all their edges are present. As ε grows from 0 to a large value, connected components merge and loops appear and eventually fill in.

Using the `ripser` backend, I compute persistent homology in dimensions H_0 and H_1 for each point cloud. The output is a pair of persistence diagrams, each a multiset of birth–death pairs (b, d) encoding when a connected component or loop appears in the filtration and when it merges or fills in. Intuitively, long bars or points far from the diagonal correspond to prominent topological features of the shot cloud: H_0 features correspond to clusters of shots, while H_1 features correspond to "holes" such as the no–man's land inside the three–point arc or regions of the court that are rarely used in a given context.

2.4 Distances and embeddings

To quantify differences between two contexts A and B (for example, two eras or two shot–clock phases) I compare their H_1 diagrams using both the bottleneck distance d_B and the 2–Wasserstein distance W_2. The bottleneck distance measures the largest discrepancy between matched features in the two diagrams, while W_2 averages discrepancies over all matched points. In both cases, standard stability theorems imply that these diagram distances are Lipschitz with respect to perturbations of the underlying point clouds, so that small geometric changes in shot locations cannot produce arbitrarily large changes in d_B or W_2.

For the three–era and three–phase experiments I collect the pairwise bottleneck and Wasserstein distances into a 3×3 symmetric matrix and then embed this matrix into two dimensions using principal component analysis (PCA) or classical multidimensional scaling (MDS). The resulting planar embeddings provide an interpretable visualization of how far each era or shot–clock phase lies from the others in TDA space. For the player experiment, I build a 290×290 distance matrix between player diagrams and apply MDS to obtain a two–dimensional "player archetype map" in which nearby players have similar shot geometries.

3 Results

3.1 NBA era geometry

Figure 1 shows the H_0 and H_1 persistence diagrams for the three NBA eras defined in Section 2.2. Qualitatively, all eras exhibit a dominant H_1 loop corresponding to the global "hole" created by the three–point line and the relative lack of midrange shots, together with several shorter–lived loops reflecting more localized gaps in shot coverage. The modern Era 3 diagrams contain slightly more medium–persistence H_1 features, consistent with a more polarized shot profile between rim and three–point attempts.

On the quantitative side, the H_1 bottleneck distances between eras are

$$d_B(\text{Era 1}, \text{Era 2}) \approx 1.49, \quad d_B(\text{Era 2}, \text{Era 3}) \approx 1.52, \quad d_B(\text{Era 1}, \text{Era 3}) \approx 1.70.$$

Thus early (2012–2014) and mid (2015–2018) offenses are slightly closer to each other than either is to the modern 2019–2024 era, but the separation is modest at the bottleneck scale. The 2–Wasserstein distances show a clearer gradient:

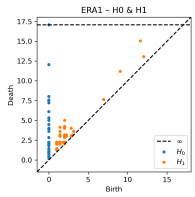
$$W_2(\text{Era 1}, \text{Era 2}) \approx 40.8, \quad W_2(\text{Era 2}, \text{Era 3}) \approx 67.1, \quad W_2(\text{Era 1}, \text{Era 3}) \approx 94.8,$$

so that the early era lies much farther from the modern era than from the mid era when averaging discrepancies over all features.

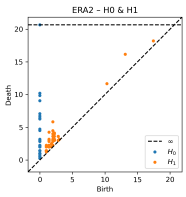
The planar PCA embedding of the 3×3 distance matrix (Figure 2) places Era 2 roughly between Era 1 and Era 3 along the first principal component, with Era 3 the most separated point. This picture supports the narrative that the modern NBA geometry did not arise abruptly: many of its topological features were already present in 2012–2014, but have intensified over time as midrange shots have been deemphasized and three–point spacing has increased.

3.2 Shot–clock phase geometry

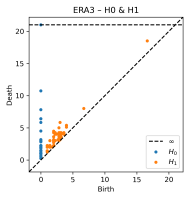
The shot–clock experiment isolates geometry within a possession. After constructing the three phases described in Section 2.2 and subsampling $n = 8000$ shots from each, I compute H_1 persis-



(a) Era 1 (2012–2014)



(b) Era 2 (2015–2018)



(c) Era 3 (2019–2024)

Figure 1: H_0 and H_1 persistence diagrams for the three NBA eras. Each point represents a connected component or loop in the Vietoris–Rips filtration of the era–level shot cloud.

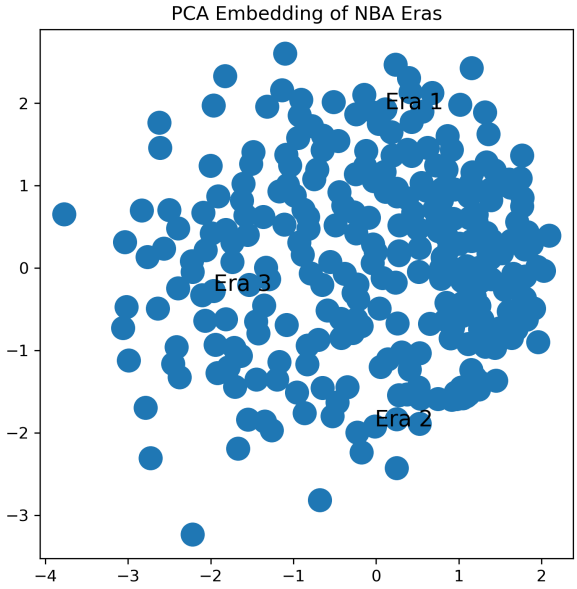


Figure 2: PCA embedding of the bottleneck distance matrix between NBA eras. Distances between points approximate the diagram distances between the corresponding eras.

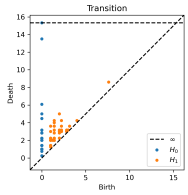
tence diagrams and their pairwise distances. The bottleneck distances are

$$d_B(\text{Transition}, \text{Set}) \approx 1.0, \quad d_B(\text{Set}, \text{Desperation}) \approx 2.49, \quad d_B(\text{Transition}, \text{Desperation}) \approx 2.49,$$

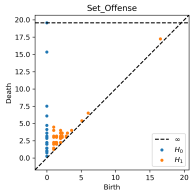
while the corresponding 2–Wasserstein distances are

$$W_2(\text{Transition}, \text{Set}) \approx 29.5, \quad W_2(\text{Set}, \text{Desperation}) \approx 44.4, \quad W_2(\text{Transition}, \text{Desperation}) \approx 52.9.$$

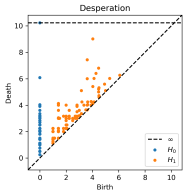
Figures 3 and 4 summarize these results. Transition and set offense have quite similar H_1 diagrams and lie close together in the MDS embedding, suggesting that, from a purely geometric perspective, early and mid–clock possessions share a common topological template. By contrast, late–clock "desperation" possessions are substantially farther away from both transition and set offense under both d_B and W_2, and occupy a separated position in the MDS plot. This supports the intuitive idea that bailout shots—often tightly contested pull–ups or forced attempts late in the clock—populate different regions of the court and induce a distinct topological signature.



(a) Transition (17–24s)



(b) Set offense (7–17s)



(c) Desperation (0–7s)

Figure 3: H_0 and H_1 persistence diagrams for the three shot–clock phases. Transition and set offense show similar loop structure, while desperation possessions exhibit more dispersed features.

3.3 College vs. professional shot geometry

To compare NCAA Division I offenses with the modern NBA, I construct a single aggregated MBB shot cloud by pooling all available college seasons (2012–2024), filter out missing coordinates, and subsample $n = 5000$ shots before computing persistence. I then compare the college H_1 diagram to the Era 3 NBA diagram.

The resulting bottleneck distance is

$$d_B(\text{NBA Era 3}, \text{MBB}) \approx 1.0,$$

with $W_2(\text{NBA Era 3}, \text{MBB}) \approx 34.9$. These values are comparable in scale to the early–vs–mid NBA distances, and substantially smaller than the early–vs–modern distance. In other words, the aggregated college geometry is topologically closer to the modern NBA than the early 2010s NBA is, at least at the level of global H_1 structure.

Figure 5 visualizes the bottleneck distance from the modern era to each comparison group: early NBA, mid NBA, and MBB. The bar for MBB lies between Era 1 and Era 2, suggesting that while college shot selection is not identical to the current NBA, its coarse topological structure already resembles that of an NBA offense with substantial spacing and three–point volume.

3.4 Player archetype map

Finally, I turn to the player–level analysis. Restricting to high–volume shooters (at least 300 attempts) in the 2024 NBA season yields 290 players. For each player I form a point cloud of shot locations, subsample $n = 1000$ points, compute H_1 persistence, and build the 290×290 bottleneck distance matrix between player diagrams. Applying MDS to this matrix produces a two–dimensional embedding that can be interpreted as a TDA–based "player archetype map" (Figure 6).

The map exhibits several coherent clusters as well as a few outliers. Players whose diagrams have many short–lived loops clustered near the diagonal tend to occupy dense regions of the map, corresponding to shooters who take a wide mix of shots across the court without strongly emphasizing any single region. In contrast, players with more pronounced H_1 features—for example, those whose shot selection heavily concentrates at the rim or in specific three–point zones—tend to lie farther from the bulk of the distribution. From a practical standpoint, this map provides a continuous notion of similarity between players based purely on the topology of their shot charts, independent of shooting efficiency, usage rate, or role labels.

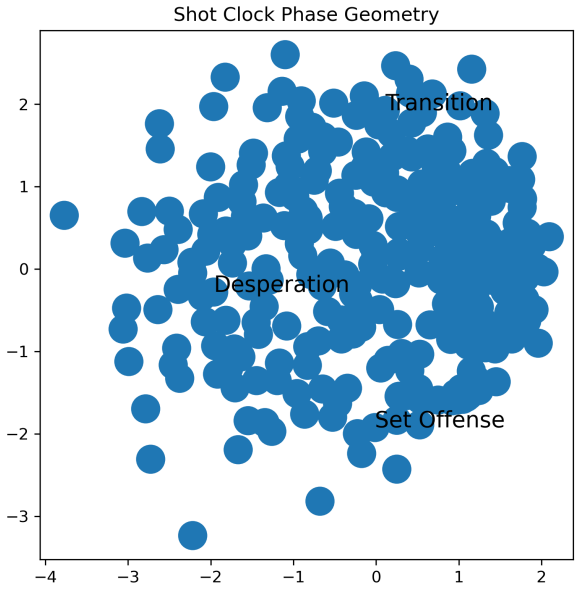


Figure 4: MDS embedding of the bottleneck distance matrix between shot–clock phases. Transition and set offense lie close together, while desperation possessions are clearly separated in TDA space.

4 Discussion and future directions

The experiments in this project suggest that persistent homology can capture meaningful aspects of offensive shot geometry at multiple levels of aggregation. At the league level, the era comparison shows that the modern NBA's "pace and space" geometry did not appear out of nowhere: the dominant H_1 loop associated with the three–point arc is present in all eras, and the bottleneck distances between eras are relatively small. At the same time, the Wasserstein distances reveal a clear progression in which the early 2010s era lies much farther from the modern era than from the mid 2010s, consistent with a gradual intensification of spacing and three–point emphasis rather than a single abrupt structural change.

Within possessions, the shot–clock phase analysis highlights that late–clock offense is topologically distinct from the rest of the shot selection landscape. Transition and set possessions share similar loop structure and lie close together in the MDS embedding, while desperation possessions are separated under both bottleneck and Wasserstein distances. This aligns with the intuitive idea that bailout shots come from different regions of the floor and fill in gaps that are rarely used when the offense is not under time pressure. From a modeling perspective, this suggests that treating all shots as exchangeable may miss systematic geometric differences across phases of the possession.

The comparison between NCAA and NBA geometry indicates that aggregated Division I men's shot selection is already reasonably close, in a topological sense, to the modern NBA. The college H_1 diagram lies at a bottleneck distance comparable to the early–vs–mid NBA distances, and much closer to the modern NBA than the early 2010s era is. One interpretation is that many of the structural ingredients of professional shot geometry—rim attacks, three–point spacing, and an underused midrange region—are already encoded in high–level college offenses, even though

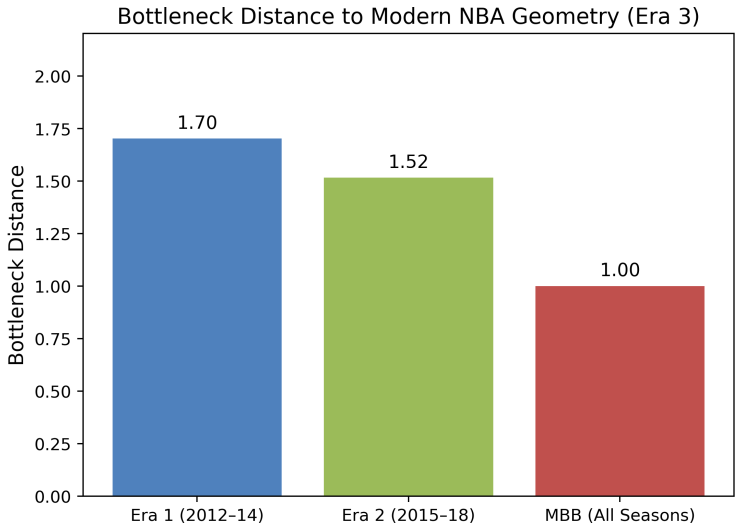


Figure 5: Bottleneck distance from the modern NBA era (2019–2024) to three comparison groups: early NBA (2012–2014), mid NBA (2015–2018), and aggregated NCAA MBB (2012–2024).

talent, efficiency, and schemes differ.

At the player level, the TDA–based archetype map provides a continuous, model–free notion of similarity that complements more conventional role labels. Unlike clustering on raw shot counts in pre–defined zones, the persistence–diagram approach is sensitive to the global arrangement of a player's shot chart, including where they *do not* shoot. In principle, this map could be used for tasks such as identifying college players whose shot geometry most closely resembles that of successful NBA role players, or tracking how an individual player's archetype evolves over time as their shot diet changes.

There are several limitations to this analysis. First, all computations rely on uniform subsampling of large point clouds to make persistence tractable. Although stability theorems guarantee that small perturbations of the underlying geometry cannot produce arbitrarily large changes in diagram distances, different random subsamples will introduce some variability in the exact numerical values. Second, I work exclusively with Vietoris–Rips filtrations on Euclidean distance and only with H_0 and H_1; alternative filtrations or higher–dimensional homology might capture additional structure, especially in richer feature spaces. Third, the analysis is intentionally outcome–agnostic: made and missed shots are treated identically, and contextual variables such as defender distance, play type, or game situation are ignored.

These caveats point to natural extensions. One direction is to study *weighted* point clouds in which shots are reweighted by efficiency or expected value, and to compare whether successful offenses or players occupy particular regions of TDA space. Another is to condition more finely on context—for example, comparing on–ball versus off–ball shooters, pick–and–roll versus isolation possessions, or different defensive schemes. Finally, one could combine TDA features with standard machine–learning models, using persistence diagrams or their vectorizations as inputs to predict outcomes such as team offensive rating, player development trajectories, or draft success. The results here indicate that persistent homology provides a flexible geometric summary of shot selection that is both interpretable and compatible with these more predictive frameworks.

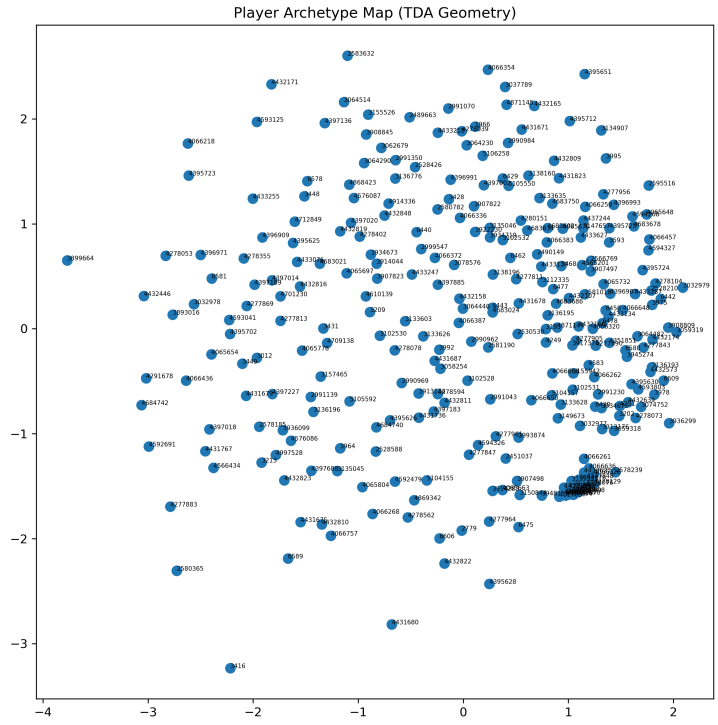


Figure 6: TDA–based player archetype map for high–volume NBA shooters in the 2024 season. Each point is a player; distances reflect bottleneck distances between their H_1 persistence diagrams.