# From Pixels to Points:
## Using Tracking Data to Measure Performance in Professional Basketball

April 23, 2018

Alexander Franks
UC Santa Barbara
afranks@pstat.ucsb.edu

# Why Sports?

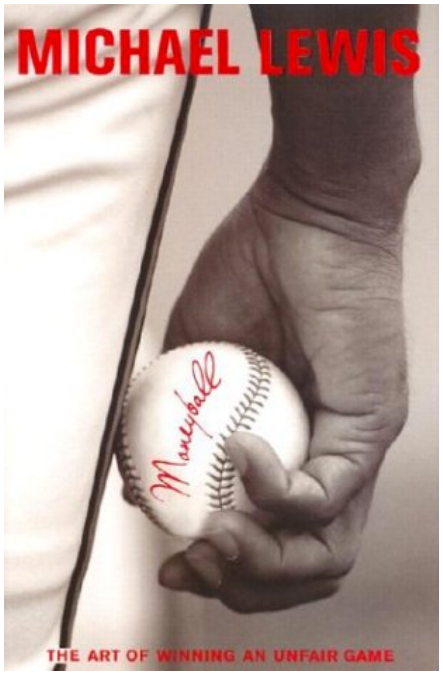
"Its not just "big data". It's something much better: rich data." --Nate Silver

http://fivethirtyeight.com/features/rich-data-poor-data/

# Why Sports?

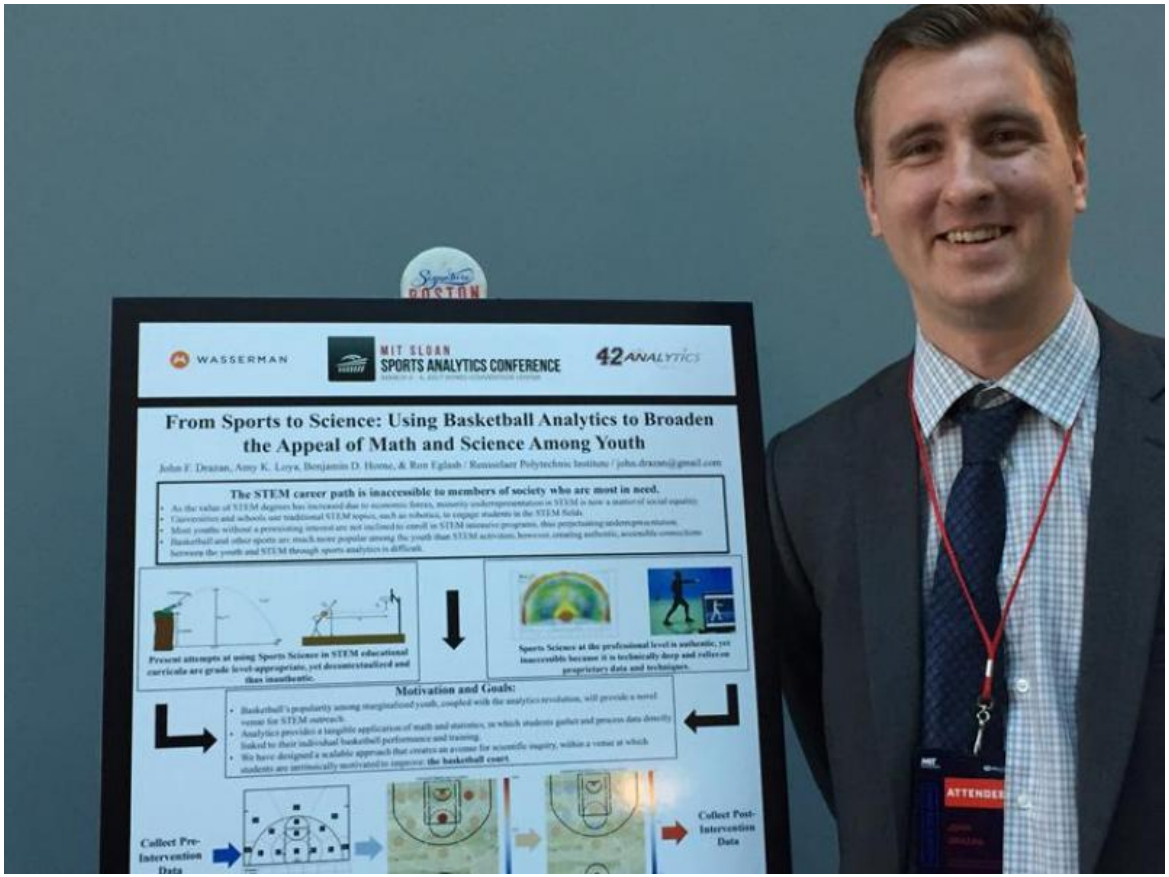
**Fundamental Ideas**

- Sample size
- Regression to the mean
- Bias-variance tradeoff
- Shrinkage and hierarchical models
- Spatial methods




**The power of analogy**

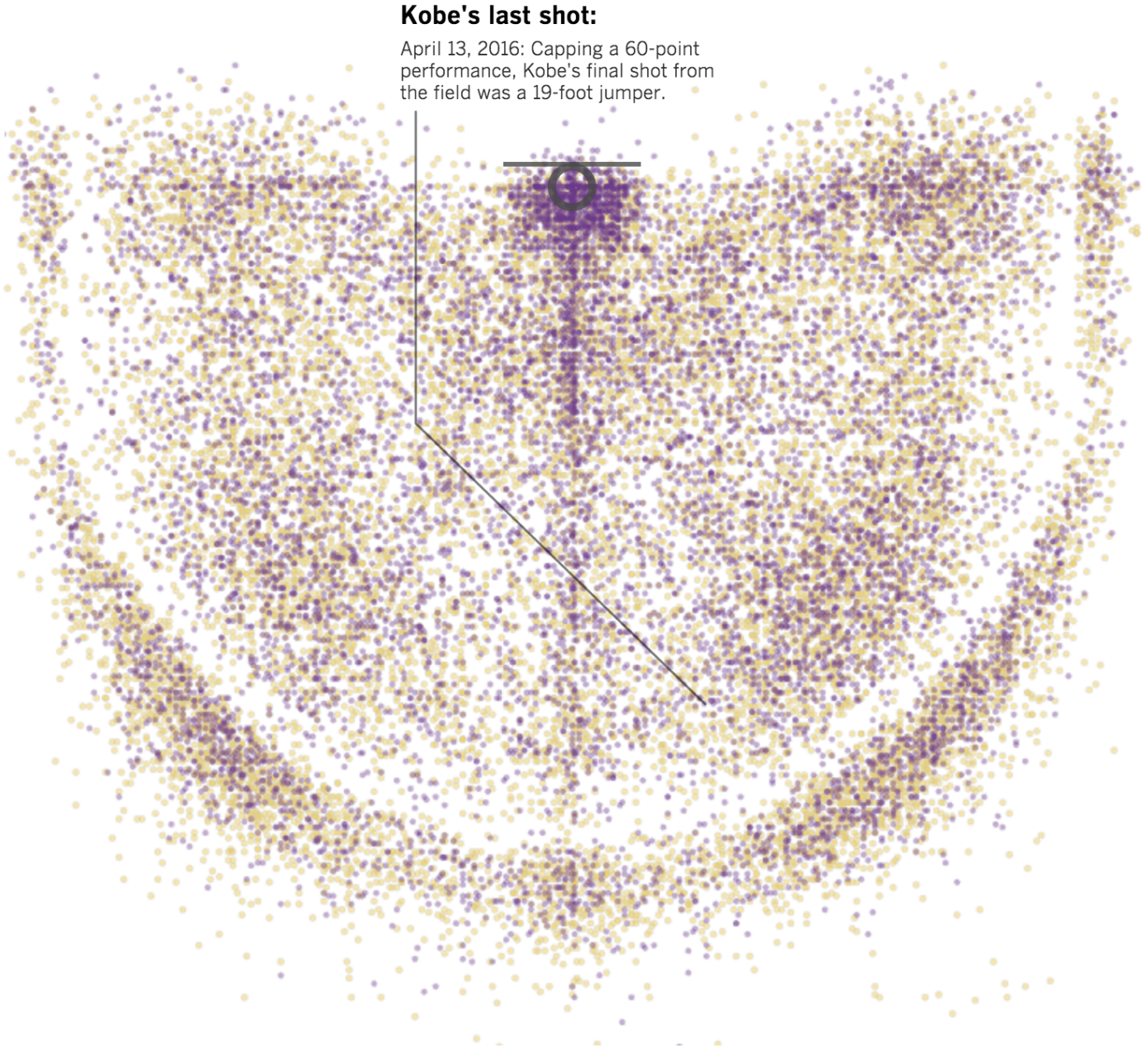
# Using Sports Analytics to Broaden the Appeal of Math and Science Among Youth



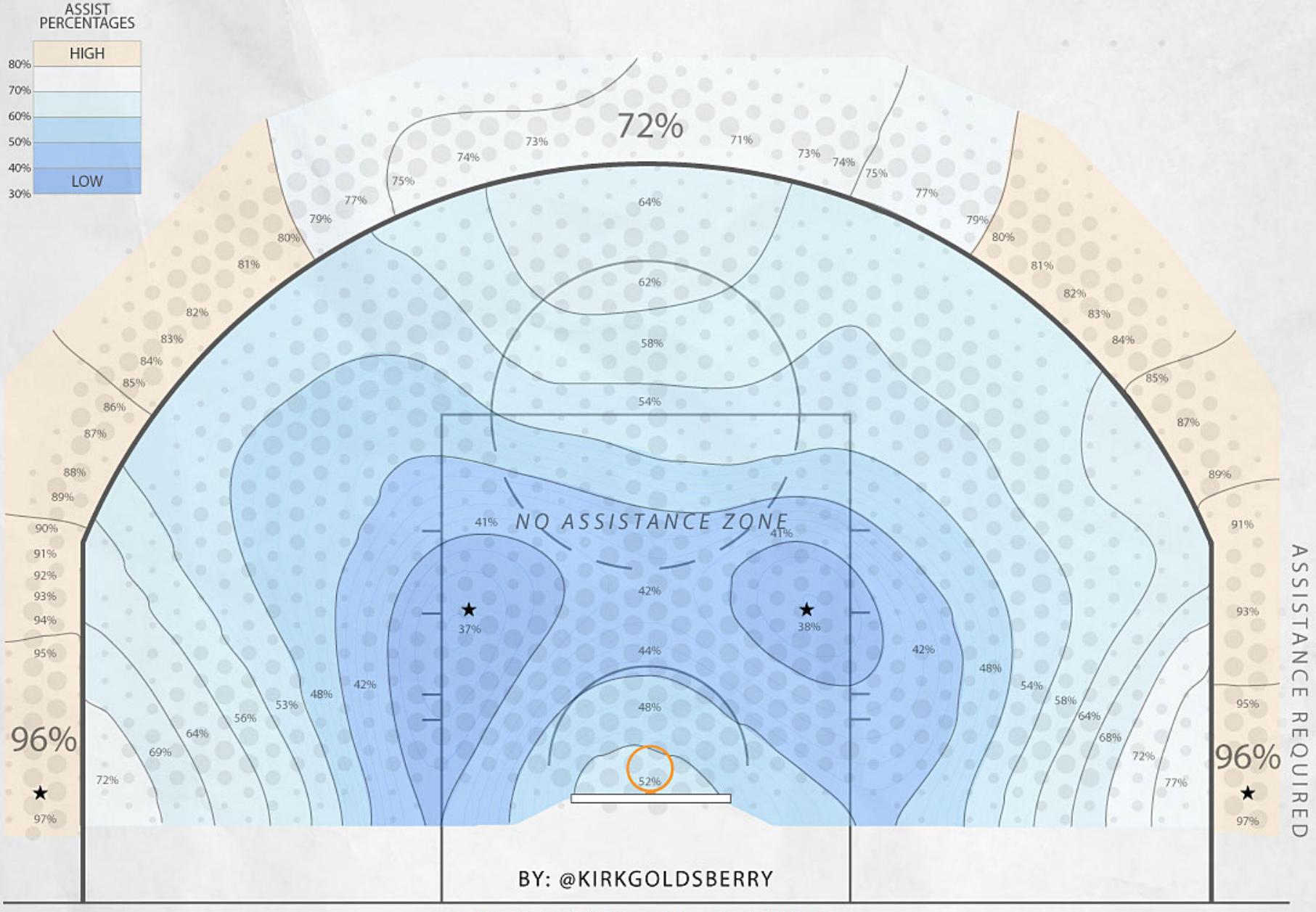

John Drazen, 2017 Sloan Research Competition Winner

Bryant attempted **30,699** shots throughout his career.

- Made
- Missed



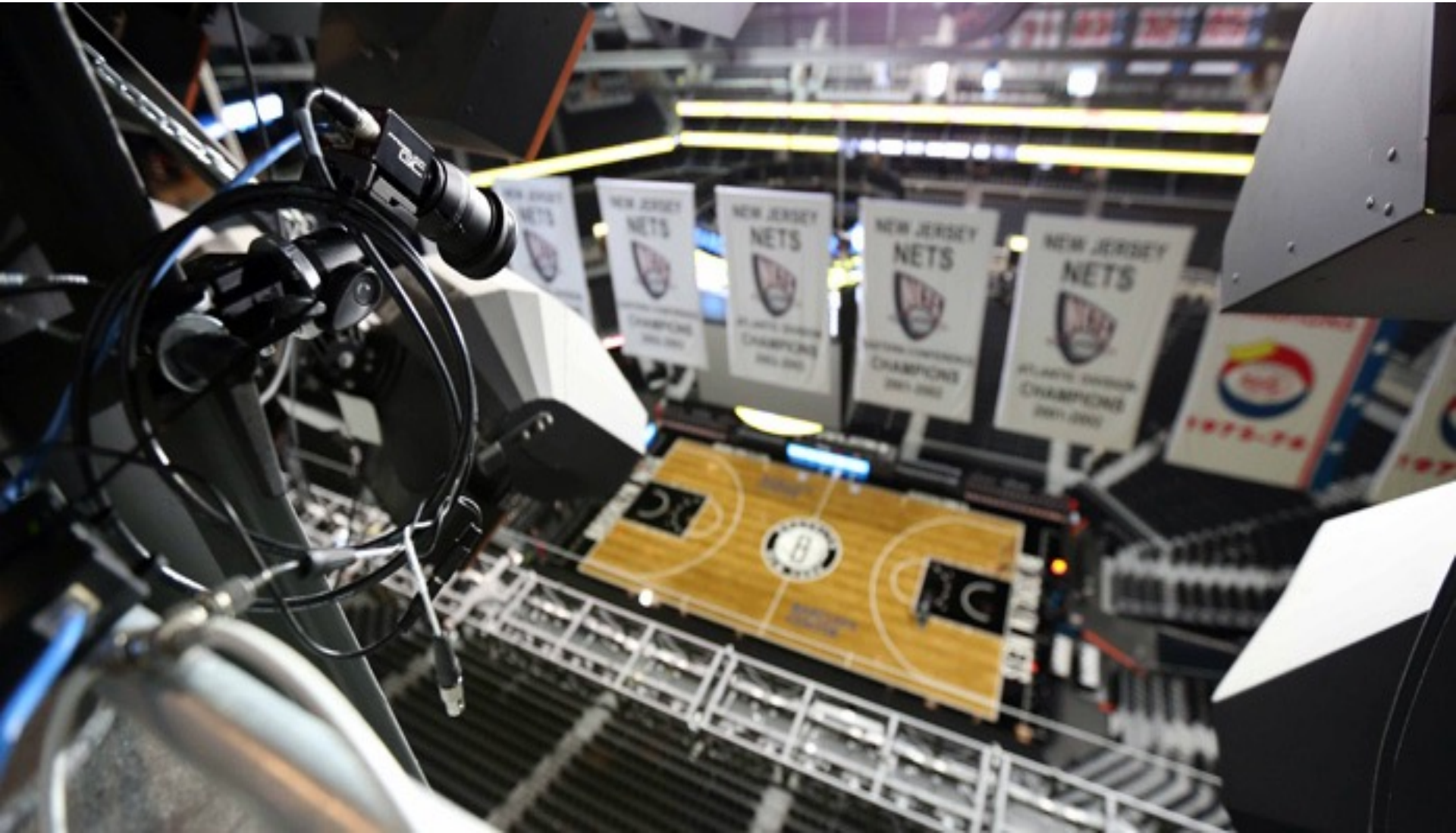
# ASSIST RATE BY LOCATION




GRANTLAND

# Optical tracking data

## A high-resolution, digital representation of the game




- With former classmates, formed a sports research group, (X,Y) Research in 2014
- New methodology at the intersection of statistics, machine learning, and sports analytics
- Projects for Arsenal (soccer), Dodgers (baseball) and San Antonio Spurs (basketball)
- Currently working with Philadelphia 76ers

# Analytics-driven Sixers ride the numbers to NBA playoffs

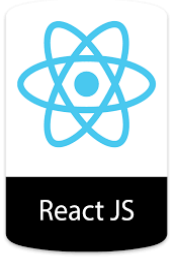
**Updated:** APRIL 18, 2018 — 11:25 AM EDT

# Data Science Tools Used by


- XML and json files produced daily after games


- Update PostgresSQL databases


- Front end: Node and React JS


- Statistical analysis: Pandas


- Version control

- Cloud computing on AWS

# Statistics and Machine Learning

- Don’t care about fancy methods: want reliable / interpretable / trustworthy results.

- When you develop a complex model, how do you check that it’s right? How would you use it to make decisions?
  - E.g. cross validation, uncertainty quantification (confidence intervals)
  - What other interpretations are consistent with your results?

# Soft Skills Matter

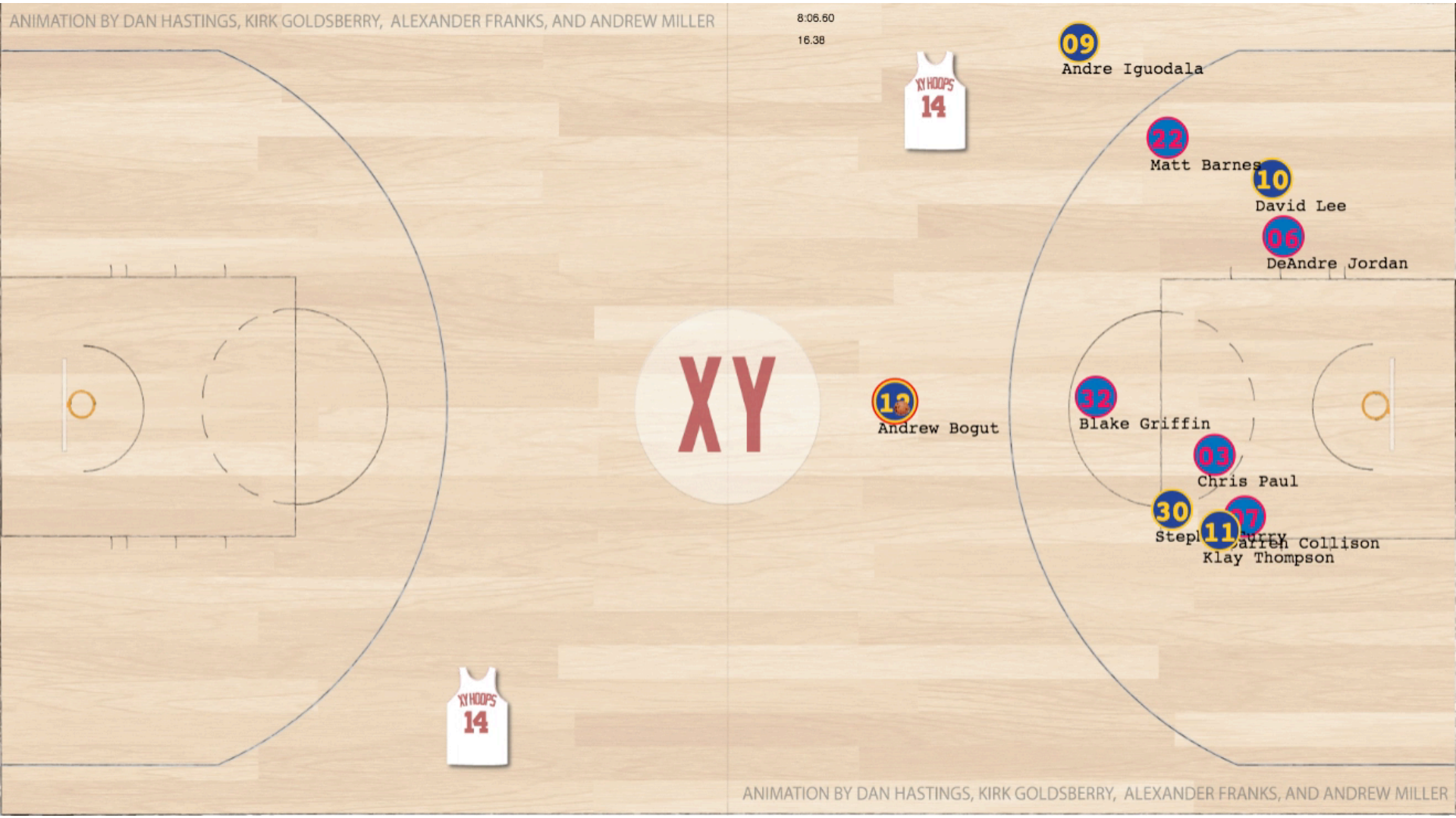
- VP of analytics and strategy: “Maybe 20 percent of my job is analytics and 80 percent is being able to collaborate with people and communicate with coaches”

- Contextualize your statistical findings! It is just as important to emphasize what the data *doesn’t* tell you as what it does.

- Can’t be an expert in everything but need to speak others’ language and be comfortable working on a team.

# Optical tracking data

## A high-resolution, digital representation of the game



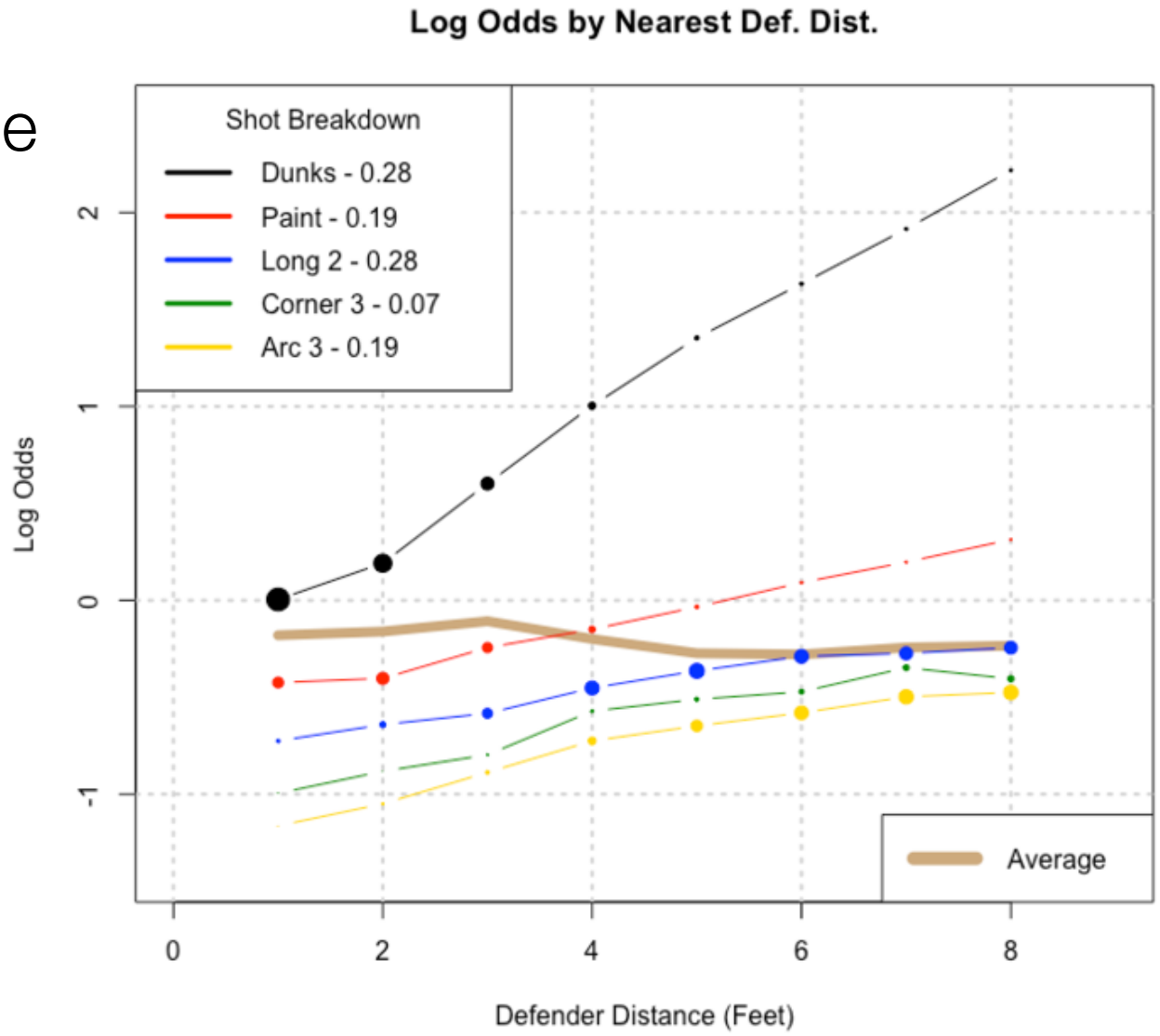
# Big data = a lot of small data

*Enormous data sets often consist of enormous numbers of small sets of data, none of which by themselves are enough to solve the thing you are interested in, and* ***they fit together*** *in some complicated way.*

— *Bradley Efron, Significance Magazine*

# Spatial Context Matters

"Does the distance of the challenging defender matter when a shot is being taken?"



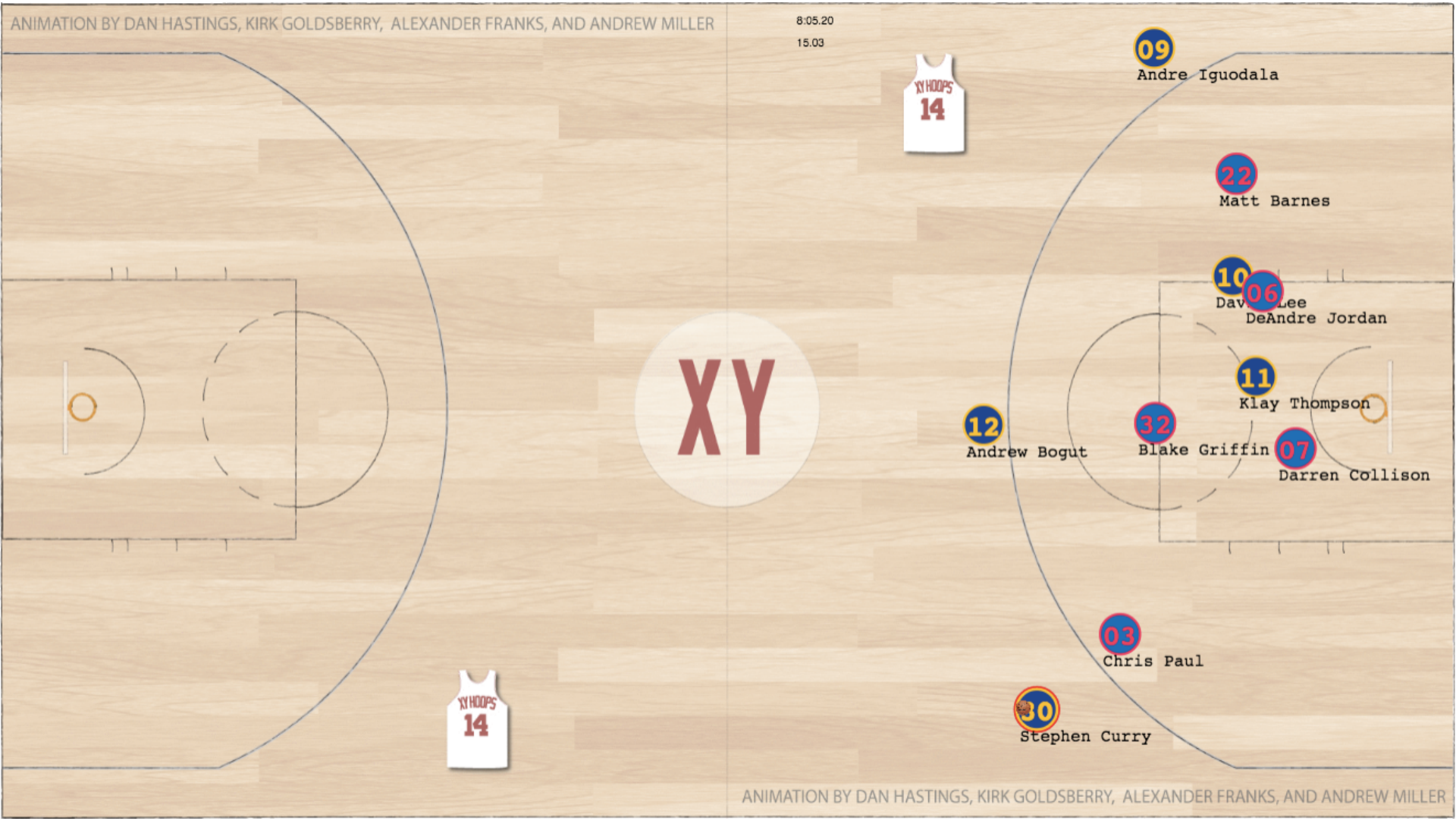
# Regression Models for Defensive Analysis

- Discretize court into regions
- Model makes/misses and shot attempts using regression
  - Offensive player
  - Defensive player, time guarding, distance
  - Court region
- Coefficients for defenders tell us about defender skill

# Shot Suppression

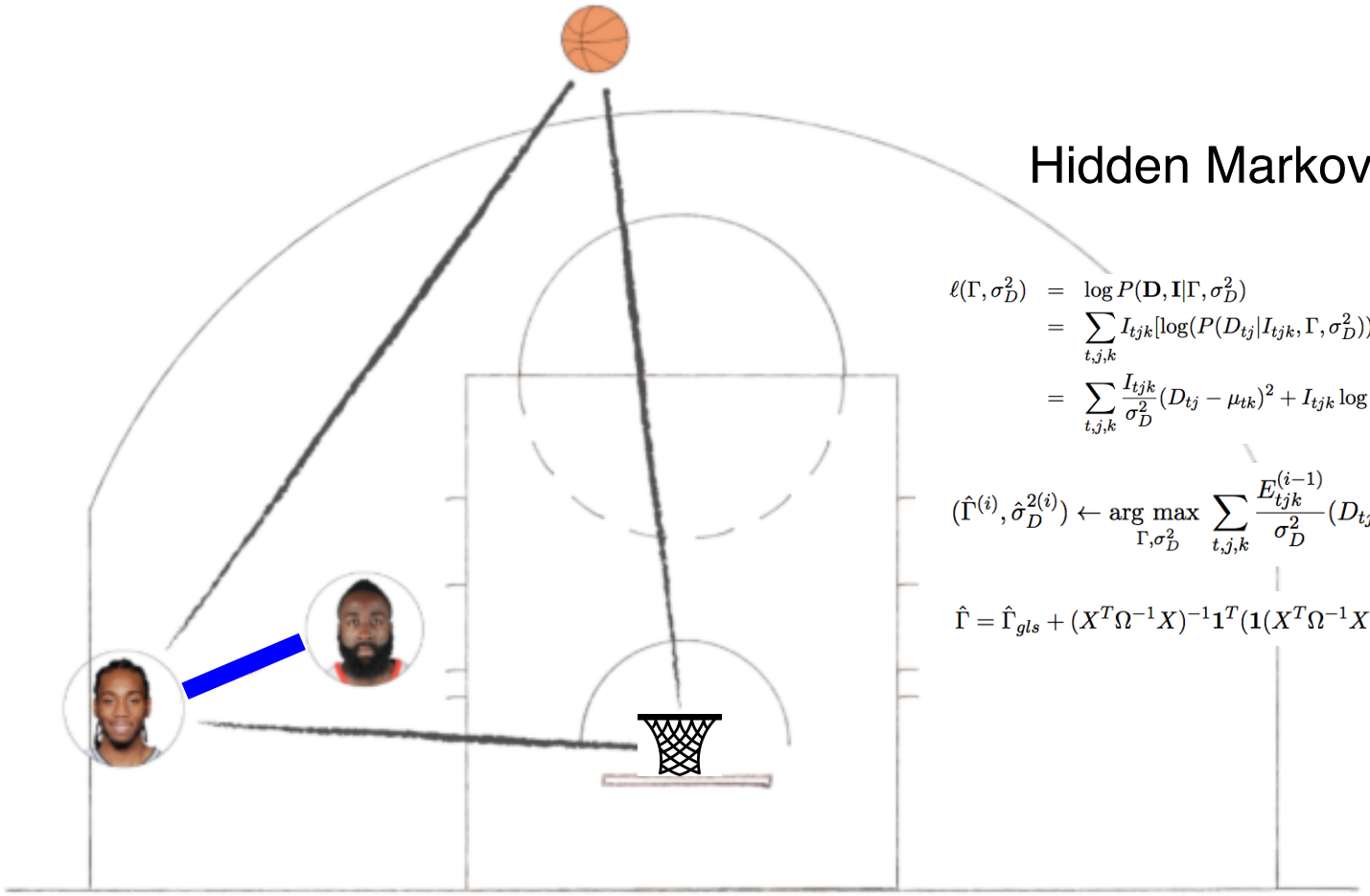
# Shot Disruption

# Deriving defense from data




# Matchups

# Matchups: who's guarding whom?



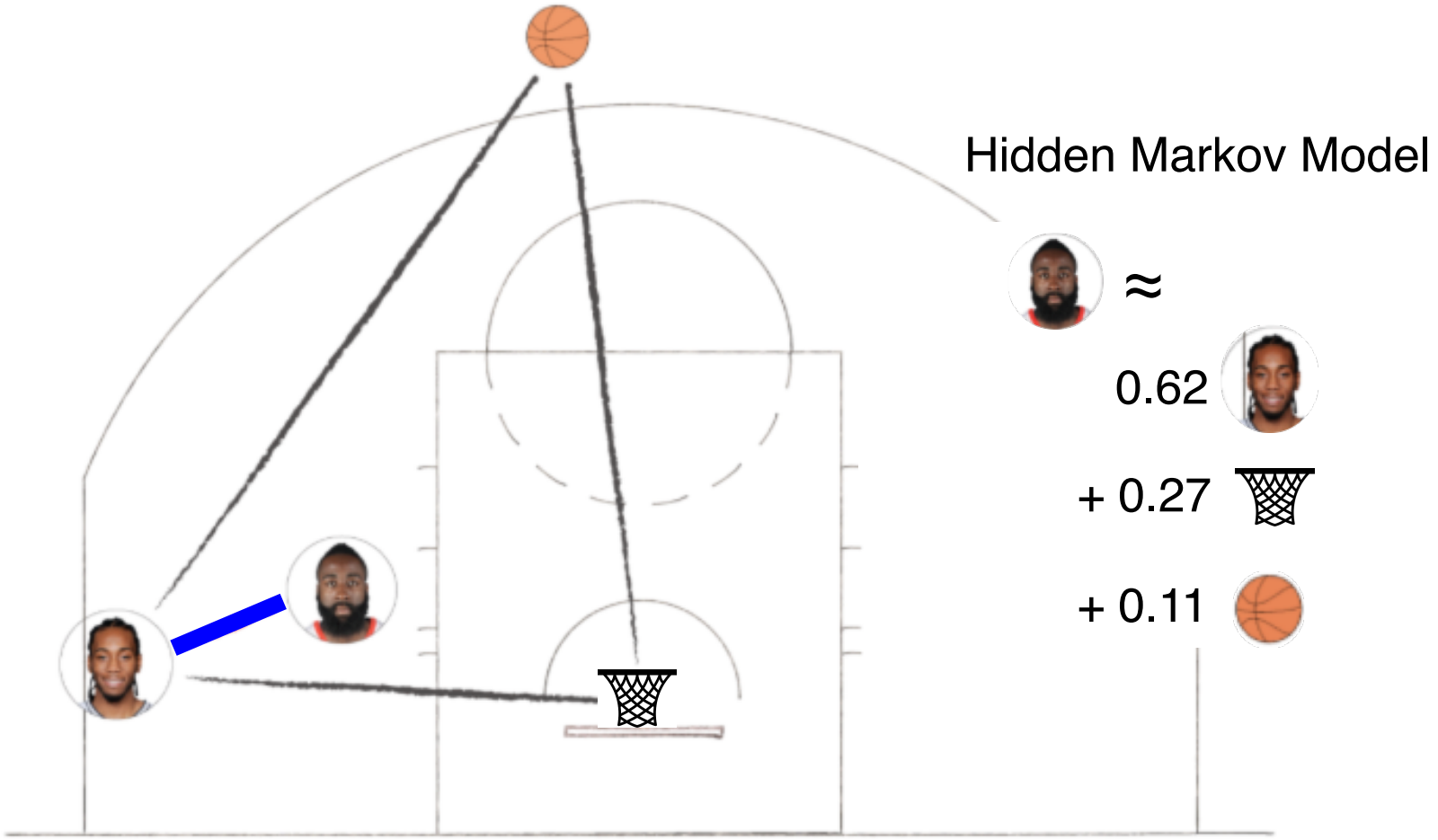
## Hidden Markov Model

$$\begin{aligned}\ell(\Gamma, \sigma_D^2) &= \log P(\mathbf{D}, \mathbf{I}|\Gamma, \sigma_D^2) \\ &= \sum_{t,j,k} I_{tjk}[\log(P(D_{tj}|I_{tjk}, \Gamma, \sigma_D^2)) + \log(P(I_{tjk}|I_{(t-1)j.}))] \\ &= \sum_{t,j,k} \frac{I_{tjk}}{\sigma_D^2}(D_{tj} - \mu_{tk})^2 + I_{tjk} \log P(I_{tjk}|I_{(t-1)j.})\end{aligned}$$

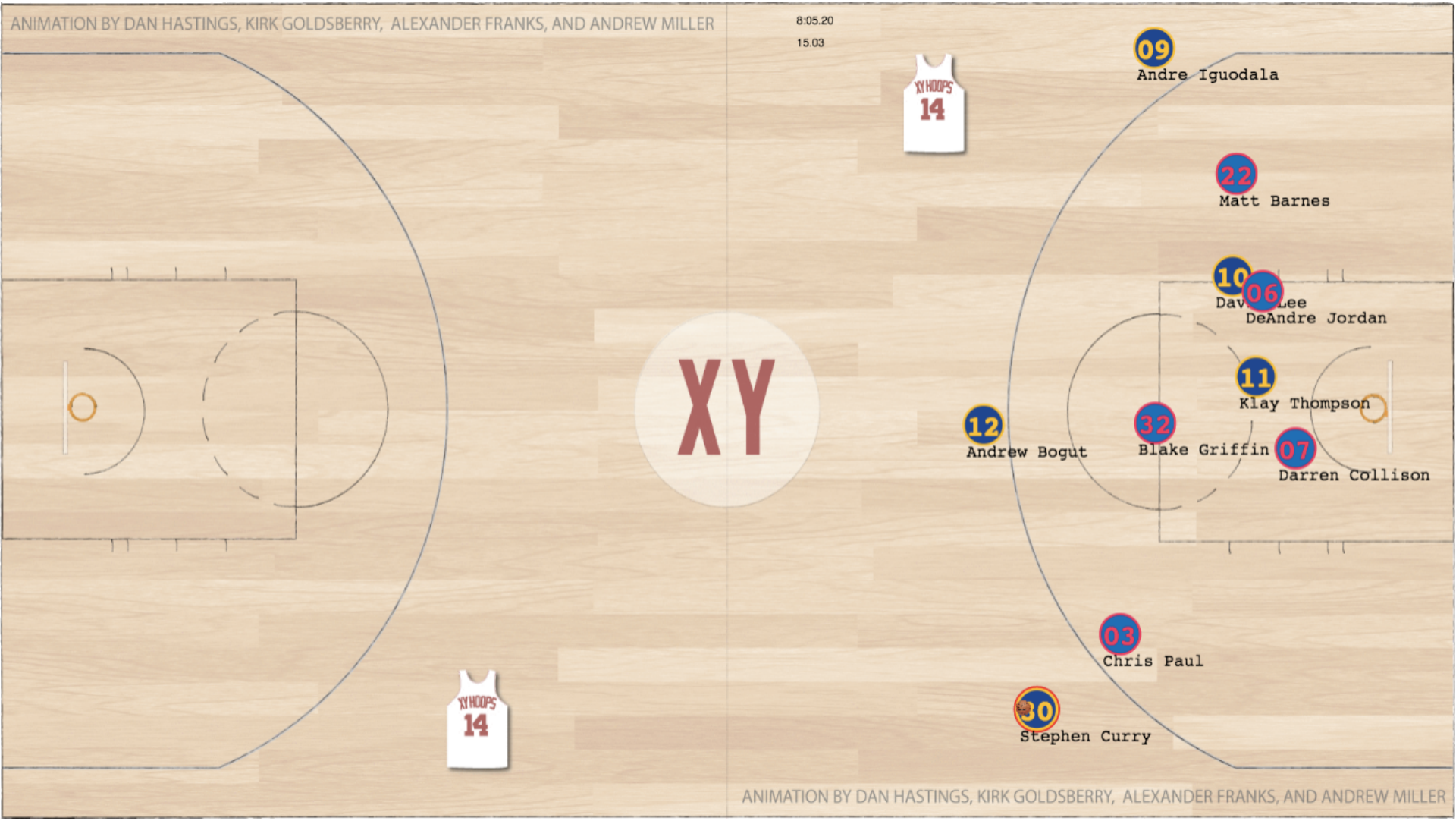
$$(\hat{\Gamma}^{(i)}, \hat{\sigma}_D^{2(i)}) \leftarrow \arg\max_{\Gamma, \sigma_D^2} \sum_{t,j,k} \frac{E_{tjk}^{(i-1)}}{\sigma_D^2}(D_{tj} - \Gamma X_{tk})^2, \ \Gamma \mathbf{1} = 1$$

$$\hat{\Gamma} = \hat{\Gamma}_{gls} + (X^T \Omega^{-1} X)^{-1} \mathbf{1}^T (\mathbf{1}(X^T \Omega^{-1} X)^{-1} \mathbf{1}^T)^{-1} (1 - \hat{\Gamma}_{gls} \mathbf{1}),$$

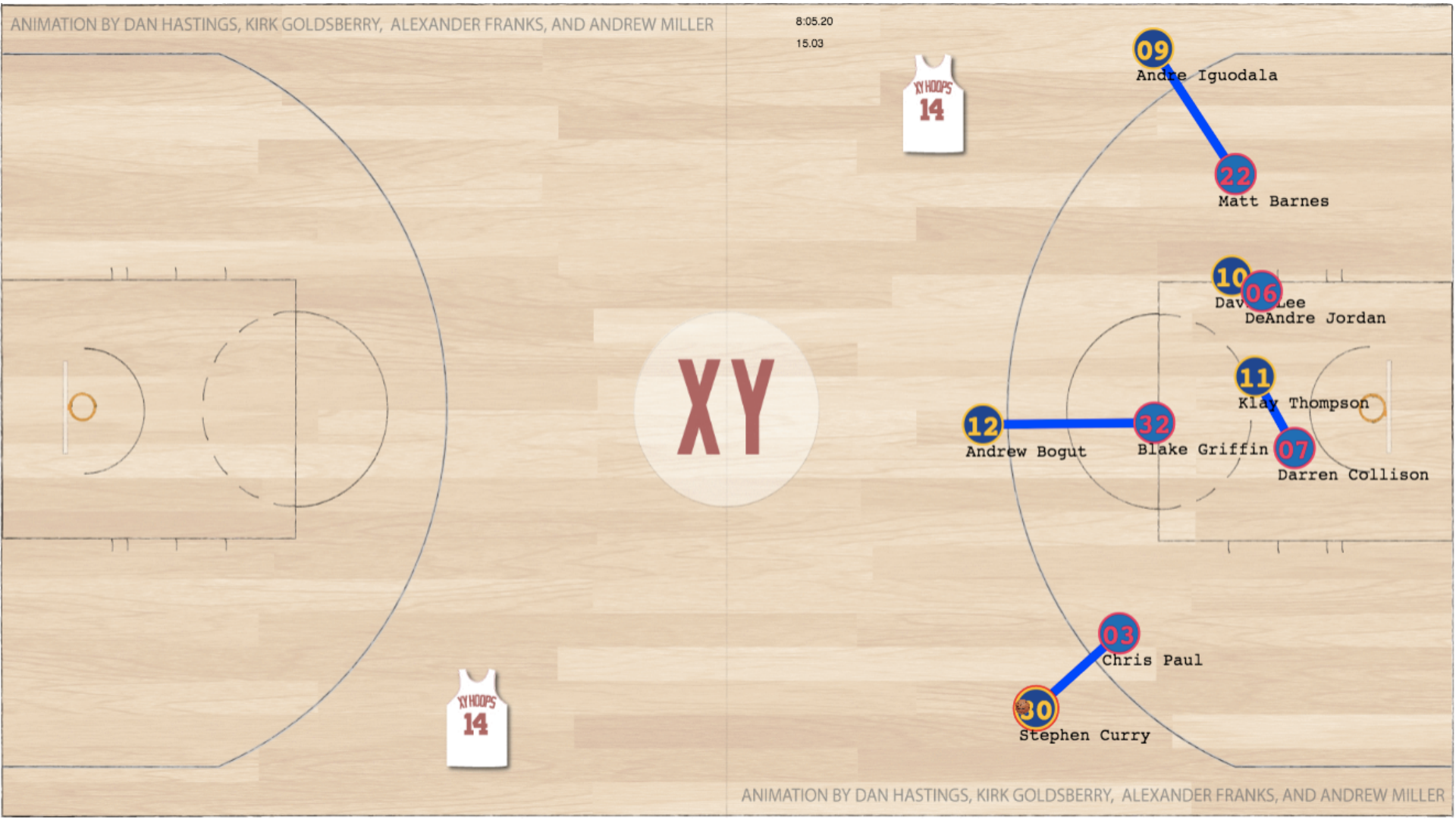
# Matchups: who's guarding whom?

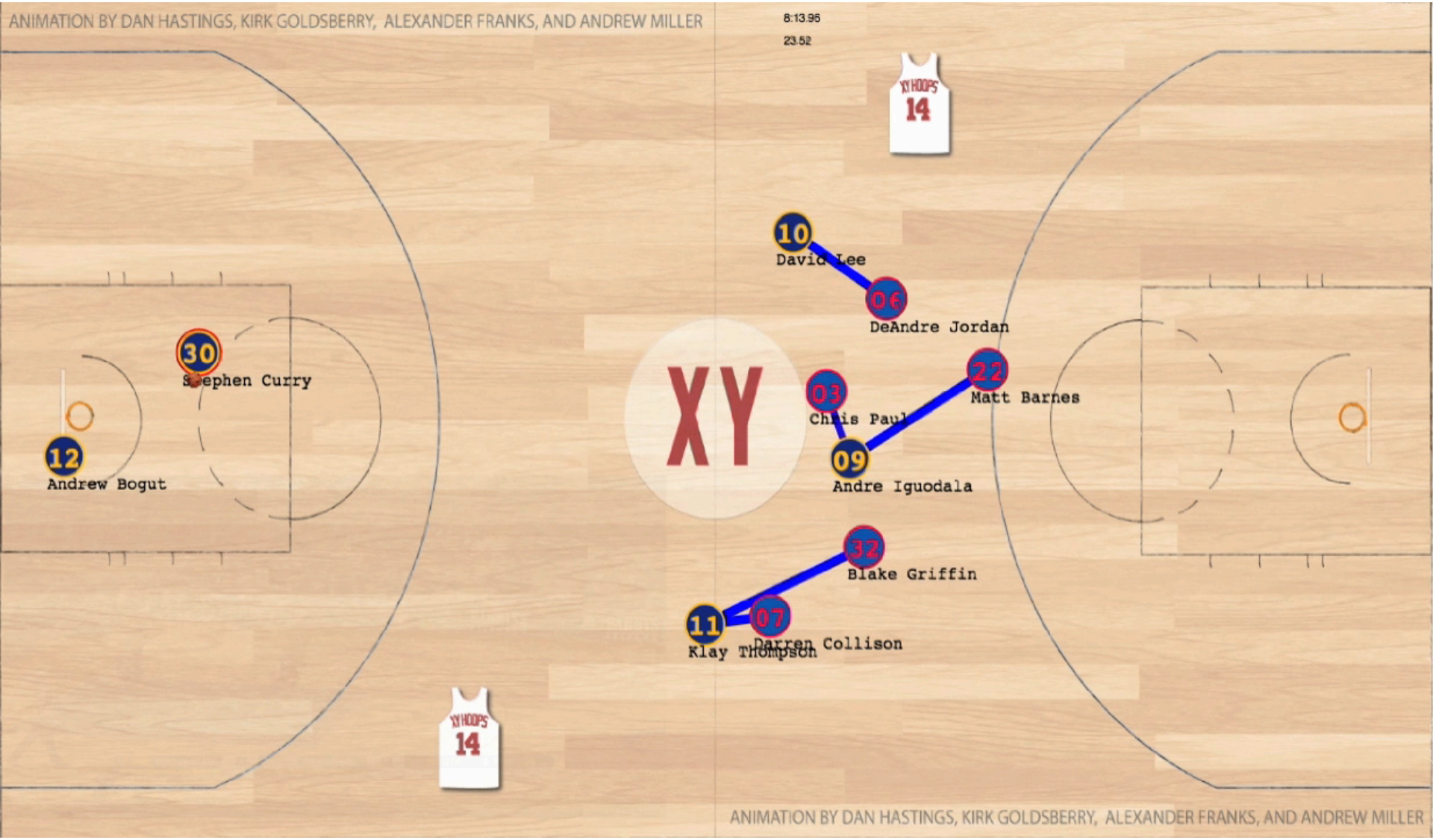


# Matchups: who's guarding whom?



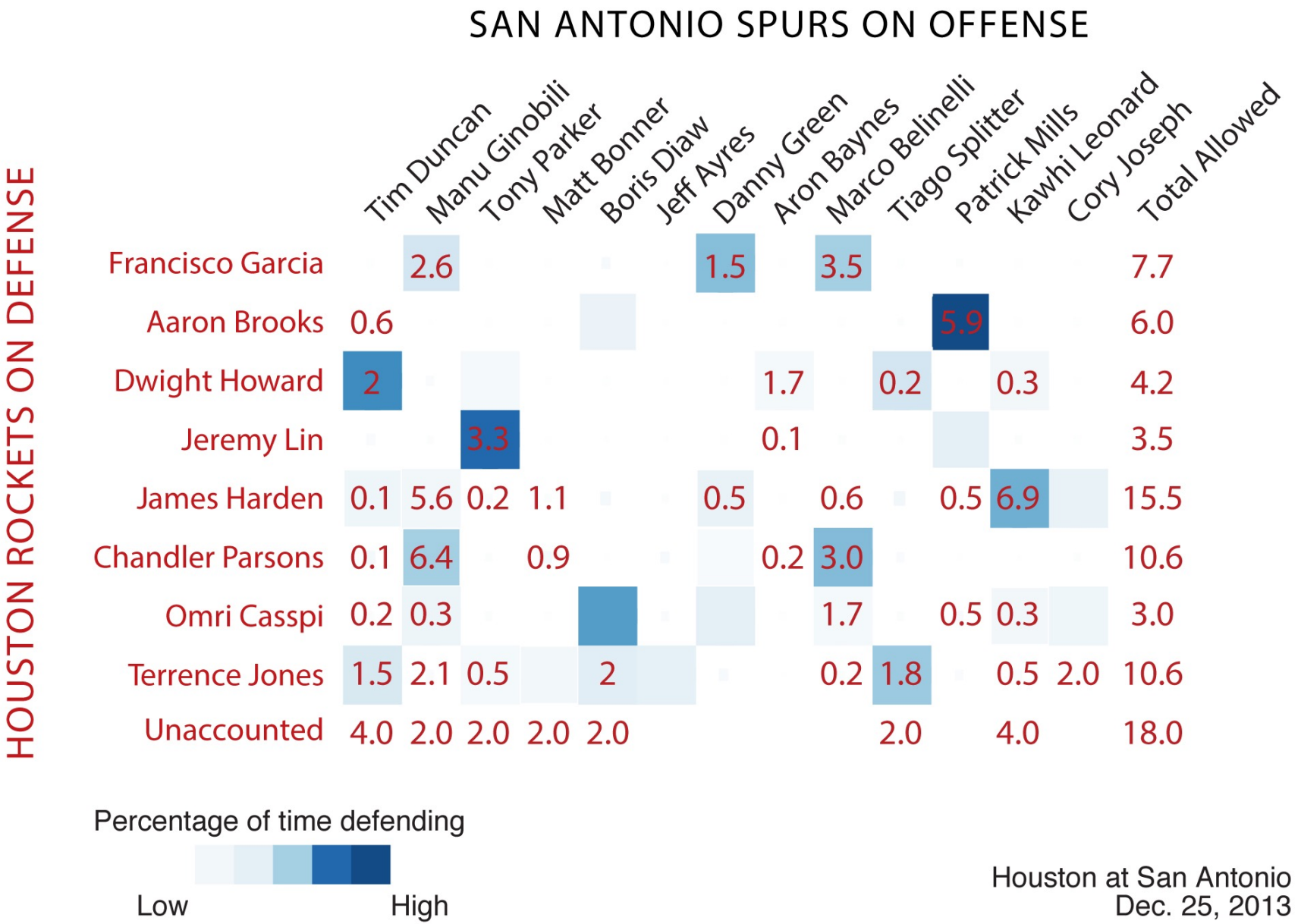
# Matchups: who's guarding whom?




ANIMATION BY DAN HASTINGS, KIRK GOLDSBERRY, ALEXANDER FRANKS, AND ANDREW MILLER
8:13.96
23.52
XY HOOPS 14
10
David Lee
06
DeAndre Jordan
30
Stephen Curry
12
Andrew Bogut
XY
03
Chris Paul
22
Matt Barnes
09
Andre Iguodala
32
Blake Griffin
11
Klay Thompson
07
Darren Collison
XY HOOPS 14
ANIMATION BY DAN HASTINGS, KIRK GOLDSBERRY, ALEXANDER FRANKS, AND ANDREW MILLER

# WHO IS GUARDING WHOM?

## AND HOW MANY POINTS DID THEY GIVE UP?




BELINELI
3
DALLAS
9


GOLDEN STATE
30
WARRIORS

# Modeling defensive skill

Defenders' impact on shot selection and disruption

$$\Pr(\text{shot attempted}|\text{data}) \propto \exp(\alpha_o + \sum_{d=1}^{5} F(o,d)\beta_d)$$

- Overall "usage"
- Defending time
- Suppression coefficient

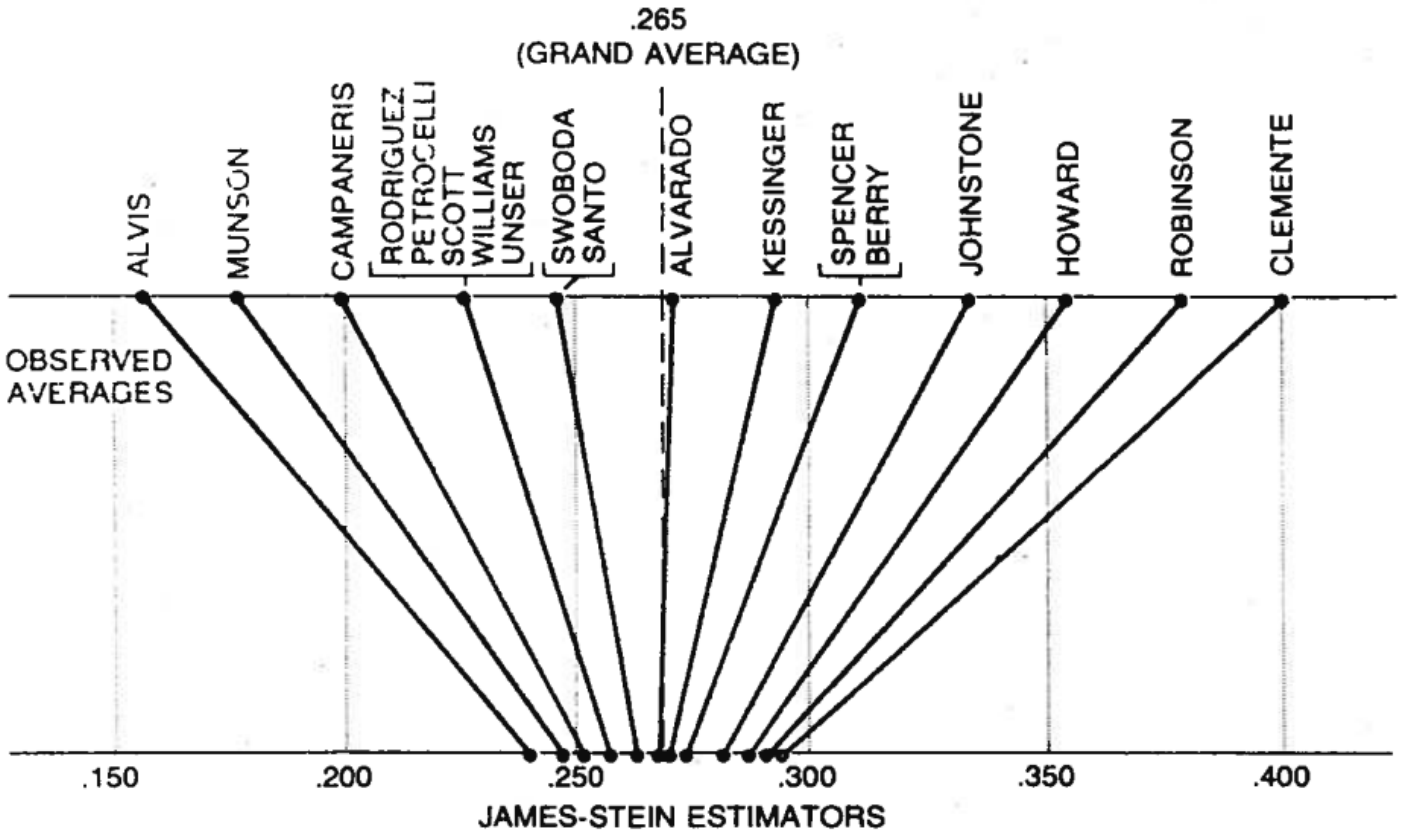
Shot selection

$$\Pr(\text{shot made}|\text{data}) \propto \exp(\theta_o + \phi_d + \xi\mathcal{D})$$

- Overall efficiency
- Disruption coefficient
- Defender distance effect

Shot disruption

# Sharing Information Across Players

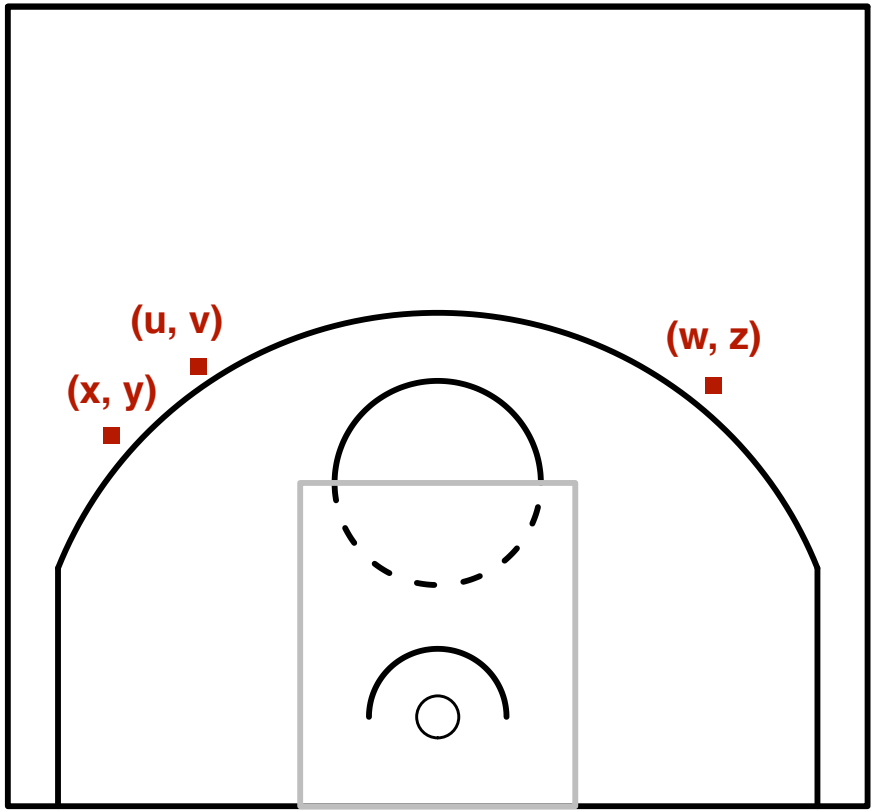



**JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.**

# Sharing Information Across Space

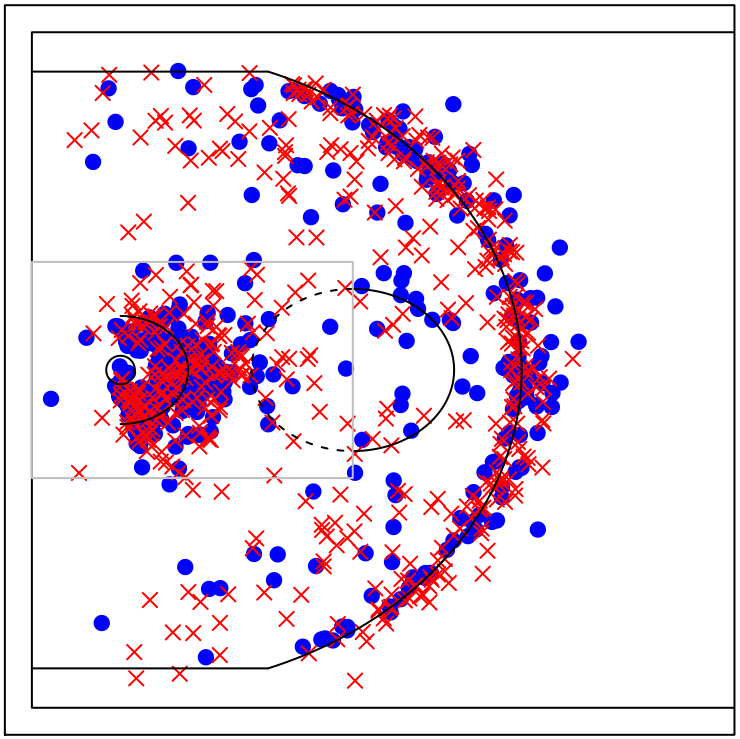
Is propensity to shoot at (x,y) correlated with the propensity to shoot at (u,v)?

What about (w,z)?

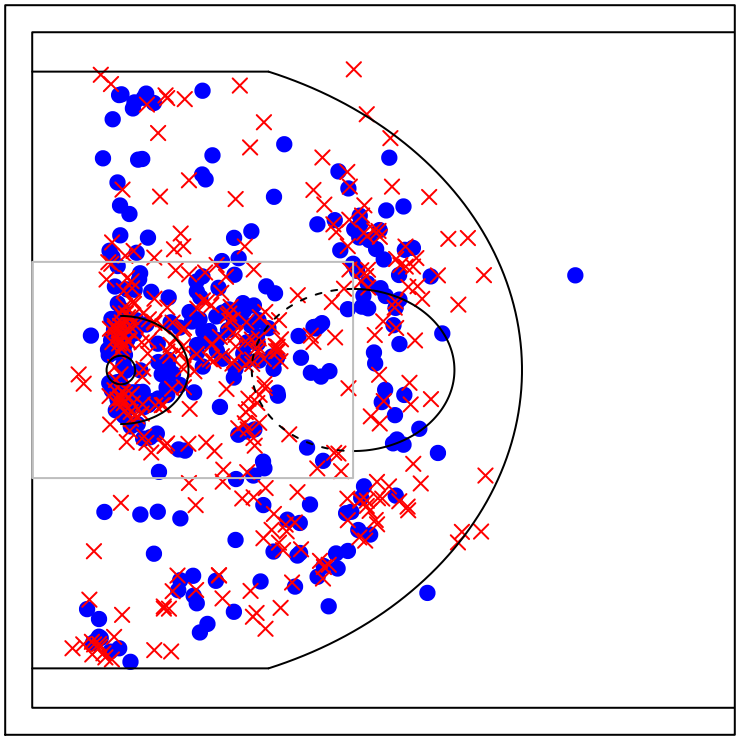


# Sharing Information Across Space

**James Harden (965 shots)**

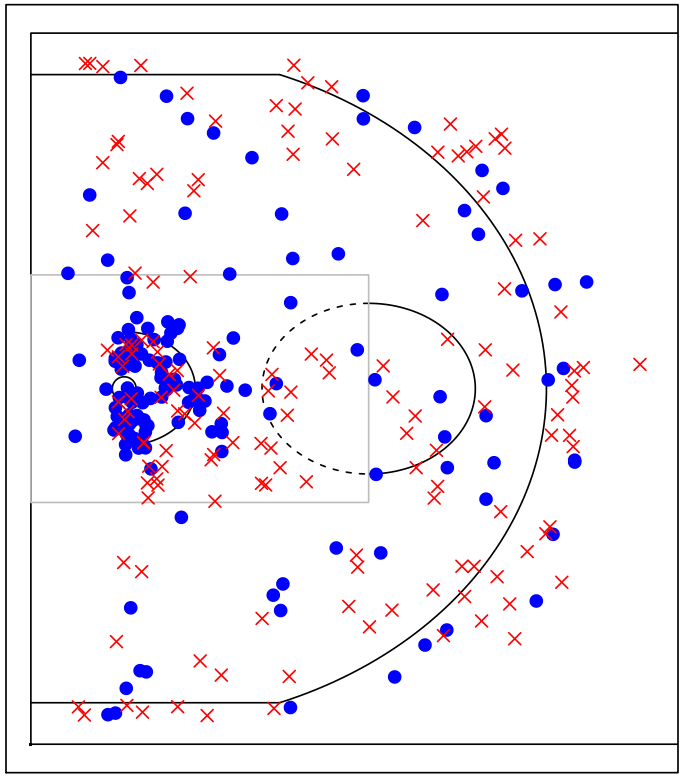


**Tony Parker (692 shots)**

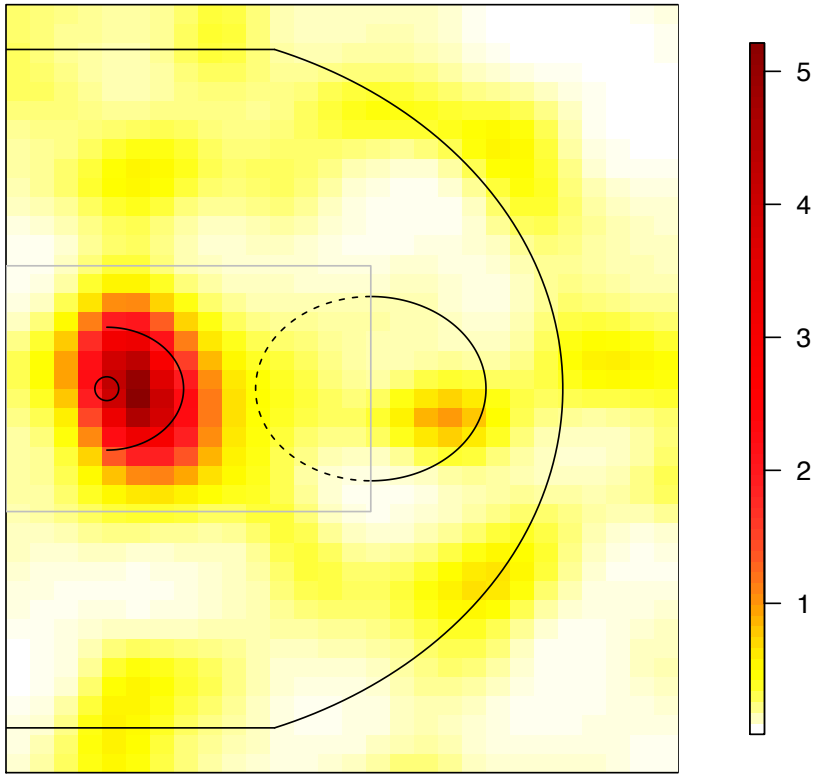


# A log-Gaussian Cox Process

**LeBron James shot chart**



**LeBron James intensity surface (posterior mean)**



# Partitioning the Court

- Each player LGCP is a V-dimensional vector (~ 1000 sq tiles)
- Represent all shot surfaces as a matrix:

$$\Lambda = [\bar{\lambda}_1, \ldots, \bar{\lambda}_N]^T$$

where all elements of $\bar{\lambda}_i > 0$, and $\sum_v \bar{\lambda}_i(v) = 1$

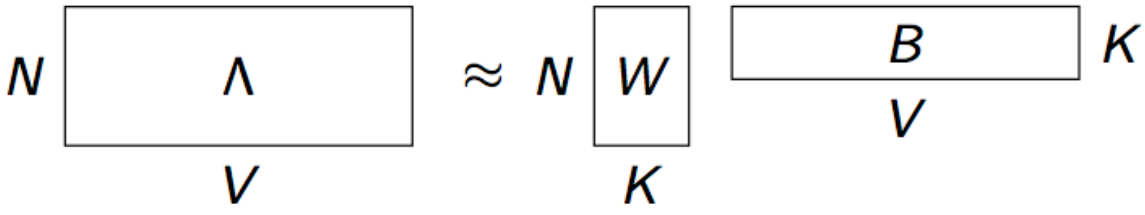
- **Goal**: find a reduce basis set to describe shot types

# Nonnegative Matrix Factorization

**Story**: Each of the $N$ players is some combination of $K$ different 'shot types' (basis surfaces):

$$\Lambda \approx WB \tag{1}$$

low rank approximation:




where

$$W \in \mathbb{R}_+^{N \times K} \text{ and } B \in \mathbb{R}_+^{K \times V}$$

**Optimize**:

$$\hat{W}, \hat{B} = \underset{W,B>0}{\arg\min}\, \ell(\Lambda, WB) \tag{2}$$

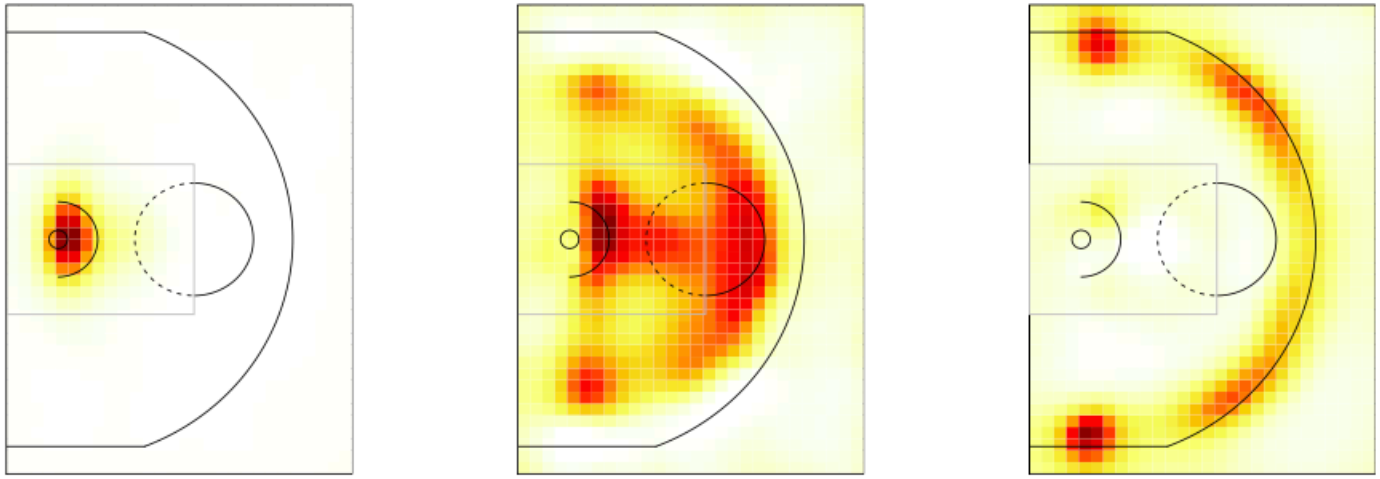
# Nonnegative Matrix Factorization

**Interpretation**: player $n$'s shooting

$$\hat{\lambda}_n = \sum_{k=1}^{K} W_{nk} B_k$$

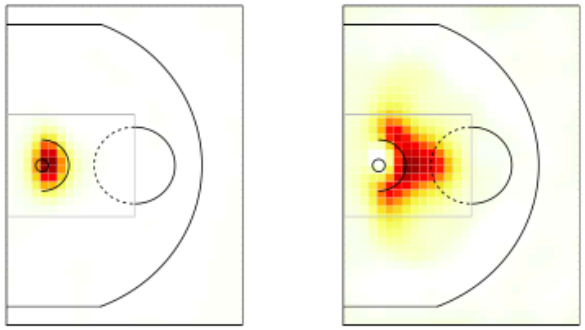
- $B_k$ is the $k$'th basis surface or 'shot type'
- $W_{nk}$ is player $n$'s loading onto $B_k$.

# NMF: 3 basis vectors

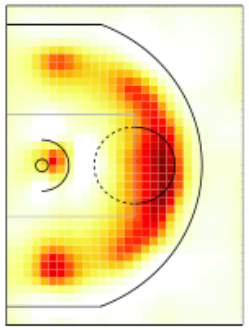


(a) close-range (b) mid-range (c) long-range

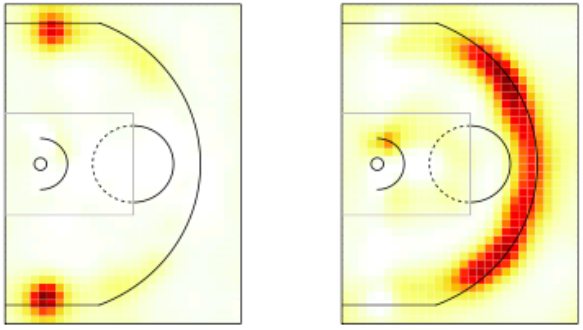
# NMF: 5 basis vectors



(a) close-range



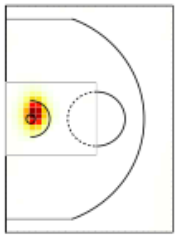
(b) mid-range

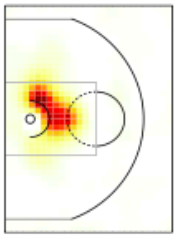


(c) long-range

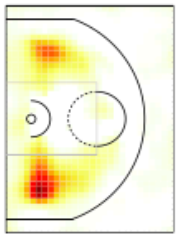
# NMF: 10 basis vectors

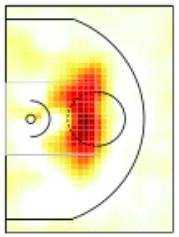
close-range

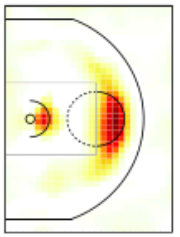


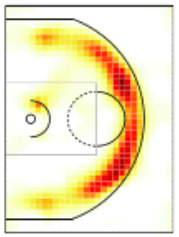


mid-range

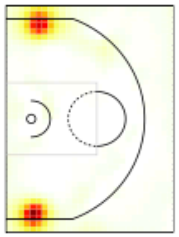


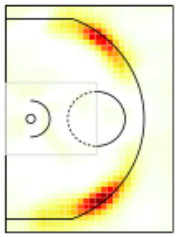


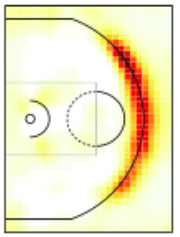




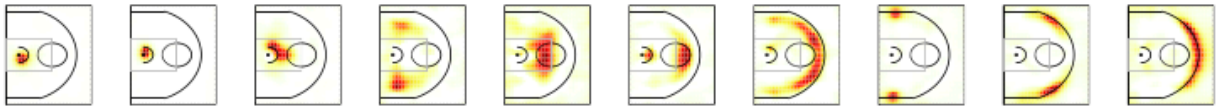
long-range





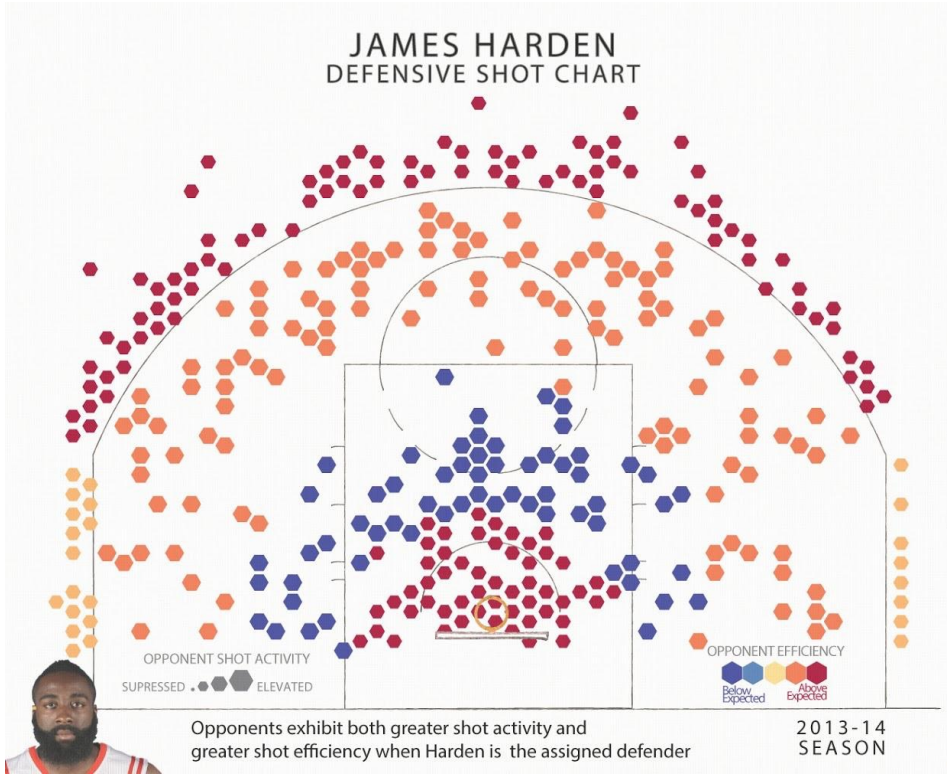


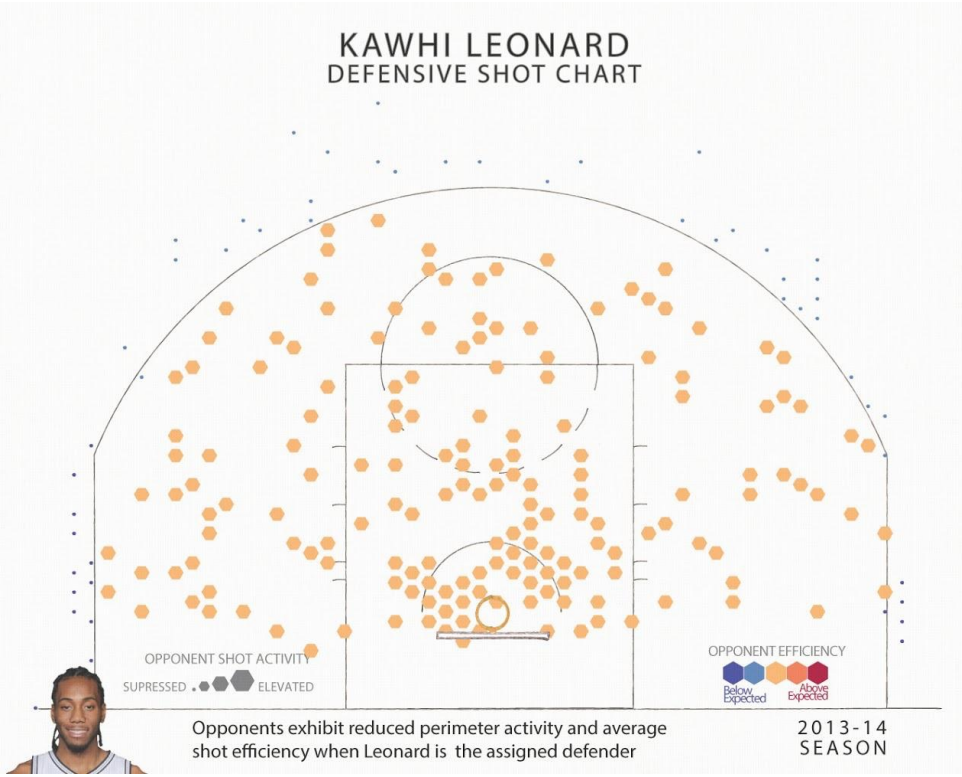
# NMF player weights



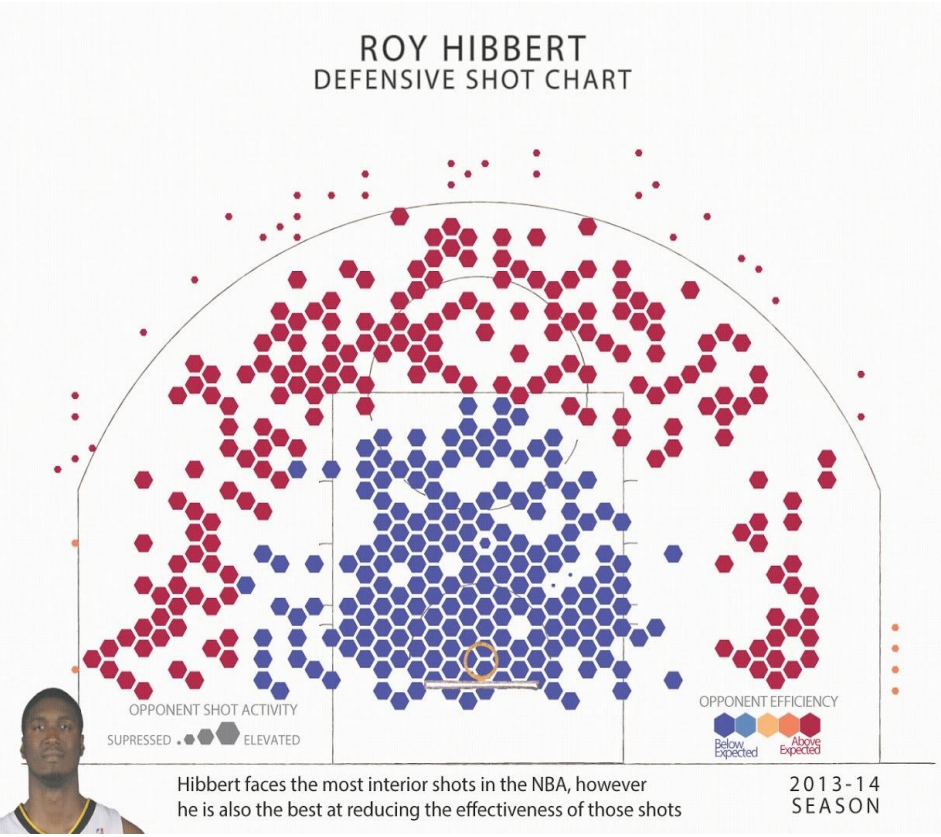
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| LeBron James | 0.21 | 0.16 | 0.12 | 0.09 | 0.04 | 0.07 | 0.00 | 0.07 | 0.08 | 0.17 |
| Brook Lopez | 0.06 | 0.27 | 0.43 | 0.09 | 0.01 | 0.03 | 0.08 | 0.03 | 0.00 | 0.01 |
| Tyson Chandler | 0.26 | 0.65 | 0.03 | 0.00 | 0.01 | 0.02 | 0.01 | 0.01 | 0.02 | 0.01 |
| Marc Gasol | 0.19 | 0.02 | 0.17 | 0.01 | 0.33 | 0.25 | 0.00 | 0.01 | 0.00 | 0.03 |
| Tony Parker | 0.12 | 0.22 | 0.17 | 0.07 | 0.21 | 0.07 | 0.08 | 0.06 | 0.00 | 0.00 |
| Kyrie Irving | 0.13 | 0.10 | 0.09 | 0.13 | 0.16 | 0.02 | 0.13 | 0.00 | 0.10 | 0.14 |
| Stephen Curry | 0.08 | 0.03 | 0.07 | 0.01 | 0.10 | 0.08 | 0.22 | 0.05 | 0.10 | 0.24 |
| James Harden | 0.34 | 0.00 | 0.11 | 0.00 | 0.03 | 0.02 | 0.13 | 0.00 | 0.11 | 0.26 |
| Steve Novak | 0.00 | 0.01 | 0.00 | 0.02 | 0.00 | 0.00 | 0.01 | 0.27 | 0.35 | 0.34 |

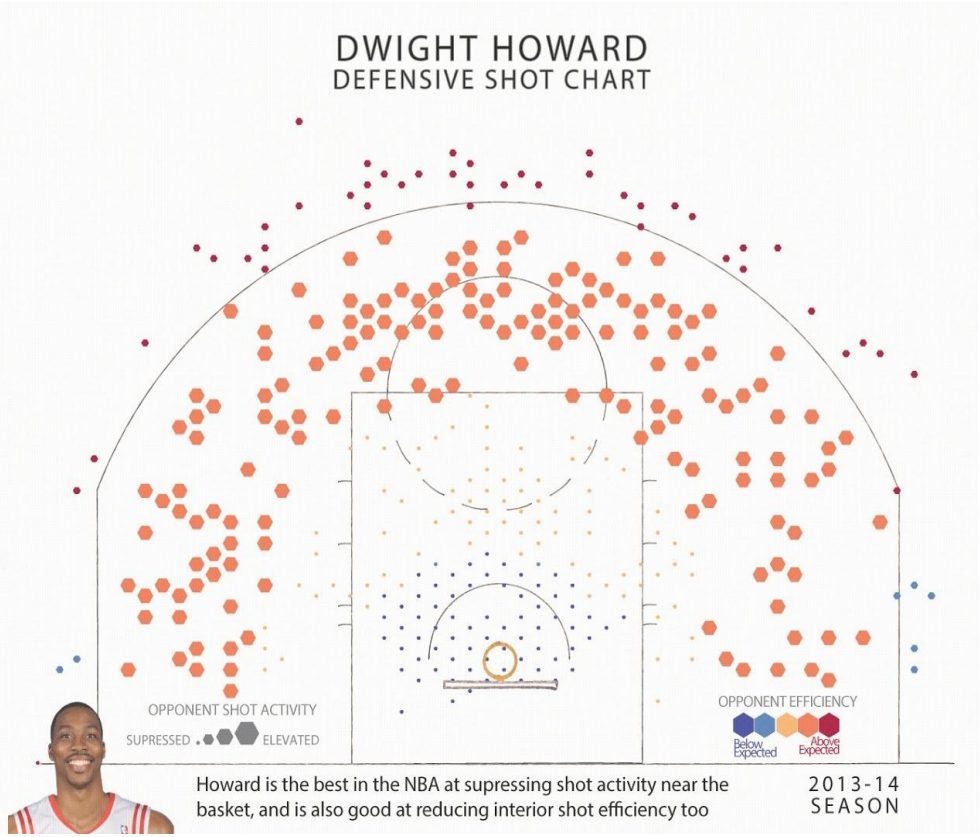
# Combine to methods to create defensive shot charts

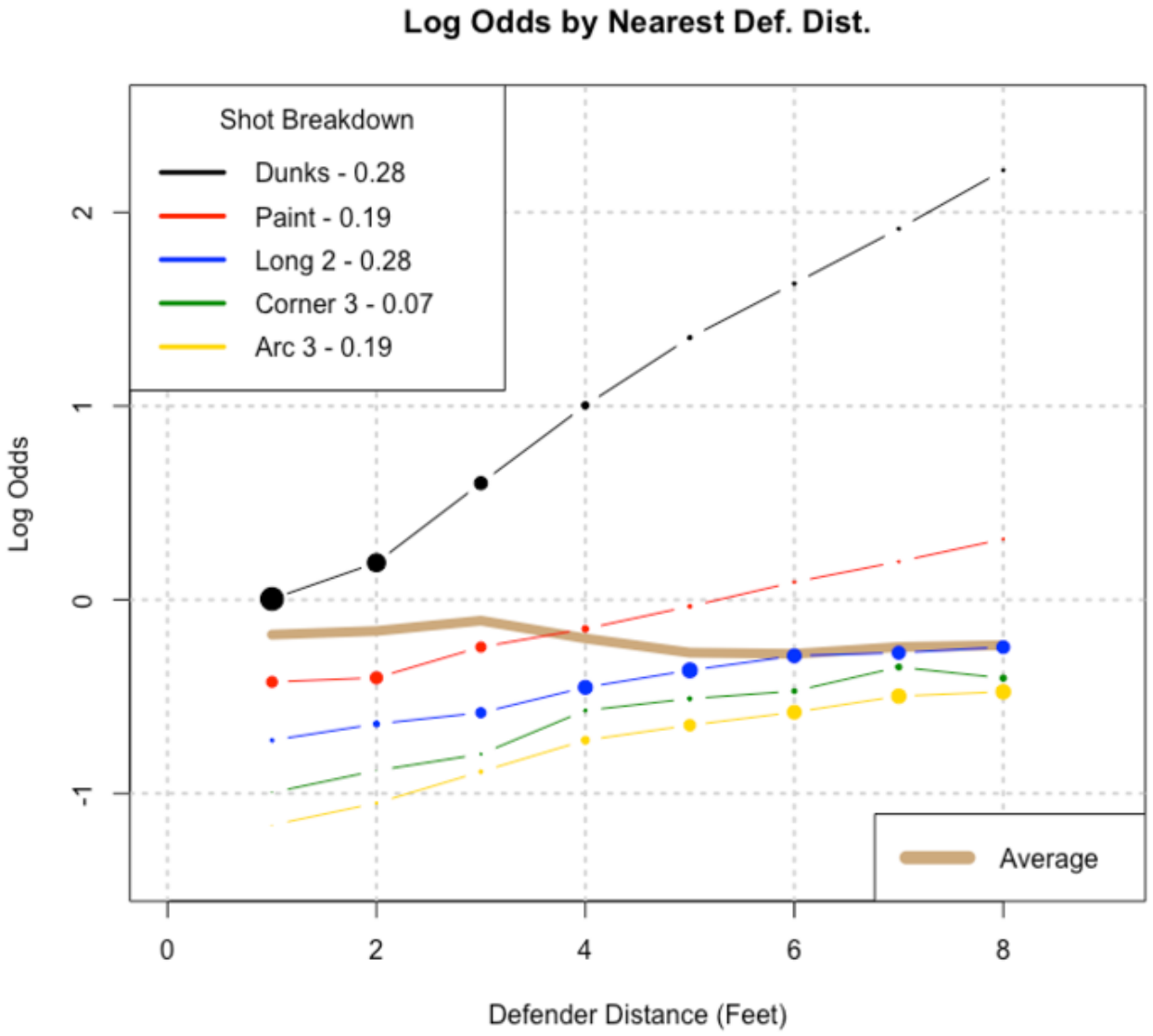


# The Big Men


ROY HIBBERT
DEFENSIVE SHOT CHART
OPPONENT SHOT ACTIVITY
SUPRESSED
ELEVATED
OPPONENT EFFICIENCY
Below Expected
Above Expected
Hibbert faces the most interior shots in the NBA, however he is also the best at reducing the effectiveness of those shots
2013-14 SEASON

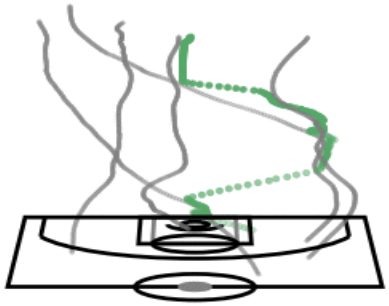

DWIGHT HOWARD
DEFENSIVE SHOT CHART
OPPONENT SHOT ACTIVITY
SUPRESSED
ELEVATED
OPPONENT EFFICIENCY
Below Expected
Above Expected
Howard is the best in the NBA at supressing shot activity near the basket, and is also good at reducing interior shot efficiency too
2013-14 SEASON


Log Odds by Nearest Def. Dist.
Shot Breakdown
Dunks - 0.28
Paint - 0.19
Long 2 - 0.28
Corner 3 - 0.07
Arc 3 - 0.19
Average
Log Odds
2
1
0
-1
0
2
4
6
8
Defender Distance (Feet)

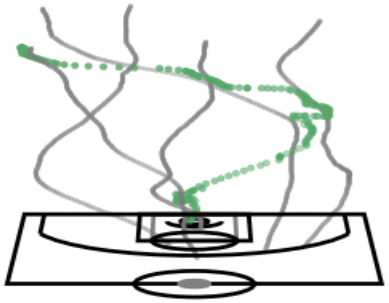
# Stealing the playbook

# Searching the data for play instances

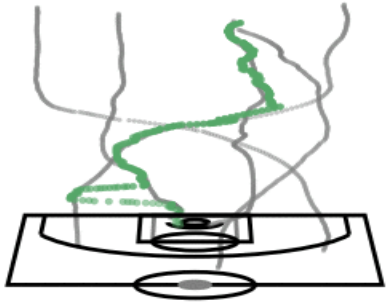
Decoy



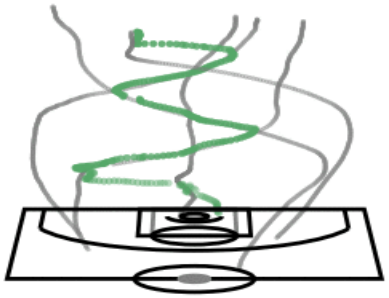
Decoy



Weave

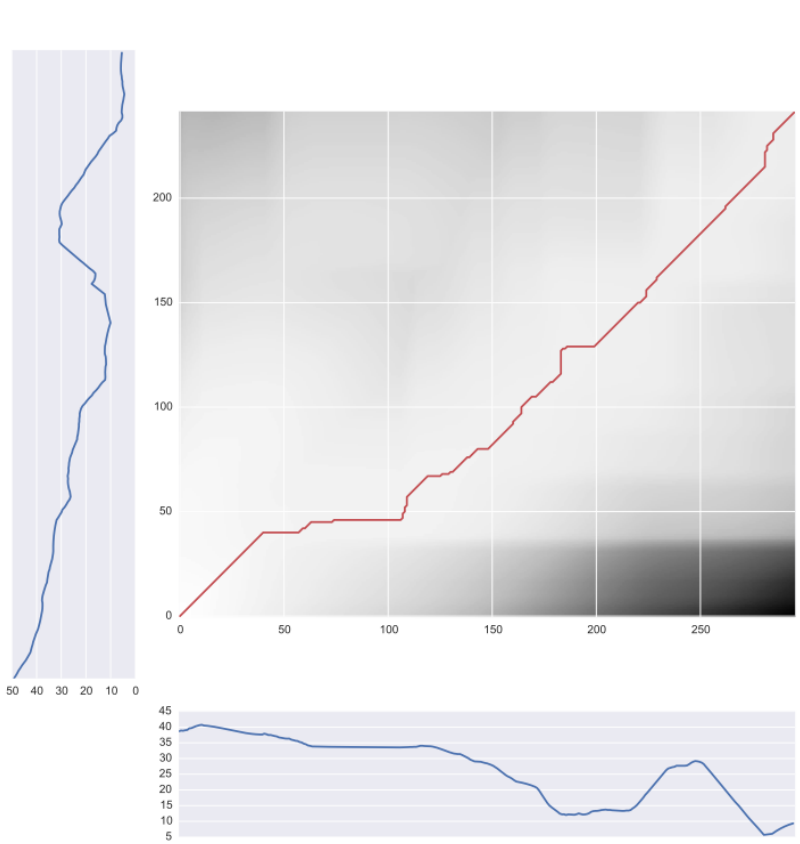


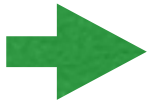
Weave

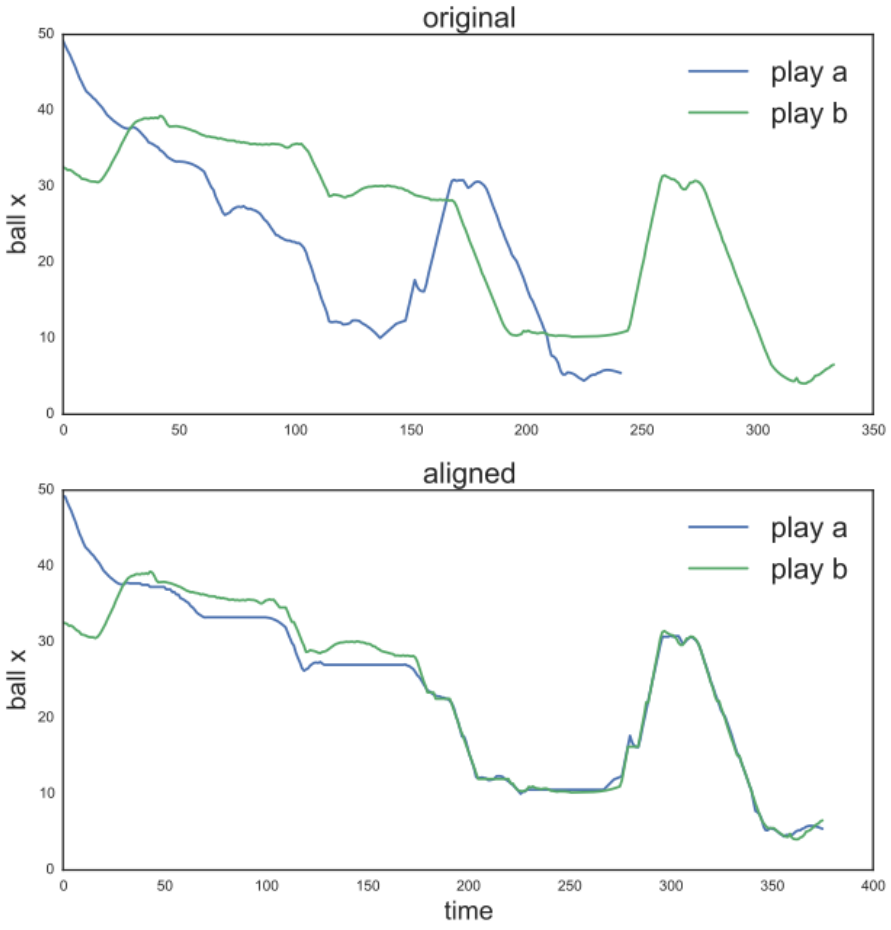


# Dynamic time warping

Matches up the same play occurring at different speeds/locations

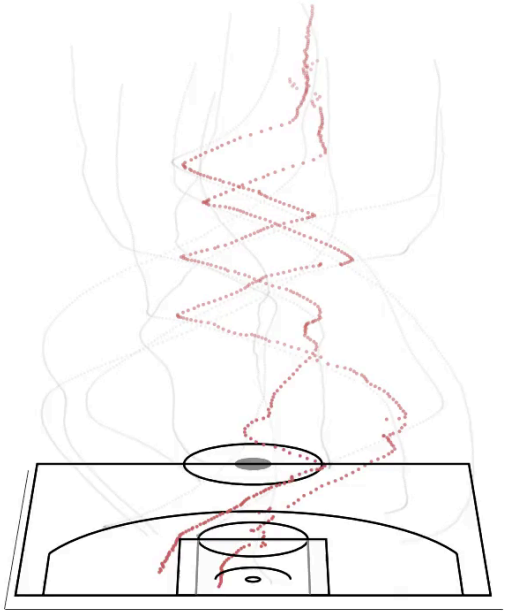


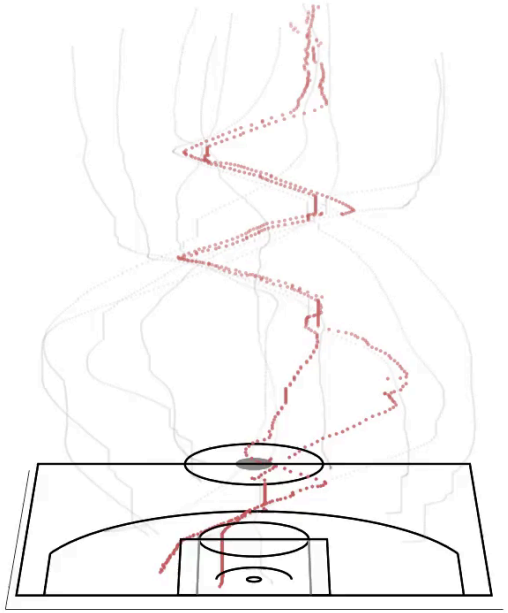


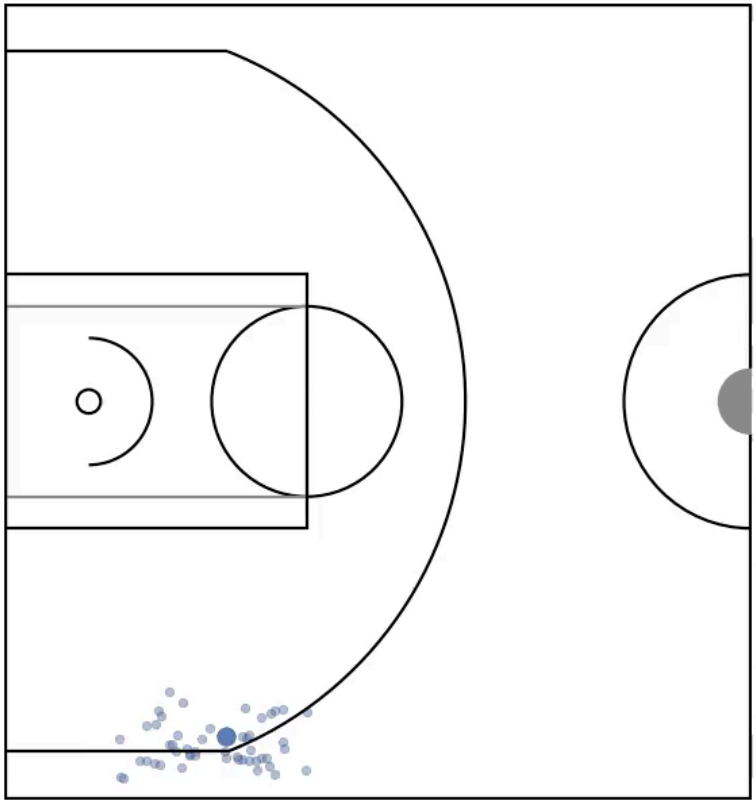
# Weave example

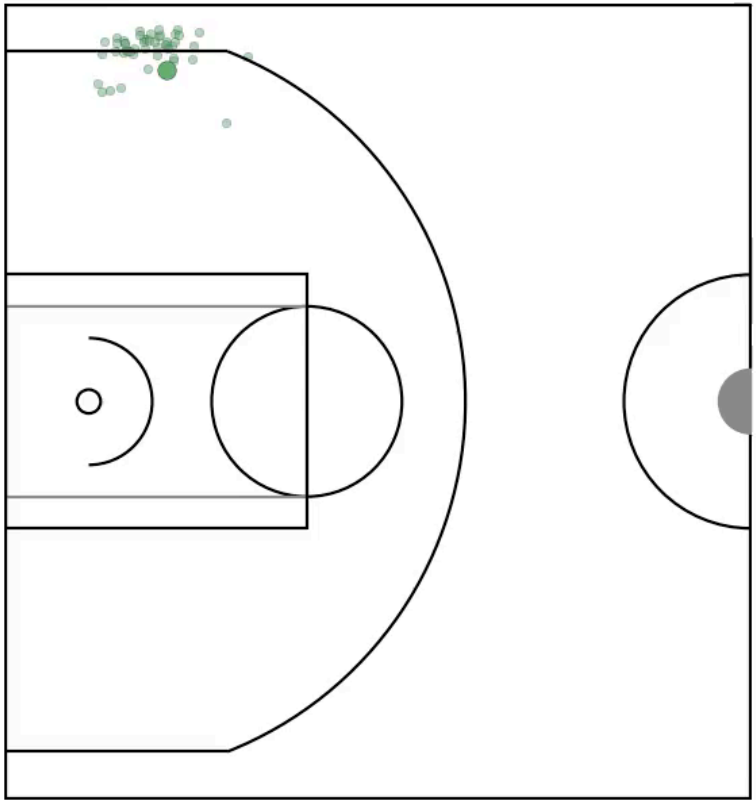
original



aligned

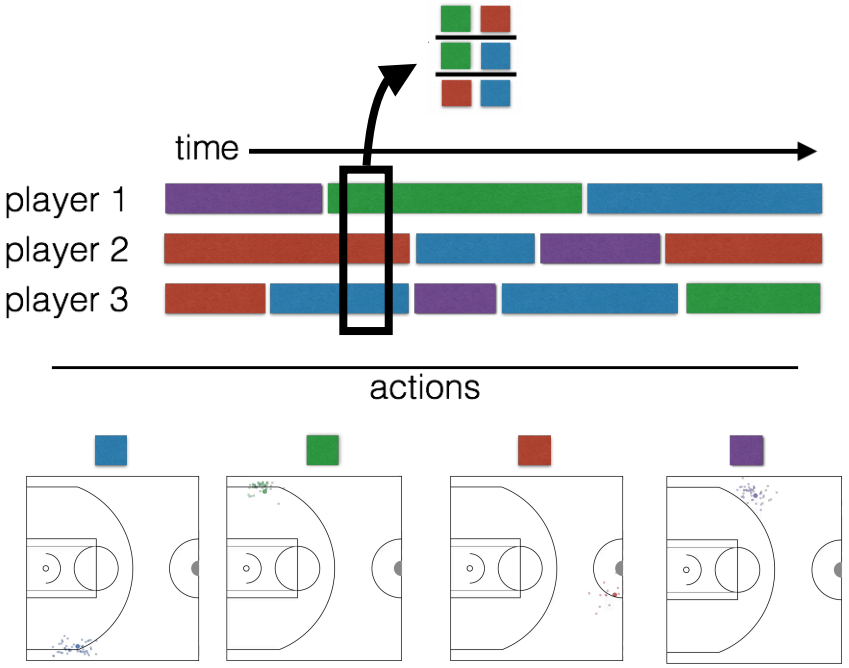




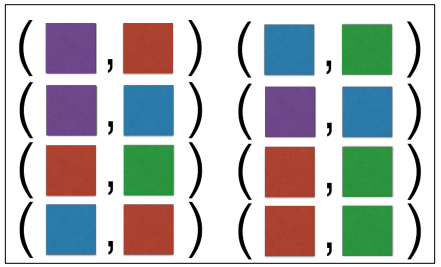


# Building “Words”

“Words” := Pairs of co-occurring actions

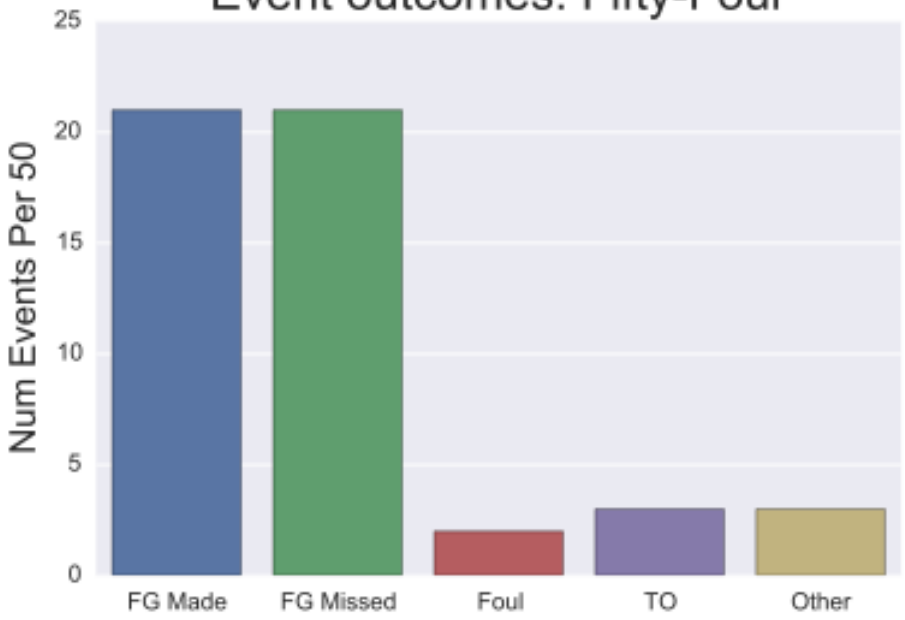


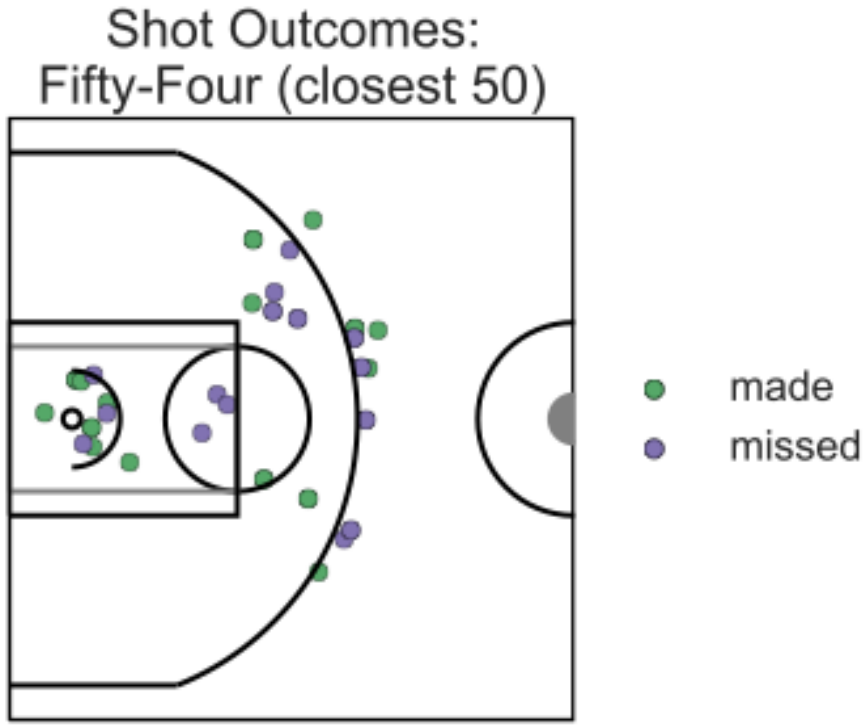

Possession “document”



# Weave (“fifty-four”) example

## Play results



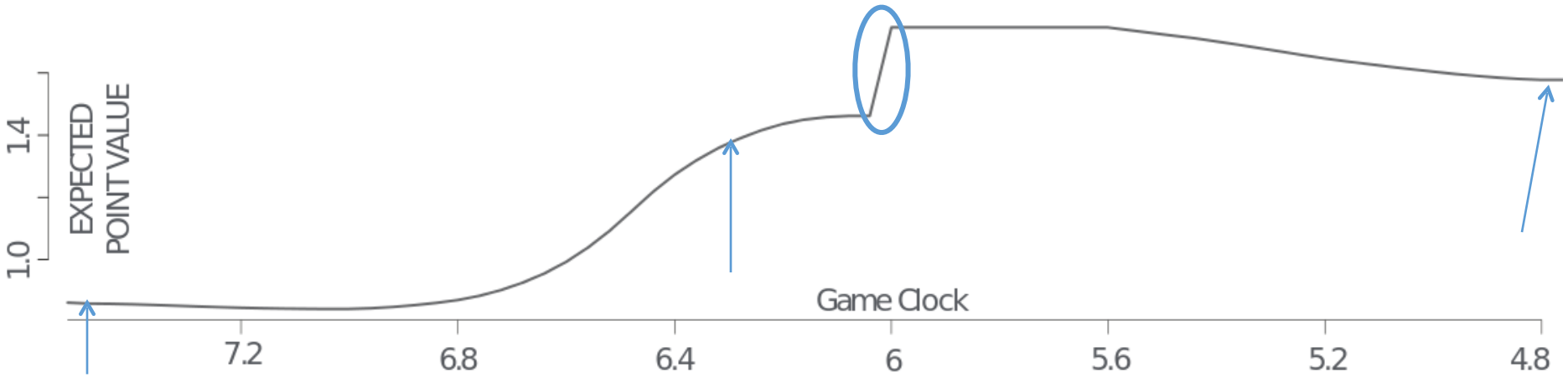



# Expected Possession Value

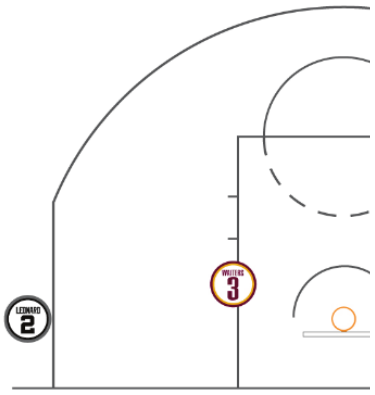
A basketball possession's stock ticker

# Expected Possession Value

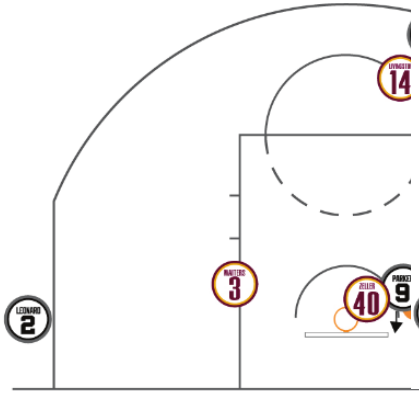
A basketball possession's stock ticker



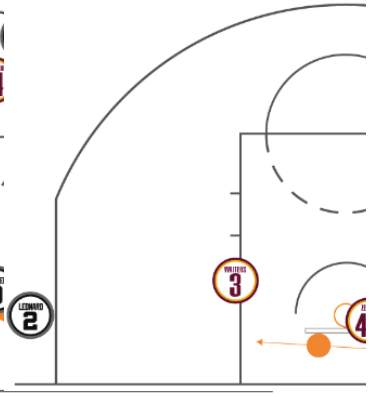

Tim Duncan Screen For Ton
Expected Points: **0.86**



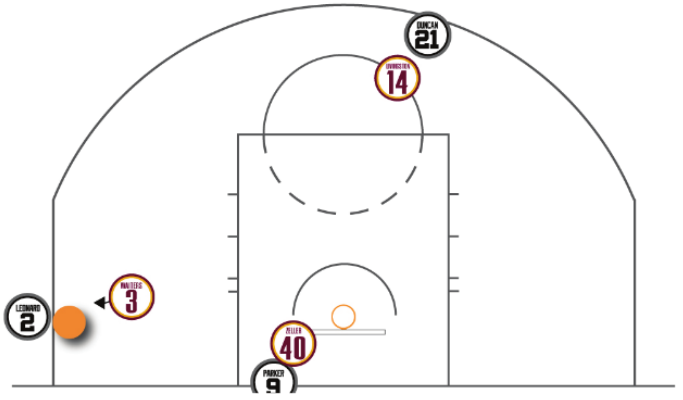
Tony Parker Enters Restricted
Expected Points: **1.36**



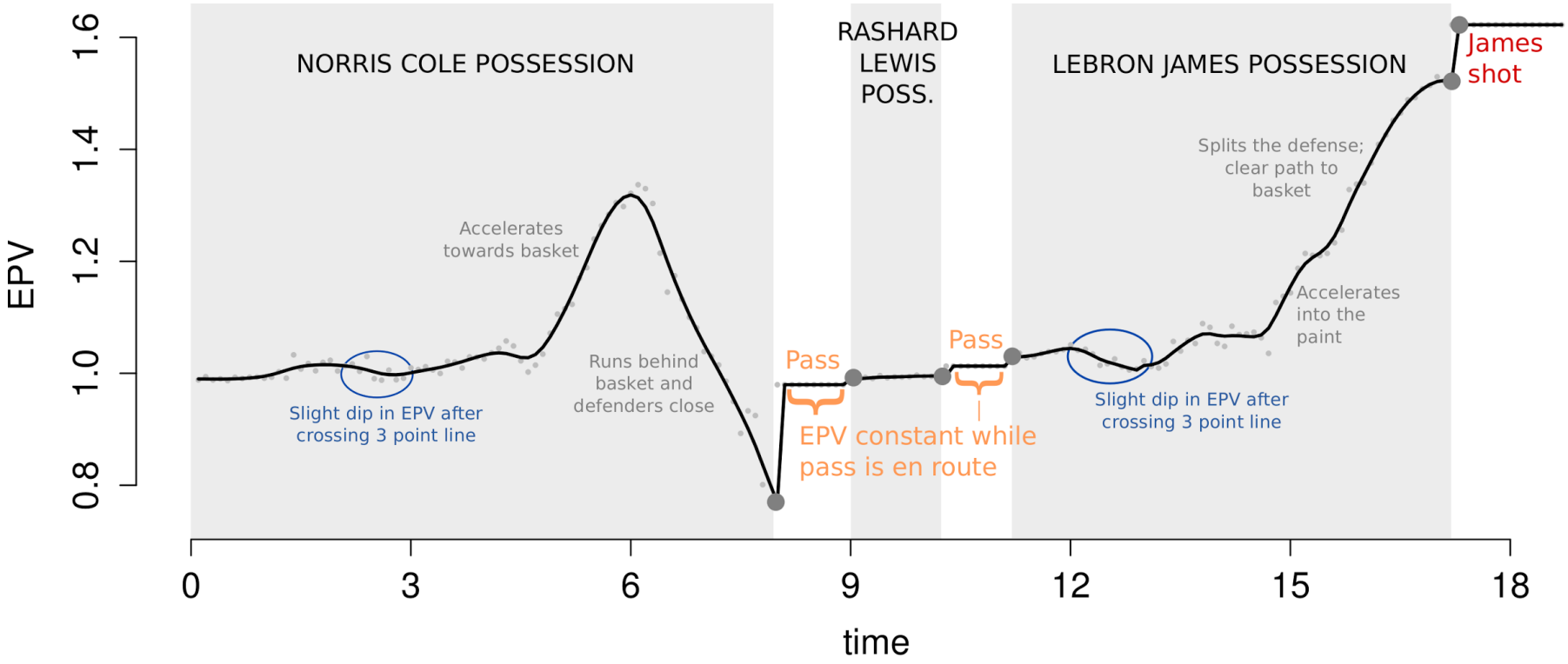
Tony Parker Passes The Ball
Expected Points: **1.46 → 1.7**



Kawhi Leonard Shoots The Game Winning Shot
Expected Points: **1.58**



# Expected Possession Value



**Expected Possession Value (EPV)** is the **number of points** the offense is **expected** to score before the end of the current possession, **given all of the information** we have up to **now.**

# Takeaways

New data allows us to answer old questions.

“Big data” is usually a bunch of small data; our questions are revealed to be more specific the more data we have.

Requirements often more complex than a naive prediction problem.

Methodology relevant to a range of modern datasets (geospatial, animal tracking, mobile health, etc)