Finding Similar Movements in Positional Data Streams

Jens Haase and Ulf Brefeld

Knowledge Mining & Assessment

brefeld@kma.informatik.tu-darmstadt.de

TECHNISCHE UNIVERSITÄT DARMSTADT

Prague, 27.9.2013

DIPF

Educational Research and Educational Information

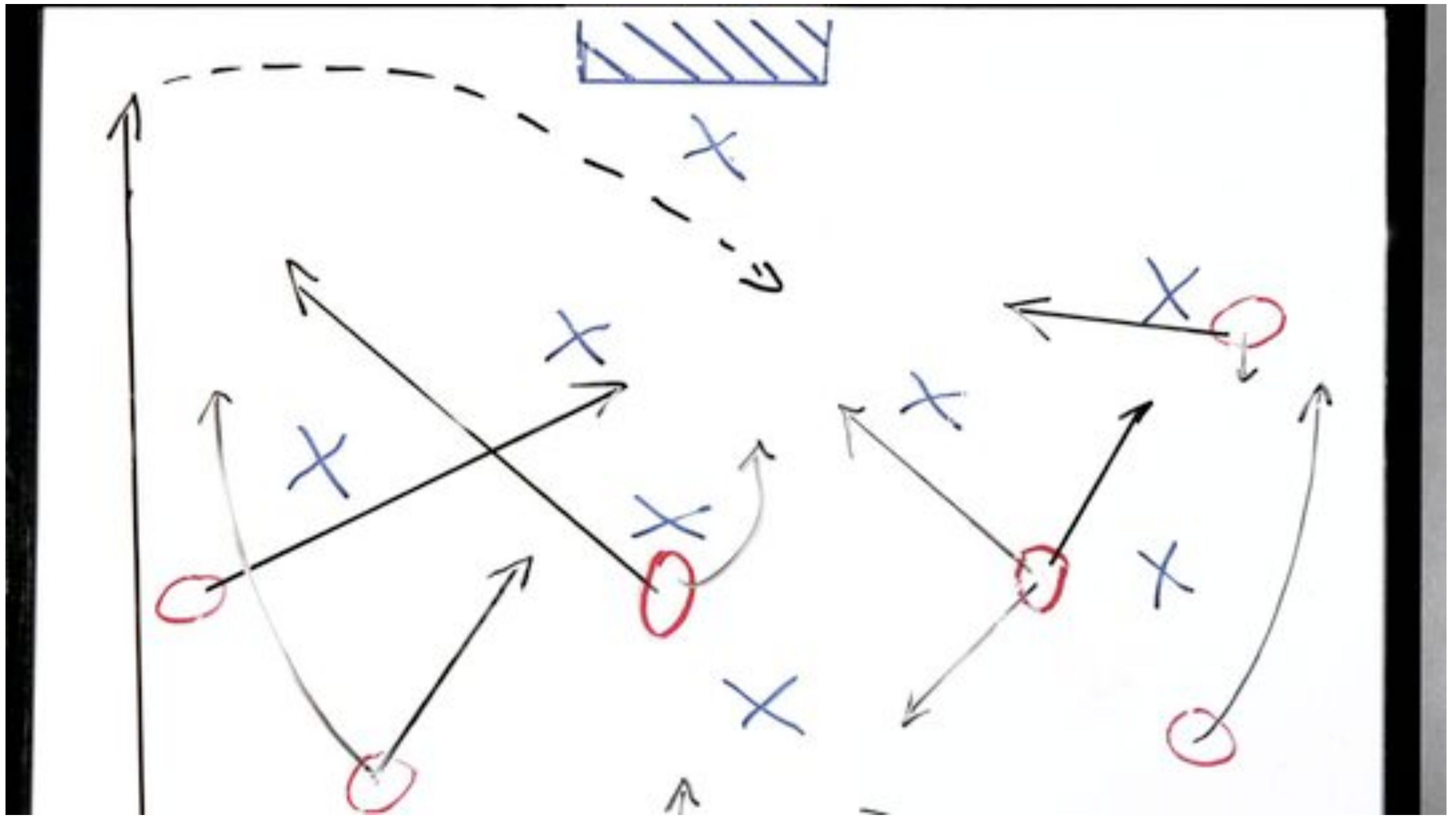
B. Charlton v F. Beckenbauer

David Marsh

1966 World Cup Final, England - W. Germany

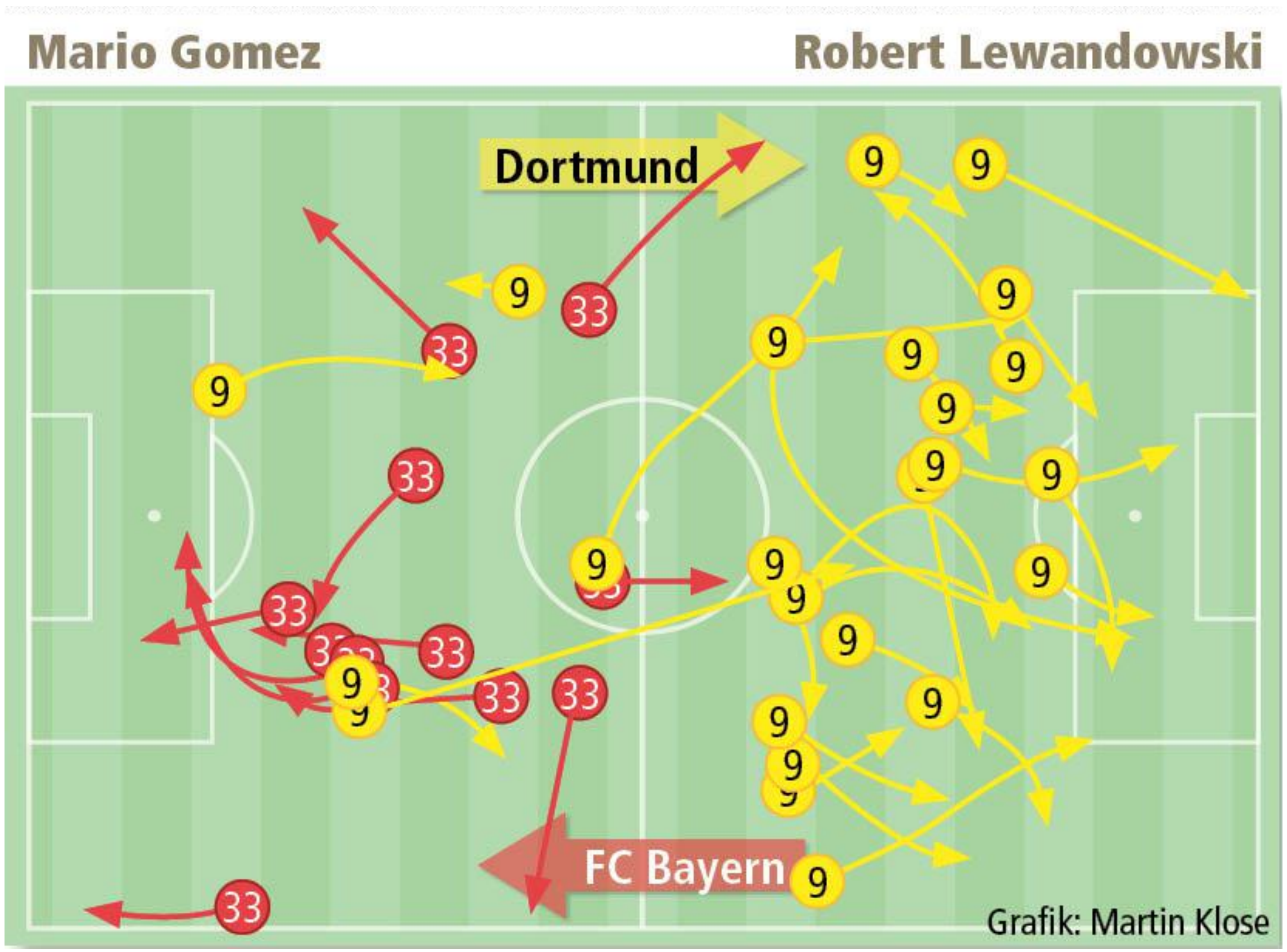
Player Briefing

(coach, before game)

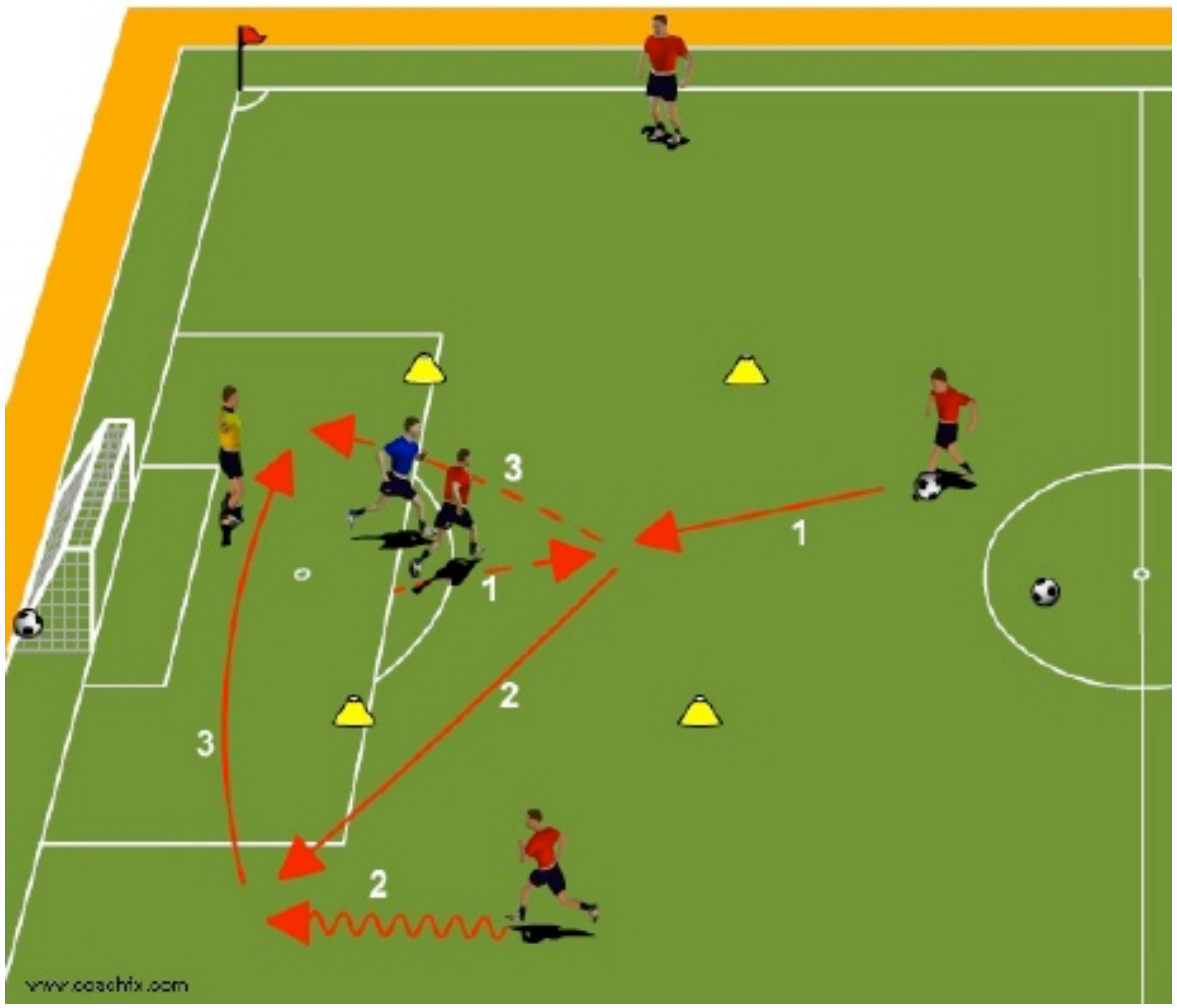


Analyses

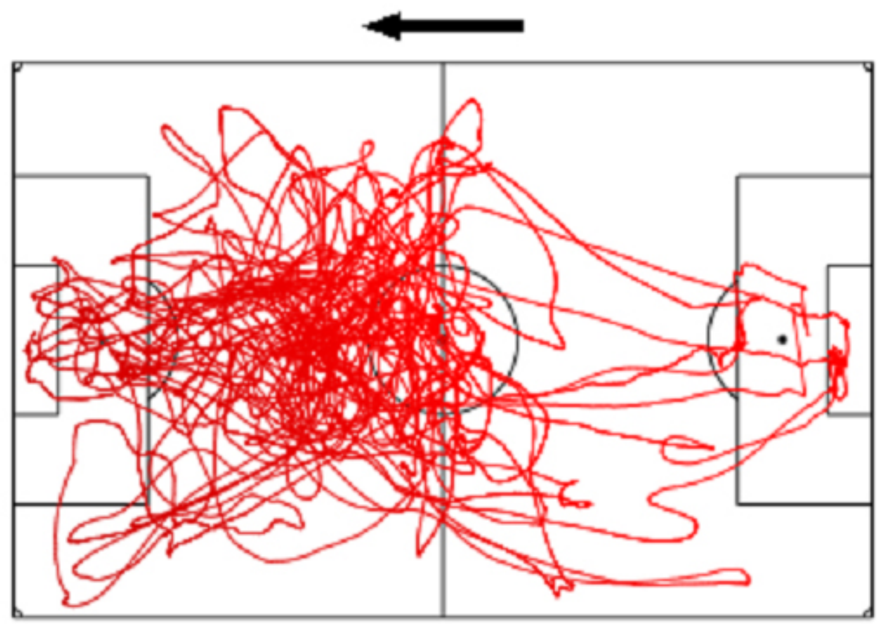
(newspapers, next day)



Youth Soccer: Tactics and Paths

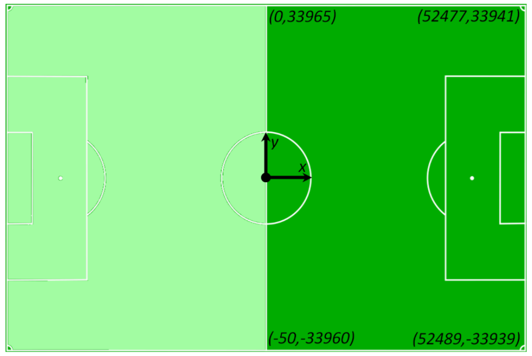


Tactics and Trajectories



- Understanding player movements precondition for analyzing tactics
- Requires efficient computation of similar movements
- **This talk**: Efficient retrieval of near-duplicate trajectories given a query movement

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

- Movements should be independent of player speed
- Dynamic time warping compensates phase shifts
- Distance measure $dist : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$
- 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)

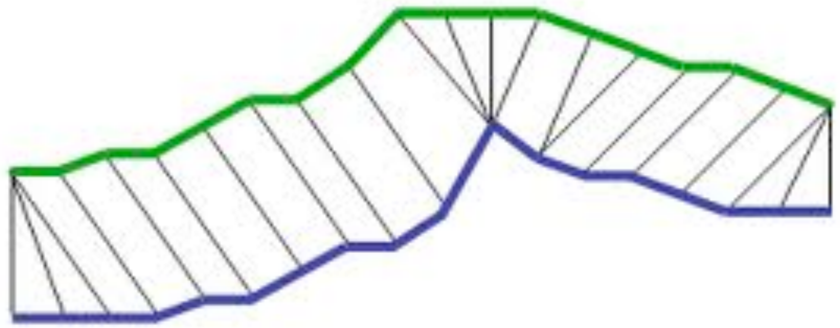
- Movements should be independent of player speed
- Dynamic time warping compensates phase shifts
- Distance measure $dist : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$
- DTW for sequences **s** and **q** defined recursively

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

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 $h : \mathcal{D} \to \mathbb{R}$

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

s_1 and s_2 randomly drawn from database

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

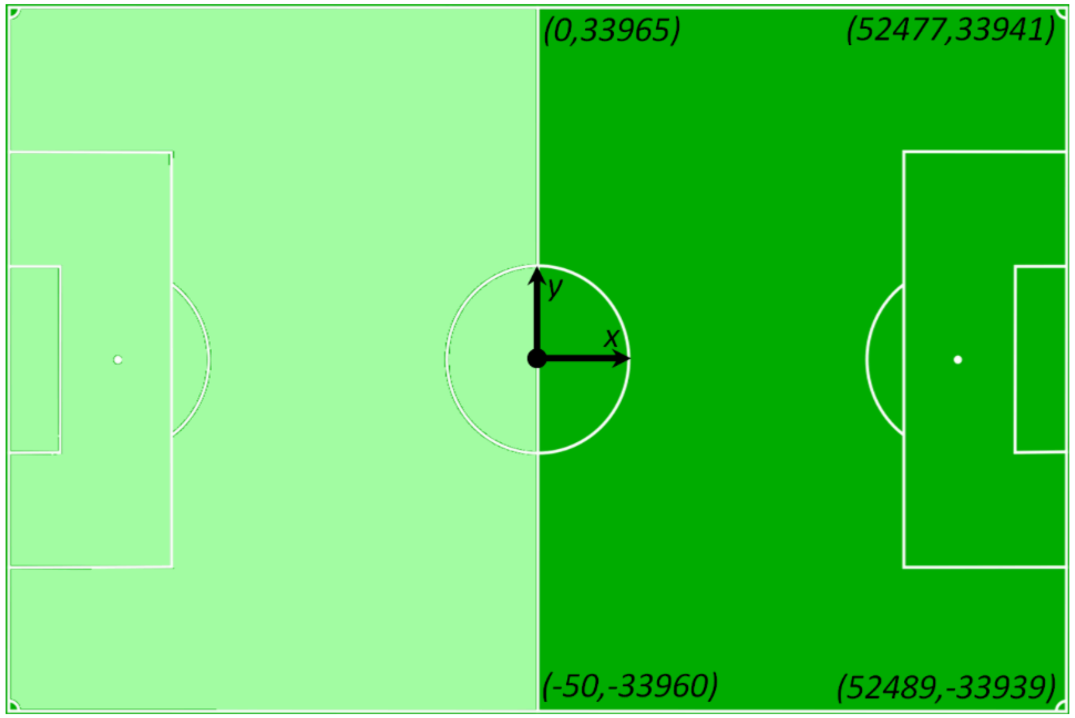
- Bucket determined by $h^{[t_1,t_2]}_{\boldsymbol{s}_1,\boldsymbol{s}_2}(\boldsymbol{s}) = \begin{cases} 1 : h_{\boldsymbol{s}_1,\boldsymbol{s}_2}(\boldsymbol{s}) \in [t_1,t_2] \\ 0 : \quad otherwise \end{cases}$
- Set of admissible intervals

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

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)

Coordinates on the Pitch

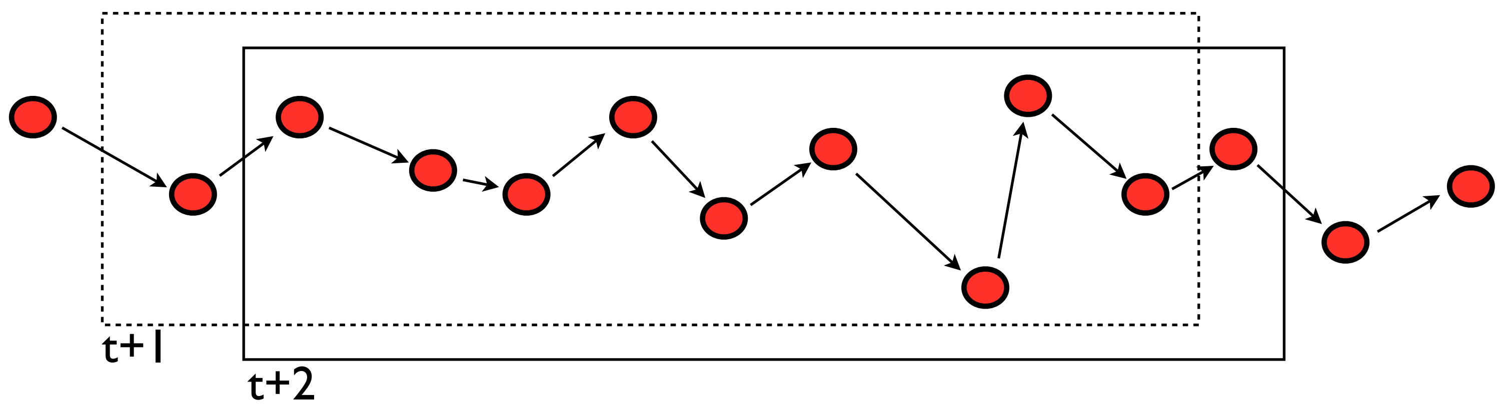


- Coordinate system, origin (0,0) is at kick-off
- Discarding additional data, players are represented by triplet:

(sensor/player id, timestamp, player coordinates)

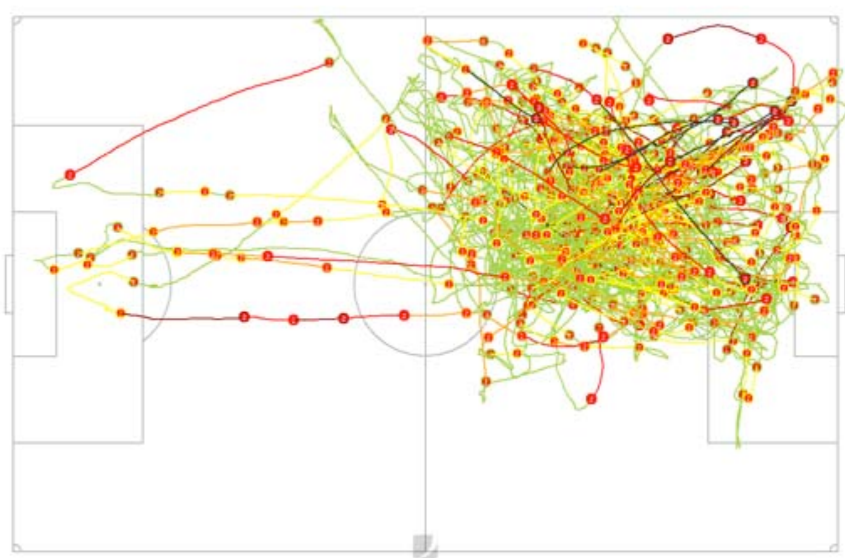
Representation

- 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

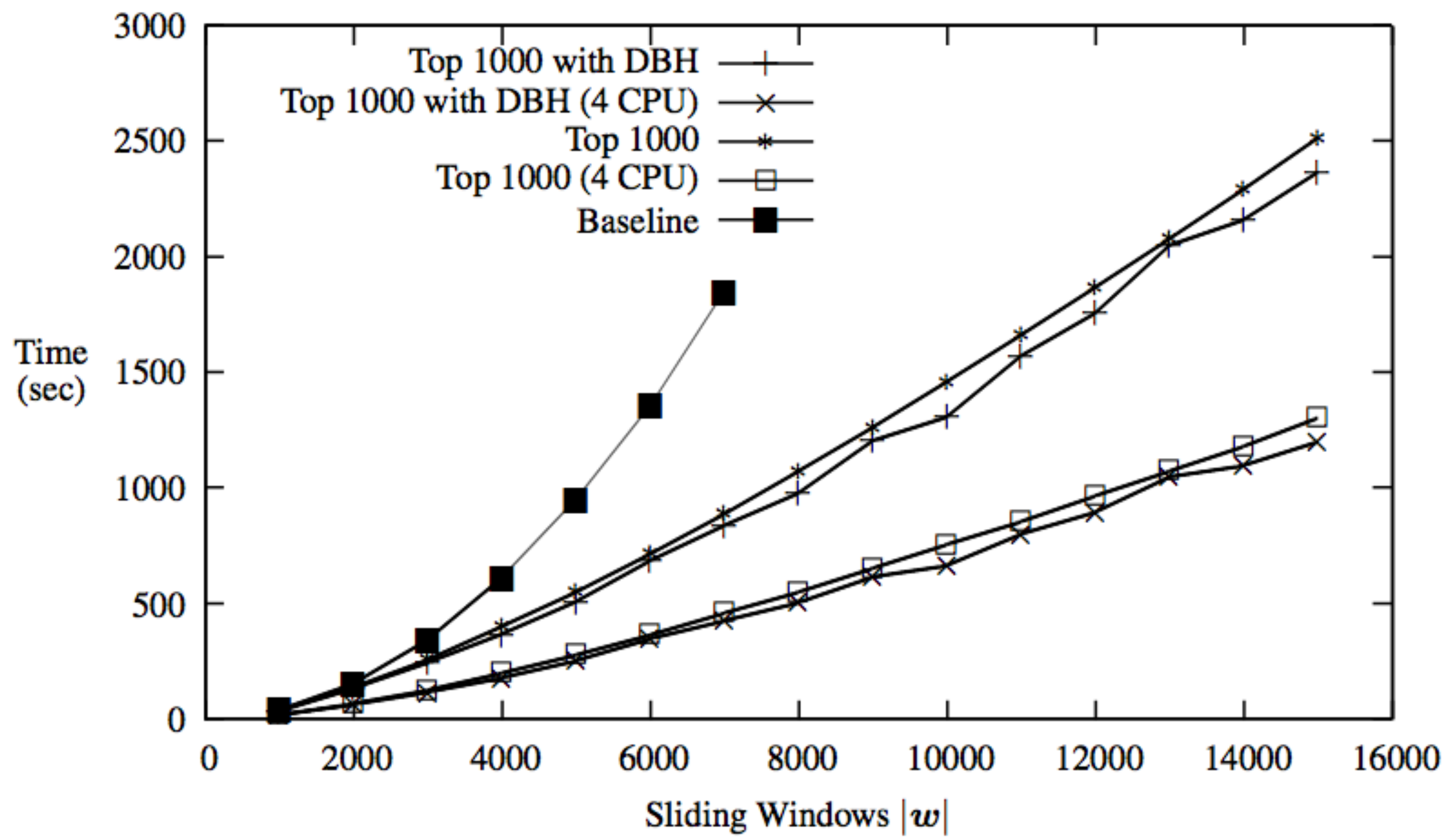


Evaluation

- 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



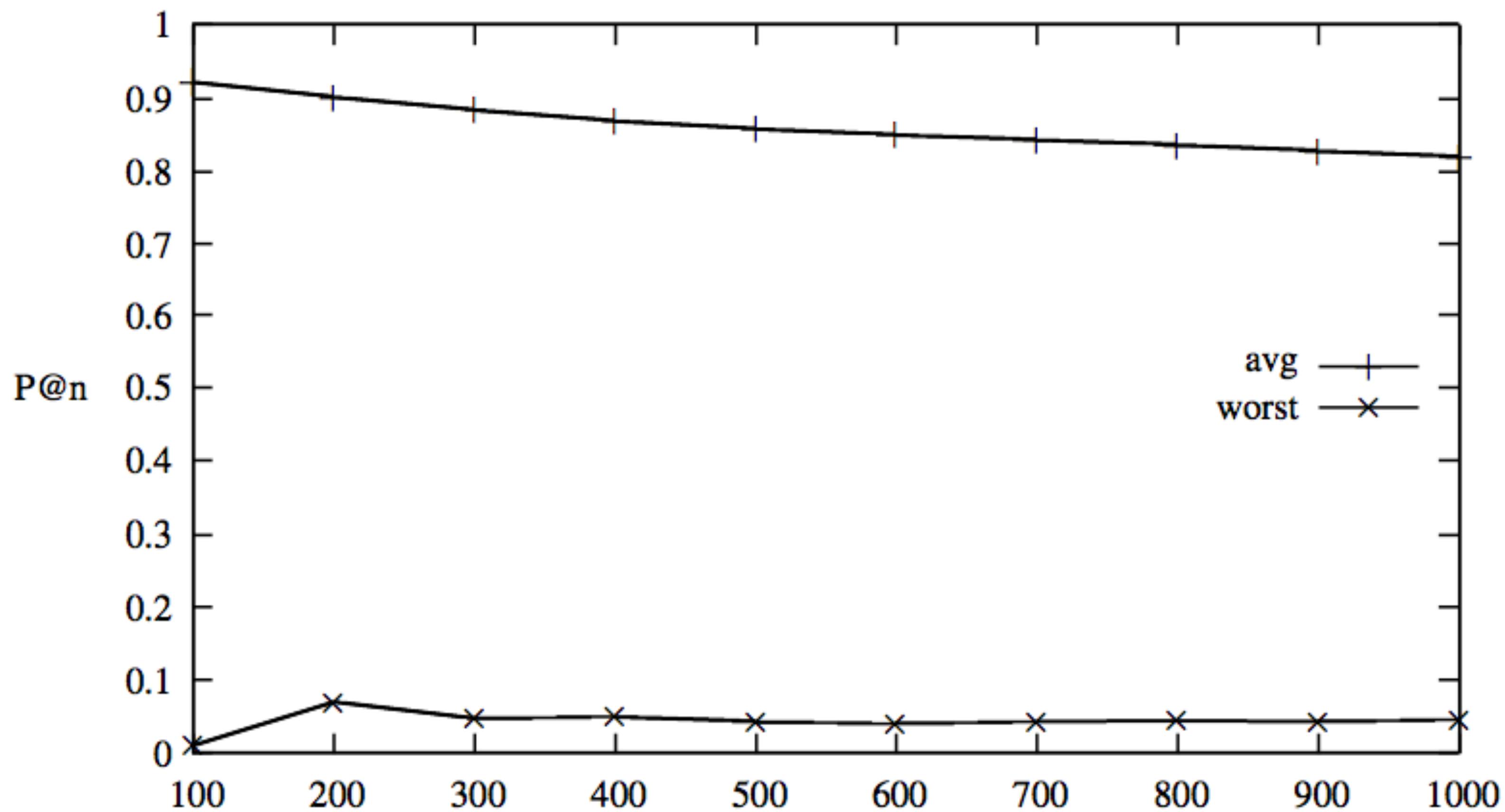
- Exact computation infeasible
- Dynamic time warping very effective
- DBH adds only little

Pruned Trajectories

nof. trajetories	Kim	Keough	DBH	**total**
1000	0.00%	0%	11.42%	**11.42%**
5000	0.28%	34.00%	16.33%	**50.61%**
10000	9.79%	41.51%	17.80%	**60.10%**
15000	17.5%	46.25%	11.82%	**75.57%**

- Effectiveness of DBH depends only on data
- Kim and Keogh effective for constant N

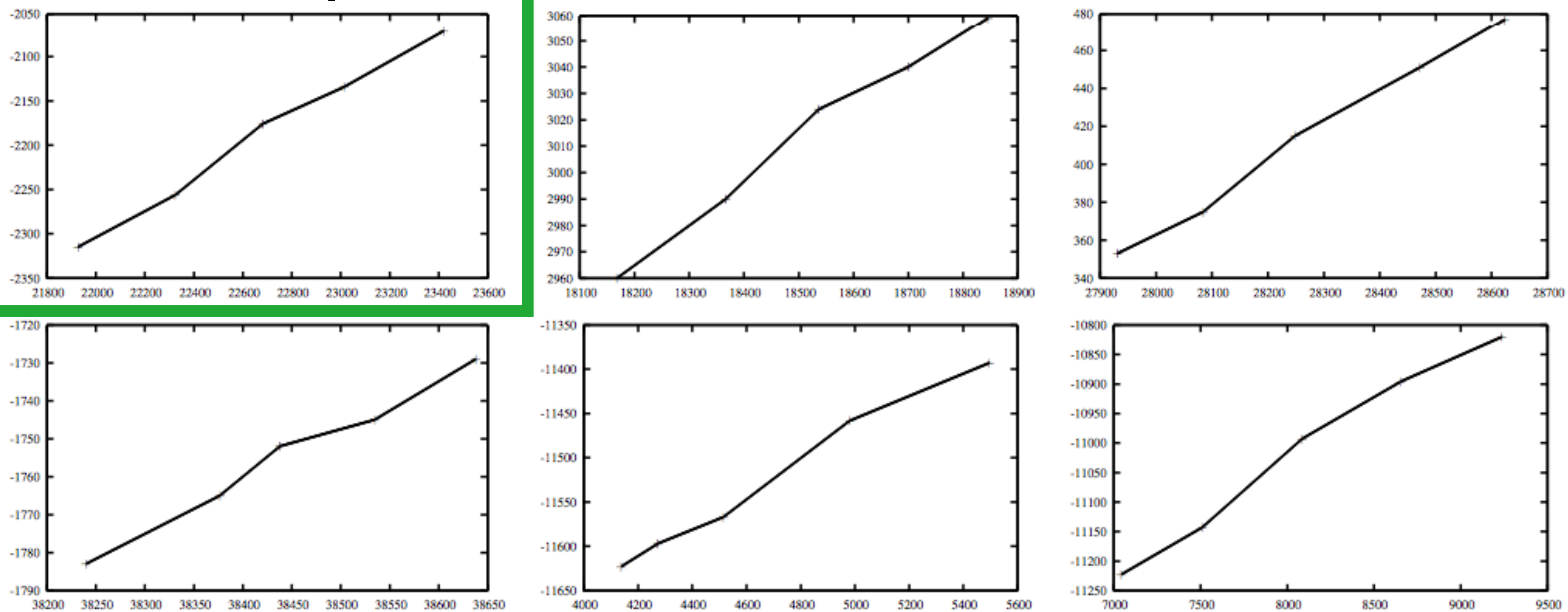
DBH Accuracy



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

Example

Query: Retrieved:



Conclusion

- Efficient computation of near duplicate movements in positional data streams
  - Dynamic time warping (DTW)
  - Distance-based hashing (DBH)
- (Super-)linear complexity
- Accurate results