# Data Mining meets Football
(soccer)

Ulf Brefeld
Knowledge Mining & Assessment
TU Darmstadt / DIPF
brefeld@cs.tu-darmstadt.de

TECHNISCHE UNIVERSITÄT DARMSTADT

DIPF
Educational Research and Educational Information

# Data Mining meets Football

(soccer)

Ulf Brefeld

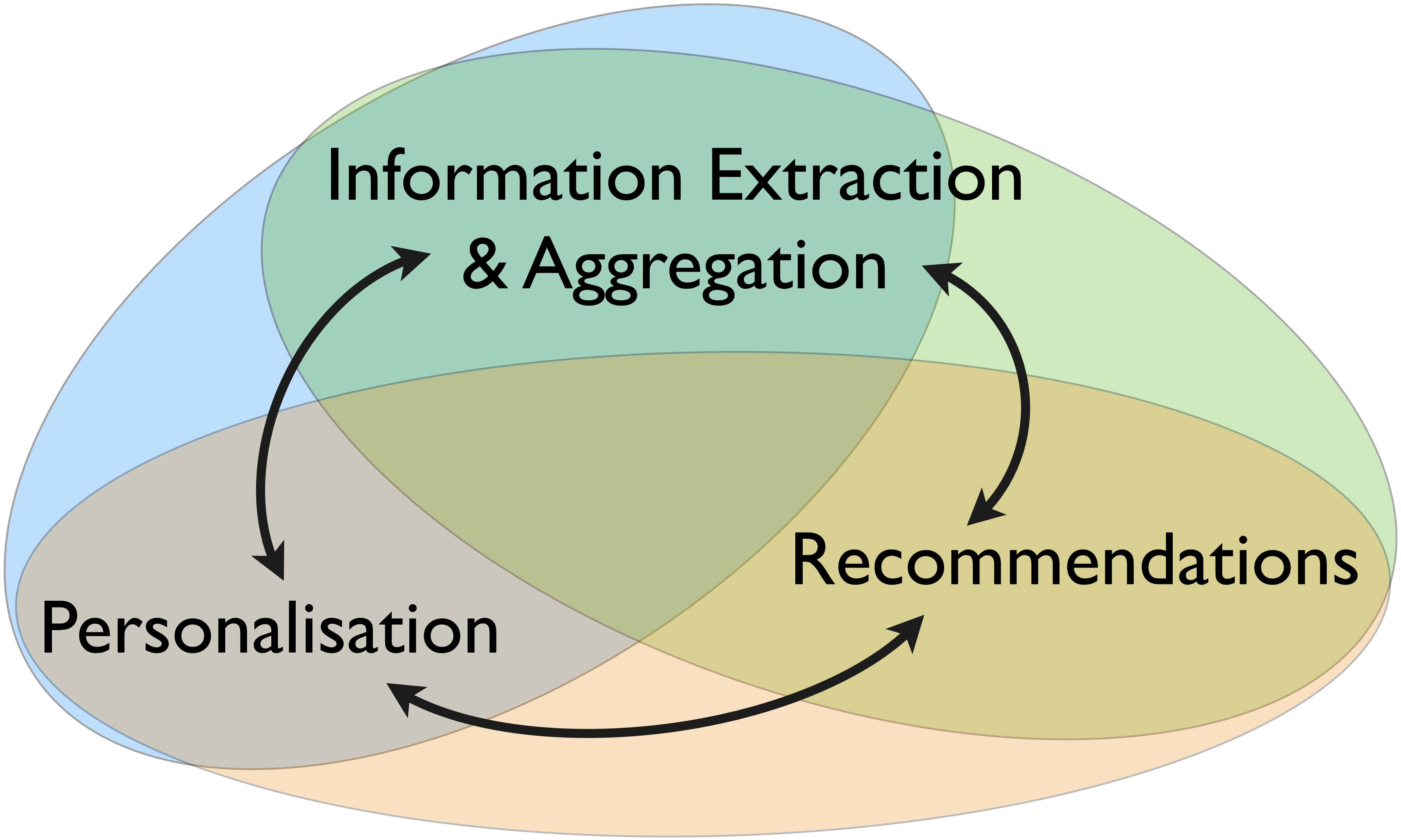
Machine Learning Group

Leuphana University of Lüneburg

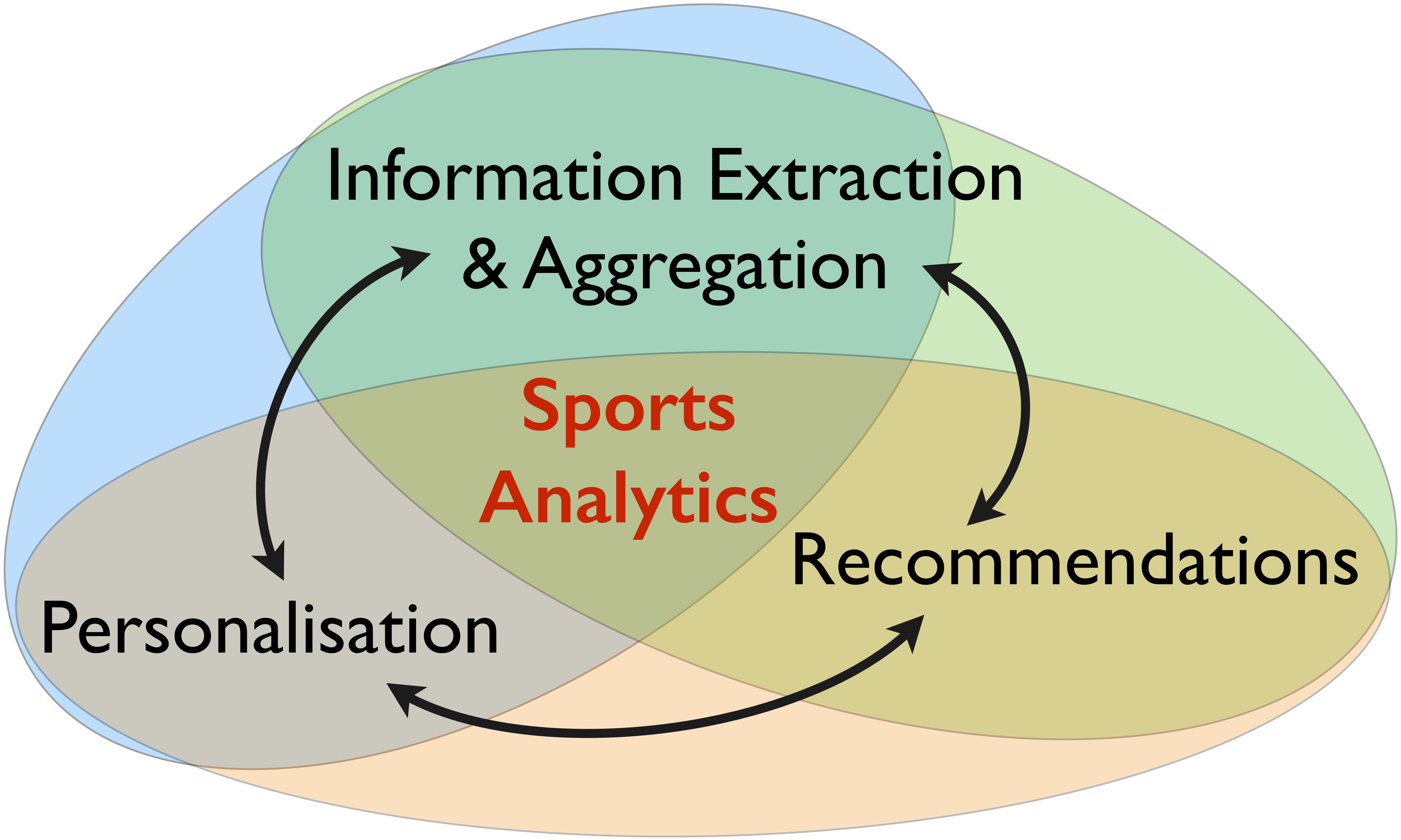
LEUPHANA

UNIVERSITÄT LÜNEBURG

# Machine Learning / Data Mining



# Machine Learning / Data Mining



# German Bundesliga

On average 43,502 attendees per game
13.31m attendees per season

http://www.ruhrnachrichten.de/storage/pic/mdhl/artikelbilder/sport/4081417_1_Bayern1.jpg?version=1387208424

# Monetary Aspects

http://www.statista.com/topics/1774/bundesliga/

| | |
|---|---|
| Revenue of European soccer market | €19.90bn |
| Revenue of German Bundesliga | €2,172.59m |
| German Bundesliga total value of player assets | €413.77m |
| FC Bayern Munich brand value | €794.60m |
| FC Bayern Munich profit after tax | €14.00m |

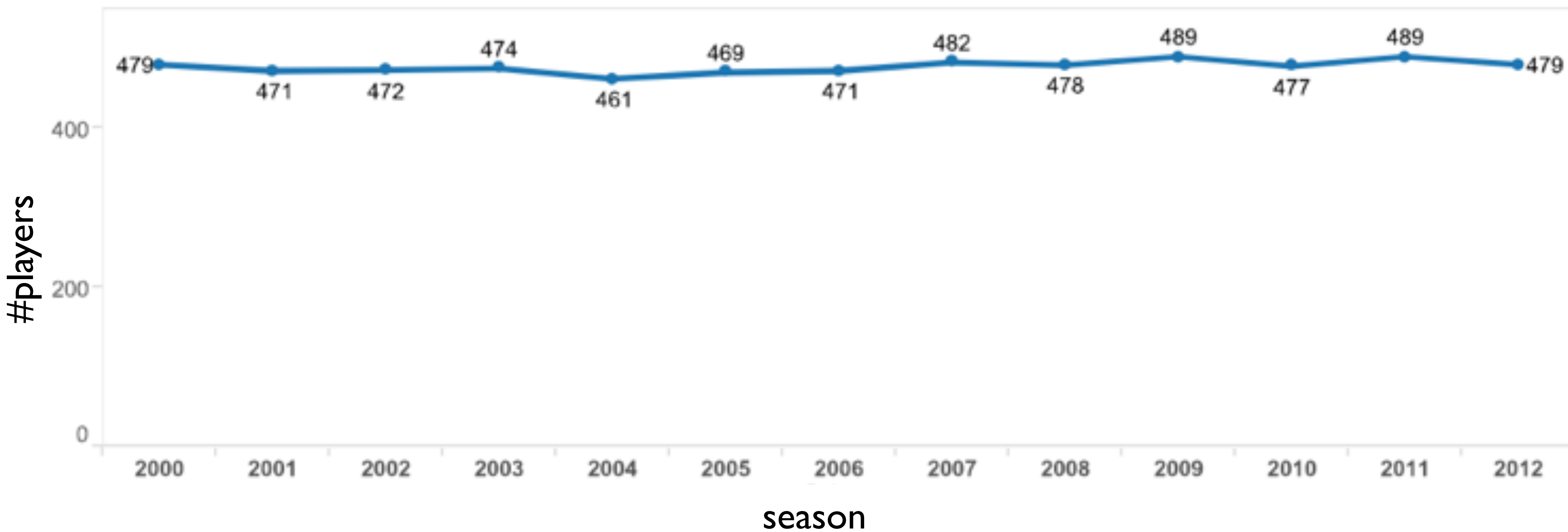
# Traditional Sports Analytics

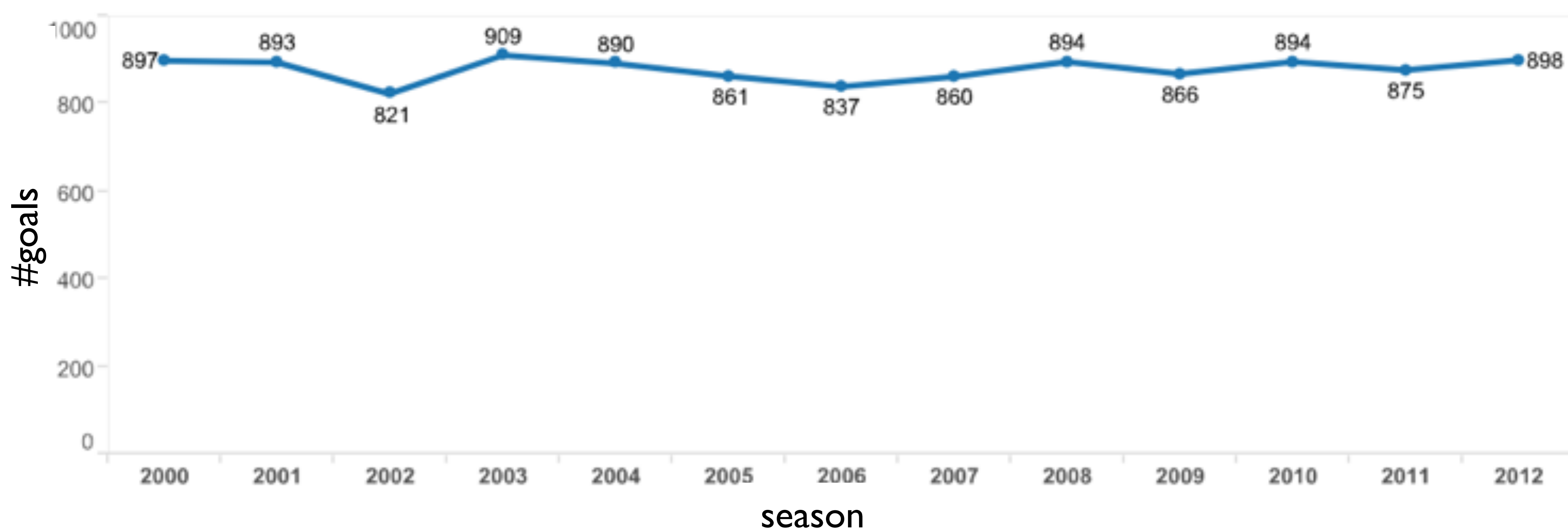
- Monetary aspects

| | | | | | |
|---|---|---|---|---|---|
| Freitag, 19.12. | | | | | |
| 20:30 | FSV Mainz 05 | Bayern München | 13 | 6,5 | 1,22 |
| Samstag, 20.12. | | | | | |
| 15:30 | Bayer Leverkusen | Eintracht Frankfurt | 1,5 | 4,4 | 6,5 |
| 15:30 | FC Augsburg | Borussia M'gladbach | 2,3 | 3,4 | 3,1 |
| 15:30 | Schalke 04 | Hamburger SV | 1,9 | 3,6 | 4,0 |
| 15:30 | VfB Stuttgart | SC Paderborn | 1,85 | 3,7 | 4,1 |
| 15:30 | Werder Bremen | Borussia Dortmund | 6,5 | 4,3 | 1,5 |
| 18:30 | VfL Wolfsburg | 1.FC Köln | 1,5 | 4,3 | 6,5 |
| Sonntag, 21.12. | | | | | |
| 15:30 | Hertha BSC | 1899 Hoffenheim | 2,5 | 3,4 | 2,8 |
| 17:30 | SC Freiburg | Hannover 96 | 2,35 | 3,4 | 3,0 |

- Statistics to serve information needs…

# Descriptive Statistics



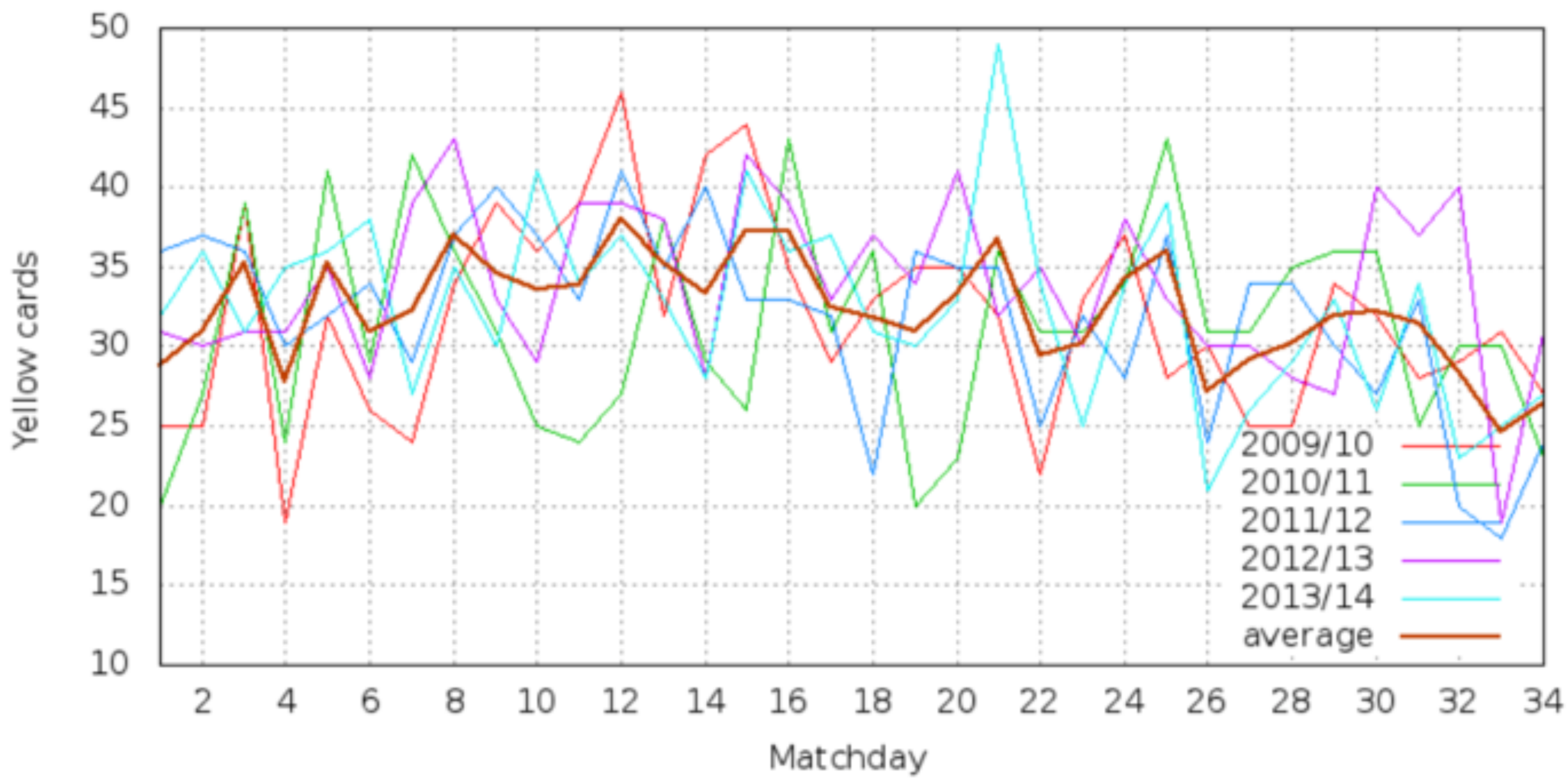



# Distribution of Goals

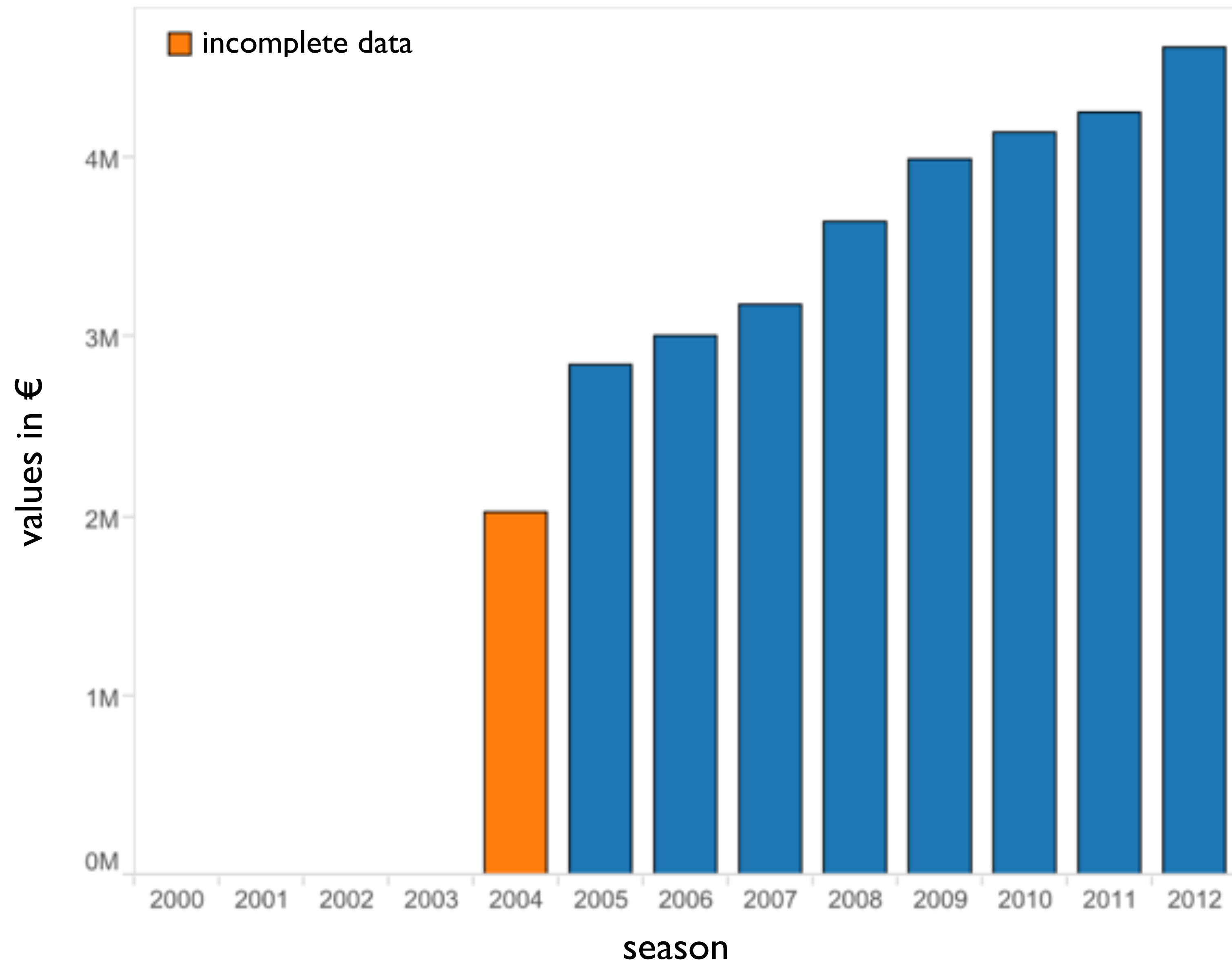
home team

| away team | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 6,81% | 8,62% | 7,59% | 5,28% | 2,31% | 0,70% | 0,28% | 0,10% | | |
| 1 | 6,03% | 11,16% | 8,85% | 4,63% | 2,51% | 0,73% | 0,23% | 0,03% | 0,03% | |
| 2 | 4,78% | 6,16% | 5,46% | 2,51% | 1,18% | 0,45% | 0,15% | | | 0,03% |
| 3 | 1,99% | 3,29% | 2,11% | 1,18% | 0,28% | 0,20% | 0,08% | | | |
| 4 | 0,78% | 1,18% | 0,88% | 0,18% | 0,13% | 0,05% | | 0,03% | | |
| 5 | 0,15% | 0,33% | 0,13% | 0,05% | | | | | | |
| 6 | 0,15% | 0,10% | 0,08% | 0,05% | | | | | | |
| 8 | | 0,03% | | | | | | | | |

# Yellow Cards



# Average Player Value



- Yeah, interesting… but what does it tell us?

"B. Charlton v F. Beckenbauer", David Marsh
1966 World Cup Final, England - W. Germany

# Trajectories and Tactics

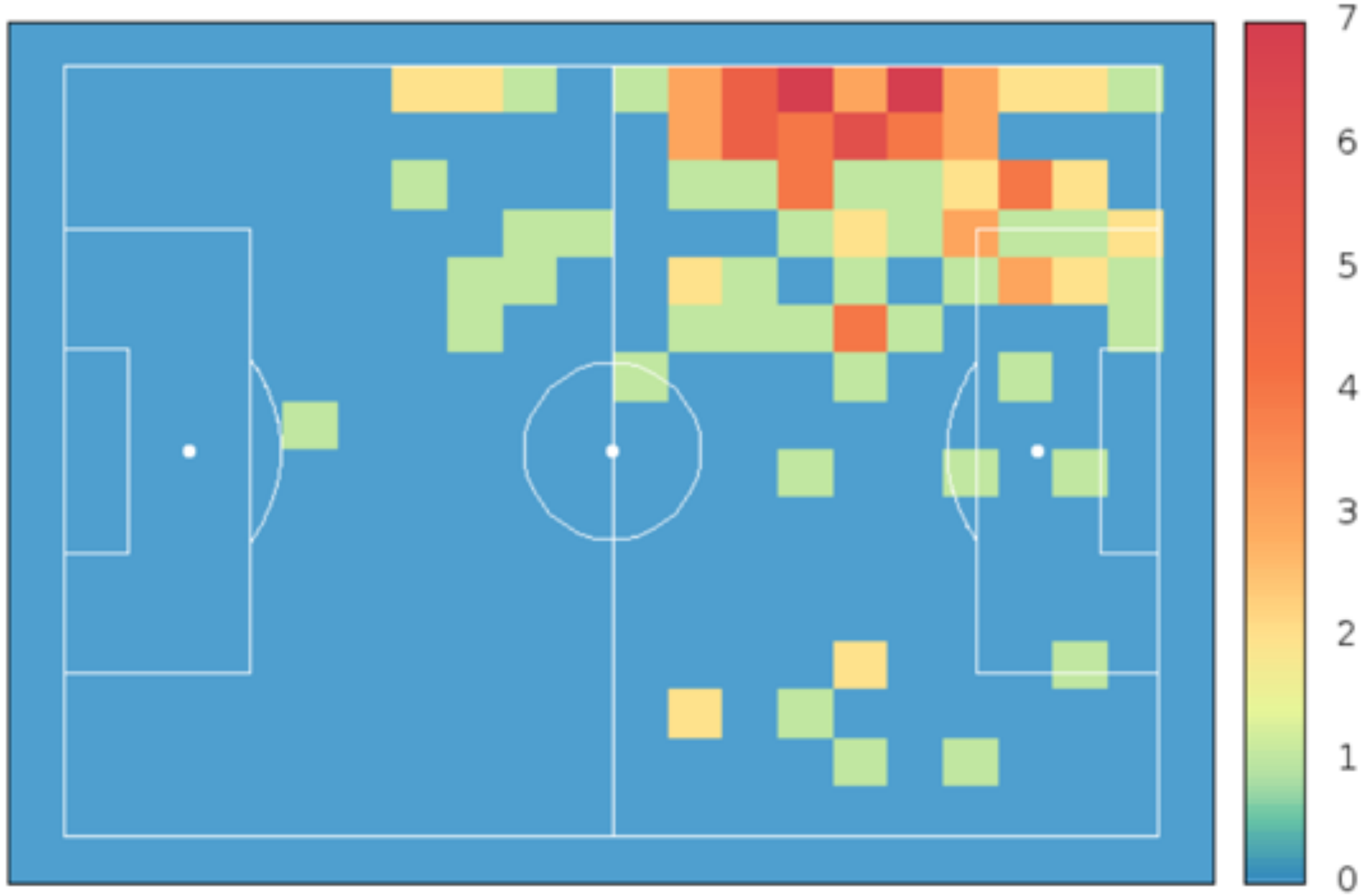
- Understanding player movements is a precondition for analysing game strategy (i.e., tactics)

# Player Trajectory Data

- Cameras capture positions of players and ball*
  * Referee also tracked and recorded but data usually kept private
- x,y,(z) coordinates
  - $\geq$24 frames p second
  - Manually denoised (corners, mass confrontations,…)
  - Players annotated
- Perfect data for analysing movements, coordination, tactics, etc.

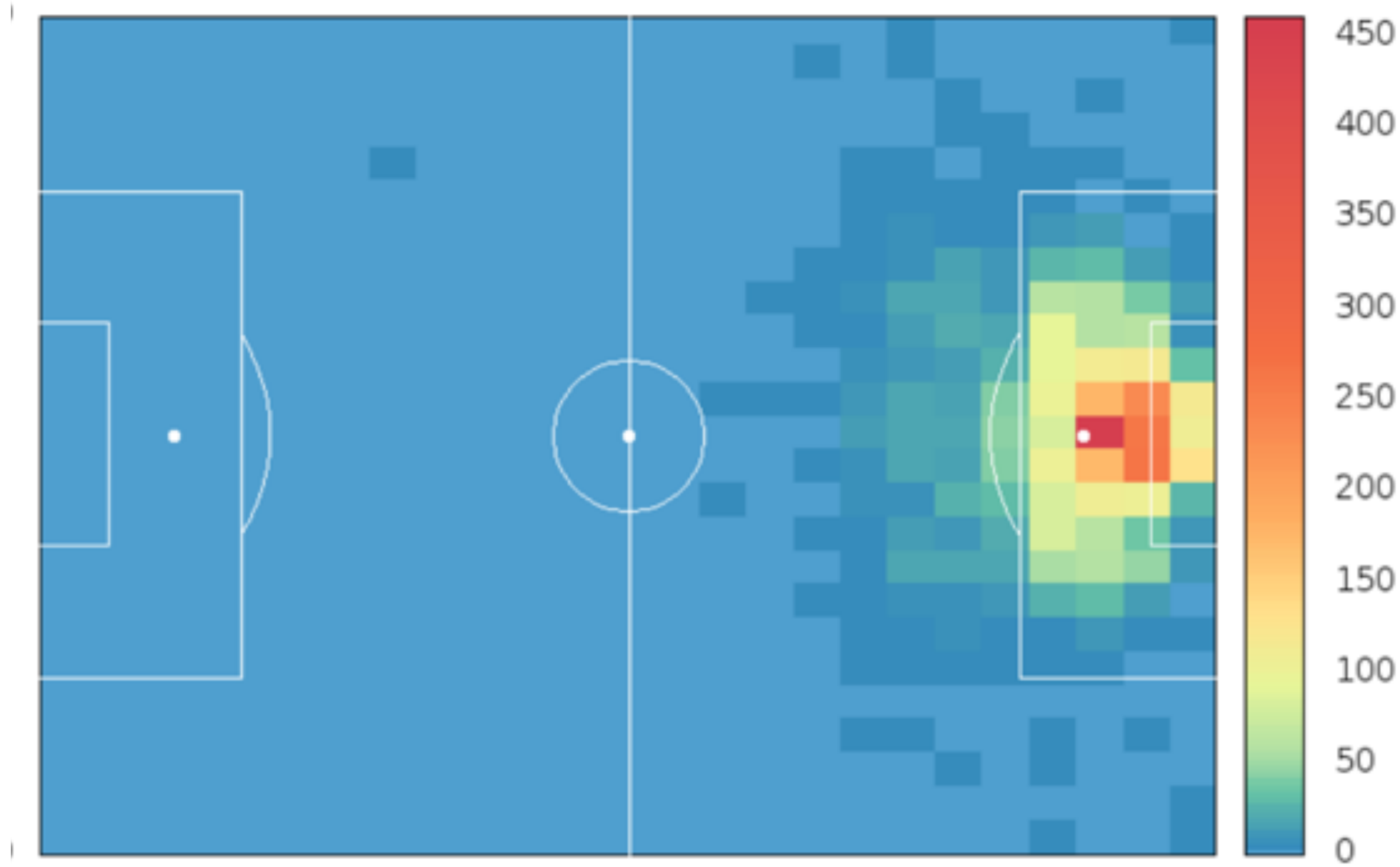
# Ball touches of Franck Ribery

(FCB vs BMG, season 2013/14)

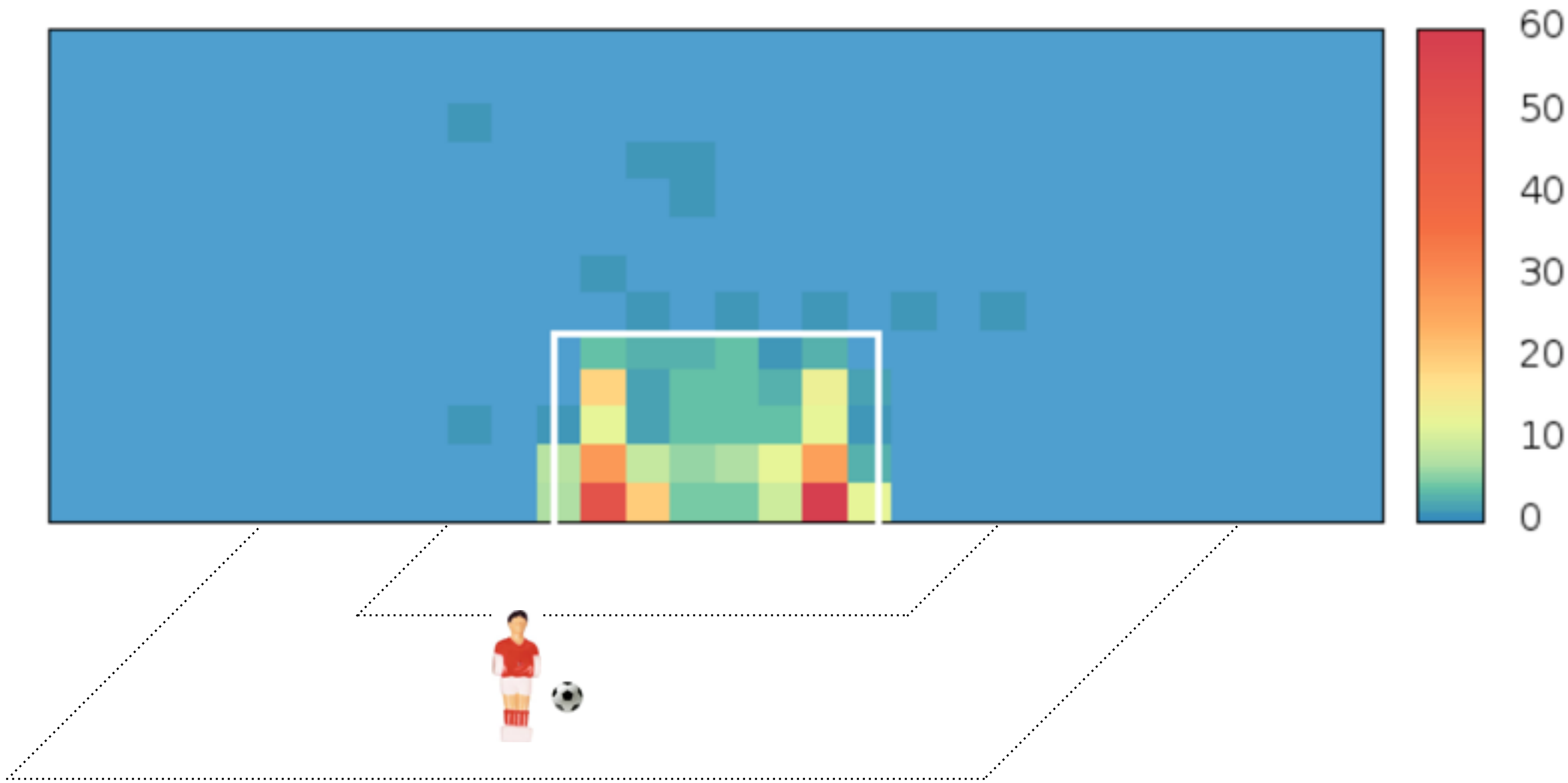


# Shots leading to Goals

(season 2009/10 - 2013/14)



# Goalmouth Coordinates (penalties)



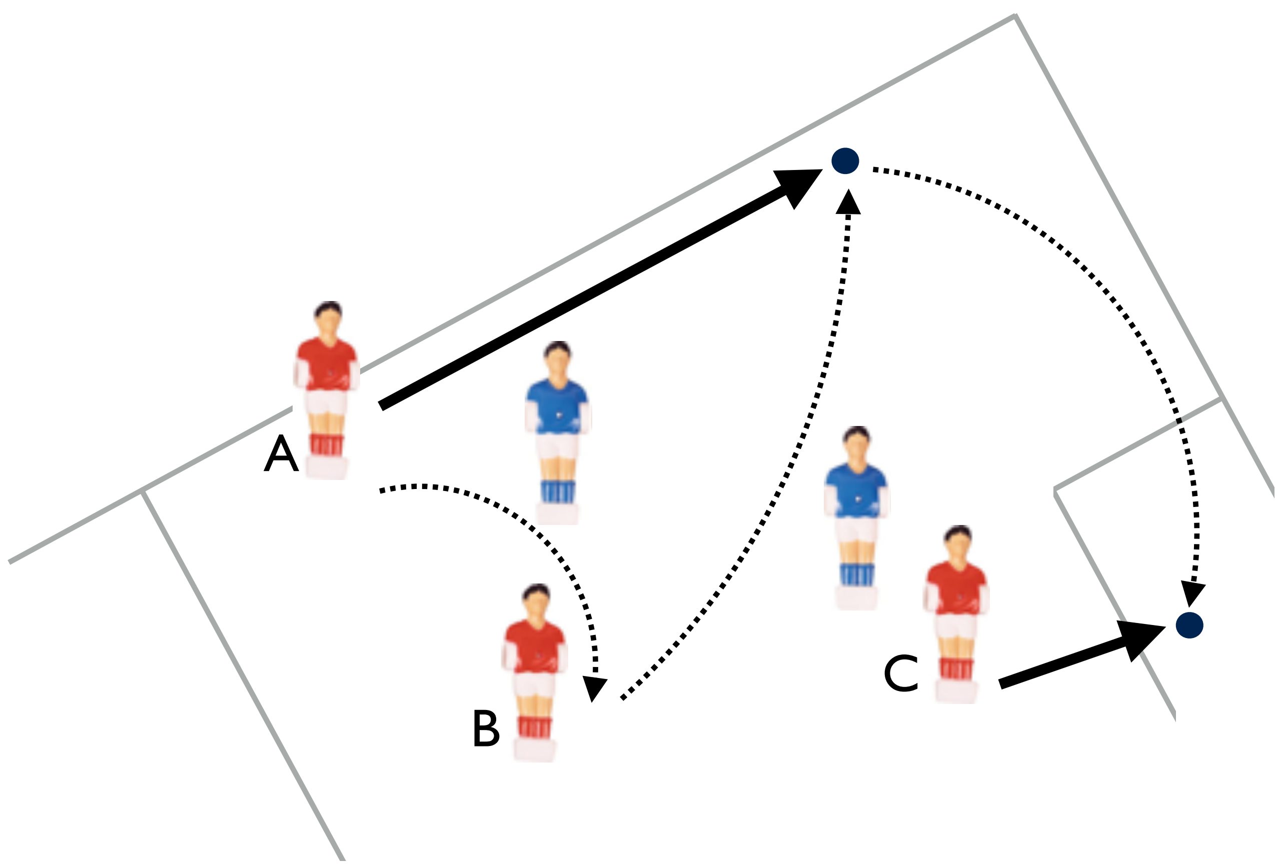
- Hm… still, what does it tell us?

# Use Cases

- Analyse opponent tactics
- Detect strengths/weaknesses in strategy
- Automatic game plans
- Serious games / training
- Player scouting
- Improved media coverage
- …

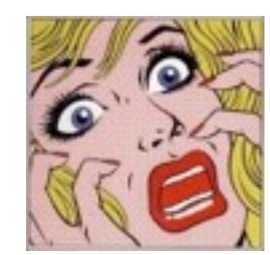
# Identifying Patterns

- Pattern = "interesting" event
- E.g., A plays 1-2 with B and crosses to C



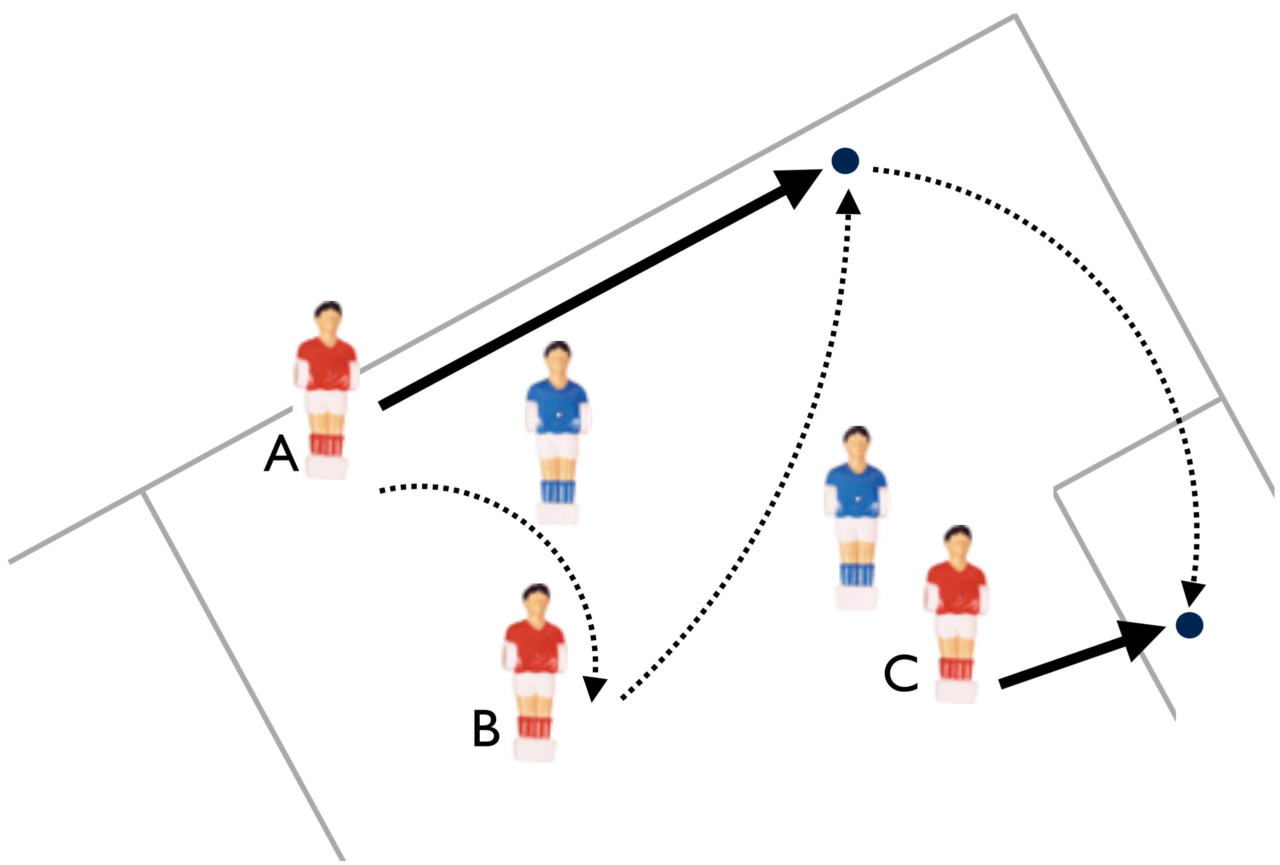
# Why is it difficult?

- >3 million positions per game
- Every player generates ≈135000 positions per game
- There are $\approx 135000^{23}$ different candidate patterns*
  * Ignoring the fact that patterns are of different lengths
- This is considerably larger than the number of atoms in our galaxy**
  ** Dark and exotic matter already included
- Explicit enumeration infeasible
- What similarity measure to use?



# Identifying Patterns

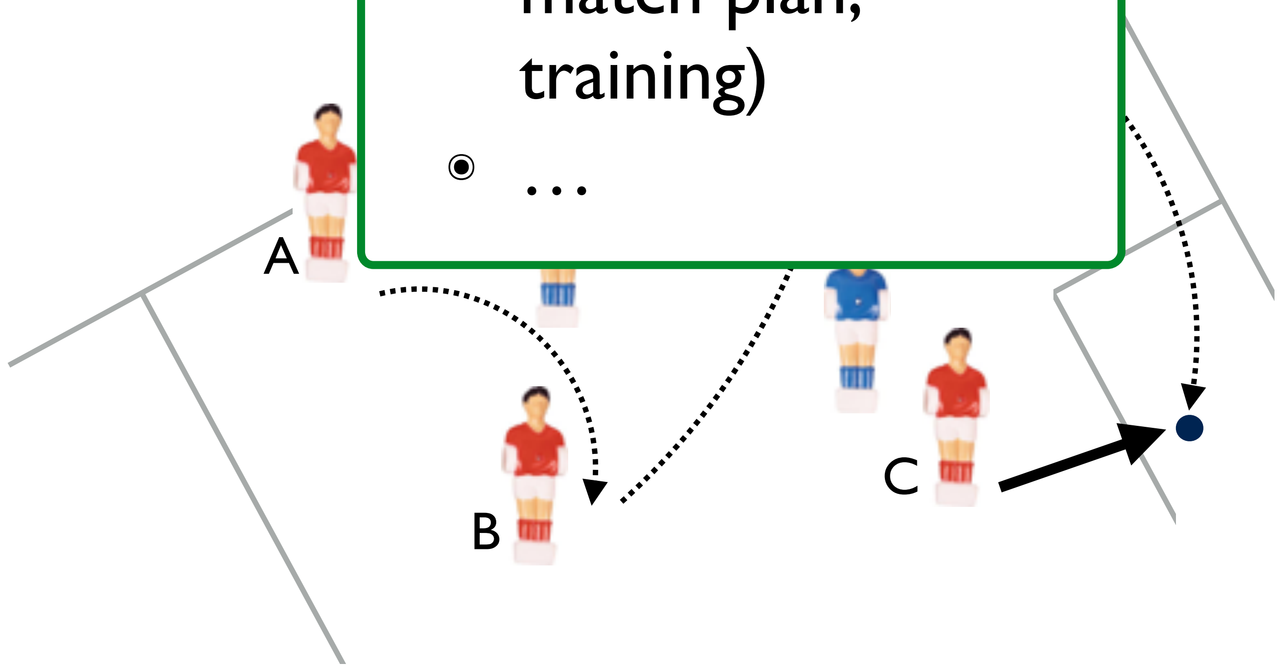
- Pattern = “interesting” event
- E.g., A plays 1-2 with B and crosses to C



# Identifying Patterns

- Pattern = "interesting" event
- E.g., A plays … sses to C

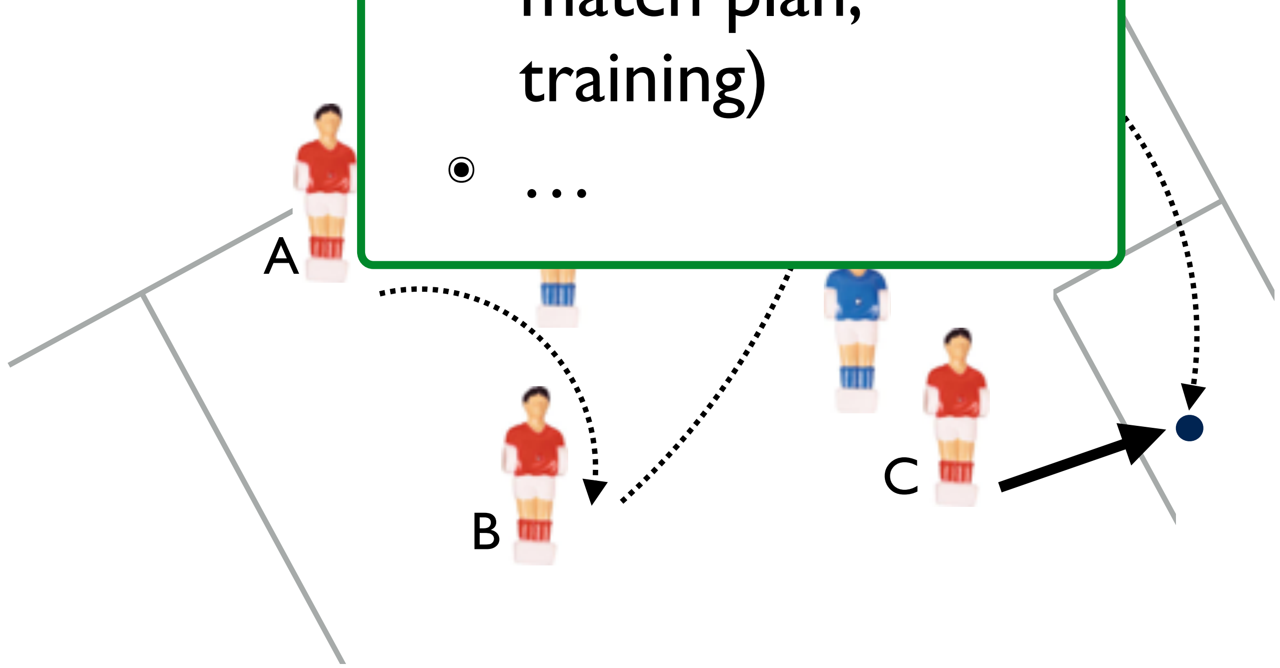
- frequent
- rare (anomalies/ outliers)
- predefined (e.g., match plan, training)
- …



# Identifying Patterns

- Pattern = "interesting" event
- E.g., A plays … sses to C

  - **frequent**
  - rare (anomalies/ outliers)
  - predefined (e.g., match plan, training)
  - …



# Representation

- Position = player coordinates on the pitch
- A game of soccer = positional data stream
- Player trajectory = sequence of consecutive positions
- Positions represented by angles wrt reference vector $\mathbf{v}_{ref}$ (translation, rotation, scale invariant)

$$\alpha_i = sign(\boldsymbol{v}_i, \boldsymbol{v}_{ref}) \left[ cos^{-1} \left( \frac{\boldsymbol{v}_i^\top \boldsymbol{v}_{ref}}{\|\boldsymbol{v}_i\| \, \|\boldsymbol{v}_{ref}\|} \right) \right]$$

Vlachos et al. (KDD, 2004)

# Dynamic Time Warping

Rabiner & Juang (1993)

- Movements should be independent of player speed
- Dynamic time warping compensates phase shifts
- Distance measure $dist : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$
- DTW for sequences **s** and **q** defined recursively

$$
\begin{aligned}
g(\emptyset, \emptyset) &= 0 \\
g(\boldsymbol{s}, \emptyset) &= dist(\emptyset, \boldsymbol{q}) = \infty \\
g(\boldsymbol{s}, \boldsymbol{q}) &= dist(s_1, q_1) + min \left\{ \begin{array}{l} g(\boldsymbol{s}, \langle q_2, \ldots, q_m \rangle) \\ g(\langle s_2, \ldots, s_m \rangle, \boldsymbol{q}) \\ g(\langle s_2, \ldots, s_m \rangle, \langle q_2, \ldots, q_m \rangle) \end{array} \right\}
\end{aligned}
$$

# Dynamic Time Warping

Rabiner & Juang (1993)

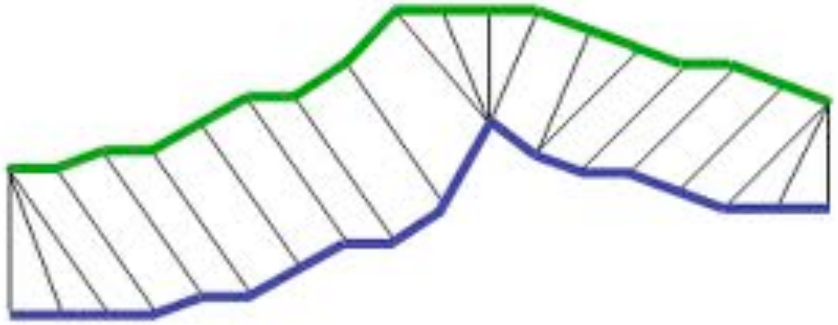
- Movements should be independent of player speed
- Dynamic time warping compensates phase shifts
- Distance measure $dist : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$
- DTW for sequences **s** and **q** defined recursively

$$\begin{aligned}
g(\emptyset, \emptyset) &= 0 \\
g(\boldsymbol{s}, \emptyset) &= dist(\emptyset, \boldsymbol{q}) = \infty \\
g(\boldsymbol{s}, \boldsymbol{q}) &= dist(s_1, q_1) + min \left\{ \begin{array}{l} g(\boldsymbol{s}, \langle q_2, \ldots, q_m \rangle) \\ g(\langle s_2, \ldots, s_m \rangle, \boldsymbol{q}) \\ g(\langle s_2, \ldots, s_m \rangle, \langle q_2, \ldots, q_m \rangle) \end{array} \right\}
\end{aligned}$$

O(|s||q|)

# Approximate DTW

- Approximate DTW by lower bounds $f(\boldsymbol{s}, \boldsymbol{q}) \leq g(\boldsymbol{s}, \boldsymbol{q})$
- Focus on characteristic values
- Kim et al. (ICDE, 2001)
  - first, last, greatest, smallest value
- Keogh (VLDB, 2002)
  - minimum/maximum values of subsequences
- Complexity in O(|**s**|)



# Locality Sensitive Hashing

Athitsos et al. (2008), Gionis et al., (1999)

- Distance-based hash function $h : \mathcal{D} \rightarrow \mathbb{R}$

$$h_{\boldsymbol{s}_1,\boldsymbol{s}_2}(\boldsymbol{s}) = \frac{dist(\boldsymbol{s},\boldsymbol{s}_1)^2 + dist(\boldsymbol{s}_1,\boldsymbol{s}_2)^2 - dist(\boldsymbol{s},\boldsymbol{s}_2)^2}{2\,dist(\boldsymbol{s}_1,\boldsymbol{s}_2)}$$

$s_1$ and $s_2$ randomly drawn from database

use Kim et al. (ICDE, 2001) as distance function

- Bucket determined by $h^{[t_1,t_2]}_{\boldsymbol{s}_1,\boldsymbol{s}_2}(\boldsymbol{s}) = \begin{cases} 1 : h_{\boldsymbol{s}_1,\boldsymbol{s}_2}(\boldsymbol{s}) \in [t_1,t_2] \\ 0 : \quad otherwise \end{cases}$
- Set of admissible intervals

$$\mathcal{T}(\boldsymbol{s}_1,\boldsymbol{s}_2) = \left\{ [t_1,t_2] \,:\, Pr_{\mathcal{D}}(h^{[t_1,t_2]}_{\boldsymbol{s}_1,\boldsymbol{s}_2}(\boldsymbol{s})) = 0) = Pr_{\mathcal{D}}(h^{[t_1,t_2]}_{\boldsymbol{s}_1,\boldsymbol{s}_2}(\boldsymbol{s})) = 1) \right\}$$

# Computing Similarities

- Remainder needs test for identity
- Use outcomes of
  - Dynamic time warping
  - Approximate DTW
  - Locality sensitive hashing (buckets)
- … together with similarity threshold

# Episode Discovery

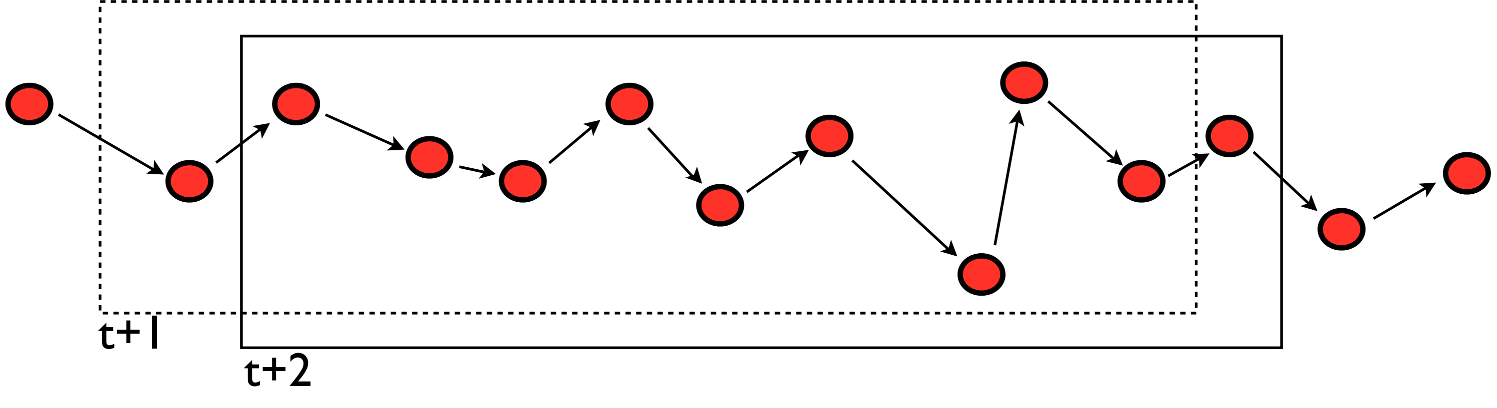
- Apriori-based algorithms
- Approach based on Achar et al. (2012)
- Distributed implementation scheme (Hadoop)
- Two phases
  - Candidate generation (Mapper)
  - Counting (Reducer)

# Empirical Evaluation

- DEBS Grand Challenge
  http://www.orgs.ttu.edu/debs2013/index.php?goto=cfchallengedetails
  - 8 vs. 8 soccer game recorded by Fraunhofer IIS
  - In total 33 sensors
    - 1 sensor per shoe (200Hz)
    - 1 sensor in the ball (2000Hz)
  - 15,000 positions per second (3 dimensional)

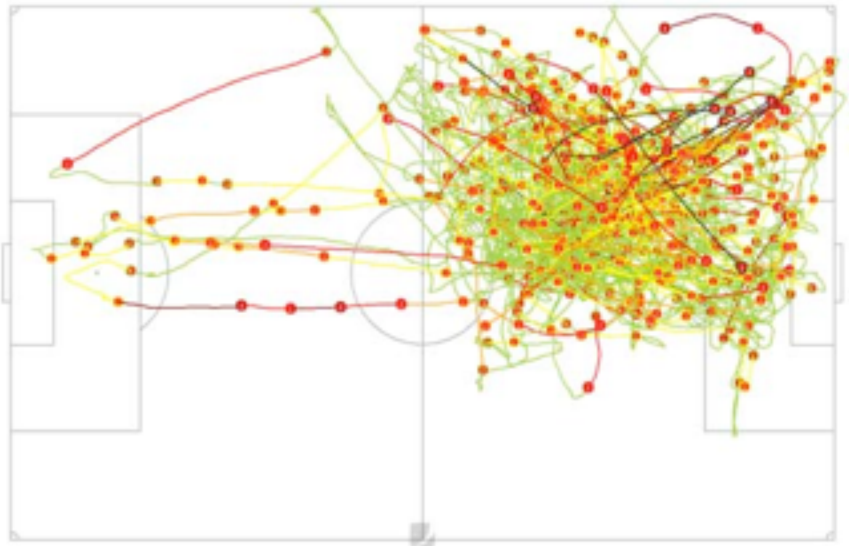
# Representation

- Further preprocessing:
  - Discarding positions outside of the pitch
  - Removing half-time effect of changing sides
  - Averaging player positions over 100ms
- Trajectory windows of size 10

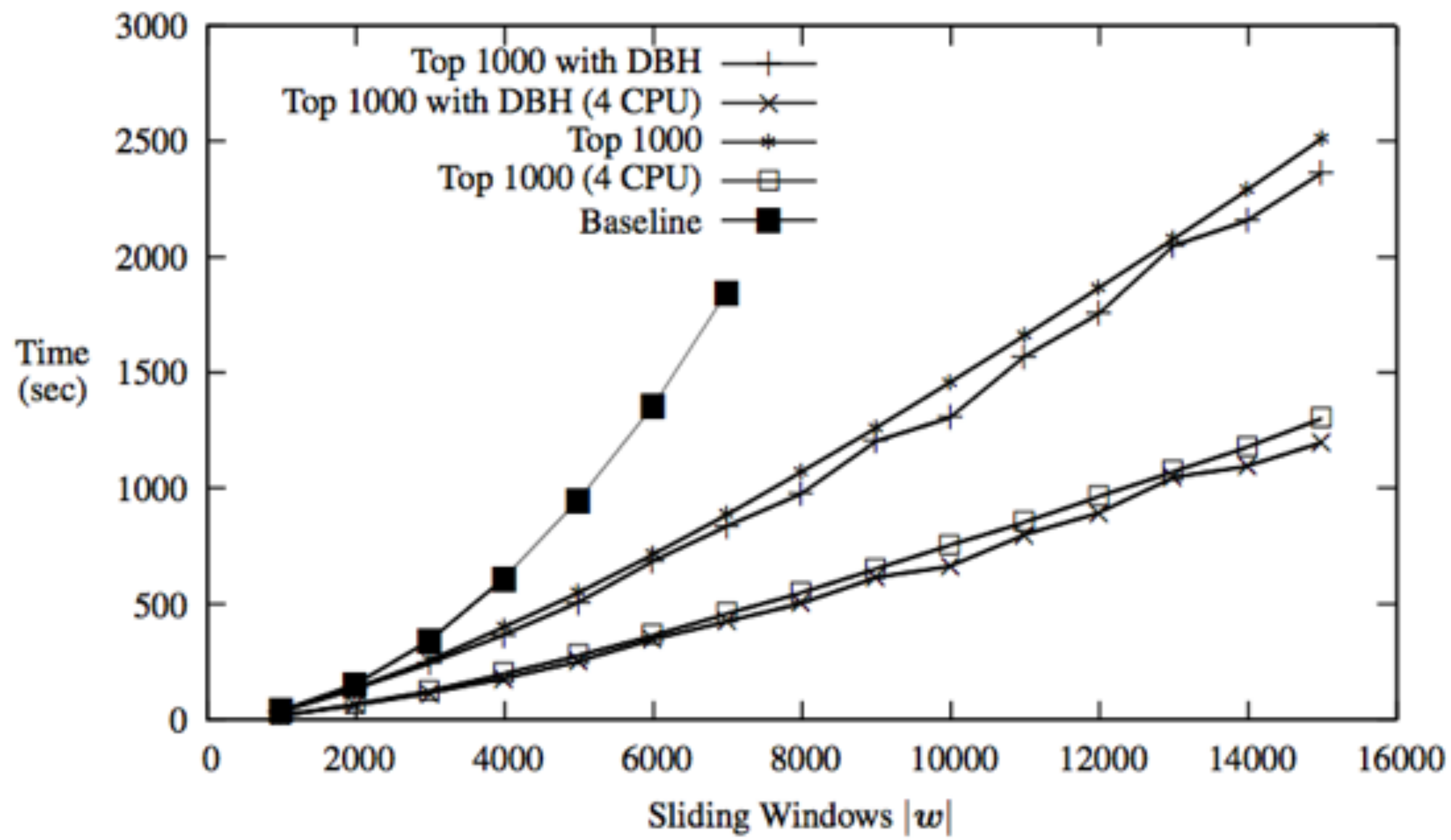


# Evaluation

- Given: a query trajectory
- Task: Find near-duplicates
  - (i.e., N=1000 most similar trajectories)
- Focus on 15k consecutive positions of one player
  - (for baseline comparisons)

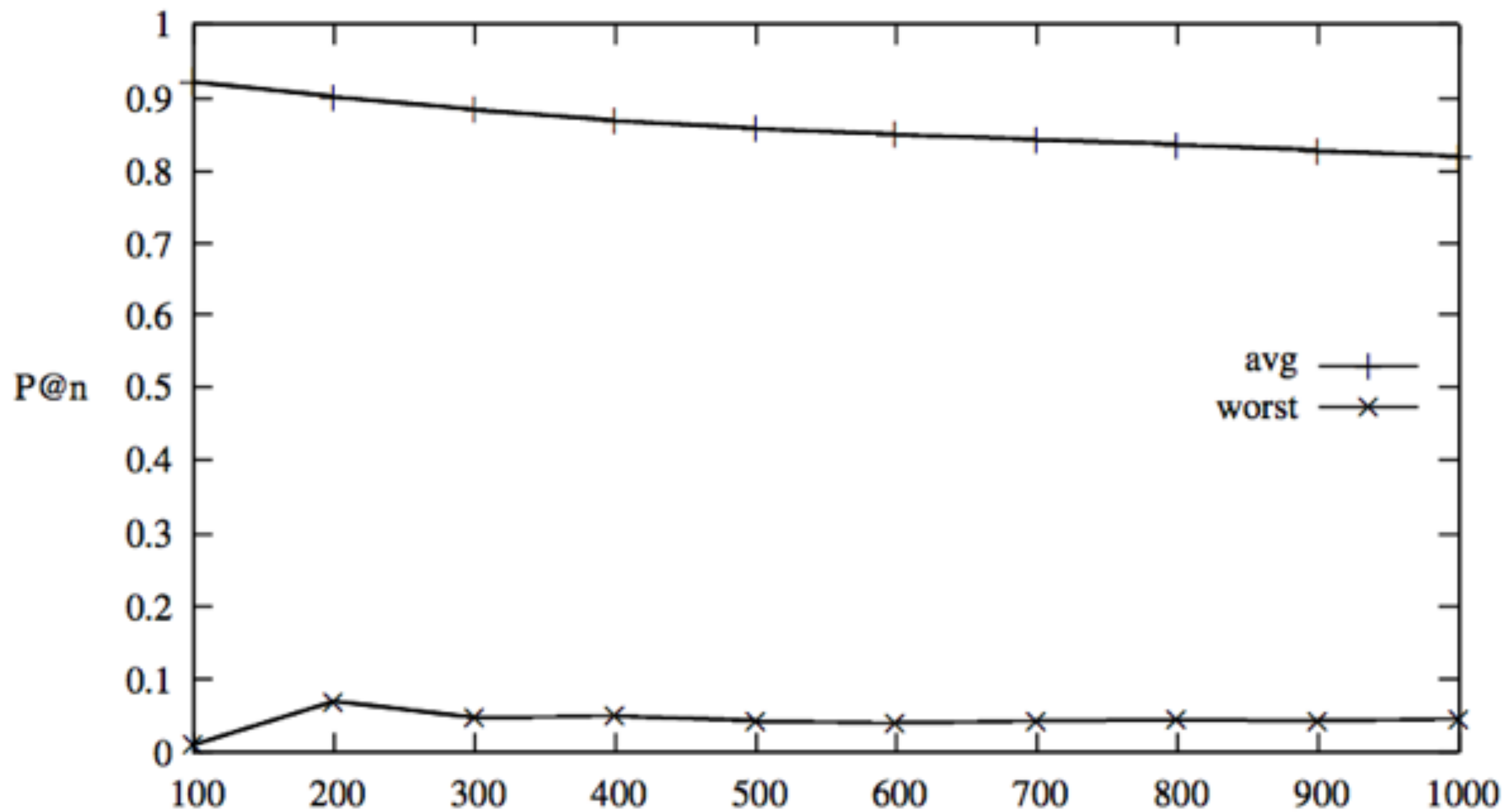


# Run-time



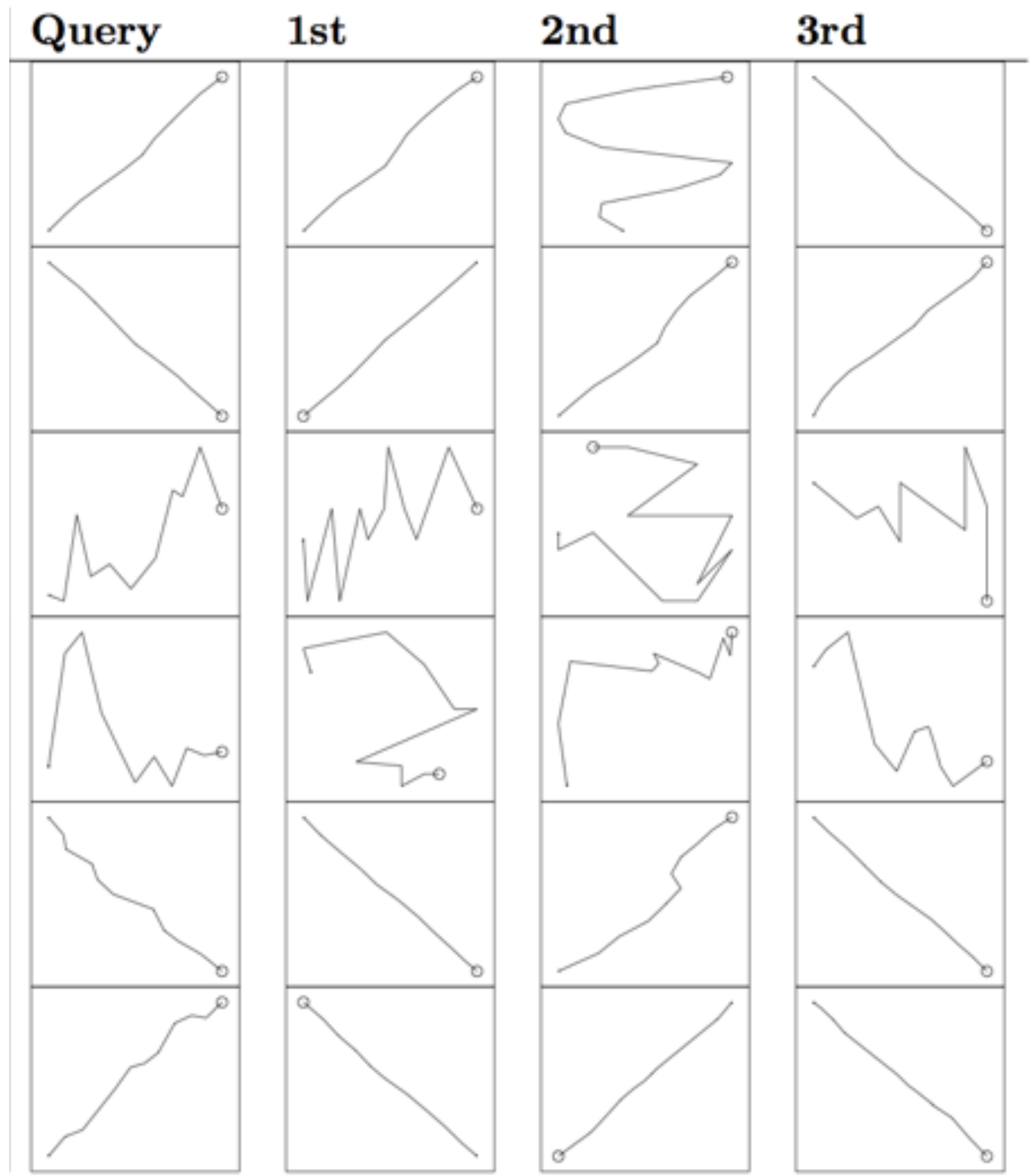

- Exact computation infeasible
- Dynamic time warping very effective
- LSH adds only little

# LSH Accuracy



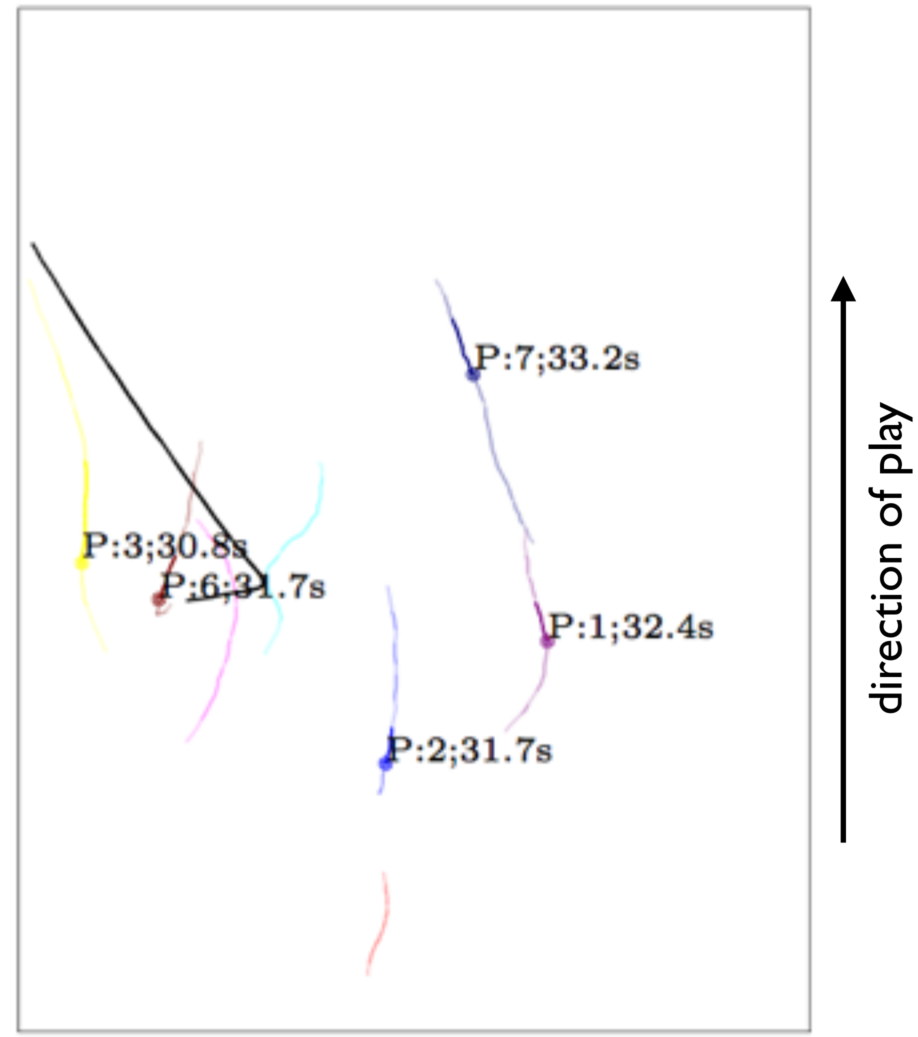

- On average LSH performs very accurate
- However, worst cases clearly inappropriate

# Exemplary Retrieval



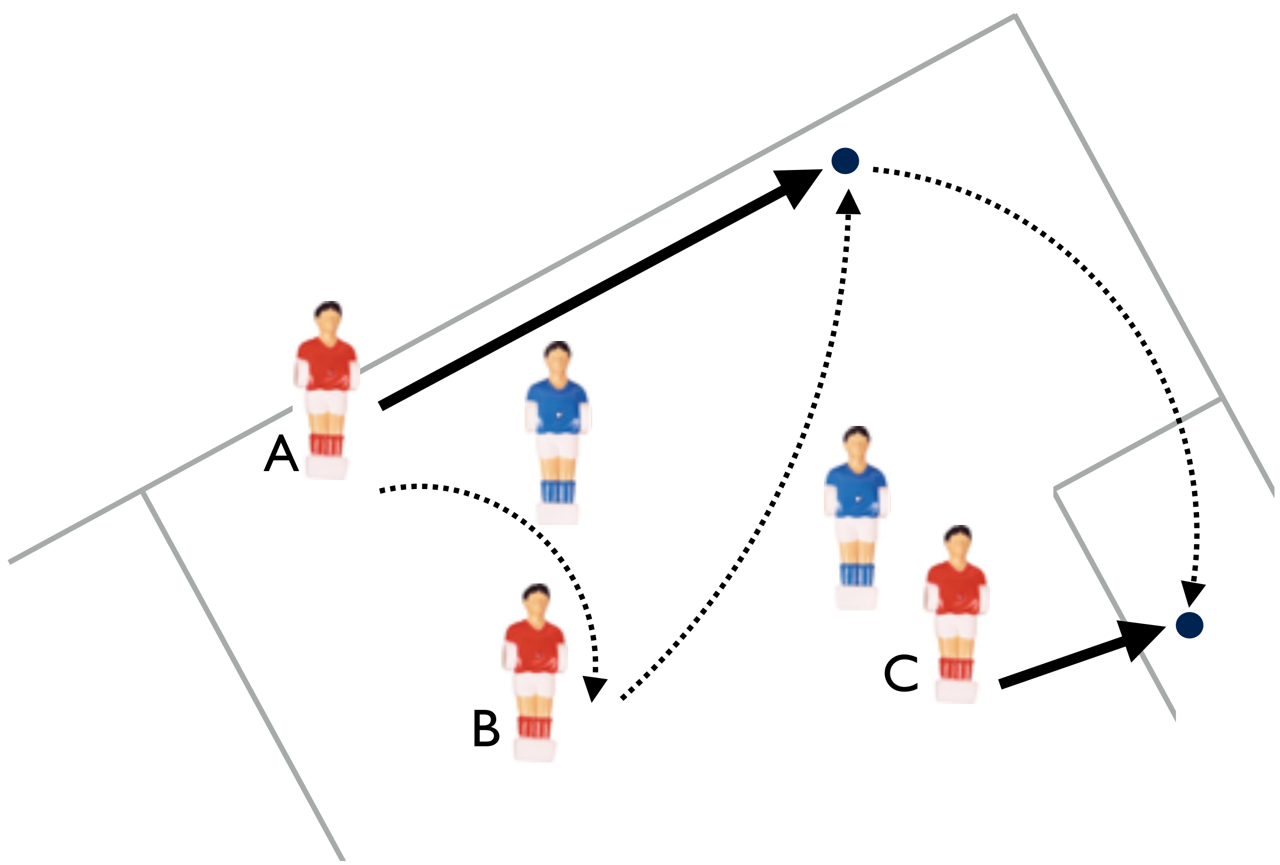
# Exemplary Pattern

- Ball is played towards opponent goal (black)
- Trajectories in pattern visualised by thick lines (dot indicates beginning)
- Players 1,2,3,6 and 7 move in direction of ball



# Identifying Patterns

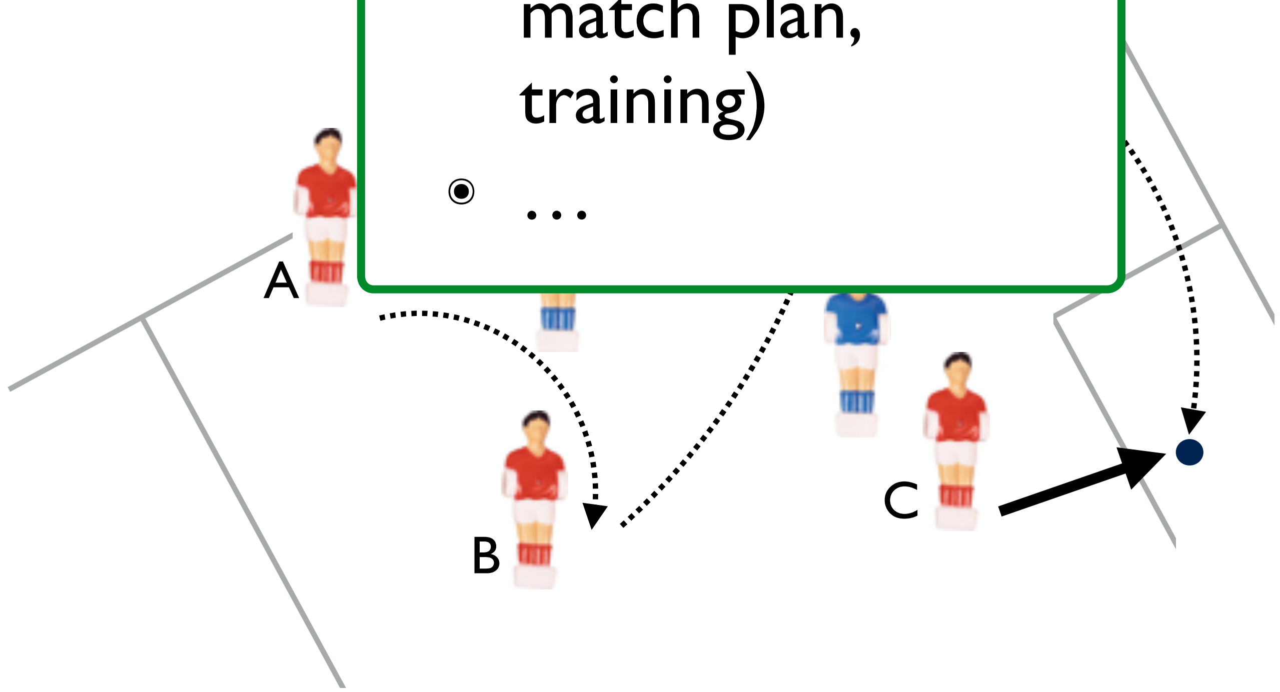
- Pattern = “interesting” event
- E.g., A plays 1-2 with B and crosses to C



# Identifying Patterns

- Pattern = “interesting” event
- E.g., A plays sses to C

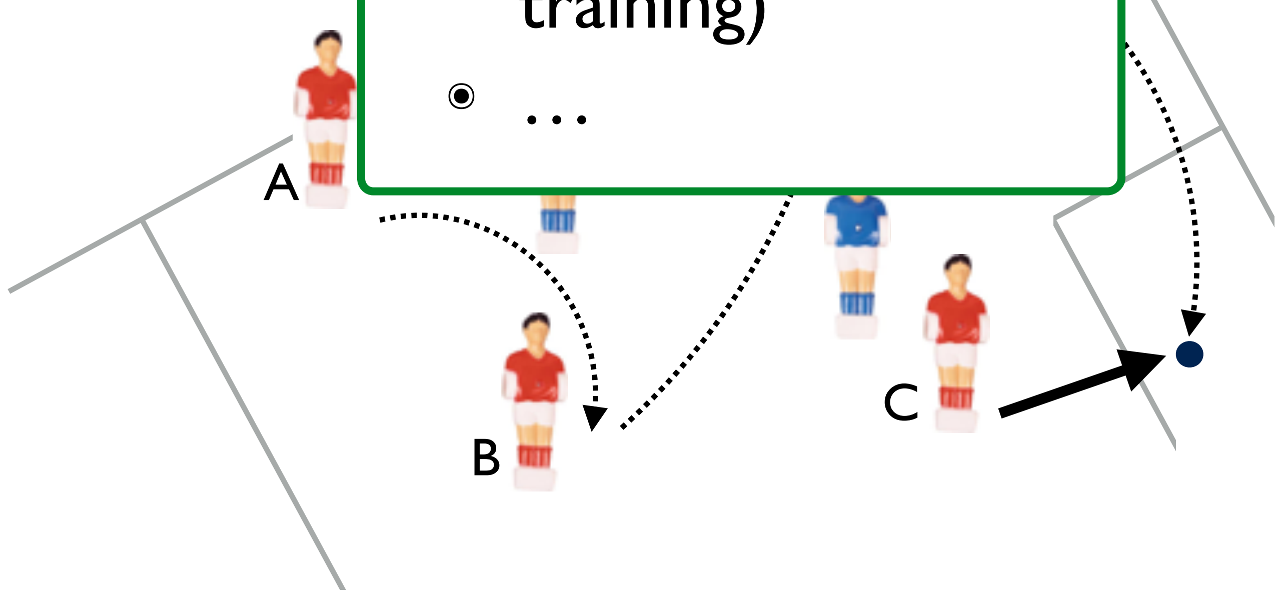
  - frequent
  - rare (anomalies/ outliers)
  - predefined (e.g., match plan, training)
  - …



# Identifying Patterns

- Pattern = “interesting” event
- E.g., A plays … sses to C

> - frequent
> - rare (anomalies/ outliers)
> - **predefined** (e.g., match plan, training)
> - …



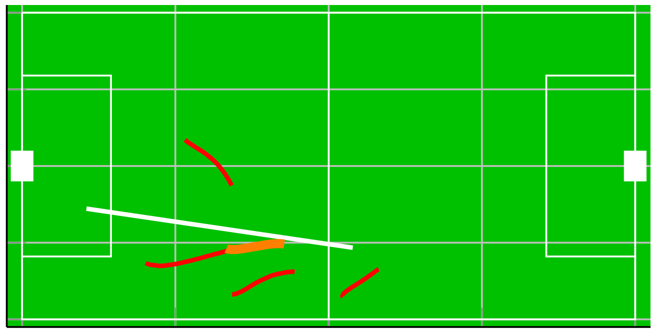
# Patterns / Events

- Individual level
- Group level
- Team level

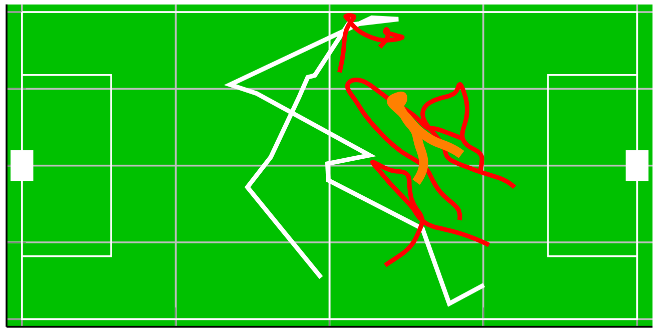
# Patterns / Events

- Individual level
- Group level
- Team level

- 4 defence players
  → game initiations



- 4 offence players
  → scoring opportunities



# Spatio-temporal Convolution Kernels

Knauf, Memmert & Brefeld, Spatio-temporal Convolution Kernels, Machine Learning Journal, 2015

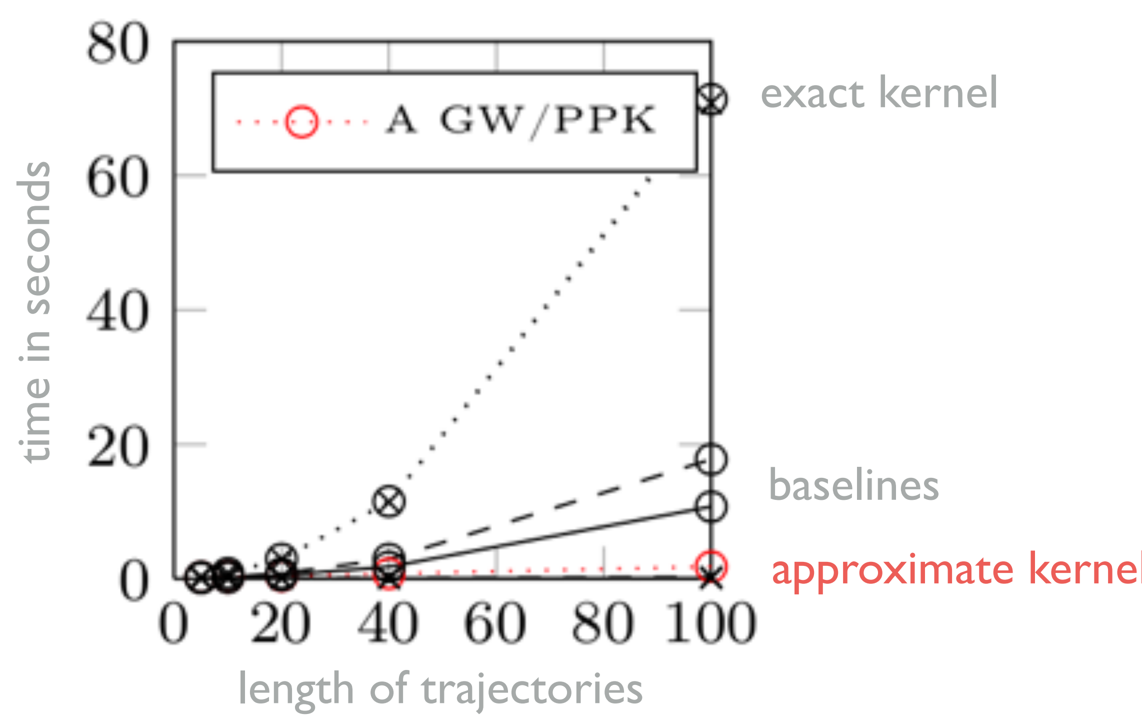
$$k(P,Q) = \frac{1}{|P||Q|} \sum_{(t,x_t)\in P,(s,y_s)\in Q} \underbrace{k_{[0,1]}(t,s)}_{\text{cheap temporal kernel}} \cdot \underbrace{k_{\mathcal{X}}(x_t,y_s)}_{\text{expensive spatial kernel}}$$

- Tailored similarity measure for multi-trajectory scenarios
- Separate data from algorithm, eg., works with every kernel machine (SVMs, kPCA, kernel kMeans, etc.)
- But: Complexity $\mathcal{O}(\underbrace{N^2}_{\text{number of trajectories}}\underbrace{L^2}_{\text{length of trajectories}})$

# Approximate STCKs

Knauf, Memmert & Brefeld, Spatio-temporal Convolution Kernels, Machine Learning Journal, 2015

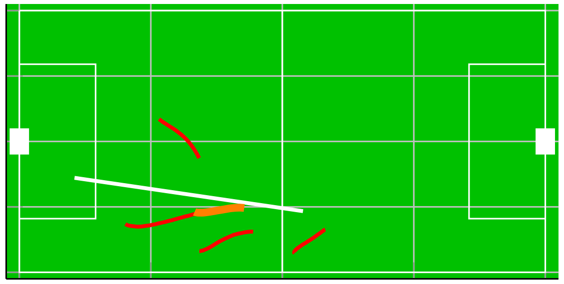
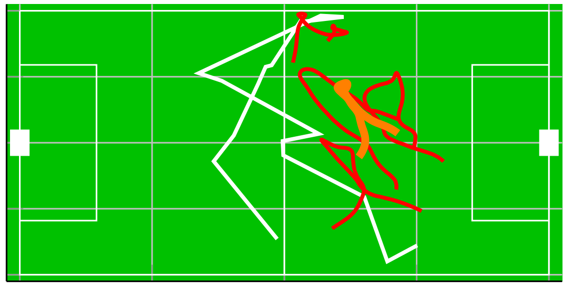
- Efficient approximation of exact kernel
- Idea: Use cheap temporal kernel as filter
- Evaluate spatial kernel by percental approximation



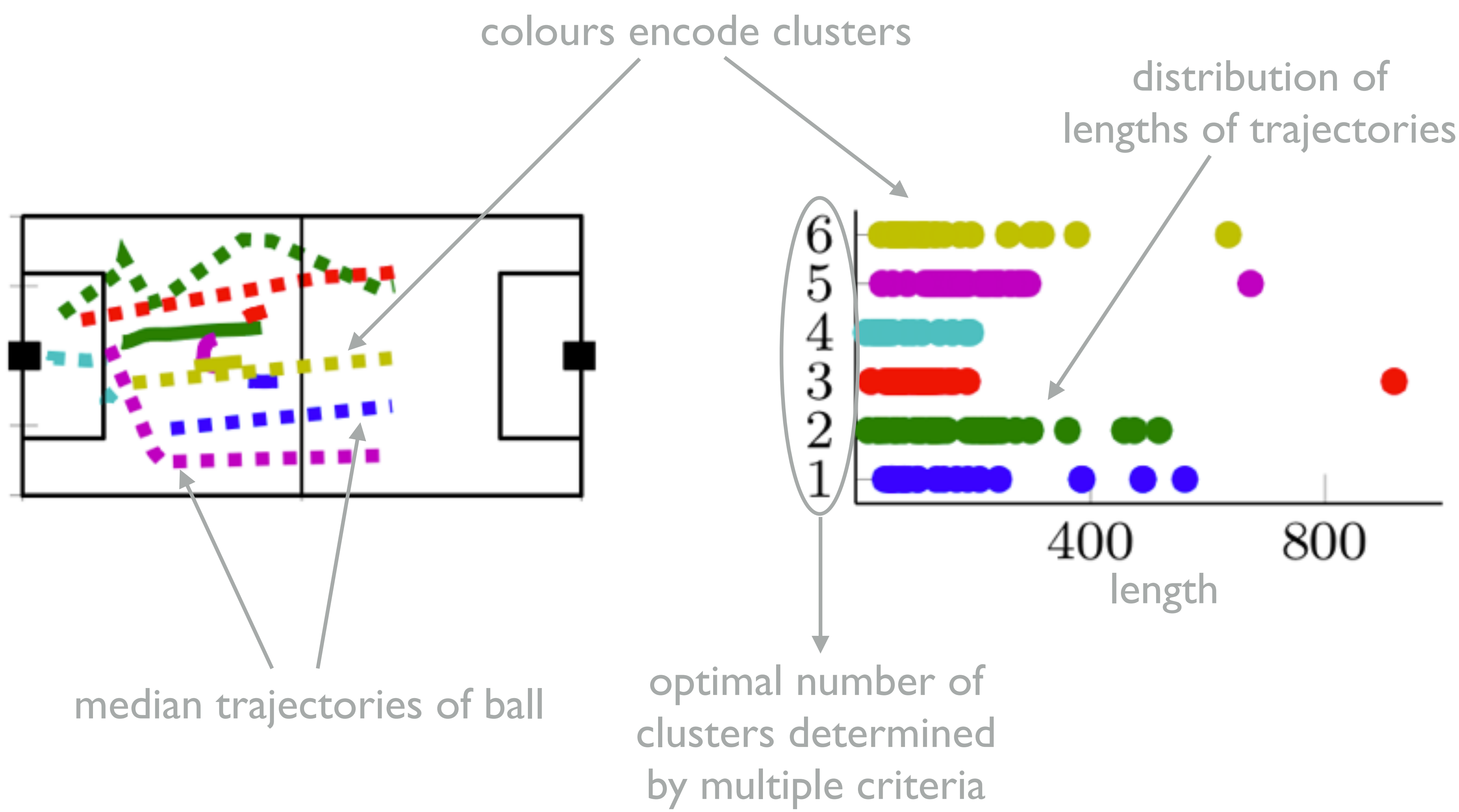
# Empirical Results

Knauf, Memmert & Brefeld, Spatio-temporal Convolution Kernels, Machine Learning Journal, 2015

- VIS.TRACK data, Bundesliga season 2011/12
- Two teams (5 games each)
- Cluster analysis w k-medoids
  - Game initiations (start: goal keeper has ball)
  - Scoring opportunities (end: ball in dangerous zone)

# Example


colours encode clusters
distribution of
lengths of trajectories
6
5
4
3
2
1
400
800
length
median trajectories of ball
optimal number of
clusters determined
by multiple criteria

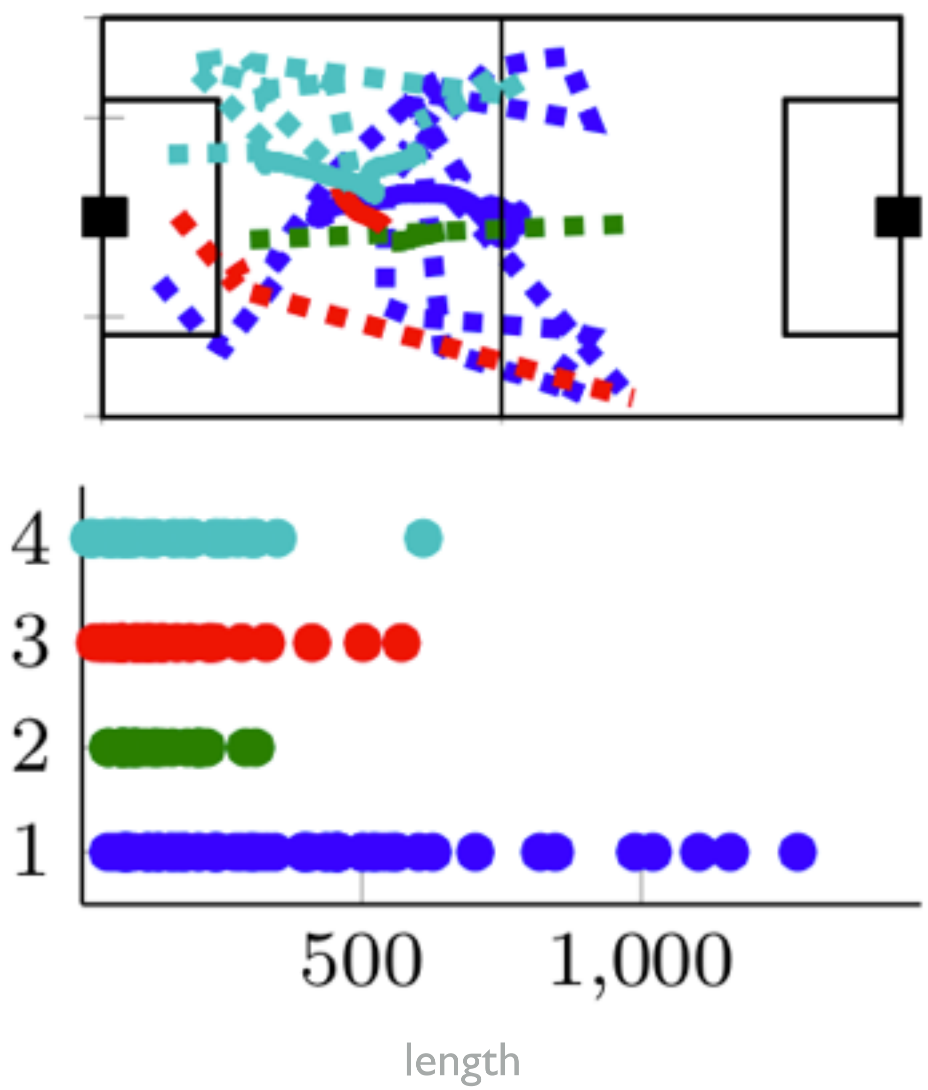
# Game Initiations

- Team A known for
  - Transporting the ball with few but rehearsed short game initiations to the opposing half
  - Many ball contacts, different players integrated
- Team B's strategy
  - Focused on increasingly long and straight balls
  - Few players involved on average

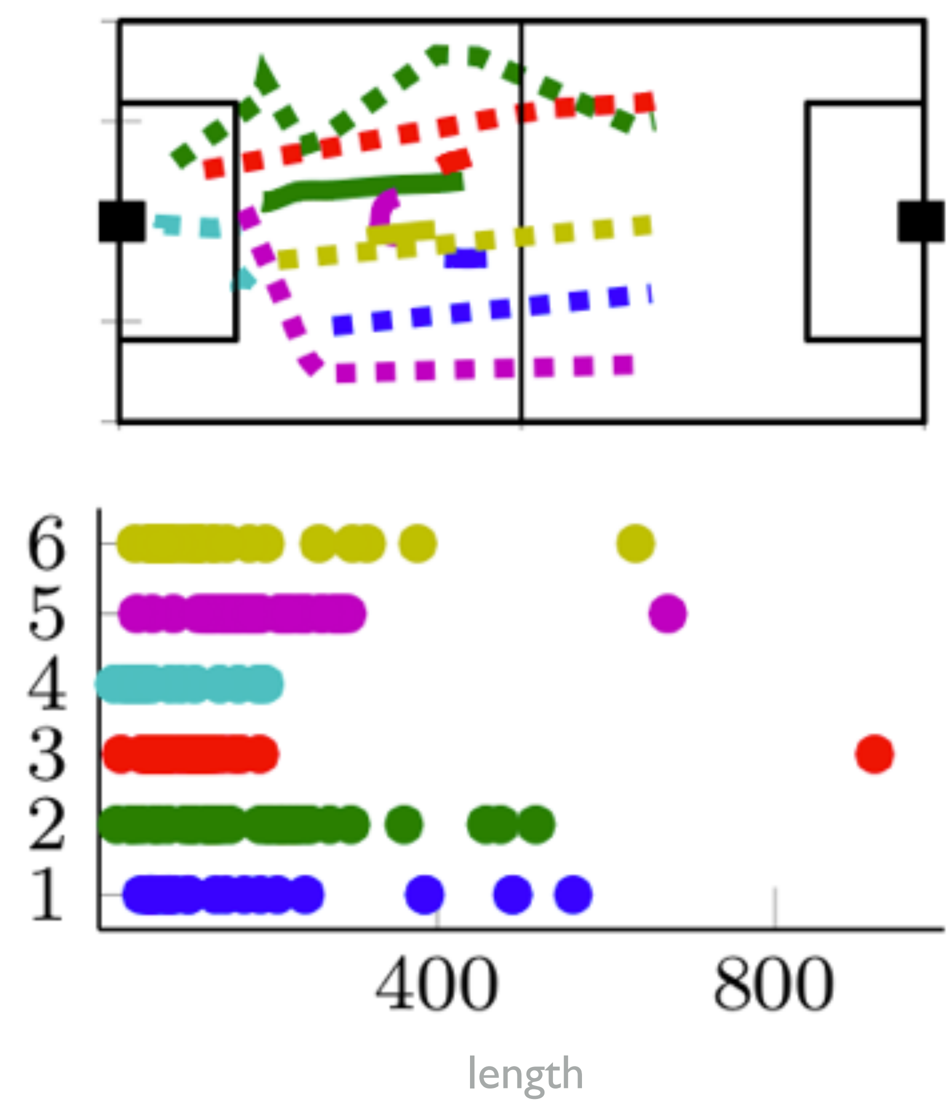
# Game Initiations

Knauf, Memmert & Brefeld, Spatio-temporal Convolution Kernels, Machine Learning Journal, 2015

Bundesliga Team A

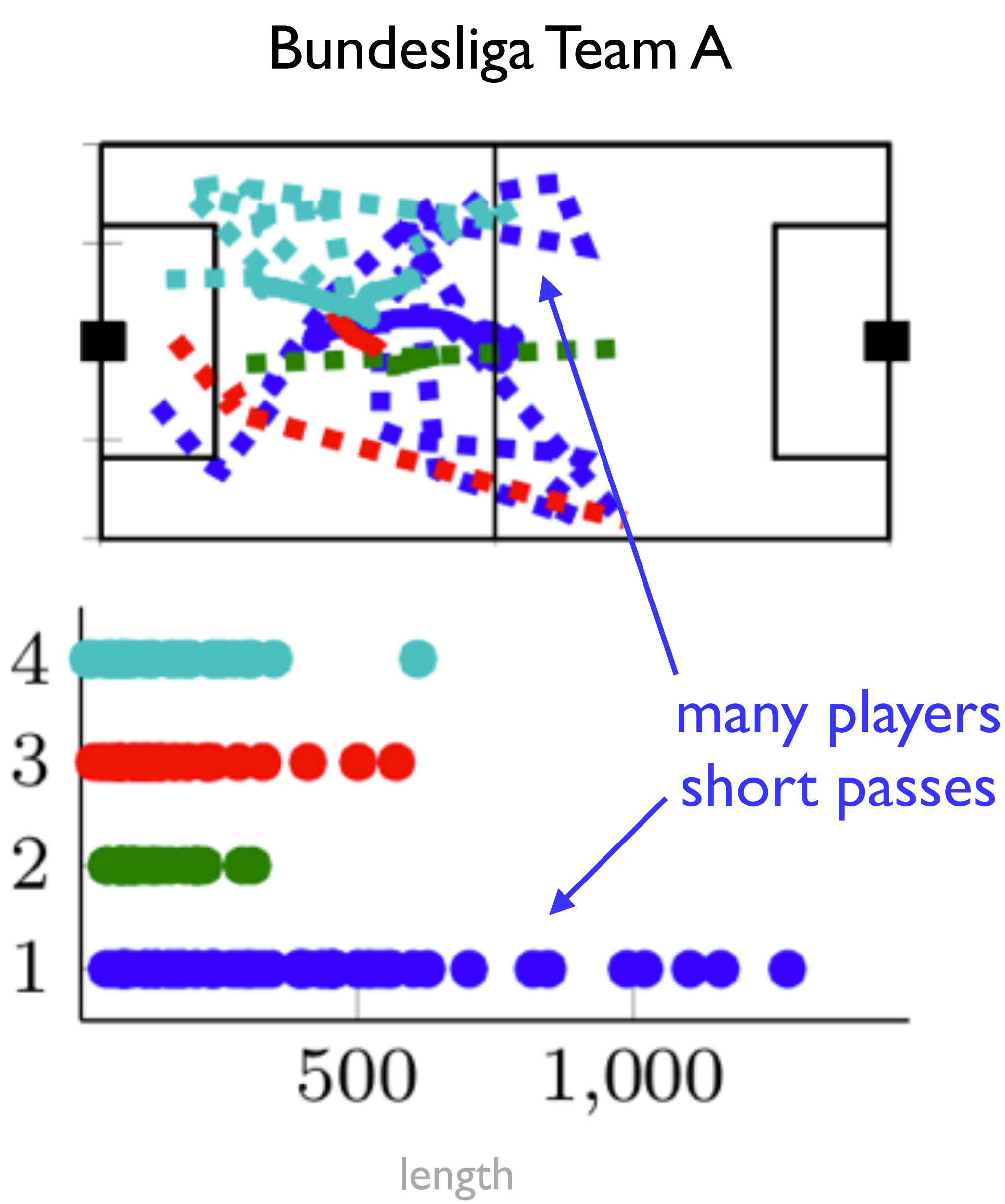


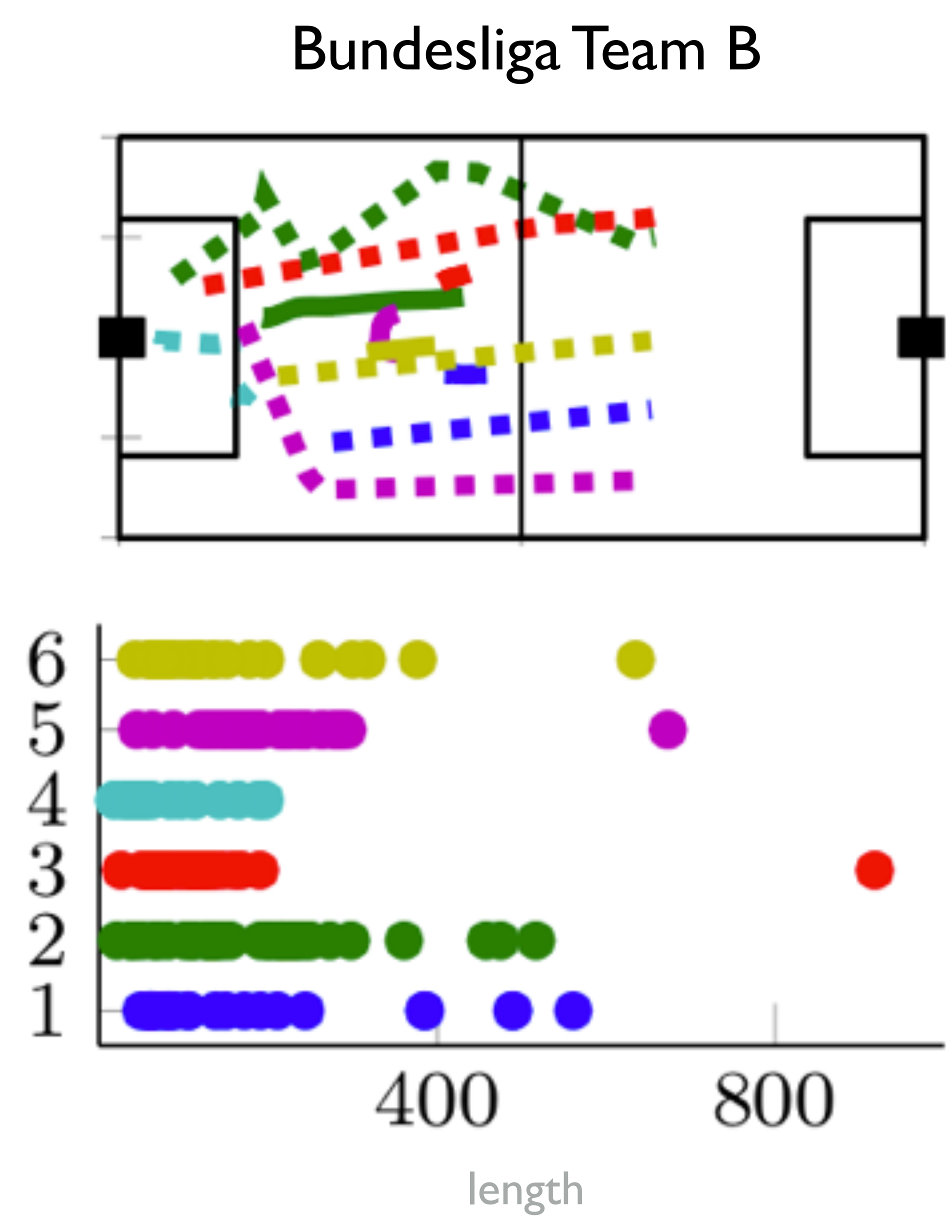

Bundesliga Team B



# Game Initiations

Knauf, Memmert & Brefeld, Spatio-temporal Convolution Kernels, Machine Learning Journal, 2015

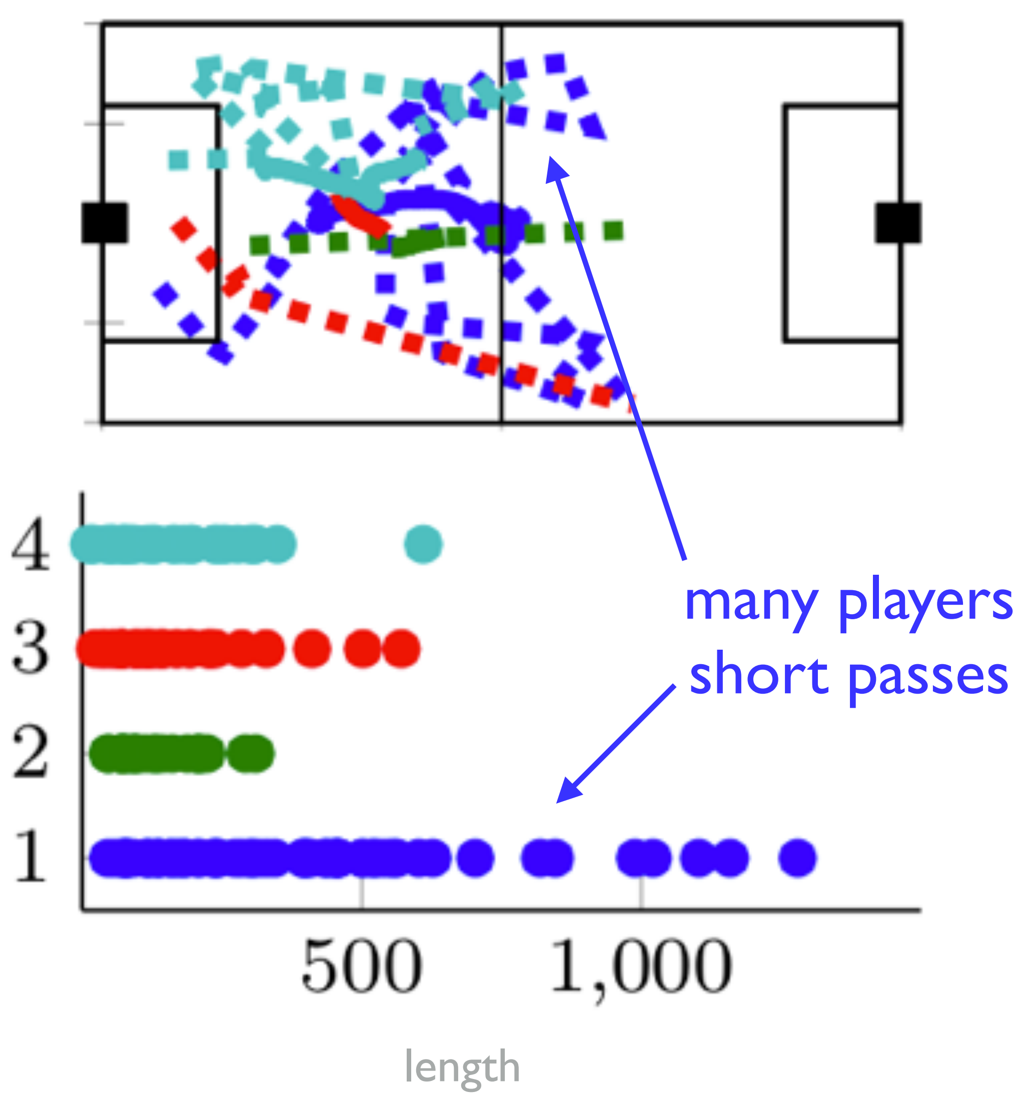


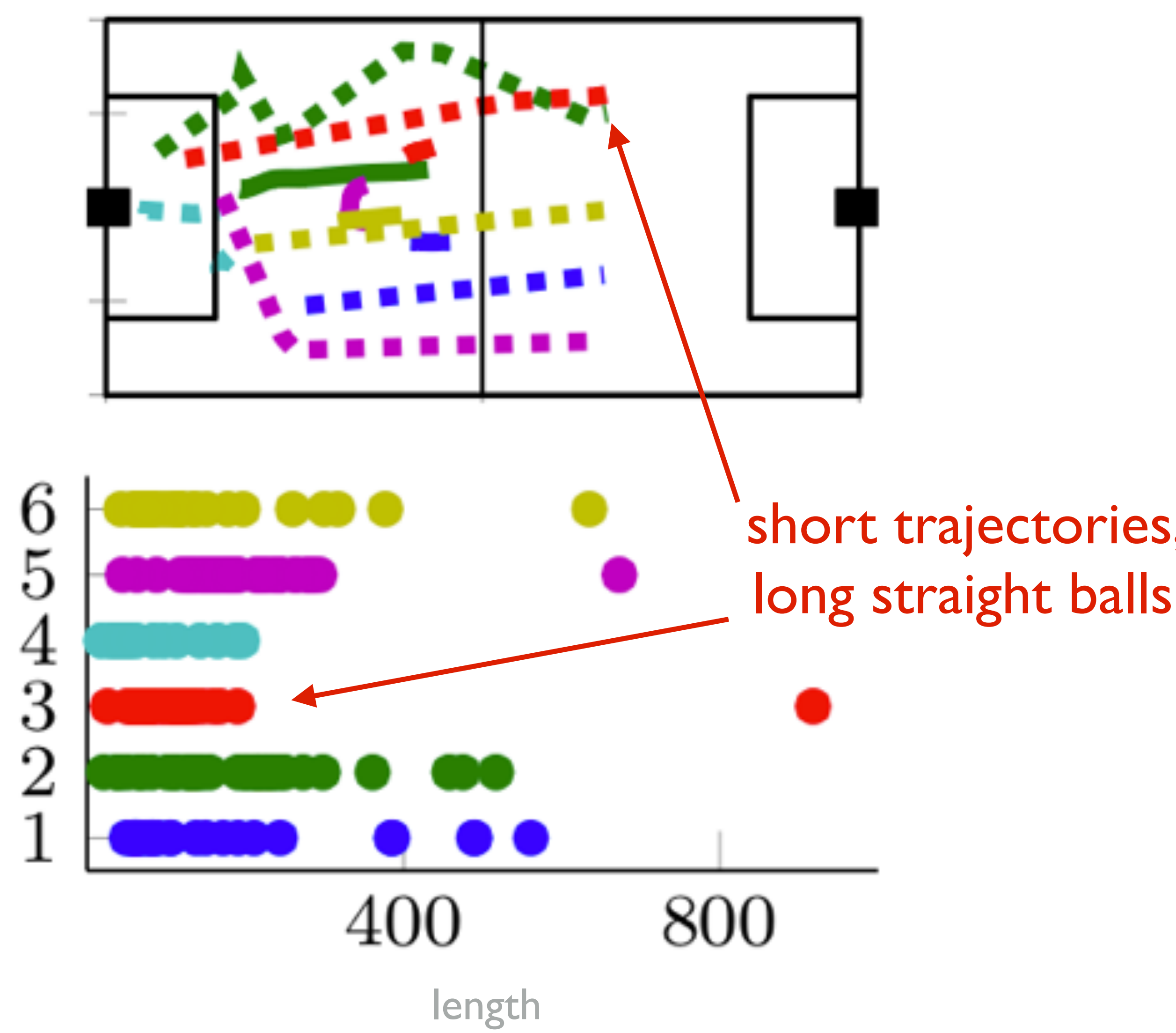
# Game Initiations

Knauf, Memmert & Brefeld, Spatio-temporal Convolution Kernels, Machine Learning Journal, 2015

### Bundesliga Team A




### Bundesliga Team B



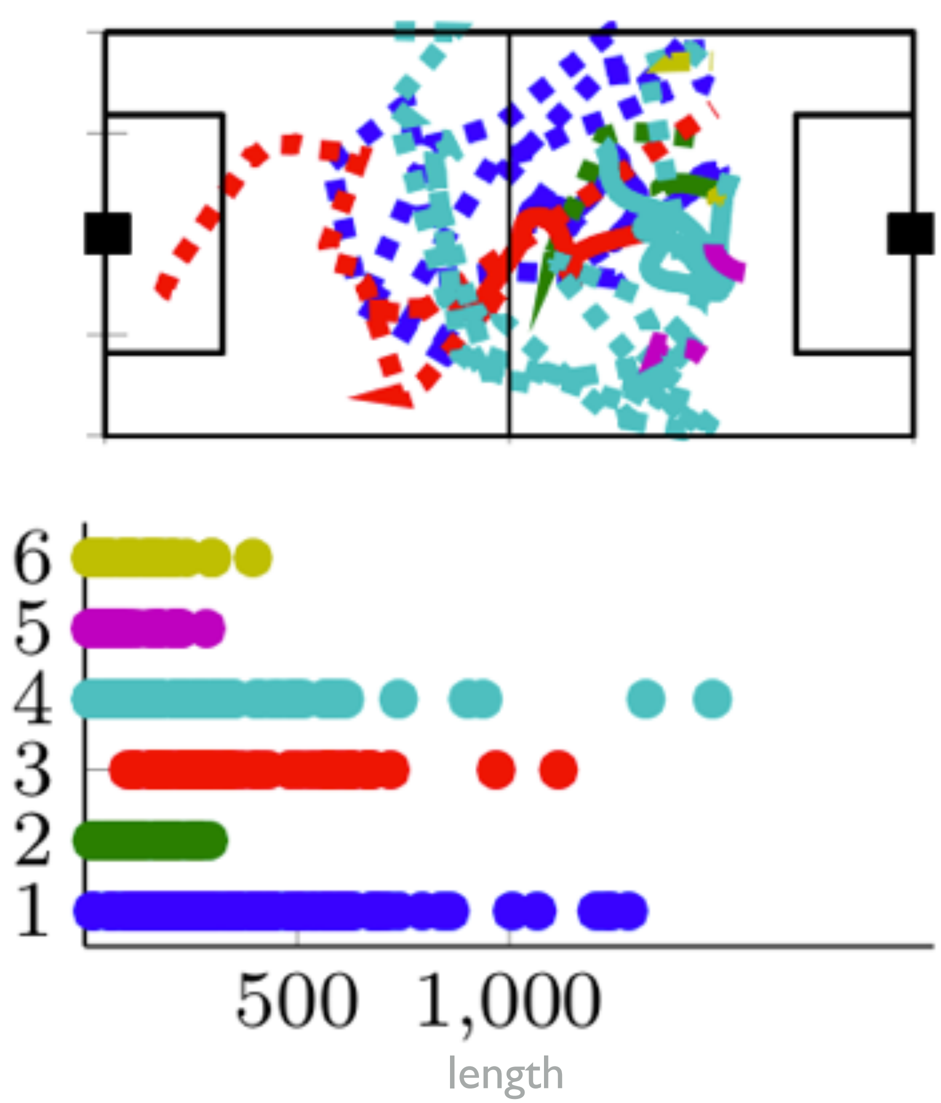
# Scoring Opportunities

- Team A:
  - Aimed at quickly scoring a goal in the opposing half, i.e., few ball contacts, faster ball transport in the zone of danger
- Team B:
  - Many ball contacts, took their time in waiting for a mistake of the opponent and only then played in the zone of danger to achieve a goal

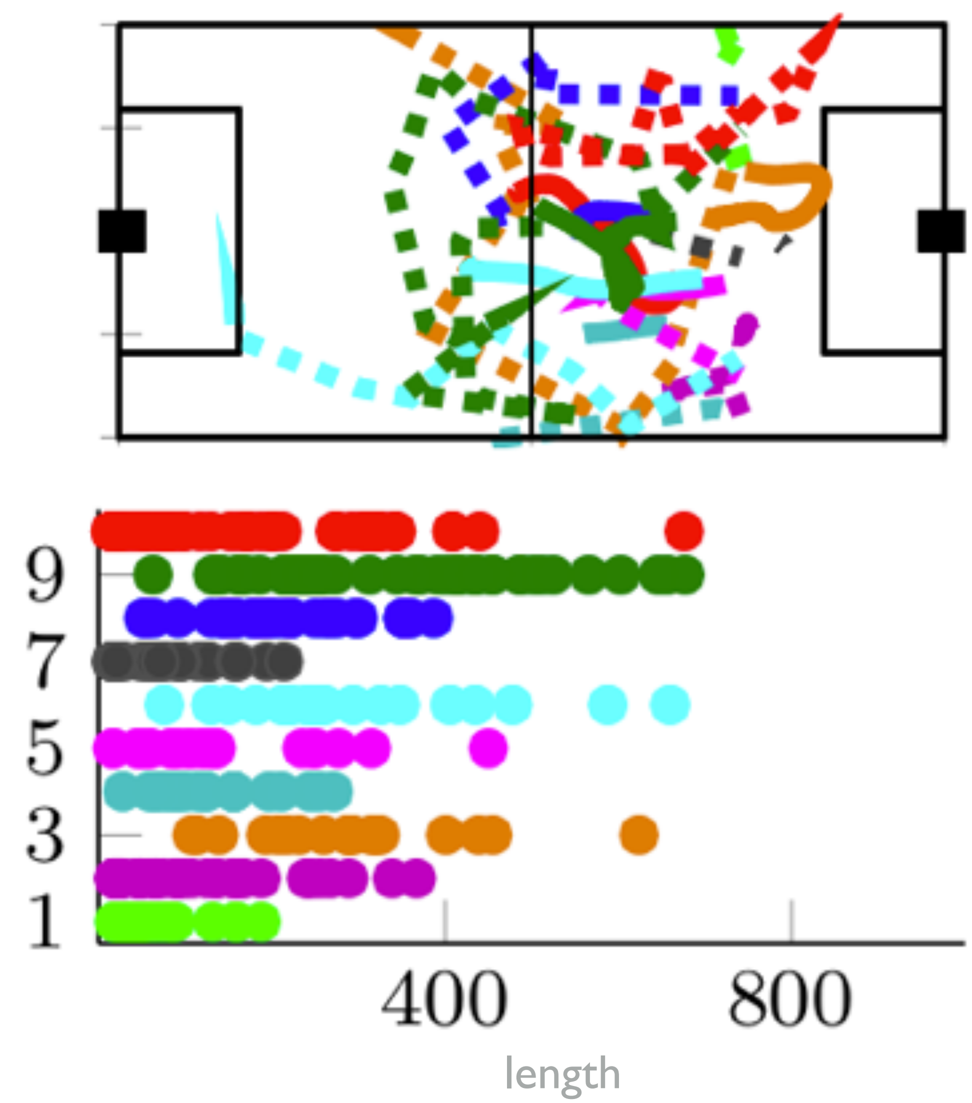
# Scoring Opportunities

Knauf, Memmert & Brefeld, Spatio-temporal Convolution Kernels, Machine Learning Journal, 2015

Bundesliga Team A



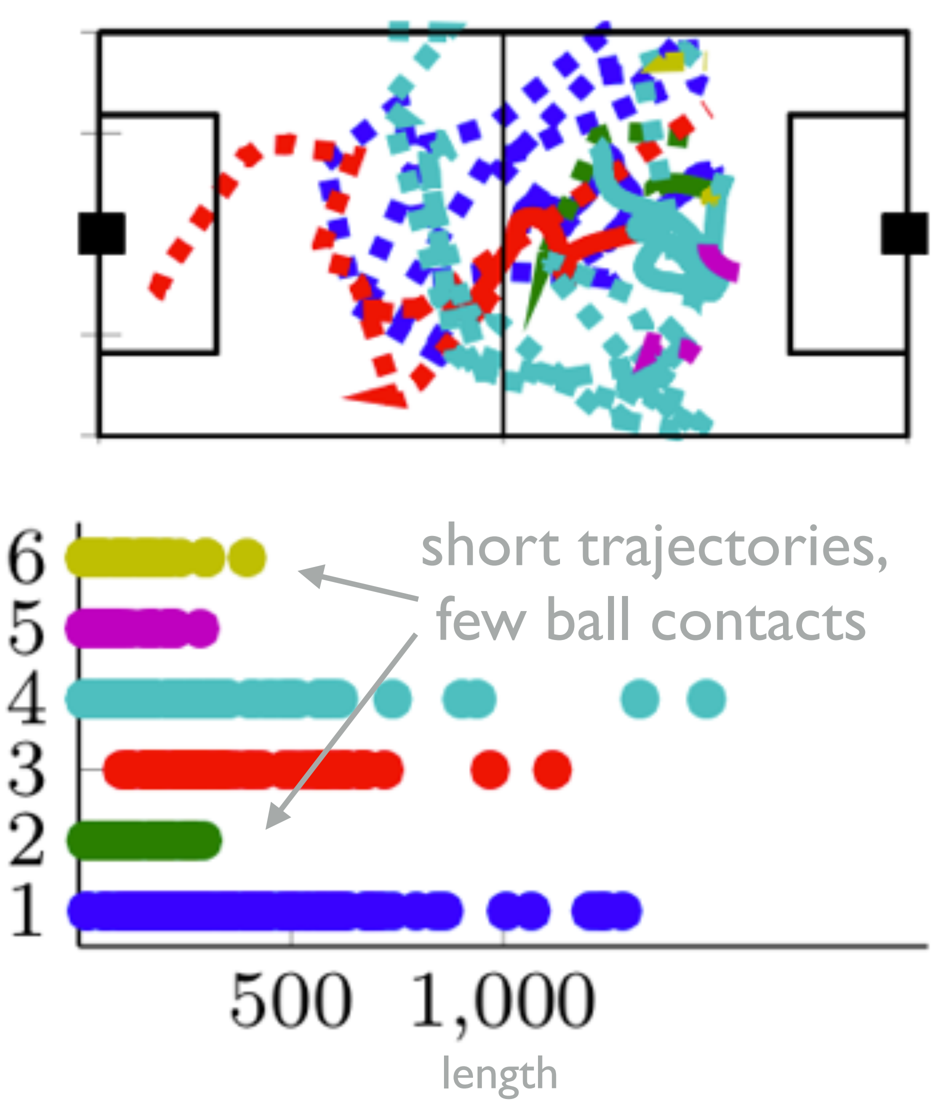

Bundesliga Team B



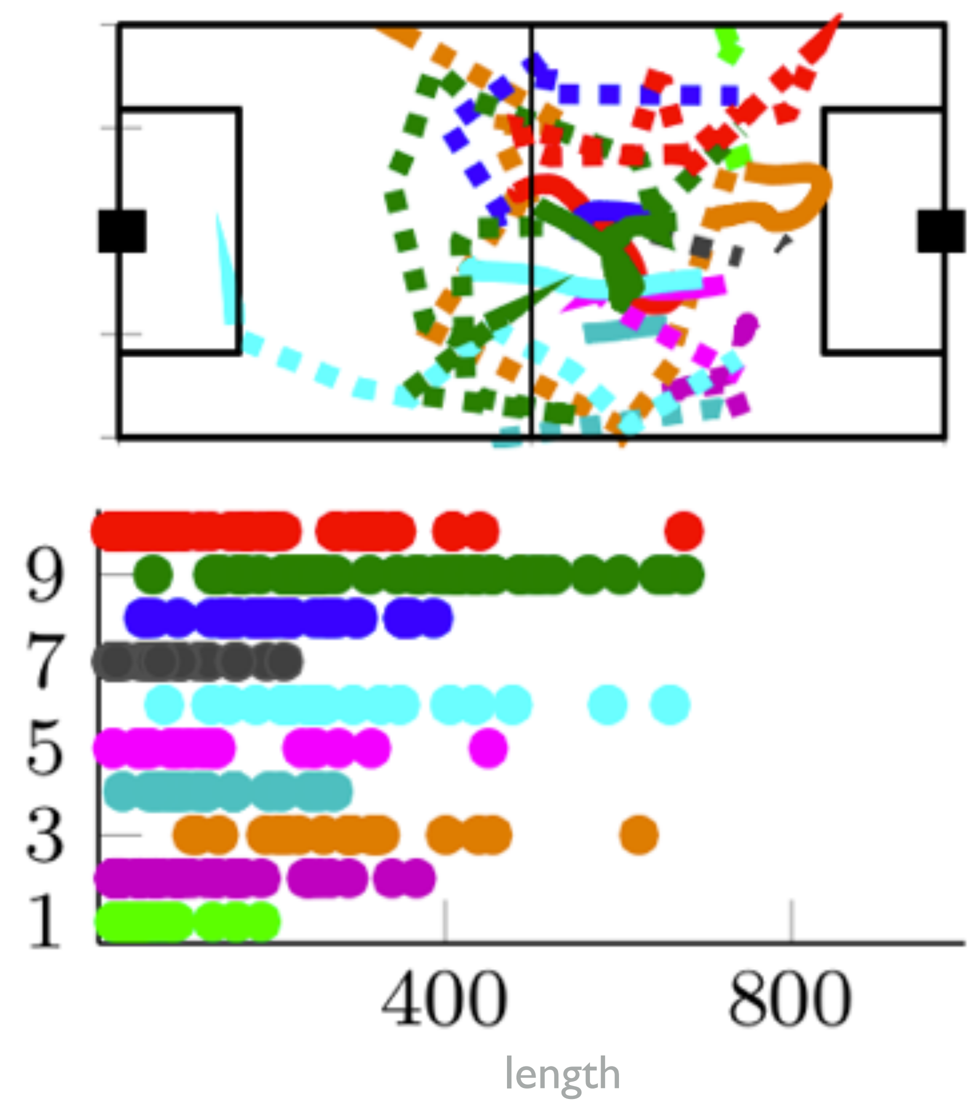
# Scoring Opportunities

Knauf, Memmert & Brefeld, Spatio-temporal Convolution Kernels, Machine Learning Journal, 2015

Bundesliga Team A



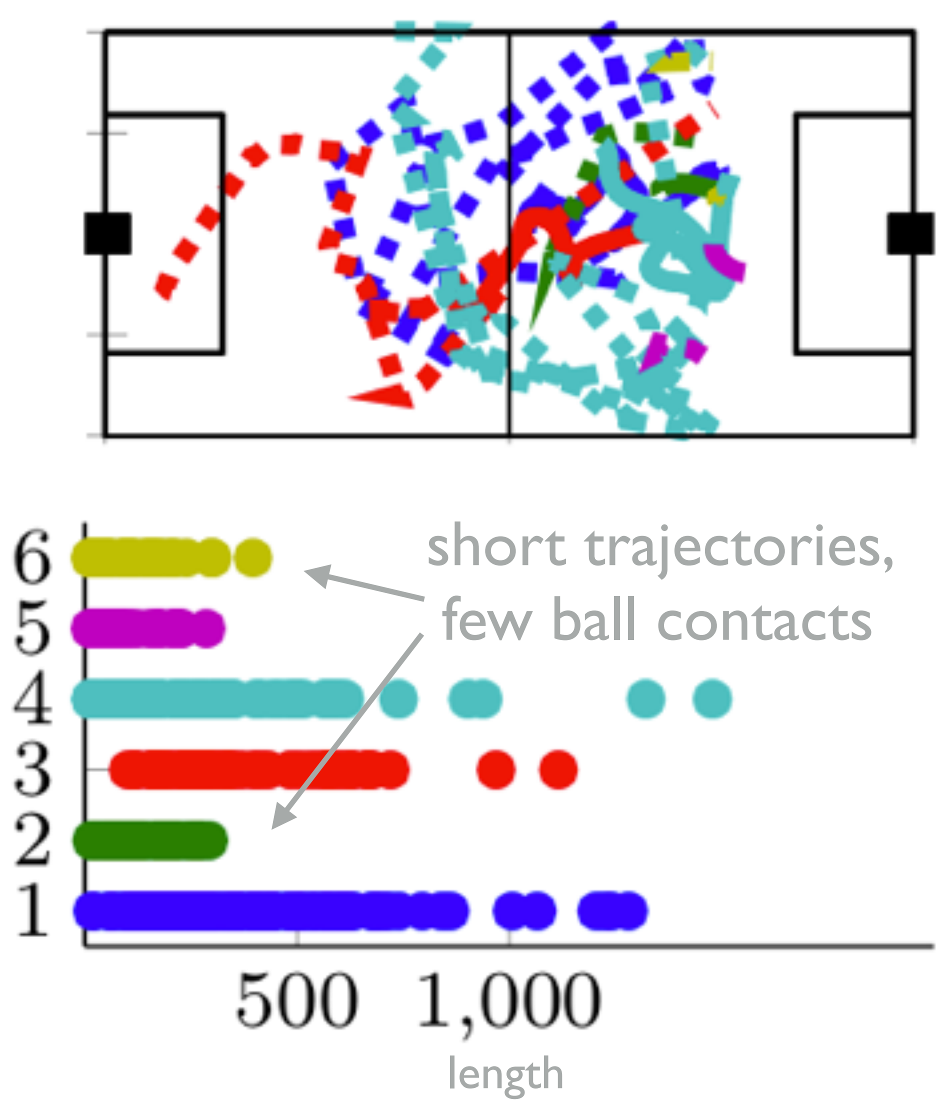

Bundesliga Team B



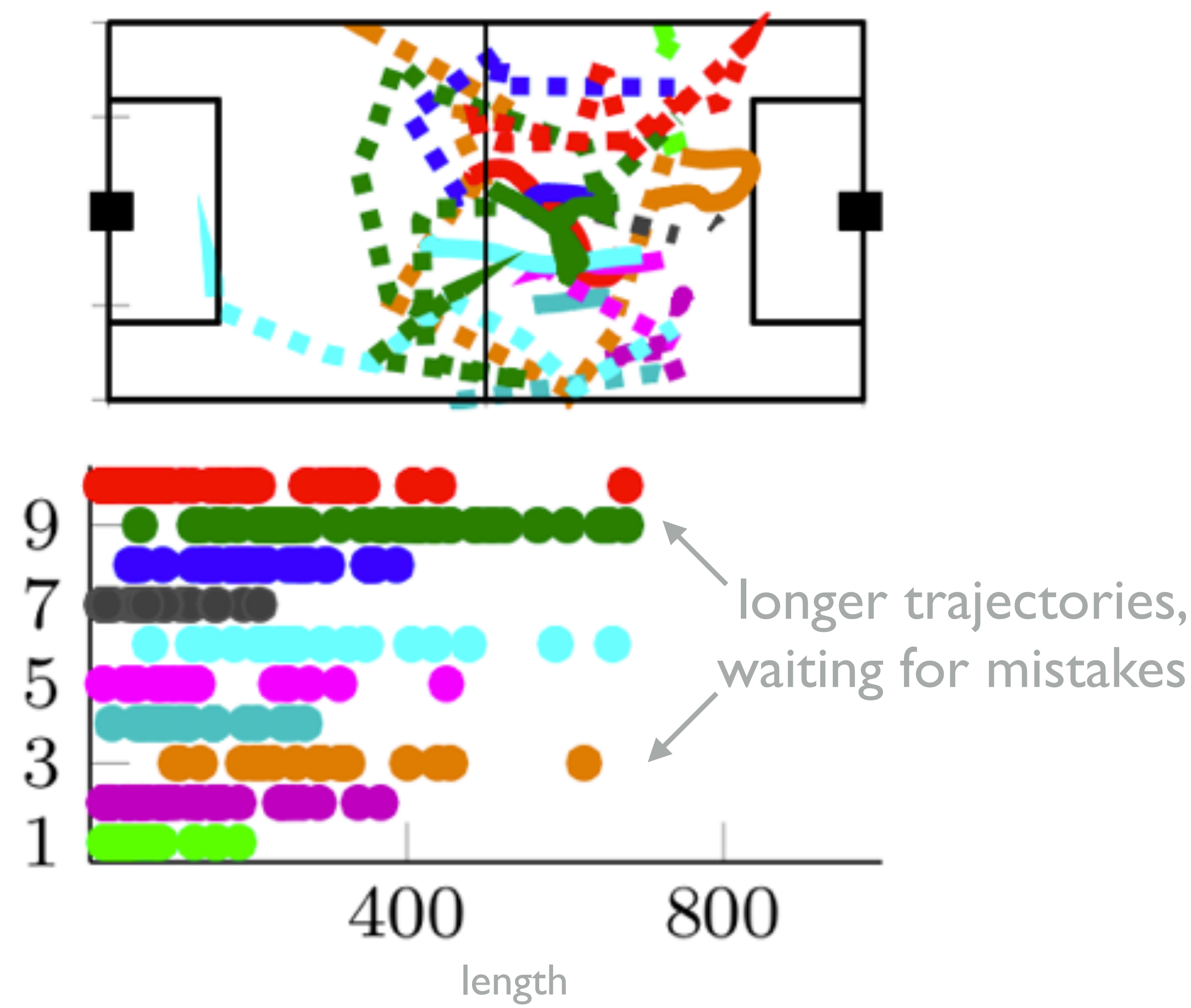
# Scoring Opportunities

Knauf, Memmert & Brefeld, Spatio-temporal Convolution Kernels, Machine Learning Journal, 2015

Bundesliga Team A



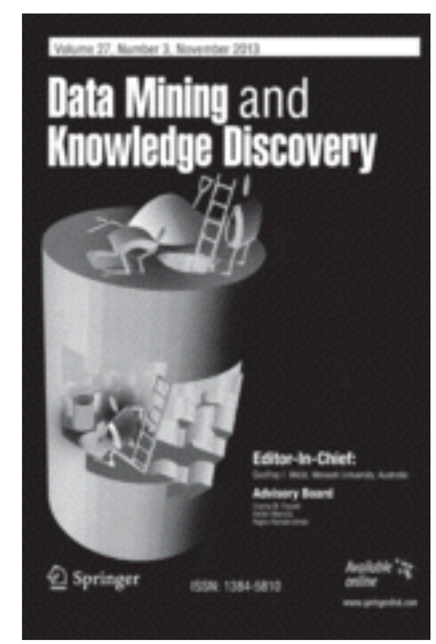

Bundesliga Team B



# DMKD Special Issue on Sports Analytics

(together with Albrecht Zimmermann)

- Goal is to publish special issue in 2016
- Cfp end of September 2015
- Submission deadline end of December 2015
- Inquiries:
  - albrecht.zimmermann@insa-lyon.fr
  - brefeld@cs.tu-darmstadt.de

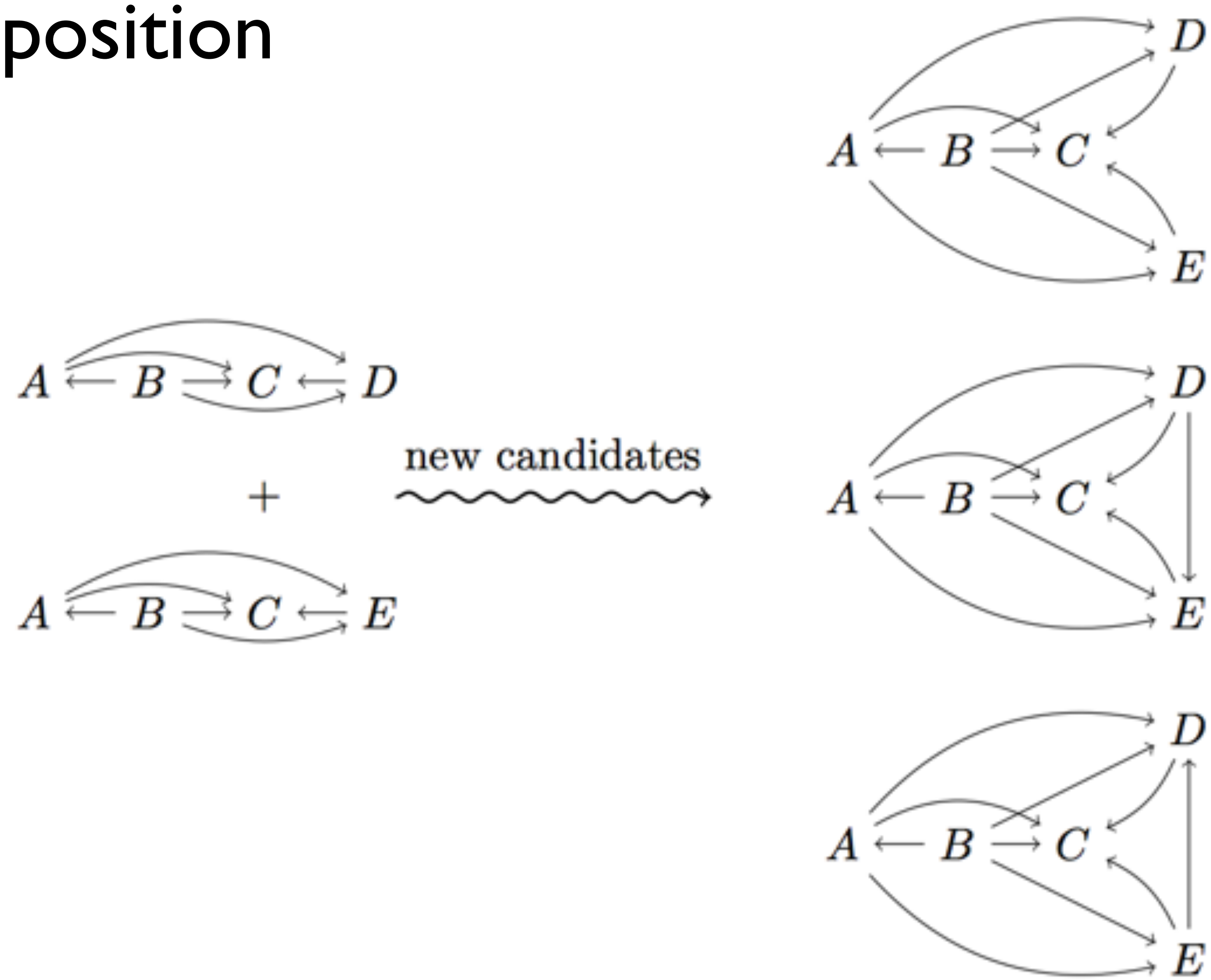


# Wrap-Up: Trajectory Data

- Analysing trajectories of players it the key to analysing coordination in team sports
- Potential use cases go far beyond heat maps
- Inherent complexity renders tasks challenging
  - Adapt existing large-scale algorithms to data
  - Exploit prior knowledge

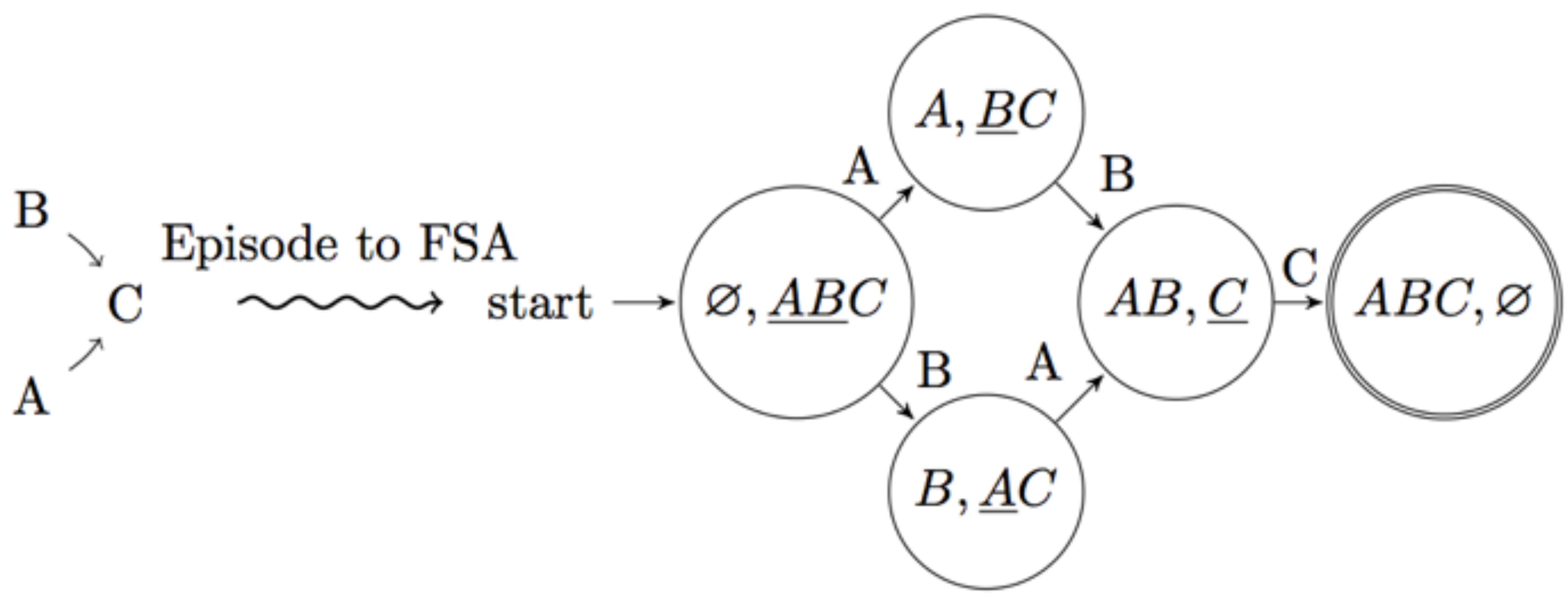
# Mapper: Candidate Generation

- Combine existing episodes that differ only in a single position



# Reducer: Counting

- FSA for every possible realisation of a known episode
- An additional FSA will always remain in initial state
- Similar to Laxman et al. (2005)

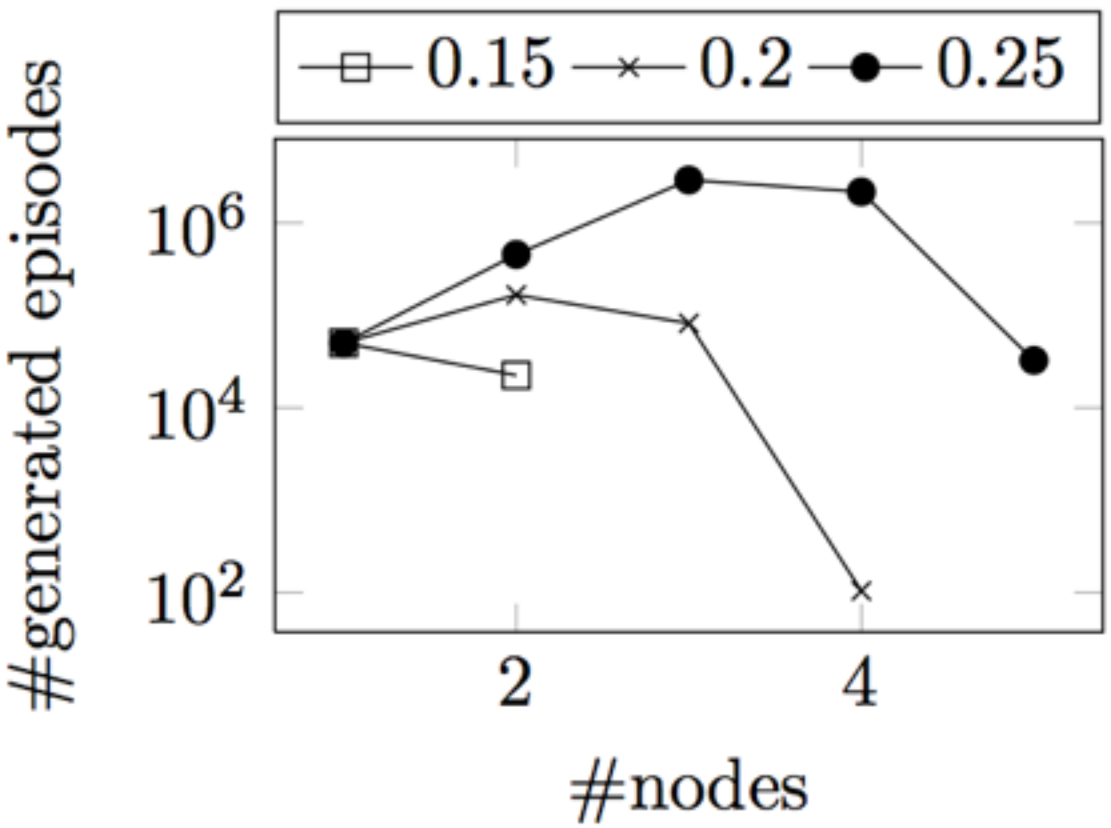


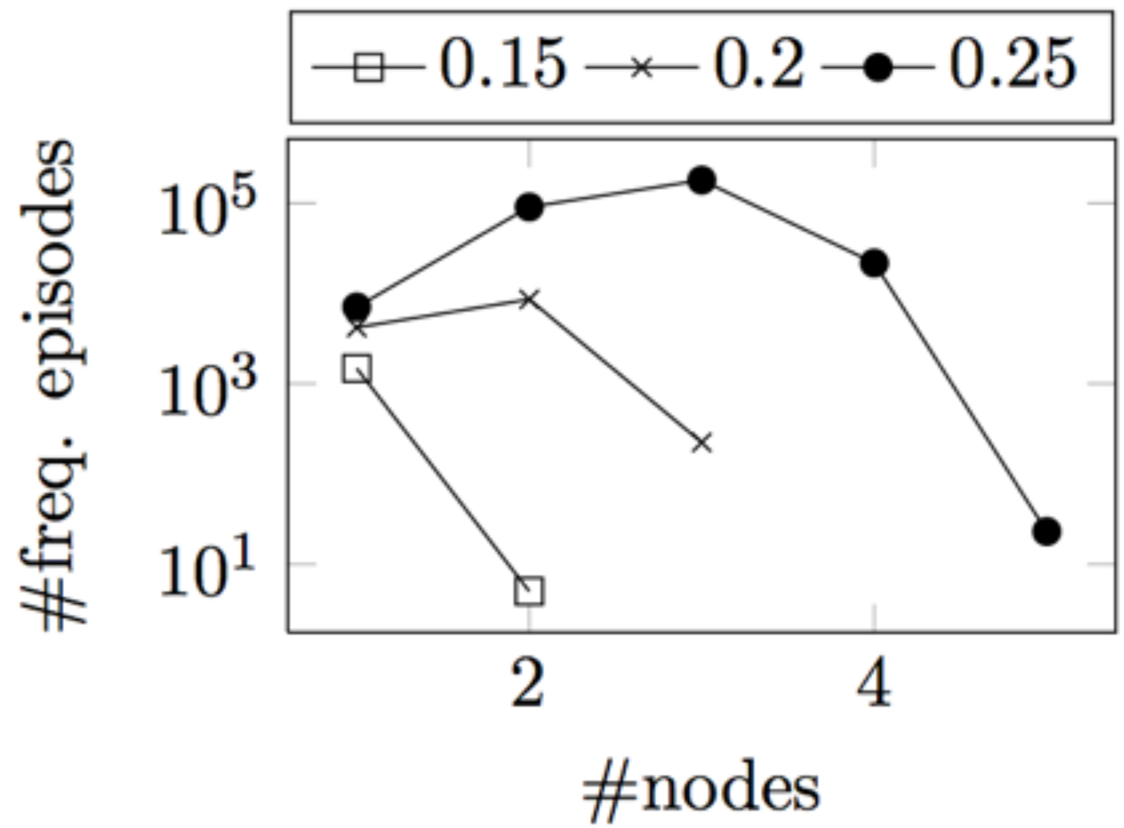
# Pruned Trajectories

| nof. trajetories | Kim | Keough | LSH | **total** |
|---|---|---|---|---|
| 1000 | 0% | 0% | 11,42% | **11,42%** |
| 5000 | 0,28% | 34% | 16,33% | **50,61%** |
| 10000 | 9,79% | 41,51% | 17,8% | **60,1%** |
| 15000 | 17,5% | 46,25% | 11,82% | **75,57%** |

- Effectiveness of DBH depends only on data
- Approximations effective for constant N

# Similarity Threshold




- Number of generated/frequent episodes depends highly on similarity threshold